Conditional Differential Expression for Biomarker Discovery in High-throughput Cancer Data

Dao Sen Wang

A thesis submitted in partial fulfilment of the requirements

for the degree of Master of Science in Microbiology and Immunology

with specialization in Bioinformatics

Faculty of Medicine

University of Ottawa

Ottawa, Ontario, Canada

Table of Contents

ABSTRACT

Biomarkers have important clinical uses as diagnostic, prognostic, and predictive tools for cancer therapy. However, translation from biomarkers claimed in literature to clinical use has been traditionally poor. Importantly, clinical covariates have been shown to be important factors in biomarker discovery in small-scale studies. Yet, traditional differential gene expression analysis for expression biomarkers ignores covariates, which are only accounted for later, if at all. We conjecture that covariate-sensitive biomarker identification should lead to the discovery of more robust and true biomarkers as confounding effects are considered. Here we examine gene expression in more than 750 breast invasive ductal carcinoma cases from The Cancer Genome Atlas (TCGA-BRCA) in the form of RNA-Seq data. Specifically, we focus on differential gene expression with respect to understanding HER2, ER, and PR biology – the three key receptors in breast cancer. We explore methods of differential expression analysis, including non-parametric Mann-Whitney-Wilcoxon analysis, generalized linear models with covariates, and a novel categorical method for covariates. We tested the influence of common patient characteristics, such as age and race, and clinical covariates such as HER2, ER, and PR receptor statuses. More importantly, we show that inclusion of a correlated covariate (e.g. PR status as a covariate in ER analysis) substantially changes the list of differentially expressed genes, removing many likely false positives and revealing genes obscured by the covariate. Incorporation of relevant covariates in differential gene expression analysis holds strong biological importance with respect to biomarker discovery and may be the next step towards better translation of biomarkers to clinical use.

ACKNOWLEDGEMENTS

I would like to sincerely thank Dr. Theodore Perkins for providing me with the opportunity for such a wonderful research experience and facilitating my growth as a scientist. Thank you for the years of guidance, knowledge, and support.

LIST OF FIGURES

1. INTRODUCTION

The use of biomarkers is fundamental in clinical research as well as in clinical practice. The World Health Organization (WHO) defines a biomarker as "any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence or outcome of disease" [1]. In simpler terms, they refer to objective, reproducible signs of a biological or disease state. With respect to cancer therapy, biomarkers often come in the form of genomic biomarkers and have important clinical use as diagnostic, prognostic, and predictive tools [2]. Most famously, the BRCA1 and BRCA2 mutations are used routinely in clinic for predicting risks of breast and ovarian cancer [3]. As well, prognostic and predictive biomarkers such as HER2 (human epidermal growth factor receptor 2), ER (estrogen receptor), and PR (progesterone receptor) statuses have major impacts on clinical decision making in breast cancer and have been approved for use by the FDA [2]. Biomarkers have long been claimed as the "key to better patient care and lower medical costs" [4]. For instance, the American Society of Clinical Oncology estimates that screening people with colon cancer for mutations in the *K-RAS* oncogene would save at least US$600 million a year [4]. Furthermore, biomarkers pave the road for personalized treatment strategies for cancer, which have much higher clinical efficacy compared to standard therapy [5]. However, biomarker advancements in healthcare are largely hindered by their poor translation from literature to use in clinic.

1.1 Traditional issues in biomarker translation

The poor translation of biomarkers is evidenced by the fact that there are more than 150,000 papers reporting thousands of disease-associated biomarkers, yet fewer than 100 have been validated for

routine clinical use [4] [6]. The biomarker pipeline can be divided into six essential processes: biomarker candidate discovery, qualification, verification, research assay optimization, clinical validation, and commercialization [6]. Success of this pipeline begins with the discovery of optimal and robust biomarkers. It is often that biomarkers discovered and evaluated in cells *in vitro* do not translate to cell *in vivo*. As well, heterogeneity – in both tissues and across patients – remains a high hurdle that biomarkers studies must overcome [7].

Flaws in the reproducibility and validity of discovered biomarkers are highlighted by a 2006 study reporting gene expression profiles for determining patient response to different forms of chemotherapy [8]. Initially, this study generated great excitement in the field of cancer biomarkers. However, the gene expression patterns in the study were discovered and tested only on a set of cancer cell lines, and subsequent studies found major flaws in this work [2]. The paper was later retracted from Nature Medicine in 2011. Lack of reproducibility and clinical validity are the two major reasons why discovered biomarkers do not make it far down the biomarker pipeline.

A 2011 Nature review cites the lack of standardization and insufficient sampling to be major impediments to biomarker research [4]. In many studies, patient characteristics in biomarker studies are not controlled properly; patients are often not matched for sex, age, weight, ethnicity, or lifestyle factors [4]. The lack of control for patient heterogeneity limits the ability to make strong correlations between biomarkers and patient conditions. This heterogeneity is further exacerbated across studies with different patient populations, which leads to poor reproducibility. Furthermore, small-scale studies lack the statistical power to infer clinically useful and robust biomarkers [9], rendering the results of the studies less compelling when it comes to validation.

With respect to cancer, it is well established that patient characteristics or clinical covariates (such as the aforementioned sex, age, weight, ethnicity, and lifestyle) play important roles in the risk and

progression of disease [10] [11] [12]. Clinical covariates can also include elements invisible to the naked eye, such as blood serum composition or the presence of other biomarkers (e.g. HER2 status in breast cancer). These covariates have confounding effects when it comes to creating prediction models, such as those for cancer diagnostics, prognostics, or treatment response. When these covariates are not controlled in the study design, it is pertinent that statistical analyses account for these possibly confounding effects [13] [14]. In small-scale studies, adjusting for clinical covariates improves the accuracy of diagnostic and prognostic biomarkers [15] [16] [17]. With respect to gene expression studies, analyses that integrate sequence information with clinical and patient data result in improved performance and sensitivity [18] [19]. Accounting for covariates can address the issue of the lack of standardization posed in the 2011 Nature review.

Furthermore, acknowledgement of covariates becomes increasingly important for large-scale high-throughput studies (as a direct solution to the issue of insufficient sampling). Covariates can introduce underlying directional biases which can skew the data and limit prediction power; and such an issue cannot be solved simply by increasing sample size. The recent explosion in data volume from high-throughput studies will require accompaniment of appropriate statistical analyses, such as controlling for covariates, before proper conclusions can be made.

1.2 Issues in differential gene expression analysis

The traditional approach to identifying candidate biomarkers is with genome-wide differential expression analyses. Differential expression refers to differences in gene expression levels between two conditions. In the context of cancer biomarkers, this can be gene expression differences in tissues between healthy and cancer patients or differences in tumours of early-stage

vs. late-stage patients, for example. Differential expression analysis has been traditionally performed using data obtained from DNA microarrays to quantify mRNA abundance [20]. RNA sequencing (RNA-Seq), a next-generation sequencing technology, offers a high-throughput alternative. Compared to microarrays, RNA-Seq has greater dynamic range (able to measure many orders of magnitude more of expression difference for the same gene), and the ability to detect point mutations or gene fusions. Many bioinformatics pipelines have been developed to accommodate RNA-Seq data for differential expression analysis, the output of which is a list of differentially expressed genes between the two conditions, which may serve as biomarkers [20]. However, despite advancements in technology and data quantity, development of gene expression biomarkers continues to experience the same pitfalls described previously.

It has been well established that patient characteristics and clinical covariates have effects on gene expression. For example, elevated gene expression levels of interleukin-6 and other proinflammatory cytokines has been correlated to age [21]. With respect to ethnicity, there is evidence to suggest that gene expression profiles differ between African Americans and European Americans in colorectal and breast cancer patients [22] [23]. Genomic changes such as single-nucleotide polymorphisms (SNPs) and copy number variants (CNVs) also impact gene expression phenotypes [24]. One of the most important findings in breast cancer research was the discovery of tumour subclass classifications based on different patterns of gene expression [25]. For example, HER2 status is a routinely tested prognostic biomarker in clinic for breast cancer; and different gene expression patterns have been found for HER2-positive and HER2-negative breast cancer tumours [26]. Numerous studies have found differential expression patterns between clinical conditions of interest. We believe there is merit for these conditions to be considered as covariates, if relevant, in further differential expression analyses.

Yet, expression biomarker studies have exclusively operated with traditional differential expression, without covariates. For example, in two highly cited and established papers for breast cancer signatures [27] [28], we believe covariates were not sufficiently considered at the stage of differential expression analysis for determining the gene signatures. In the studies, gene signatures were selected to predict either survival [27] or recurrence [28] without consideration of patient characteristics. Only later, when assessing survival outcomes or recurrence risks, are covariates considered in the prediction models. As a result of the omission of covariates at the stage of differential expression analysis, it is quite possible that valid genes for the signature are missed and less valid genes creep in. A list of possibly suboptimal gene candidates has downstream consequences for analysis and validation. And even if clinical covariates are considered later, it may be too late.

Statistical tools developed for differential expression analysis (DESeq2 [29], edgeR [30]) do allow for inclusion of covariates, however such features have been used mostly to account for batch effects or other technical artifacts in RNA-Seq data. For example, a 2014 study by Li et al. found an eight-miRNA signature as a potential prognostic biomarker for lung adenocarcinoma [31]. In this study, although clinical data was considered later on in receiver operating characteristic (ROC) analysis, differential expression analysis using edgeR only accounted for the batch effects and the sampling properties of RNA-Seq data.

We conjecture that covariate-sensitive biomarker identification should lead to the discovery of more robust, and true biomarkers as confounding effects are considered. There is a lack of evidence that the inclusion of patient characteristics and clinical covariates are being considered at the level of differential expression analysis in high-throughput cancer studies. We believe this may be a

major flaw in the biomarker discovery pipelines which needs to be addressed in order for better biomarker translation.

In this thesis, we evaluate incorporation of clinical covariates in differential expression analysis of high-throughput cancer data. We seek to answer the fundamental question of how the inclusion of covariates changes what appear to be relevant biomarkers. We hypothesize that 1) covariate models have increased robustness and consistency across patient sets, 2) covariate models remove covariate-associated genes and reveal covariate-obscured genes, and 3) use of correlated over uncorrelated covariates has stronger effects on differential expression analysis of the condition of interest. We examine gene expression in more than 750 breast invasive ductal carcinoma patients from The Cancer Genome Atlas (TCGA-BRCA) in the form of RNA-Seq data. Specifically, we focus on differential gene expression with respect to understanding HER2, ER, and PR biology - the three key receptors in breast cancer. We explore methods of differential expression analysis, including non-parametric Mann-Whitney-Wilcoxon analysis, generalized linear models with covariates, and propose a novel categorical method for covariates. Furthermore, we test the influence of common patient characteristics, such as age and race, and clinical covariates such as HER2, ER, and PR receptor statuses. Here, we present strong evidence of the importance of accounting for covariates in differential expression analysis of high-throughput data. With covariates and with more data from large-scale studies, we believe optimized biomarker discovery can overcome hurdles that previously limited its translation to clinical use.

2. BACKGROUND

Large-scale collaborative efforts and advancements in technology have greatly improved the field of cancer research. Initiatives such as The Cancer Genome Atlas (TCGA), led by the National Institutes of Health (NIH), and the International Cancer Genome consortium have produced a wealth of publicly available high-throughput cancer data. Such large-scale and regulated collaborations are needed to overcome the challenges often faced by investigator-initiated research models, such as limited sample size and standardization [4]. Since its conception in 2005, TCGA has been used extensively by researchers producing many distinguished publications in high-impact journals such as *Cell* or *Nature* [32]. Examples include the identification of the glioma-CpG island methylator phenotype (G-CIMP), which correlated with improved survival in glioblastoma patients [33], and the discovery of novel breast cancer oncogenes and potentially druggable targets [34]. Overall, TCGA provides data on 33 different tumour types from over 11,000 cases across Canada and the United States [32]. The different data types include RNA-Seq for transcriptome profiling, microRNA sequencing, DNA sequencing for mutation profiling, genotyping arrays, and methylation arrays. TCGA and other similar large-scale initiatives continue to be an invaluable source of knowledge for cancer genetics. However, translation of the enormous amounts of data to clinical therapeutics and diagnostics still requires proper bioinformatic and statistical analyses. In this thesis, we intend to properly account for covariates in differential expression analysis of high-throughput breast cancer data from TCGA with focus on HER2, ER, and PR biology – the three key receptors in breast cancer.

2.1 Breast invasive carcinoma

Breast cancer is one of the most prevalent cancers worldwide. In the U.S., it is reported that there is a 1 in 8 chance that a woman would develop breast cancer in her lifetime [35]. Currently, breast cancer is the second leading cause of cancer death in women, following only lung cancer [35]. Most breast cancers begin in the either the lobules, the milk-producing glands, or the milk ducts, which carry milk from the lobules to the nipple. Breast cancer can also be either non-invasive (cancer remains in the lobules or ducts), or invasive (cancer grows into normal healthy tissues) [36]. Invasive ductal carcinoma (IDC) is the most common form of breast cancer, which accounts for about 80% of all breast cancer cases [37]. In this type of cancer, tumours have broken through the milk duct walls and have begun to invade surrounding breast tissue. In this thesis, we focus on data from invasive ductal carcinoma cases.

There are three crucial biomarkers that are routinely examined in clinic for invasive breast cancer: 1) HER2, the human epidermal growth factor receptor 2, encoded by the gene ERBB2; 2) ER, the estrogen receptor α isoform, encoded by the gene ESR1; and 3) PR, the progesterone receptor, encoded by the gene PGR [37]. These three biomarkers are prognostic of cancer progression and predictive of treatment efficacy. The statuses of these receptors are determined by immunohistochemistry (IHC) of primary tumours, with positive status indicating elevated protein expression.

HER2-positive breast cancer occurs for 1 in 5 women with breast cancer [37]. In these cases, the HER2 protein is overexpressed on the surface of cancer cells which leads to aggressive growth and spread. HercepTest^{TM} (Agilent) is one of several tests for HER2 status used in-clinic and is approved by the U.S. FDA [2]. With HER2-positive breast cancers, targeted therapies are more

effective when compared to general chemotherapy [37]. HER2+ cancer confers drug sensitivity to trastuzumab, a monoclonal antibody against HER2 [38], and lapatinib, a kinase inhibitor.

With respect to the hormone receptors, 2 out of 3 breast cancer cases test positive for ER or PR status [37]. For over three decades, ER has been the most clinically important biomarker evaluated for breast cancer, as ER-positive tumours respond significantly better to hormone therapy than ER-negative tumours [39]. Tamoxifen, claimed as one of the most important advances in oncology [39], is a selective ER modulator (SERM) which blocks the estrogen receptor on cell membranes. For ER+ breast cancer, treatment with tamoxifen results in substantial improvement in survival [40] [41]. Aromatase inhibitors are a more recent form of hormone therapy which block conversion of precursors to estrogenic molecules, particularly estradiol and estrone [42]. Both tamoxifen and aromatase inhibitors are effective and are routinely used in clinic for ER+ breast cancers. However, hormone therapy is largely ineffective for women with ER- breast cancer [40]. As well, hormone therapy has significant side effects such as increased risk of thromboembolic events and uterine cancers [39].

Although PR status is routinely measured alongside ER, the exact role of PR in breast cancer remains rather unclear [39]. The PR gene, PGR, is controlled by estrogen, thus PR expression might be indicative of the estrogen response pathway [43]. As such, ER expression is often correlated with PR expression; we see this in the TCGA-BRCA cohort as well (see section 4.1.3 Figure 3). Several studies presented the importance of PR as a prognostic biomarker, as ER+/PR- patients have worse prognosis overall compared to ER+/PR+ patients [44] [45]. Some studies saw increased sensitivity to tamoxifen in ER+/PR+ patients however subsequent studies were unable to replicate the results [39] [41] [46]. Thus, the value of PR as a predictive biomarker remains controversial.

Further understanding of HER2, ER, and PR in cancer biology is important in order to fully elucidate the underlying biochemical pathways and other possible drug targets. In this thesis, we emphasize the importance of covariates in differential expression analysis, in the context of HER2, ER, and PR status in breast cancer as an example. Genes differentially expressed between receptor positive and receptor negative conditions could act as possible drug targets. Furthermore, the genes affected could have implications for understanding the mechanisms of side effects of therapies targeting these receptors. Better discovery of biomarkers by incorporating covariates would lead to more clinically valid biomarkers for prognostics or therapy. The validity of covariate-sensitive differential expression analysis would not be limited to the context of breast cancer exclusively.

2.2. Differential gene expression

Differential gene expression refers to differences in mRNA transcript abundance of each gene between two conditions. In the context of breast cancer, this can be a difference in expression of an oncogene between tumour and non-tumour cells or differences in gene expression between ER+ and ER- tumours, for example. Differential expression has traditionally been accomplished using hybridization-based microarray technology, but since the advent of next-generation sequencing (NGS), massively parallel RNA sequencing (RNA-Seq) has largely replaced microarrays for transcriptomics, as it offers higher dynamic range and higher sensitivity [47]. The high-throughput nature of NGS technologies, including RNA-Seq, continues to be the driving force behind the large-scale studies and collaborations, and as well, enables researchers to pursue big data analytics. In this section, we will briefly discuss RNA-Seq protocols and tools for differential expression analysis.

2.2.1 RNA-Seq protocols and pipelines

The RNA-Seq workflow first requires defining biological conditions of interest and appropriate library design for sequencing. Factors to consider for sequencing include read length, library type (single-end vs. paired-end), depth of coverage, and number of technical replicates [48]. Library preparation begins with extraction and isolation of total RNA from the samples. For example, this could be RNA extraction and isolation from breast primary tumours and non-tumour cells. From total RNA, the low quantity of mRNA is enriched using poly(A) selection of mRNA or by depletion of the abundant ribosomal RNA (rRNA) [48]. Fragments of appropriate size (according to selected read length) are generated by fragmentation of the enriched RNA followed by size selection. Complementary DNA (cDNA) is then generated from the RNA template by reverse transcription. The first strand cDNA is made double stranded by DNA polymerase. The last steps before sequencing include end repair and adaptor ligation [48].

The double stranded cDNA can now be sequenced using high-throughput short-read NGS, which includes methods such as sequencing by synthesis (Illumina), pyrosequencing (Roche-454), and sequencing by ligation (SOLiD sequencing) [49]. RNA-Seq data from TCGA-BRCA is generated using Illumina platforms. In short, sequencing by synthesis is a cyclic DNA elongation process in which one fluorophore labelled deoxynucleotide (dNTP) with a reversible terminator is added to the strand each cycle. Following the addition of the four different dNTPs (one colour for each nucleobase), the fluorescent label is read and the reversible terminators are removed, completing one cycle [49]. The Illumina HiSeq 2500 system, used in TCGA for example, can produce 300 million to 4 billion reads per run with a max read length of 250 base pairs.

After sequencing, the reads must be either aligned to a reference genome or assembled *de novo*. Bowtie [50] and TopHat [51] are two well cited bioinformatic tools for genome-guided alignment.

In simple terms, the sequences of the reads are matched to that of the reference genome and an optimal alignment is determined using dynamic programming. In the TCGA RNA-Seq pipeline [52], reads are mapped to the GRCh38 reference genome using STAR – another popular software package for RNA-Seq alignment [53]. From the aligned reads, gene expression can be quantified using the software package HTSeq [54] which outputs read counts for each gene. The raw read counts can be further transformed by normalizing for library size, turning into reads per million mapped reads, and normalizing by gene length, resulting in fragments per kilobase of transcript per million mapped reads (FPKM). From here, genes can be evaluated for differential expression using software tools such as DESeq2 [29] and edgeR [30].

2.2.2 Differential expression analysis

In this thesis, we evaluate differential gene expression using DESeq2, edgeR, and a non-parametric Mann-Whitney-Wilcoxon (MWW) test. DESeq2 and edgeR are two highly cited and commonly used packages in R (an environment for statistical computing) designed for differential expression analysis of RNA-Seq data. The MWW test is a statistical test which compares differences in expression of a gene by ranking samples, then tests whether there is a difference in ranking between the two conditions of interest. The MWW test is not specifically designed for RNA-Seq data but should be evaluated as a non-parametric alternative. In the latter parts of the thesis, we mainly focus on using DESeq2 for differential expression analysis. In this section, we will discuss the features and statistical frameworks of each method for differential expression analysis. Methods for incorporating covariates will be addressed in section 3.2, as we propose a novel method for covariate analysis.

Both DESeq2 and edgeR are designed for the analysis of count data. Testing for differential expression is dependant on the log2 fold change (LFC) in expression strength of each gene, and the dispersion (representing variance). The inner workings of DESeq2 and edgeR are quite similar and benchmark tests have suggested there is no direct advantage of one over the other [55]. In both packages, read counts in count matrix K_{ij} (row for gene i, and column for sample j) are modeled following a negative binomial distribution with mean μ_{ij} and dispersion α_i.

$$K_{ij} \sim \mathrm{NB}(\mathrm{mean} = \mu_{ij}, \mathrm{dispersion} = \alpha_i)$$

$$\mathrm{Var}\, K_{ij} = \mu_{ij} + \alpha_i \mu_{ij}^2$$

Counts are normalized according to sequencing depth using the median-of-ratios method in DESeq2 [56], and the trimmed means of M-values (TMM) method in edgeR [57]. For the normalized counts of each gene q_{ij}, a generalized linear model (GLM) with a logarithmic link is fit with design matrix elements x_{jr} with coefficients β_{ir} and a constant offset β_0 representing the average log expression.

$$\log_2 q_{ij} = \sum_r x_{jr}\beta_{ir} + \beta_0$$

The design matrix represents features of interest for the analysis. For example, in the simplest case of a direct comparison of HER2+ vs. HER2- for differential expression, the design matrix would consist of only one element which is HER2 status. The GLM fit, determined by maximum likelihood estimation, returns coefficients β_{ir} for each design matrix element x_{jr}. In the HER2 example, the coefficient of HER2 status would represent the overall expression strength of each gene and the LFC between HER2 groups. Essentially, these coefficients are an indication of the

differential expression of each gene, however, further corrections must be made before hypothesis testing and the generation of p-values.

In both DESeq2 and edgeR, corrections are made to the dispersion estimates. First, gene-wise dispersion estimates (based on only data from each gene) are made using maximum likelihood. Empirical Bayes is used to shrink gene-wise dispersions towards a consensus value (prior distribution). The strength of shrinkage depends on 1) estimates of how close true dispersion values match the fit and 2) degrees of freedom (larger sample size results in weaker shrinkage). DESeq2 differs from edgeR here, as the prior distribution is estimated from the data in DESeq2 whereas edgeR requires a user-adjusted parameter [29].

DESeq2, by default, also corrects for fold change using Empirical Bayes estimation. It is observed that in high-throughput sequencing count data, noise is higher for genes with low counts [29]. Thus, there is strong variance of LFC estimates for genes with low expression. The empirical Bayes procedure fits a zero-centered normal distribution to observed maximum likelihood estimates of LFC for all genes. Using this approach, LFC of genes with low information (low count, high dispersion, or few degrees of freedom) are more strongly shrunk towards zero. Essentially, the LFC of genes with weakest expression are corrected, with strong bias towards zero. This procedure is not performed by default in edgeR but can be parameterized by the user using one of the edgeR functions.

For hypothesis testing for differential expression, DESeq2 employs a Wald test whereas edgeR uses an exact test, analogous to Fisher's exact test, adapted for negative binomial data. In the Wald test, a shrunken estimate of the LFC is divided by its standard error producing a z-statistic which is then compared to the normal distribution. P-values are further adjusted for multiple testing by the Benjamini and Hochberg procedure. In the edgeR exact test, the adjustment is that

hypergeometric probabilities in Fisher's exact test are replaced with negative binomial. The full mathematical details can be found in Material and methods sections of the 2014 DESeq2 paper [29] and 2010 edgeR paper [30].

In contrast to the Wald test and the exact test, which are both parametric in nature, the Mann-Whitney-Wilcoxon test is a non-parametric test for ranks. It tests the null hypothesis that the difference in the distribution of ranks is zero. In this case, all samples are ranked for each gene based on expression strength, and the distributions of the ranking of samples between the two conditions are compared. MWW is evaluated by the *U* statistic, whose distribution can be approximated by the normal distribution for sample sizes above 20 [58]. The MWW test makes several assumptions: 1) observations from both groups are independent, 2) responses are ordinal, and 3) the shape of the distribution of both groups must be similar. Although not specifically designed for RNA-Seq count data, MWW can be a valid non-parametric alternative and should be evaluated.

3. METHODS

In this thesis, our work can be chronologically split into two phases – the exploratory phase and the hypothesis testing phase.

First, in the exploratory phase, we tested the ability of DESeq2 and edgeR for incorporating covariates by generalized linear models (GLMs) using simulated gene expression data. Next, using real gene expression data from TCGA-BRCA, we compared the methods for differential expression analysis, namely DESeq2, edgeR, and MWW. At this stage, we focused on differential expression based on the HER2 status of primary tumours (HER2+ vs. HER2-). We then proceed to evaluate covariate-dependant differential expression analysis of TCGA-BRCA data using GLMs in DESeq2. We considered covariates including age, race, ER status and PR status.

In the hypothesis testing phase, we primarily focused on using DESeq2 for analysis as the default functions in DESeq2 made it advantageous over edgeR. We hypothesized that 1) covariate models have increased robustness and consistency across patient sets, 2) covariate models remove covariate-associated genes and reveal covariate-obscured genes, and 3) use of correlated over uncorrelated covariates has stronger effects on differential expression. We first examined HER2 focused differential expression with and without covariates (age, race, ER, PR), followed by cross-validation with a mutually exclusive similarly sized dataset. For improved validation metrics, we performed 20-fold non-mutually exclusive cross-validation of HER2 focused analysis (with and without ER/PR as covariates). We compared ER focused differential expression analysis (with and without HER2/PR as covariates) using the same approach.

Lastly, we evaluated ER focused differential expression in depth. Using the 20-set, we compared the non-covariate model, the GLM covariate model, and our proposed categorical covariate model.

We looked at pairwise comparisons of resulting gene lists of each method and the respective validation consistencies. Resulting genes of interest were further examined by expression boxplots, Gene Ontology enrichment analysis [59], and KM-plotter [60].

3.1 Datasets for analyses

Gene expression data was simulated for 60 samples of 20,000 genes based on a negative-binomial model. Expression levels of genes were then influenced by 3 simulated binary discrete covariates and 3 simulated continuous covariates. Our first dataset strictly followed a negative binomial distribution with covariate influence and the second with added noise. The negative binomial distribution was generated in R using the `rnbinom` command with a base mean μ of 4 and a dispersion α of 0.1. For covariate influence, 2000 genes for each covariate were randomly chosen to have altered mean μ. For binary discrete covariate influence, the mean was squared. For continuous covariates, the mean was exponentiated to the continuous variable – which ranged from 1 to 3. Noise was generated using the `rnbinom` command with a base mean μ of 2 and a dispersion α of 0.05 then added to the starting data.

For real gene expression data, we downloaded RNA-seq data processed by Illumina HiSeq 2500 in the form of raw HTSeq counts from TCGA-BRCA using the TCGAbiolinks package in R [61]. As well, we downloaded the clinical data associated with each case. In total, there were 1,092 cases of breast invasive carcinoma primary tumours with RNA-Seq data. We focused on the 814 cases of breast invasive ductal carcinoma from female patients.

For the comparison of methods for HER2 focused analysis in the exploratory phase, we randomly selected n = 154 cases (77 HER2+, 77 HER-). We did not sample from cases with equivocal or indeterminant HER2 status. For covariate-sensitive analysis, the same 154 cases were evaluated.

Race was considered as a categorical covariate with 3 levels (White, Black or African American, and Asian). Age was treated as a categorical covariate with 5 bins (27-45, 46-52, 53-59, 60-68, 69-90) with ranges chosen for equal sample sizes in each bin. ER and PR status were either positive or negative.

In the hypothesis testing phase, for HER2 cross-validation, a mutually exclusive dataset of n = 156 cases (78 HER2+, 78 HER-) was randomly sampled. Covariate criteria was the same as the previous 154 cases.

For the generation of the 20-set, 140 cases were randomly selected without replacement for each set from the total 814 cases. After 20 resamplings for 140 cases each, a total of 786 distinct cases were selected which almost entirely covered the total 814 cases. The three factors considered here were HER2, ER, and PR status. Only positive or negative statuses (not equivocal or indeterminant) were considered for selection. No restrictions were placed on the number of samples from each category; the distribution of HER2, ER, and PR status in the each of the 20 set mirrors that of the entire cohort.

3.2 Frameworks for covariates

In DESeq2 and edgeR, covariates can be incorporated using GLMs. GLM functionality is present in both packages, however, GLMs are often only used to account for batch effects or technical artifacts of RNA-Seq and not clinical covariates as confounding factors [62]. As discussed earlier, a GLM with a logarithmic link can be fit with design matrix elements x_{jr} with coefficients β_{ir} and a constant offset β_0 representing the average log expression. The coefficients from the GLM fit represents the overall expression strength of each gene and the LFC.

$$\log_2 q_{ij} = \sum_r x_{jr}\beta_{ir} + \beta_0$$

Clinical covariates as well as batch effects can be included as elements in the design matrix. In GLMs with multiple elements, the assumption of linearity must be made as the elements are combined in a linear fashion. As a simple example, a possible design matrix could include just one binary variable, the HER2 status. In this model, for each gene, the constant coefficient β_0 would capture the average log expression, while β_1 would capture the relationship between HER2 status and the gene's expression, if any. If β_1 is statistically significantly different from zero, it means that the gene is differentially expressed in HER2+ vs. HER2- samples. An example with covariates could have HER2 and ER in the design matrix. In this covariate model, β_0 would capture the average log expression, β_1 would represent the variability explained by ER, and then β_2 would capture the remaining variability attributing it to HER2. In this model, the total variability beyond the average log expression is explained by HER2 and ER. The ER effect is removed if it contributes linearly, otherwise the log-linear model performs suboptimally for non-linear relations.

We propose a novel categorical method to account for covariates which directly removes the covariate effect. In this method, the dataset is partitioned based on the covariates, then differential expression analysis is performed within each sub-category. As per the previous example, with ER status as the covariate, the dataset can be divided into ER+ and ER- group. HER2 focused analysis can then be performed within only the ER+ group and only the ER- group. The resulting gene lists from each sub-category can be combined by Fisher's method [63] or Stouffer's method [64] of combining p-values. Fisher's method combines p-values, p_i, into one test statistic X^2 by the following formula:

$$X_{2k}^2 \sim -2\sum_{i=1}^{k} \ln(p_i)$$

The test statistic X^2 follows a chi-squared distribution with $2k$ degrees of freedom, where k is the number of tests being combined. This test statistic X^2 can be evaluated to obtain a p-value for the global hypothesis.

Stouffer's method (also called the weighted Z-method) is based on Z-scores weighted on the sample size of the combined categories. Individual Z-scores are represented by $Z_i = \Phi^{-1}(1 - p_i)$, where Φ is the standard normal cumulative distribution function. The combined Z-score can be evaluated by:

$$Z \sim \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$

where weight w_i can be the square root of the sample sizes of each category [65]. The combined Z-score follows a standard normal distribution under the null hypothesis which can be evaluated to obtain an overall p-value.

The two major advantages of our proposed categorical method are that 1) the effect of the covariate is directly removed, and 2) the assumption of linearity is not required as there is no combination of features (i.e. the design matrix would always include only the factor of interest). One limitation of this method is that the sample size of sub-categories become increasingly small with more covariates, thus analysis may require large datasets or may be limited to few covariates. We believe this may be another suitable avenue to tackle covariate effects. In section 4.2.3, we evaluate performance of the GLM method and our categorical method for ER focused differential expression with HER2 and PR as covariates.

4. RESULTS AND DISCUSSION

4.1 Exploratory phase

The exploratory phase consists of our first steps in the project of this thesis. We present here: 1) testing of covariate differential expression analysis using edgeR and DESeq2 on simulated gene expression data, 2) comparison of DESeq2, edgeR, and MWW for non-covariate analysis focused on HER2 status for real TCGA-BRCA data, and 3) covariate vs. non-covariate analysis of HER2 status using DESeq2 for real data. Findings in our exploratory phase led to our hypotheses and directed our efforts in the hypothesis-testing phase.

4.1.1 Testing using simulated gene expression data

We first generated simulated gene expression data without noise and with noise to test the capacity of edgeR and DESeq2 to account for covariate influence. Gene expression data was simulated for 60 samples for 20,000 genes. Three binary discrete covariates and three continuous covariates were designed to each have influence on expression of 2,000 genes. The genes which are influenced by each covariate are chosen independently and at random, with the result that most genes are unaffected by any covariate, some by one, and a few by more than one covariate. Our main criterion here is how well the algorithms identify which genes are affected by which covariate.

Figure 1 shows the receiver operating characteristic (ROC) curves of edgeR and DESeq2 in cases without noise and with noise. Three discrete and three continuous covariates were evaluated in each case. Performance of edgeR and DESeq2 was quite similar with respect to area under curve (AUC) values, as shown in the bottom right corners of each plot. DESeq2 performed better for continuous covariates than discrete covariates in data without noise. As data with noise is more

representative of real count data, we placed more emphasis on those results. For edgeR and DESeq2, AUC values were extremely similar at 0.88 and 0.87 respectively with performance better for discrete than continuous covariates. From the results of simulated testing, we decided to proceed to real gene expression data with both edgeR and DESeq2.

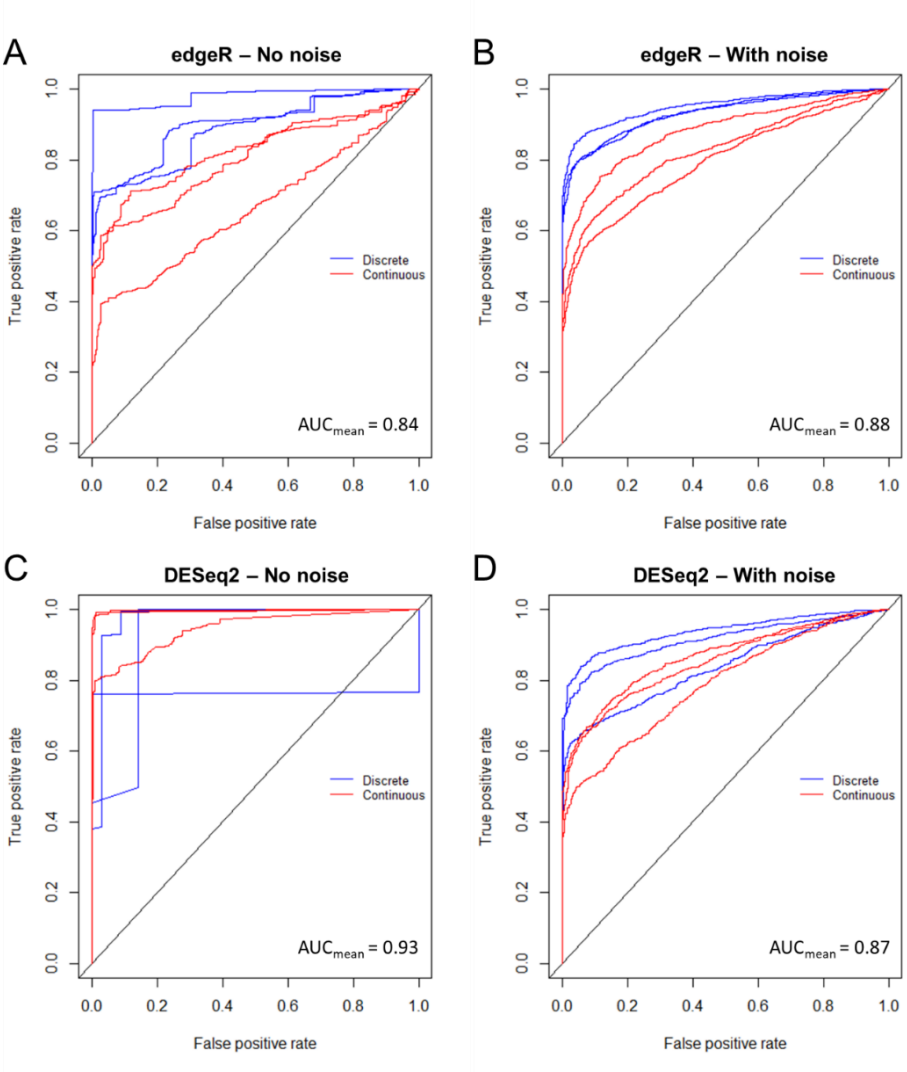


Figure 1. ROC curves of results of differential expression analysis using simulated gene expression data evaluated for edgeR and DESeq2. Gene expression count data for 20000 genes and 60 samples was simulated following a negative binomial distribution without noise and with noise (to better mirror real data). 3 binary discrete covariates and 3 continuous covariates each influenced expression of 2000 genes. Differential expression tested for genes influenced by discrete (blue) and continuous (red) covariates with each line representing one covariate influence. Mean area under the curve values are shown in the bottom right. (A) Analysis of noiseless data by edgeR, (B) analysis of noisy data by edgeR, (C) analysis of noiseless data by DESeq2, (D) analysis of noisy data by DESeq2.

4.1.2 Evaluation of methods for analysis of real data

We evaluated differential expression analysis without covariates by edgeR, DESeq2, and the MWW test of HER2+ vs. HER2- primary tumours in 154 cases of breast invasive ductal carcinoma. Raw HTSeq counts were taken as input by each method and default parameters were used for edgeR and DESeq2. In this case, we did not know which genes are truly differentially expressed between the two conditions. Of interest, however, was how well different approaches agree. Also, this analysis set a point of comparison for the covariate-dependent analyses that follow.

The pairwise comparisons of resulting gene lists between the three methods are shown in **Figure 2**. The Venn diagram in **Figure 2A** presents genes that are considered significant, at a threshold of p-value < 0.001, by each method and across methods. There was a sizeable overlap of 915 genes between edgeR and DESeq2. MWW agreed with DESeq2 more than edgeR. However, due to the very large number of genes that edgeR considers significant, comparison by simply overlap in significant genes may be limited.

In **Figure 2B, C, D**, the ranking of each gene (from 1 to 57,251) by adjusted p-values of the x-axis was plotted against that of the y-axis. In **Figure 2B**, edgeR results agreed with MWW results but not strongly as shown by the faint diagonal. DESeq2 results were most consistent with MWW results as shown by high density of agreement in the bottom left corner (top gene ranks) in **Figure 2C**. edgeR results partially agreed with DESeq2 results, demonstrated by the curved trend in **Figure 2D**. They agreed heavily on low ranking, non-significantly differentially expressed genes, represented by the high-density cluster on the right.

The gene of interest ERBB2 (HER2) ranked first in DESeq2 and MWW while it ranked 600th in edgeR. Many of the genes ranked above ERBB2, in edgeR, were genes with extremely low or zero read counts. As, DESeq2 strongly corrects the dispersion and LFC for low/zero count genes by default, DESeq2 is not as strongly affected by this issue. MWW, being a non-parametric test based on ranks, also seemingly avoids this issue. edgeR could be further parameterized by the user for similar functionality as DESeq2. However, based on default functionality, we decided to proceed with further analysis of real gene expression data using DESeq2. And although MWW seemed to perform reasonably well on this data, it does not have a direct mechanism for accounting for covariates, thus it was primarily of interest as a comparison to edgeR and DESeq2.

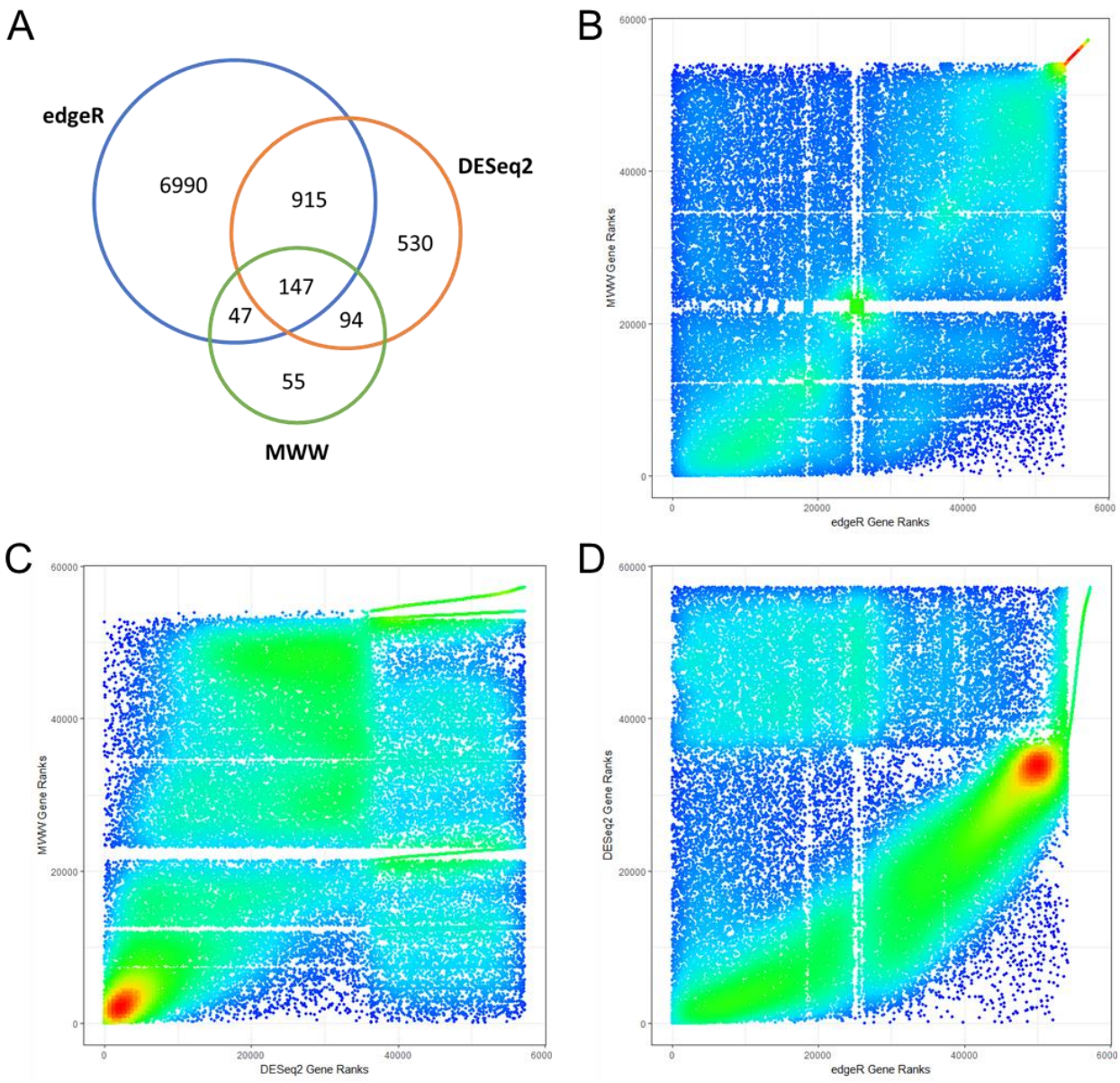


Figure 2. Comparison of gene lists between MWW, edgeR, and DESeq2. Gene lists were generated by differential expression analysis by each method between HER2+ and HER2- samples. (A) Venn diagram comparing significant genes at a threshold of p-value < 0.001. (B, C, D) Gene lists from differential expression analysis were ranked by adjusted p-values. Genes were then plotted based on their ranking (from 1 to 57,251) in each pairwise method comparison. (B) comparison between edgeR and MWW, (C) comparison between DESeq2 and MWW, (D) comparison between edgeR and DESeq2. Differential expression was focused on HER2 status in n=154 patients from TCGA-BRCA.

4.1.3 Covariate-dependent differential expression using generalized linear models

We evaluated covariate-sensitive differential expression analysis focused on HER2 using GLMs in DESeq2. Using the same 154 cases as before, we considered age, race, ER status, and PR status as covariates.

As the receptor statuses (HER2, ER, and PR) are now being considered, we wanted to further explore their correlations in the TCGA-BRCA cohort. In clinic, ER and PR status are often considered in conjunction. As the PR gene (PGR) is controlled by estrogen, ER expression is often correlated with PR expression. HER2 expression and status has not shown to be correlated to either ER nor PR expression and status.

These expectations were verified in the TCGA-BRCA cohort, as shown in **Figure 3**. The receptor statuses determined by IHC of 786 breast invasive ductal carcinoma primary tumours are shown in **Figure 3A**. 84% of ER+ cases are PR+ and 93% of ER- cases are PR-, thus ER and PR status have strong positive correlation. HER2 status does not seem to have correlation to PR nor ER status. Fisher's exact test confirms this correlation between PR status and ER status ($p < 0.0001$) but not HER2 ($p = 0.48221$). Gene expression levels of HER2 (ERBB2), PR (PGR), ER (ESR1) is shown in **Figure 3B**, represented in log2(reads per million). There is a clear correlation in expression of PGR and ESR1. **Figure 3C** compares the IHC receptor statuses with the expression levels for each receptor gene. High gene expression of HER2, PR, or ER tends to be correctly identified as receptor-positive (green), however there are some identified as receptor-negative (orange) and vice-versa. This presents a possible discordance between receptor identification by IHC and gene expression by RNA-Seq, as gene expression levels do not directly represent receptor statuses – possibly disrupted by post-translational regulation and protein turnover.

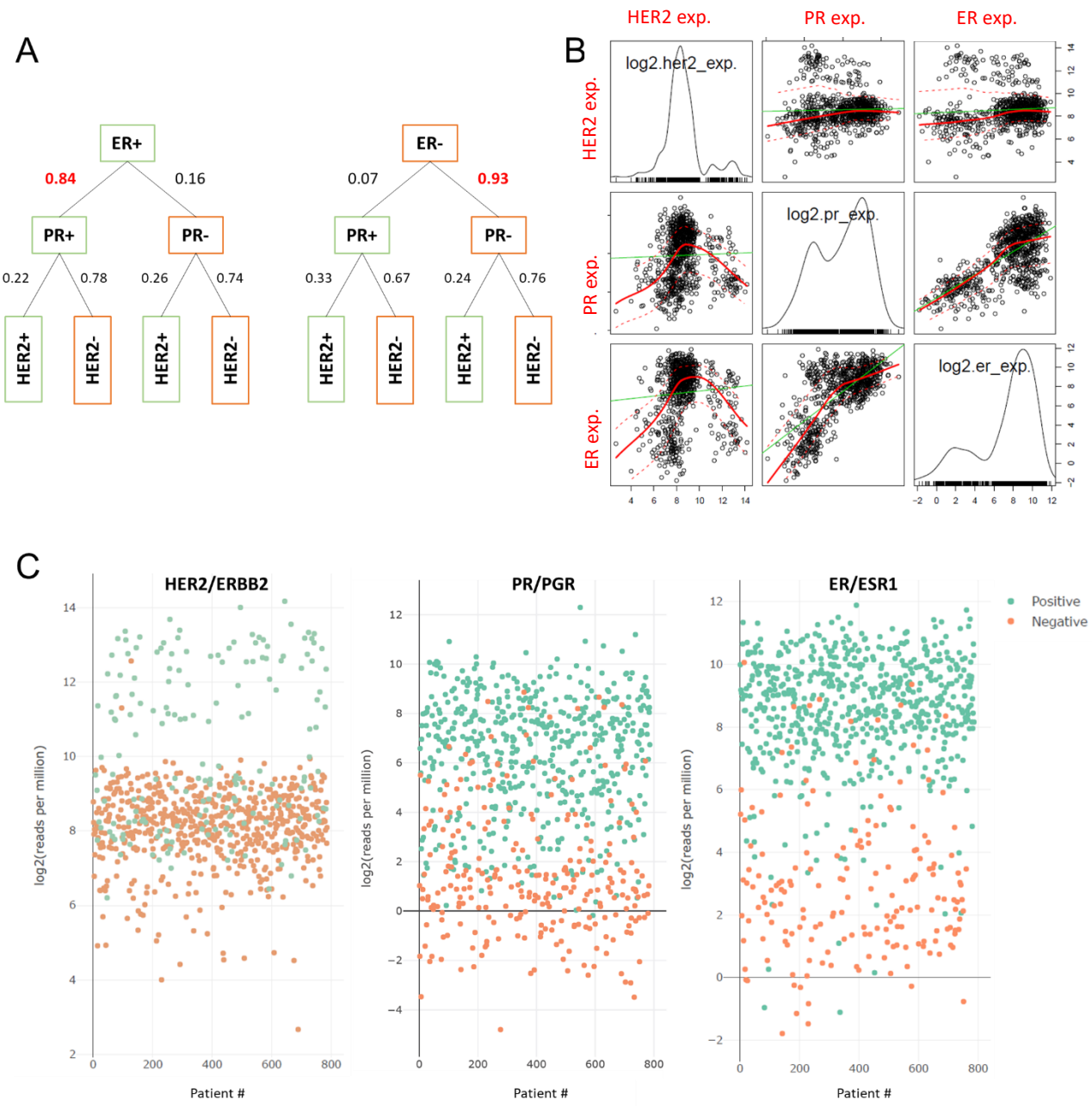


Figure 3. Examination of patient IHC receptor statuses and gene expressions of HER2 (ERBB2), PR (PGR), and ER (ESR1). Receptor statuses of HER2, PR, and ER were evaluated by IHC in clinical data obtained from TCGA-BRCA. Gene expression data from TCGA-BRCA represented in log2(reads per million). Patients here represent the n = 786 patients in the later 20-set analysis. (A) Receptor statuses of patients for HER2, PR, ER. Branches represent proportion of patients belonging to next category from the previous. (B) Gene expression levels of patients represented in log2(reads per million) cross-compared for ERBB2, PGR, ESR1. (C) Gene expression levels compared to IHC receptor statuses; positive status in green, negative status in orange.

The comparison of gene rankings between analysis with no covariates (NCV) and with all covariates (ACV; age, race, ER status, PR status) can be visualized in **Figure 4A**. There was agreement in gene rankings between NCV and ACV with slight variability as demonstrated by the width of the diagonal. For top genes between NCV and ACV, there was an overlap of 85% in the top 20 genes, and a 52% overlap in the top 100 genes.

At a significance threshold of p-adj ≤ 0.05, the number of significant genes is shown by the Venn diagram in **Figure 4B**. 61% of the significant genes in NCV analysis did not appear significant in ACV analysis; these genes could possibly be false positives. 27% of genes significant in ACV analysis did not originally appear in NCV analysis, possibly obscured by the covariates. These differences were exciting to us as it demonstrated, as a preliminary result, that covariate-sensitive analysis 1) changed the resulting list of differentially expressed genes and 2) could lead to downstream benefits for biomarker discovery as it possibly removes false positives and reveals true positives.

We looked to expression boxplots of these genes for validation. Some examples are shown in **Figure 4C** where EN1 fell out of significance and SOX11 appeared significant with the inclusion of covariates. ERBB2, our gene of interest, appeared significant with or without covariates, as expected. In the boxplots, we present ER as a covariate for illustration of the covariate effect. In the case of EN1, expression was higher in HER2- samples compared to HER2+. However, the higher expression was largely attributed to HER2-, ER- samples, which skewed the expression of the overall HER2- group. Accounting for the effect of ER, this large difference in the expression of EN1 between HER2- and HER2+ samples was reduced, thus EN1 fell out of significance with the inclusion of covariates.

In the case of SOX11, the differences between HER2- and HER2+ groups overall were minimal thus in analysis without covariates, SOX11 was not considered significant. However, when we considered the covariate effect of ER, HER2 was shown to have a slight effect on SOX11 expression as HER2-/ER- and HER2-/ER+ were lower than their respective HER2+ counterparts (green vs. green and purple vs. purple boxplots). ER, in this case, proved to have strong effects on SOX11 expression, as expression in ER- samples was higher compared to ER+ samples. Due to the strong effect of ER, the HER2 effect may have been obscured. After accounting for the effect of the covariates, the differences in SOX11 expression with respect to HER2 may have appeared to be more significant. In both cases of EN1 and SOX11, we show that there is a covariate effect that should be considered.

The results of this exploratory phase led us to two of our main hypotheses. From the differences between NCV and ACV gene lists, we hypothesized that covariate models remove covariate-associated genes and reveal covariate-obscured genes. From the correlations between ER and PR and not HER2, we hypothesized that use of correlated covariates over uncorrelated covariates would have stronger effects on differential expression. In the next section, we present our findings to address these hypotheses.

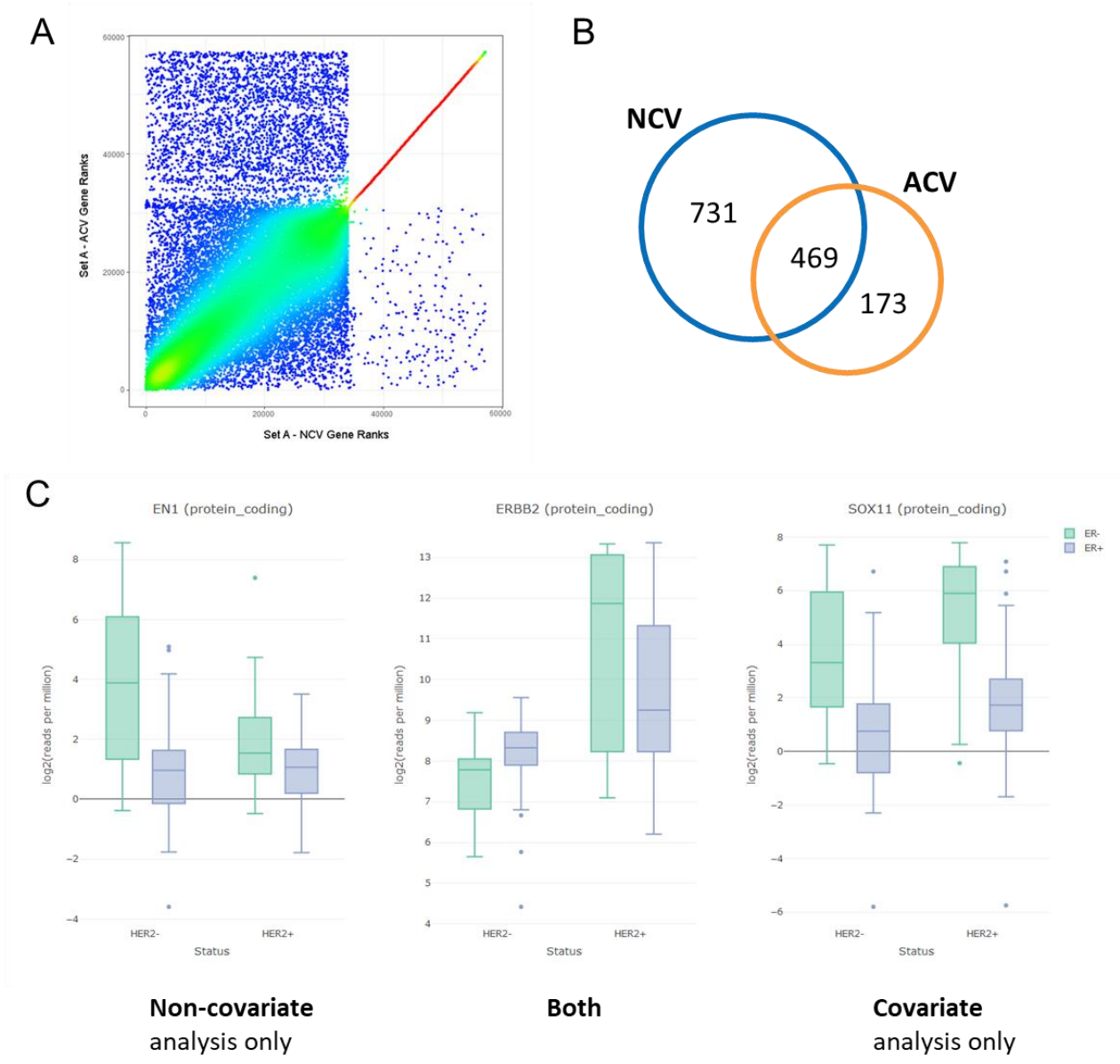


Figure 4. Results of covariate-based HER2 differential expression analysis using DESeq2. Differential expression was evaluated for 77 HER2+ vs. 77 HER2- patients with no covariates (NCV) and with all covariates (ACV; age, race, PR status, ER status). Covariates were included using GLMs in DESeq2. (A) ranking comparison of gene lists from NCV analysis vs. ACV analysis. (B) Venn diagram of significant genes at adjusted p-value ≤ 0.05 for NCV analysis and ACV analysis. (C) gene expression boxplots of genes found in only NCV analysis (EN1), or ACV analysis (SOX11), or both (ERBB2). ER as a covariate for illustration of covariate effect in boxplots. ER+ in purple, ER- in green.

4.2 Hypothesis testing phase

In the hypothesis testing phase, we addressed our three main hypotheses: 1) covariate models have increased robustness and consistency across patient sets, 2) covariate models remove covariate-associated genes and reveal covariate-obscured genes, and 3) use of correlated over uncorrelated covariates has stronger effects on differential expression.

We first present covariate and non-covariate HER2 focused differential expression with two-fold cross validation. For improved validation, we proceeded with 20-fold non-mutually exclusive cross-validation of HER2 analysis with ER/PR as covariates. We compared ER analysis with HER2/PR as covariates using the same approach. Lastly, we present an in-depth evaluation of ER focused differential expression. Using the 20-set, we show analysis with the non-covariate model, the GLM covariate model, and our proposed categorical covariate model.

4.2.1 Two-set HER2 focused differential expression analysis

We wanted to validate our observations in the previous section with a mutually exclusive patient set from TCGA-BRCA. We randomly sampled patient Set B with 78 HER2+ and 78 HER2- patients mutually exclusive from patient Set A (77 HER2+, 77 HER2-). Likewise, we considered covariates age, race, PR status, and ER status.

The gene rank comparisons of the results of non-covariate (NCV) and covariate analysis (ACV) for Set A and Set B are shown in **Figure 5**. As shown previously, the correlation of gene rankings between NCV and ACV was high within Set A. When comparing Set A analysis and Set B analysis, the overlap of resulting gene lists was extremely low in non-covariate analysis (**Figure**

5B) as well as covariate analysis (**Figure 5C**). The overlap between NCV and ACV analysis within Set B had considerable overlap (**Figure 5D**) similar to Set A.

We took a deeper look at the top-ranking genes in **Figure 6**. We evaluated overlap of gene lists between Set A and Set B for NCV and ACV analysis using the Jaccard index (gene list intersection divided by the union) for genes above each gene rank. For NCV analysis between Set A and Set B: 67% genes overlapped in the top 20, 22% in the top 100, 13% in the top 500. For ACV analysis: 74% in the top 20, 24% in the top 100, 11% in the top 500. The Jaccard index compared at each gene rank can be visualized in **Figure 6A** for the top 1000 genes. ACV analysis showed more overlap at earlier ranks and falls slightly below NCV at around rank 100. Both NCV and ACV validated equally well and ACV results could be considered just as valid, yet the two produced different gene lists.

The overlap of genes considered significant at p-adj ≤ 0.05 is shown in **Figure 6B**. Many genes considered significant in Set A did not validate in Set B. With respect to non-covariate and all-covariate analysis, neither validated well; NCV 9.7% slightly outperforms ACV 5.5%. Out of 3,524 genes from all four groups, only 145 (4.1%) appeared significant in all four. Although a low amount, most of these genes rank highly in all cases which may represent truly differentially expressed HER2 related genes. This leads us to believe that differential expression analysis is very dependent on the patients selected, as the patient population could be very heterogenous. We hypothesized that incorporation of covariates would lead to better validation but in this case, this did not appear to be true when comparing genes by significance. We speculate that the differences between patient Set A and Set B are substantial, thus trumping the differences due to covariates.

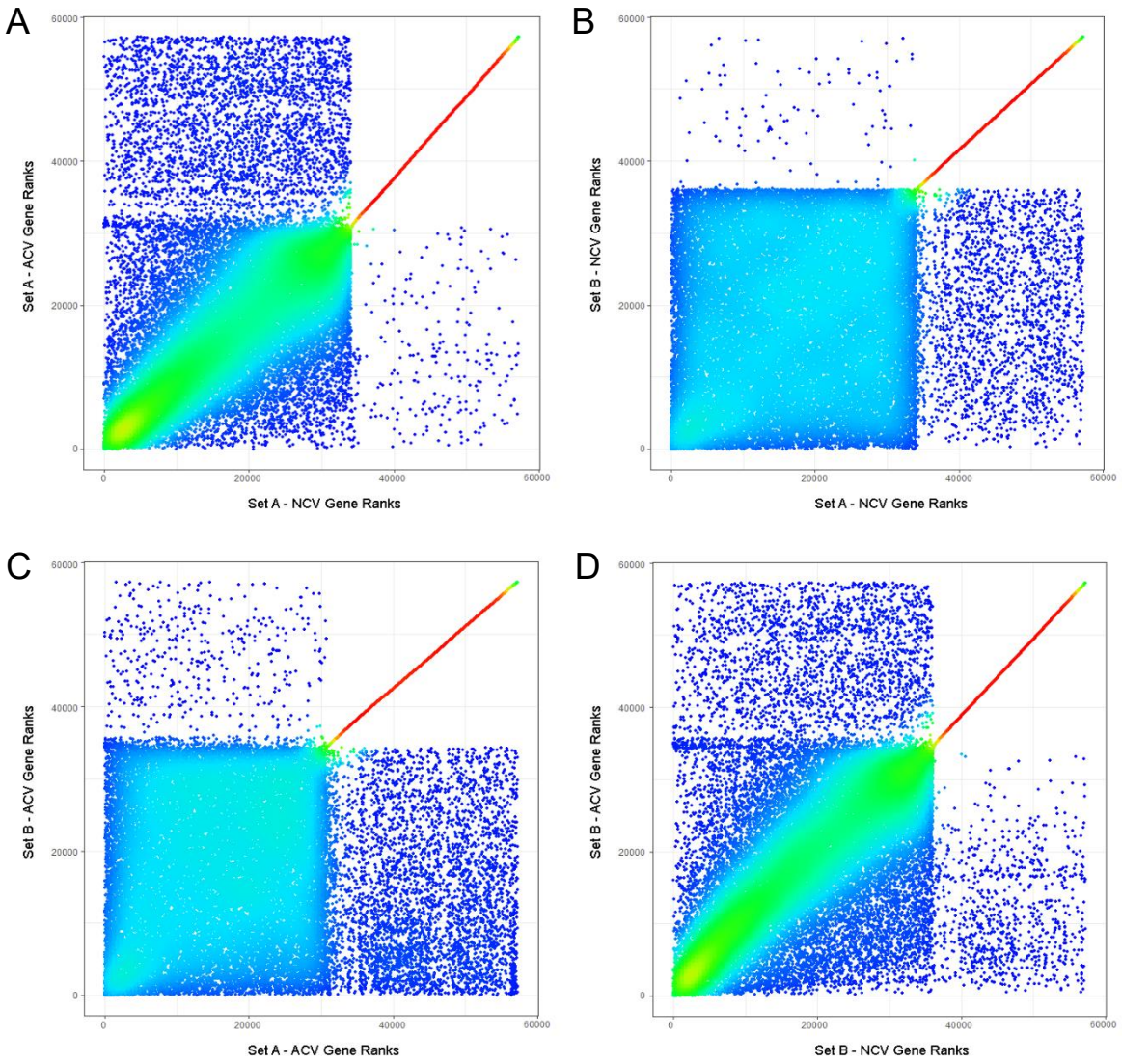
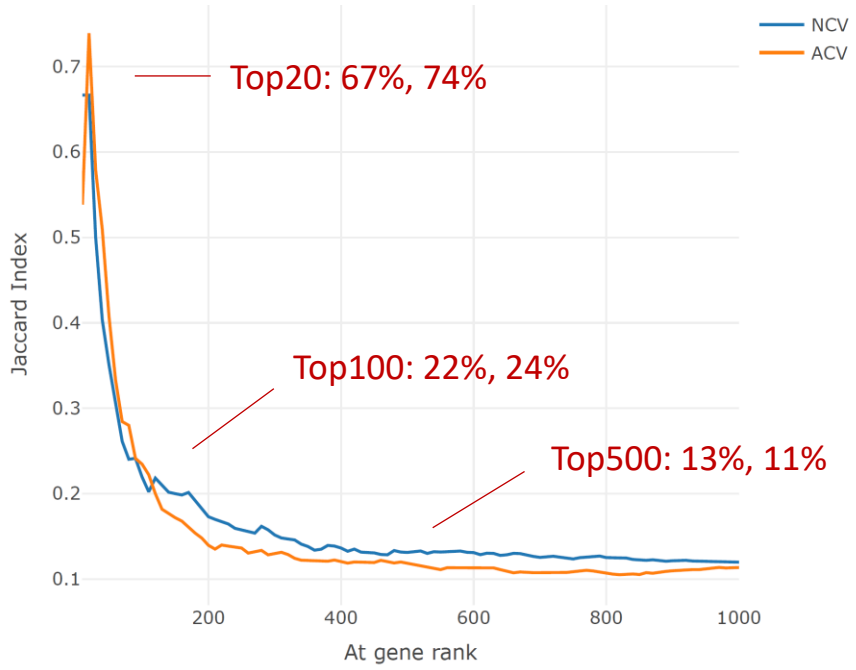


Figure 5. Comparison of full gene ranks generated by DESeq2 between patient Set A and Set B, with or without covariates. Gene lists from differential expression analysis were ranked by adjusted p-values. Genes were then plotted based on their ranking (from 1 to 57,251) in each pairwise method comparison. Patient Set A (n=154) and Set B (n=156) are mutually exclusive and randomly selected from TCGA-BRCA. Differential expression analysis with no covariates (NCV) was compared with analysis with all covariates (ACV; age, race, PR status, ER status). (A) comparison between Set A NCV and Set A ACV, (B) Set A NCV and Set B NCV, (C) Set A ACV and Set B ACV, (D) Set B NCV and Set B ACV.



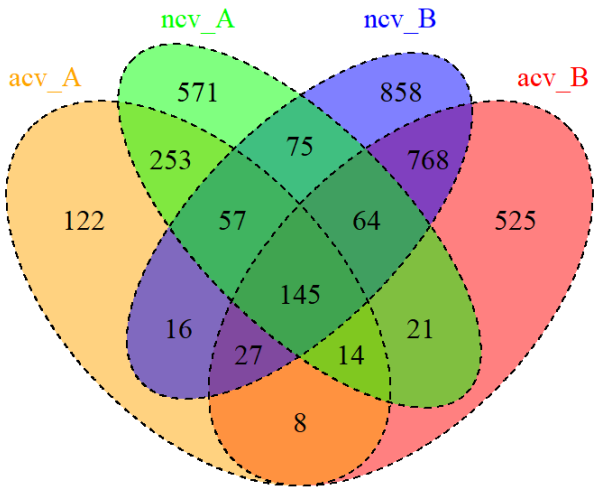


Figure 6. Comparison of gene lists between patient Set A and Set B, with or without covariates. Gene lists from differential expression analysis were ranked by adjusted p-values. Patient Set A (n=154) and Set B (n=156) are mutually exclusive. Differential expression analysis with no covariates (NCV) was compared with analysis with all covariates (ACV; age, race, PR status, ER status). (A) Jaccard index of overlap (gene list intersection divided by union) between Set A and Set B gene lists compared above each gene rank. Jaccard index presented for the top 20, 100, 500 genes for NCV and Acv respectively. (B) overlap of significant genes at p-adj ≤ 0.05 in Set A ACV, Set A NCV, Set B NCV, and Set B ACV respectively.

4.2.2 20-set HER2 focused vs. ER focused analysis

To strengthen our validation process, we proceeded with 20-fold non-mutually exclusive cross-validation. We randomly sampled 140 patients from TCGA-BRCA twenty times forming 20 patient sets, totalling 786 cases of primary tumours of breast invasive ductal carcinoma. As described previously, ER and PR status are correlated, while HER2 does not correlated with either. At this stage, we wished to evaluate the differences in influence of incorporating correlated covariates versus uncorrelated covariates in the GLM model. We began by testing differential expression with uncorrelated covariates in HER2+ vs. HER2- samples with ER and PR as covariates (HER2 does not correlate with either ER nor PR). We compared this with analysis of ER+ vs. ER- samples with PR and HER2 as covariates (ER strongly correlates with PR).

Using a significance threshold of p-adj ≤ 0.05, we looked at the resulting gene lists as shown in **Figure 7**. In **Figure 7A**, the mean number of significant genes across the 20-set of NCV (non-covariate) and GLM (covariate) analysis is shown for HER2 analysis and ER analysis. The value in parentheses represents the standard deviation in number of significant genes across the 20-set. In HER2 analysis, there was a sizable overlap between NCV and GLM of 803 genes on average. In ER analysis, DESeq2 called substantially more significant genes at p-adj ≤ 0.05. The overlap between NCV and GLM was 6191 genes on average. **Figure 7B** represents these counts in terms of proportion of genes. In ER analysis compared to HER2 analysis, a significantly larger proportion of genes belonged to the NCV only category (blue bar). Genes in this category represent those that fell out of significance with the incorporation of covariates. As covariates were not considered in NCV analysis, the confounding effects of covariates could influence the detection of differentially expressed genes. Thus, genes in the NCV only category could represent likely false positives, which are genes that are not truly differentially expressed for the condition of

interest. Presuming these NCV only genes could possibly represent some false positives, more genes in this category would imply more false positives. As ER analysis (with correlated covariates) results in more genes in the NCV only category than that of HER2 analysis (with uncorrelated covariates), this can be thought as one of the effects of the use of correlated covariates vs. uncorrelated covariates – the removal of more possible false positives.

As expected, a significantly smaller proportion of genes belonged to covariate analysis only (orange bar), when comparing HER2 analysis and ER analysis. A correlated covariate, compared to an uncorrelated covariate, would be more similar to the factor of interest in terms of influencing gene expression. Accounting for the effect of a correlated covariate, thus, would reveal fewer covariate-obscured genes, as more of the difference in expression can be explained by the main factor of interest. If we were to presume these covariate-only genes represented true positives, then differential expression analysis with a correlated covariate vs. an uncorrelated covariate is likely to reveal fewer true positives. However, a smaller list can imply that these genes are more truly differentially expressed for the factor of interest. As well, a smaller list of candidate genes is more manageable and can be more feasible for biological validation.

The consistency of significant genes is shown by the boxplots in **Figure 7C**. The Jaccard index was evaluated by pairwise comparisons of significant genes between each of the 20-set. This was performed for NCV and GLM methods for HER2 and ER analysis. In HER2 analysis, fewer genes were significant between datasets with NCV and GLM methods when compared to that of ER analysis. In ER non-covariate analysis, the Jaccard index approached 0.5 indicating that many genes were found significant consistently across datasets. However, since ER analysis called substantially more genes significant at p-adj ≤ 0.05 than HER2 analysis, this higher Jaccard index

may be due to the size of the ER gene list. In this case, comparison using the Jaccard index has its limitations.

In HER2 and ER analysis, covariate models led to a reduction in the number of genes called significant across datasets. From one perspective, this may be viewed as a decrease in validation consistency due to the inclusion of covariates. However, under the assumption that the inclusion of covariates removes possible false positives, the smaller number of consistently validated genes would be more likely to be true positives. This reduction in Jaccard index in covariate models can be seen as false positives being filtered out by accounting for covariates. The smaller list can imply a stronger list of candidate genes that are truly differentially expressed for the factor of interest.

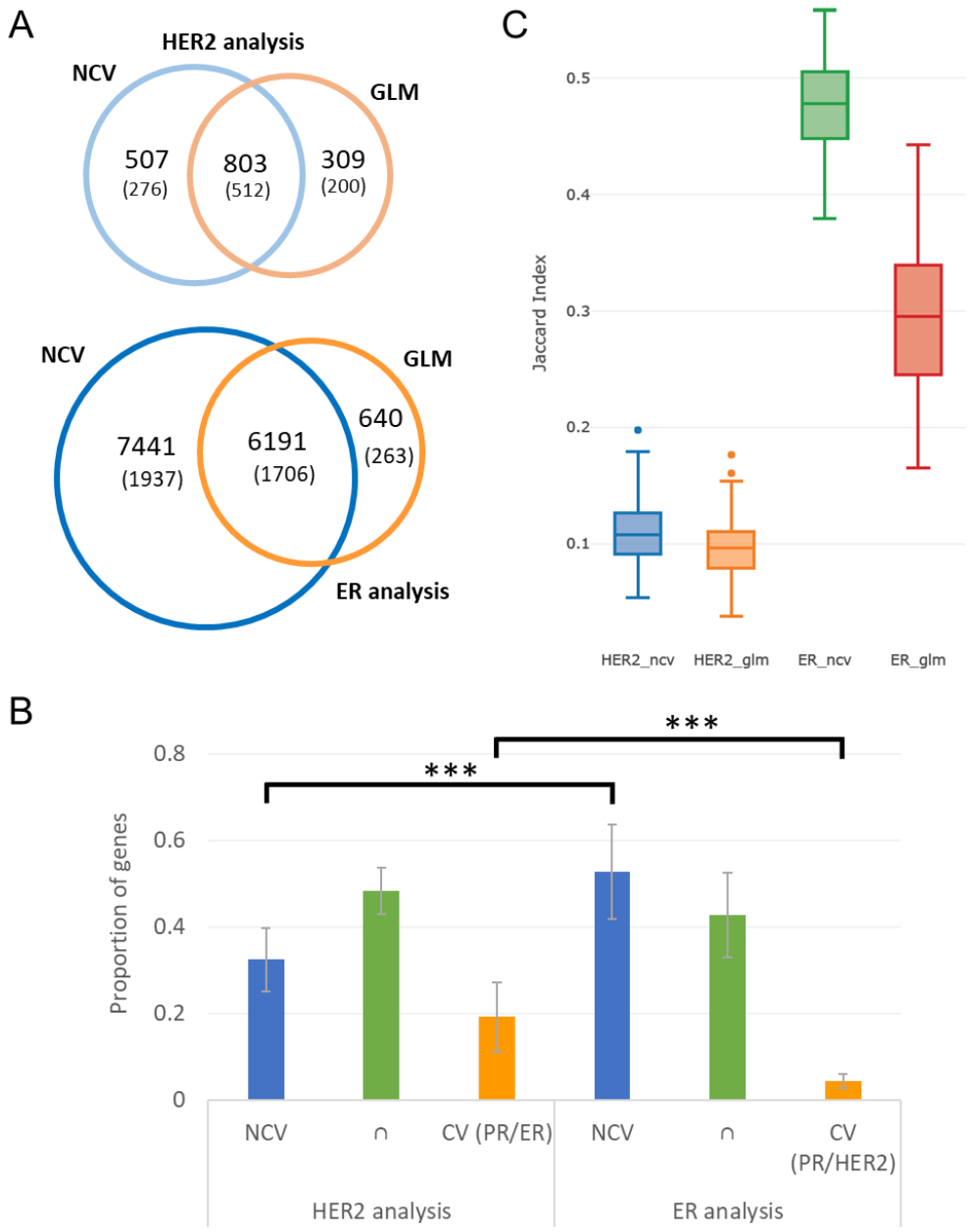


Figure 7. Comparison between HER2 and ER focused differential expression analysis. Differential expression analyses using DESeq2 for the 20 patient sets with covariates (HER2_glm, ER_glm) and without covariates (HER2_ncv, ER_ncv). In HER2 analysis, PR and ER status were used as covariates. In ER analysis, PR and HER2 status were used as covariates. Resulting gene lists used p-adj ≤ 0.05 as the significance threshold. (A) Venn diagrams of gene lists in HER2 and ER analysis with and without covariates. Mean number of genes shown with standard deviation in parentheses. (B) Proportion of genes found significant only in non-covariate analysis (blue), only in covariate analysis (orange), or in both (green); both HER2 and ER focused analysis are shown. Significant differences between bars tested using a paired t-test of proportion of genes for each type of analysis for each set of the 20-set. (C) Consistency of significant genes for each of the four methods evaluated by the Jaccard Index of pairwise comparisons of gene lists between the 20 patient sets.

4.2.3 20-set ER focused differential expression analysis

In the previous sections, we saw major differences between non-covariate and covariate analysis with GLMs. GLMs have been established as a method to account for covariates, however we do not know if it is the best method for optimally capturing covariate effects. One concern was the linear combination of features by the GLM method and whether it would be appropriate for these covariates. As demonstrated in the expression boxplots previously in **Figure 4C**, the assumption of linearity poses issues for determining differentially expressed genes. If the covariate does not contribute linearly to the main feature of interest, then the GLM cannot optimally capture the non-linear relation. Particularly, for a gene like EN1 in **Figure 4C**, the covariate effect is non-linear as the difference in expression is about 3 log-fold changes in ER- samples and almost zero in ER+ samples. To see if the differences between non-covariate and covariate analysis were due to the method of accounting for covariates, we evaluated differential expression with our proposed categorical method. Our categorical method averts the linearity assumption and can be another avenue to tackle covariate effects.

In this section, we tested non-covariate models (NCV), covariates in a GLM, and our categorical method for ER differential expression with PR/HER2 as covariates across the 20-set. In our categorical method, we compared Fisher's method and Stouffer's method for combining p-values. Lastly, in this section, we aimed to delve further into the biological importance of our findings.

Figure 8A shows the mean number of genes obtained by each method. In ER focused analysis, inclusion of covariates greatly reduced the number of significant genes at p-adj ≤ 0.05. The categorical method seemed to be most strict, with Stouffer slightly more so than Fisher. This can be explained by the way in which covariate effects are considered. In GLMs, covariate effects are considered by allocating a coefficient to the covariate in the linear model. In our categorical

method, the effect of the covariate is directly removed by partitioning into sub-categories. The strength of covariate consideration seems to negatively correlate with the number of significantly differentially expressed genes. **Figure 8B** presents the biotype distributions with respect to each method. Overall, a large majority of significantly differentially expressed genes were protein coding, hovering around 60-70%. This is followed by lincRNA and antisense transcripts. Results from the categorical analyses seemed to have the highest proportion of protein coding genes. We believe this may be explained by the fact that they also called the fewest number of significant genes. As many of the top rankings genes are protein coding, they would represent a larger proportion in the smaller gene lists called by categorical analyses. Overall, there does not seem to be any actionable differences in biotype distributions between the four methods.

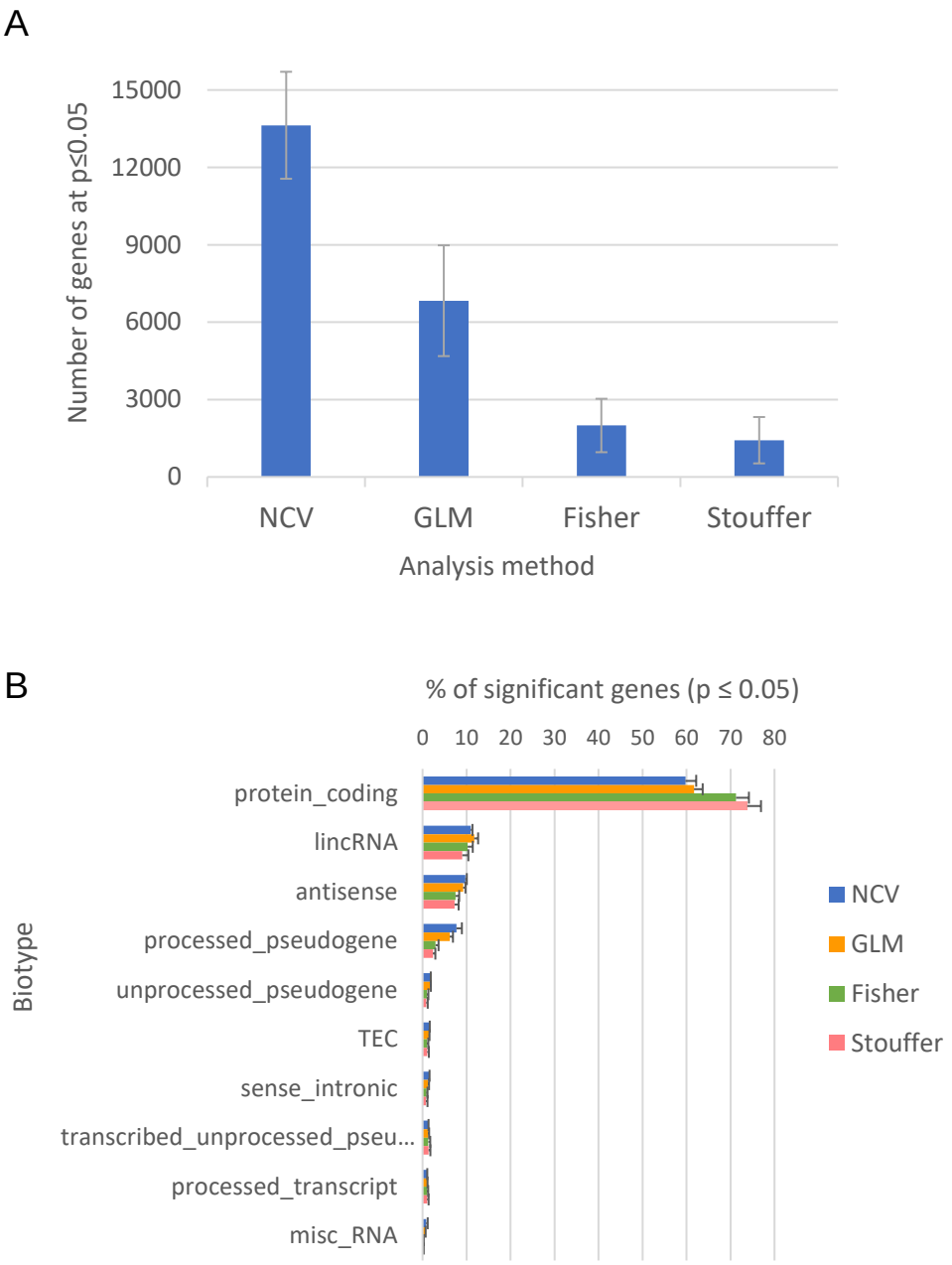


Figure 8. Overview of gene lists of ER focused differential expression analysis. Differential expression analyses using DESeq2 for the 20 patient sets without covariates (NCV), with PR/HER2 as covariates (GLM), and using our categorical method for PR/HER2 covariates with Fisher's and Stouffer's method for combining p-values. Resulting gene lists used p-adj ≤ 0.05 as the significance threshold. (A) Mean number of genes at significance threshold of p-adj ≤ 0.05 between 20 patient sets for each method; error bars represent standard deviation. (B) Biotypes of significant genes for each method, expressed as a mean percentage of genes between 20 patient sets; error bars represent standard deviation.

The direct comparisons between each method are shown in **Figure 9**. The pairwise comparisons of gene lists were evaluated by the Jaccard index using significant genes by p-adj ≤ 0.05 or the top ranking 1000 genes (**Figure 9A**). The Fisher-Stouffer overlap was the highest as they shared the same results from differential expression analysis and only differed in how p-values are combined. Interestingly, categorical covariate analysis overlapped significantly more with GLM than NCV analysis. In a similar analysis for HER2 with PR/ER as covariates, there was no significant difference in overlap between categorical-GLM and categorical-NCV. We believe that this is due to a correlated covariate of PR on ER analysis, absent in HER2 analysis.

Consistency of significant genes evaluated by the Jaccard index of pairwise 20-set comparisons of gene lists is shown in **Figure 9B**. Covariate analysis seemed to reduce the number of consistently significant genes between patient sets in all three methods GLM, Fisher, and Stouffer. As before, a reduction in the number of genes consistent across sets could be beneficial. If the number of truly differentially expressed genes is in reality small, then covariate methods would be narrowing down the number of true positives. However, it is also possible that the inclusion of covariates may be overfitting to each patient set, leading to poor generalization over other patient sets. This is evidenced by the fact that Stouffer's method (most strict, more patient set specific genes) has the lowest Jaccard index. Further examination of these smaller lists of genes is required to determine whether they are actually true positives or if covariate models lead to more false negatives.

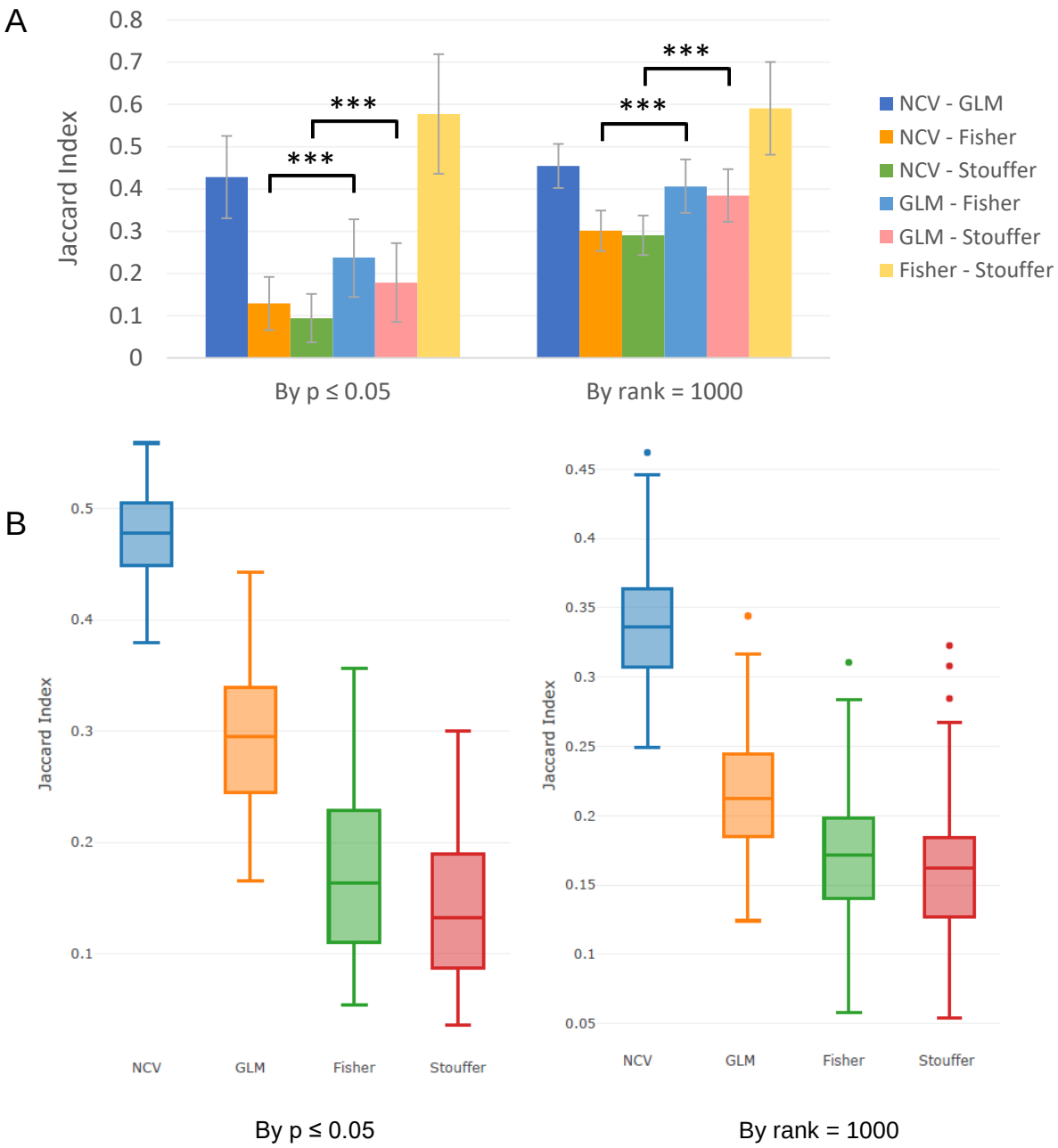


Figure 9. Comparisons between different methods for covariates in ER focused differential gene expression analysis. Differential expression analyses using DESeq2 for the 20 patient sets without covariates (NCV), with PR/HER2 as covariates (GLM), and using our categorical method for PR/HER2 covariates with Fisher's and Stouffer's method for combining p-values. (A) Pairwise method comparisons of gene lists significant at p-adj ≤ 0.05 and top ranking 1000 genes. Overlap evaluated by the mean Jaccard Index between 20 patient sets; error bars represent standard deviation. Significant differences between bars tested using a paired t-test of Jaccard indices. (B) Consistency of significant genes for each of the four methods evaluated by the Jaccard Index of pairwise patient set comparisons of gene lists; both comparisons by p-adj and by rank are shown.

Similar to **Figure 7** in the previous section, **Figure 10** compares gene lists of covariate analysis with that of non-covariate analysis. Shown first, on the left, is the comparison of the top 1000 ranking genes. By this metric, GLM was somewhat similar to categorical methods and there was no significant difference between Fisher and Stouffer. This can be interpreted as the top 1000 genes being differentially expressed strongly enough for consistency across the three methods. Across the 20-set, the variability was also low for the top 1000 genes as shown by the small error bars.

At a significance threshold of p-adj ≤ 0.05, about 50% of genes presented in the NCV only category, falling out of significance with the inclusion of covariates. This was true to a greater extent in categorical analyses with values upward of 80%. We speculate that this high value may imply that the covariate method could be too strict and could lead to false negatives.

The actual genes in each group are shown in **Figure 11** represented by word clouds. The frequency of significance of each gene was totalled over the 20 sets for each of the 3 methods (GLM, Fisher, Stouffer). Thus, the maximum frequency would be 60, which is represented by the largest font. For example, ESR1 encoding ER is significant in all 60 cases showing up in both covariate and non-covariate analysis as expected. In the non-covariate analysis only category (red) are genes for which differential expression could not be solely explicable by ER status. The genes shown in the largest fonts, in both covariate and non-covariate analysis (purple), are those that consistently appear significant across the 20 sets. These genes should represent the most likely truly differentially expressed genes between ER+ and ER- cases. 20-fold cross validation with 140 samples each is powerful. For a gene to be significant across almost all sets for 3 methods, there is compelling evidence that the gene would belong to its respective category. These genes should make for the most suitable candidates for further biomarker validation.

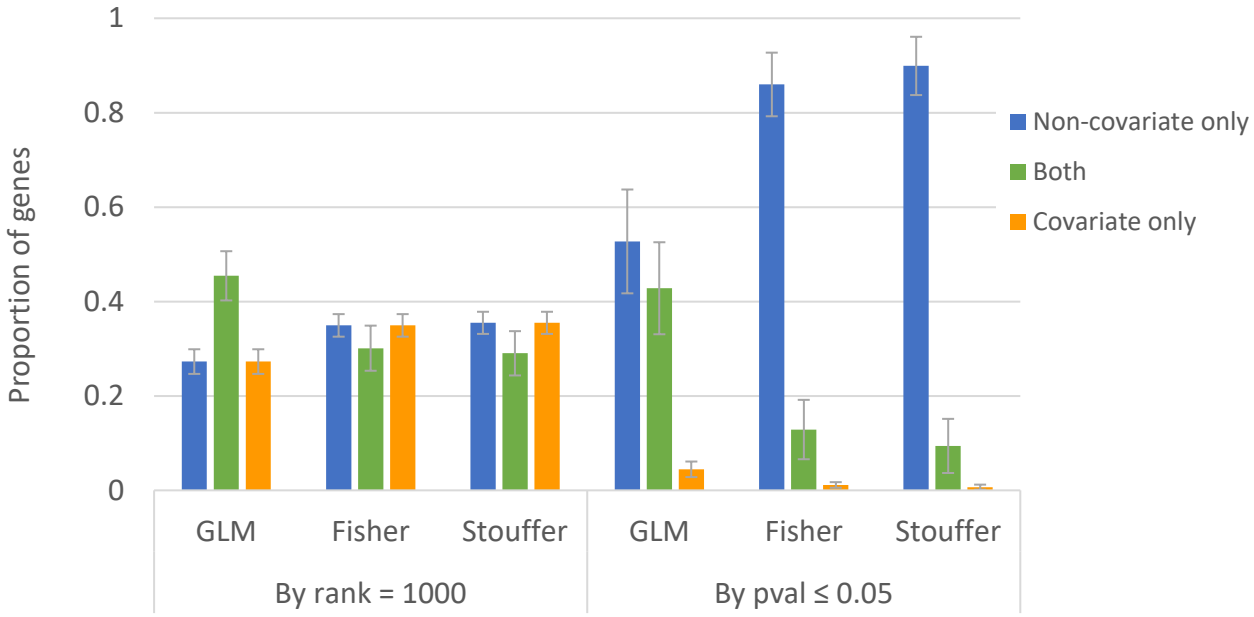


Figure 10. Gene discoveries and disappearances in ER focused differential expression analysis. Gene lists from differential expression methods using covariates (GLM, Fisher, Stouffer) were compared to the non-covariate gene list. Shown are the proportion of genes found significant only in non-covariate analysis (blue), only in covariate analysis (orange), or in both (green). Both comparisons of gene lists by p-adj ≤ 0.05 and by top ranking 1000 genes are shown. Columns represent mean proportion of genes between 20 patient sets and error bars represent standard deviation.

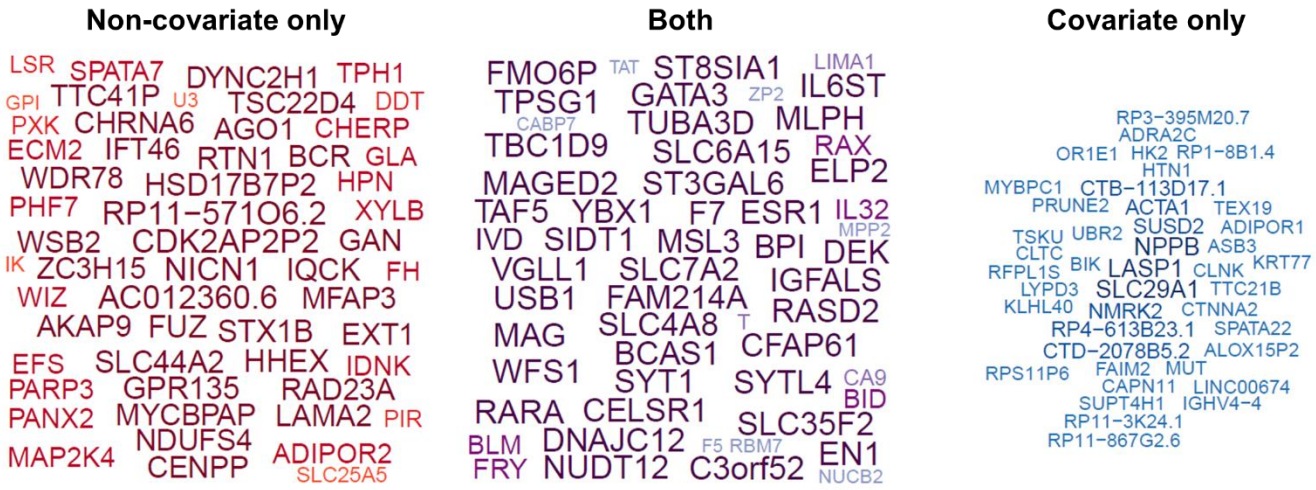


Figure 11. Word clouds of gene symbols found in only non-covariate analysis, only covariate analysis, or both. Resulting gene lists used p-adj ≤ 0.05 as the significance threshold. Genes found significant only in non-covariate analysis (red), only in covariate analysis (blue), or in both (purple). Shown are most consistently significant genes (appear most frequently in the 20 patient sets) between the three methods combined (GLM, Fisher, Stouffer). The largest font represents 60 appearances (in all 20 patients sets for all 3 methods). Smaller fonts represent lower frequency of significance.

We further looked to gene expression boxplots to understand why some genes may fall out of significance or become significant with the inclusion of covariates (**Figure 12**). In the non-covariates only group, HHEX is an example of a gene that consistently fell out of significance (in 58/60 cases) with inclusion of covariates. Comparing across covariate groups between ER- and ER+ (i.e. green to green, red to red, etc.), there is very little difference in expression. There is, however, a strong PR/HER2 effect as shown by trend of differences in expression between covariate groups (green, red, purple, pink). This trend is consistent in both ER- and ER+ groups. In non-covariate analysis, the detected difference in expression for HHEX may be due to the covariate effects. Thus, when covariates were accounted for, HHEX fell out of significance. HHEX (Hematopoietically Expressed Homeobox) encodes a member of the homeobox family of transcription factors involved in developmental processes [66].

On the other hand, LASP1 was a gene that becomes significant only in analysis with covariates. Each of the boxplots are slightly higher individually in ER+ compared to ER-. However, this difference is partially negated by the high expression in ER-, PR-, HER2+ cases (red). Accounting for the effect of PR/HER2 revealed the differential expression of LASP1. The high variance may explain why LASP1 was revealed with covariates in only some of the cases. LASP1 encodes for the MLN50 protein, the overexpression of which has been linked to metastatic breast cancer [67].

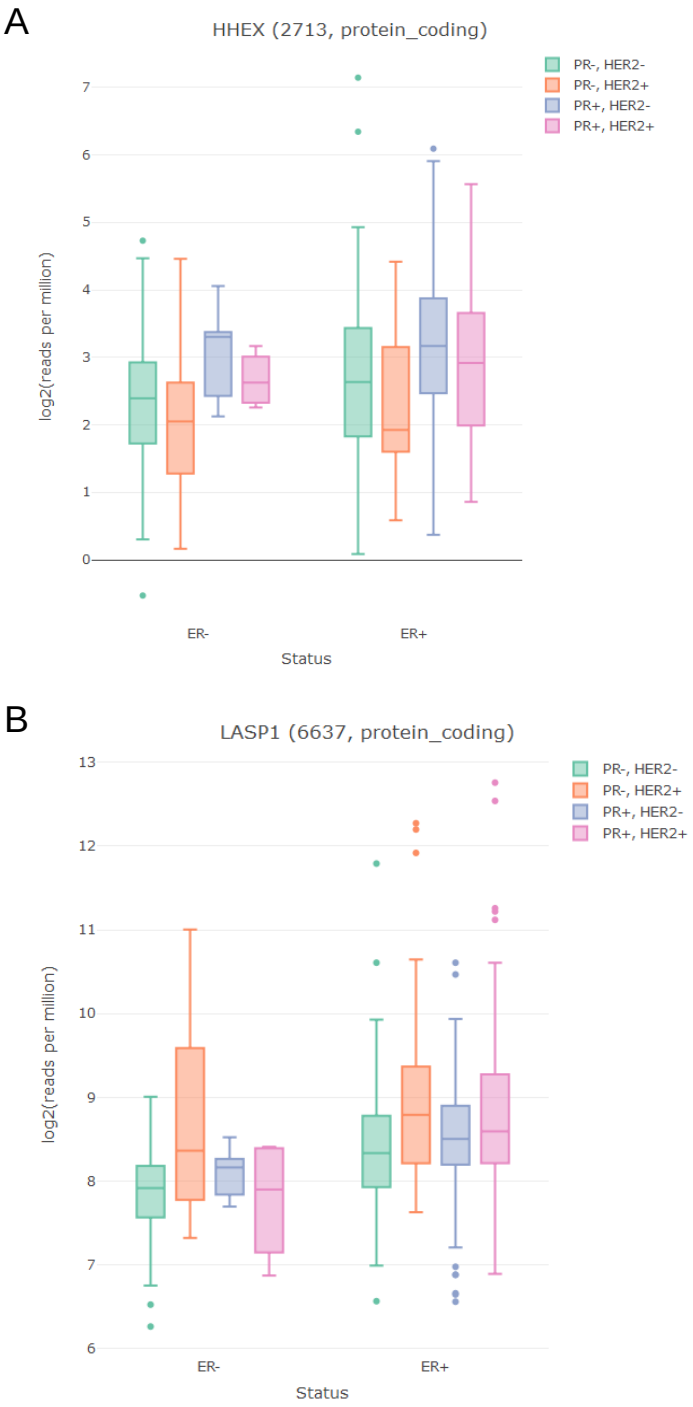


Figure 12. Gene expression boxplots of non-covariate only and covariate-only genes. Gene expression is represented in log2(reads per million). Boxplots are grouped by ER- (left) vs. ER+ (right), then further divided by PR and HER2 status to highlight their effect as covariates. (A) HHEX falls out of significance at p-adj ≤ 0.05 when PR/HER2 covariates are considered. It is a member of the homeobox family of transcription factors in developmental processes. (B) LASP1 becomes significant at p-adj ≤ 0.05 when PR/HER2 covariates are considered. The LASP1 protein (MLN50) has been linked to metastatic breast cancer and other cancers.

Gene Ontology enrichment analysis was performed (**Figure 13**) to assess the biological importance of using covariates in differential expression analysis. The top ranking 1000 genes from non-covariate analysis and covariate analysis were submitted to GO [59]. The gene list from covariate analysis showed enrichment for several more biological processes, which can be seen as evidence for the biological importance of covariates in differential expression analysis.

As a preliminary assessment of impact on breast cancer survival, we used KM-plotter [60] [68] to plot Kaplan Meier survival curves (**Figure 14**). This tool is based on microarray data from 3951 breast cancer patients. ESR1 was our gene of interest in the differential expression between ER+ and ER- cases and the impact of its expression on survival is shown for reference. USB1 was consistently differentially expressed in non-covariate models and in covariate models (with PR/HER2). NPPB was significantly differentially expressed only when covariates were considered. For all three genes, high expression correlated with improved overall survival. For ESR1 (ER), this can be explained by greater efficacy of treatment for ER+ tumours. The trend of USB1 and NPPB were very similar to that of ESR1, thus there is evidence that USB1 and NPPB could translate well as prognostic biomarkers or possibly surrogate biomarkers for ER status. Although the goal of this thesis was not focused on finding survival biomarkers, these KM-plots serve as preliminary evidence that accounting for covariates in biomarker discovery does indeed have biological importance. In this case, NPPB would not have been identified without the consideration of PR/HER2 as covariates.

We believe it is evident that covariates in differential expression analysis play an important role for the detection of truly differentially expressed genes. In covariate models, covariate-associated genes – which may be false positives – are removed and covariate-obscured genes are revealed, both of which have downstream benefits for biological validation. We believe optimized biomarker discovery, with covariates, is the next step for better biomarker translation.

A

Displaying only results with P<0.05; click here to display all results

GO biological process complete	Homo sapiens (REF) #	upload_1 (Hierarchy NEW!) #	expected	Fold Enrichment	+/-	P value
nervous system development	2228	119	78.72	1.51	+	3.44E-02
Unclassified	3579	93	126.45	.74	-	0.00E00

B

Displaying only results with P<0.05; click here to display all results

GO biological process complete	Homo sapiens (REF) #	upload_1 (Hierarchy NEW!) #	expected	Fold Enrichment	+/-	P value
central nervous system development	888	56	28.62	1.96	+	1.88E-02
↳nervous system development	2228	114	71.82	1.59	+	5.10E-03
↳system development	4199	185	135.35	1.37	+	2.34E-02
↳anatomical structure development	5141	221	165.72	1.33	+	8.87E-03
↳developmental process	5506	234	177.49	1.32	+	8.37E-03
↳multicellular organism development	4786	205	154.28	1.33	+	3.57E-02
ion transport	1314	75	42.36	1.77	+	1.38E-02
regulation of multicellular organismal process	2791	136	89.97	1.51	+	4.96E-03
Unclassified	3579	75	115.37	.65	-	0.00E00

Figure 13. Gene Ontology enrichment results of biological processes of gene lists from ER focused differential expression without and with covariates. From differential expression analysis of the 20 patient sets, the median ranking of each gene was calculated. The top 1000 median ranked genes from non-covariate and covariate (PR/HER2) were submitted to Gene Ontology. (A) Output of GO enrichment analysis with gene list from non-covariate analysis. (B) Output of GO enrichment analysis with gene list from covariate analysis.

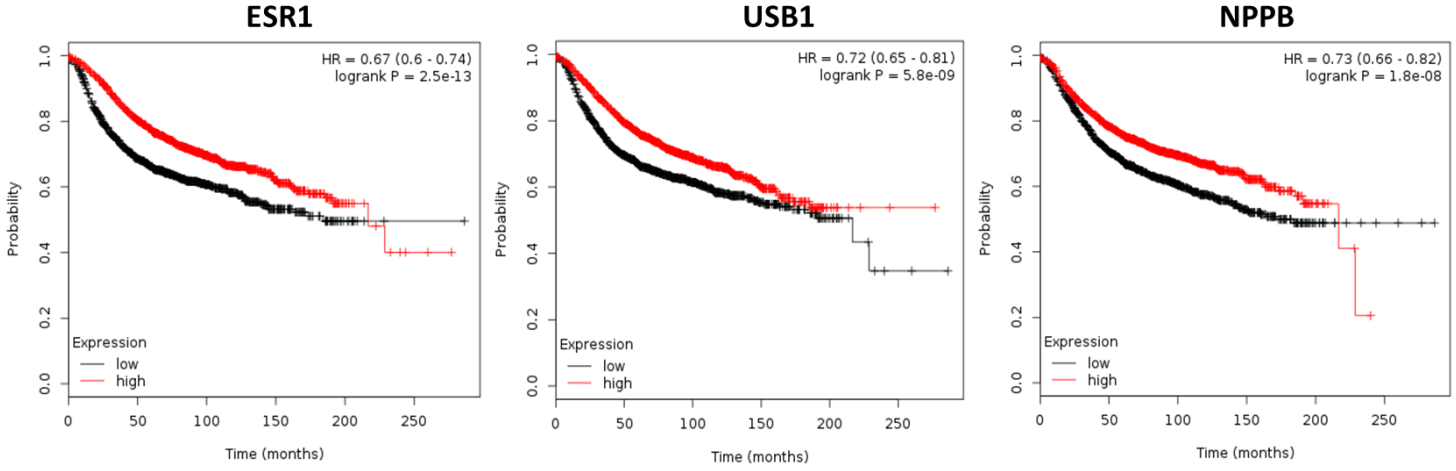


Figure 14. Kaplan-Meier survival curves for 3951 breast cancer patients. Generated using the KM-plotter tool for breast cancer microarray data [60] [68]. ESR1, USB1, NPPB were consistently significant genes selected from the word cloud in Figure 11. ESR1 is our gene of interest in the differential expression between ER+ and ER- cases. NPPB is a differentially expressed gene, only significant when PR/HER2 covariates are considered. USB1 is a significantly differentially expressed gene in both covariate and non-covariate analysis.

5. CONCLUSIONS AND FUTURE WORKS

Valid clinical biomarkers have tremendous impacts for patient care in the emerging world of personalized medicine. The problem with biomarkers lies in the poor translation from literature to clinical use. Biomarkers that are discovered in early stage or small-scale studies often do not validate in any subsequent steps beyond the study. This ultimately results in hundreds of thousands of claimed biomarkers failing to make it into clinical practice. We believe better biomarker translation begins with better biomarker discovery – that is, to use high-throughput data from large-scale studies and to account for clinical covariates at the stage of differential expression analysis.

Many scientists and agencies have recognized the issue of small-scale, investigator-initiated research models. Initiatives such as The Cancer Genome Atlas (TCGA), led by the National Institutes of Health (NIH), and the International Cancer Genome consortium have been established to tackle this problem. In the field of cancer genomics, TCGA has produced a wealth of publicly available high-throughput cancer data which continues to be a valuable resource for researchers worldwide. With respect to gene expression biomarkers, RNA-Seq offers a high-throughput alternative to microarrays for assessing gene expression. Since the advent of next-generation sequencing, many scientists and consortia – including TCGA – have adopted RNA-Seq for high-throughput transcriptomic analysis.

However, despite advancements in technology, translation of gene expression cancer biomarkers remains relatively poor [69]. One glaring issue, we found, was the lack of inclusion of clinical covariates at the level of differential expression analysis in high-throughput cancer studies. Covariates are often neglected until later, such as when answering prediction problems. Bioinformatic tools for differential expression analysis (DESeq2 and edgeR) offer GLM

functionality for inclusion of covariates, however it is often used to only account for batch effects or technical artifacts and not clinical covariates. Covariates can introduce underlying directional biases which can skew the data – something that cannot be solved simply by increasing sample size. Thus, with larger-scale studies such as ones in TCGA, the consideration of covariates remains important. By considering clinical covariates and with more data from large-scale studies, we believe optimized biomarker discovery can overcome some of the previous limitations to clinical success.

In this thesis, we evaluated incorporation of clinical covariates in differential expression analysis of high-throughput cancer data. Overall, we examined 786 cases of primary tumours of breast invasive ductal carcinoma from TCGA-BRCA. Our overarching objective was to answer the fundamental question of how the inclusion of covariates changes what appear to be relevant biomarkers. We had three main hypotheses: 1) covariate models have increased robustness and consistency across patient sets, 2) covariate models remove covariate-associated genes and reveal covariate-obscured genes, and 3) use of correlated over uncorrelated covariates has stronger effects on differential expression analysis of the condition of interest. To contextualize our analyses in field of cancer biology, we focused on differential gene expression with respect to HER2, ER, and PR status – three key receptors in breast cancer.

In our exploratory phase, we saw that edgeR with default parameters had limitations when dealing with low-count or zero-count genes. DESeq2, by default, strongly moderates low or zero-count genes and did not face this issue. The Mann-Whitney-Wilcoxon test was demonstrated to be a possible non-parametric alternative for differential expression analysis. However, the model was not designed for RNA-Seq data and there is no direct method for incorporating covariates.

In our initial covariate-sensitive HER2 differential expression analysis, we considered age, race, ER status, and PR status as covariates. As expected from literature, ER and PR were positively correlated in both receptor status and gene expression. At this stage, we found preliminary evidence to support our second hypothesis that covariate models do indeed remove covariate-associated genes and reveal covariate-obscured genes. We attempted to validate these findings in a mutually exclusive dataset. We found that the agreement between datasets was low in both covariate and non-covariate analyses. In this two-fold cross validation, contrary to our first hypothesis, covariate models did not objectively improve validation consistency. As well, in our 20-set analysis, validation consistency was not improved by covariate models. Covariate models, both GLMs and our categorical method, had fewer genes that consistently appeared significant across the 20-set. However, there seemed to be a core list of genes that do remain significant in a majority of the 20-set. It could be that very few genes are truly differentially expressed and what we see as a reduction in similarity between gene lists could be the removal of genes not in the core list. Certainly, the question of what are truly HER2/ER/PR-associated genes has not been answered and without knowing which genes are true positives and true negatives, it remains difficult to evaluate whether covariate models produce a more accurate candidate list of differentially expressed genes

In HER2 and ER 20-set analysis, we observed that covariate models do produce substantially different gene lists compared to non-covariate models. Many genes called significant in non-covariate analysis fell out of significance after the inclusion of covariates. Some genes which did not appear in non-covariate analysis became significant with covariates. There is strong evidence to support our second hypothesis as well as our overarching claim that covariates change what appear to be relevant biomarkers.

In the evaluation of correlated vs. uncorrelated covariates in the case of ER and HER2 analysis respectively, we found evidence to support our third hypothesis that correlated covariates have stronger effects on differential expression analysis. In ER analysis with PR/HER2 as correlated covariates, significantly more genes fell out of significance (genes only found in non-covariate analysis) when compared to HER2 analysis with ER/PR uncorrelated covariates. When comparing discovered genes (genes only found in covariate analysis), significantly fewer genes were found in the case of correlated covariates vs. uncorrelated covariates. However, as mentioned previously, a smaller list may imply a higher likelihood of being true positives and may be beneficial.

For ER focused differential expression analysis, we evaluated GLMs and our categorical method for covariates. There was a degree of agreement between the GLM method and the categorical method with better agreement for top ranking genes. The categorical method may be a stronger alternative for covariate analysis as it directly removes the covariate effect and does not make the assumption of linearity as in GLMs. We saw in some cases, that the categorical method may produce strict gene lists which led to few genes consistently called significant across the 20-set. It also removed from significance, a substantial number of genes that were found significant in non-covariate analysis. As the categorical method strongly removes the covariate effect, it is possible that this leads to an increase in false negatives. However, keeping in mind that most validated gene signatures for cancer consist of fewer than 100 genes, the roughly 200 genes consistently found by categorical analysis could truly be valid biomarkers.

From Gene Ontology enrichment analysis, we saw that covariate models of ER differential expression analysis resulted in genes representing more biological process categories. This analysis provided strong evidence for our main objective, as it showed that the inclusion of

covariates led to gene lists with distinct biological differences. Furthermore, covariate models revealed genes which may impact survival in breast cancer patients, such as NPPB.

A critical question in the biomarker discovery process is what are true positives and true negatives? Without biological validation or clinical validation of these biomarkers, true positives cannot be definitively ascertained by differential expression analysis – rather a list of likely gene candidates is produced. Hence for improving biomarker discovery, it would be beneficial to 1) refine the list of candidate genes to improve likelihood, and 2) reduce the number of likely candidate genes to reduce cost of downstream biological validation. In this thesis, we saw avenues for covariate analysis to address these issues.

To our first hypothesis, we saw that covariate analysis did not improve validation consistency. If we solely place value on validation consistency or similarity of gene lists, then the use of covariate models could be unnecessary or even detrimental to differential expression analysis. In this sense, it could also be a reason why there is a lack of use of covariates in current publications for differential expression analysis. From our results, it may be the case that inclusion of covariates in analysis produces a list of genes that is overly reflective of each specific dataset. This could explain why there was wide disparity between the resulting gene lists and reduced validation consistency. However, we did see that there is a small core list of genes that are consistently significantly differentially expressed in both covariate and non-covariate methods. It would be logical to believe that these genes are the more likely candidates for being a true biomarker for the condition of interest. Comparing covariate and non-covariate methods, covariate methods results in a smaller core list of genes. This may be beneficial with respect to downstream biological validation.

To our second hypothesis, we saw many genes that were called significantly differentially expressed only with covariate models. As well, we saw genes that fell out of significance with the

inclusion of covariates. Without knowing the true positives and negatives, it is difficult to evaluate whether covariate models improve the list of candidate genes. It could be possible that covariate models are removing true biomarkers and adding in false positives. It could also be the case that false positives are removed and true positives are identified. We did see some evidence in our expression boxplots that truly differentially expressed genes are revealed by covariate models and some false positives are removed. Without biological validation, the benefit of covariate models over non-covariate models cannot be definitively claimed. However, the genes lists are indeed different, and it would be appropriate to claim that without the inclusion of covariates, some of these candidate genes would have gone unconsidered.

There is evidence to support the general notion that covariate models produce different candidate gene lists compared to non-covariate models. In this sense, covariate models offer the opportunity to explore differential expression from another perspective. Whether covariate models improve upon non-covariate models with respect to enhancing candidate gene lists for biomarker discovery remains ambiguous without biological validation. Covariate models do, however, produce smaller lists of core candidate genes which can be more feasibly validated downstream.

Several avenues exist for future work. The first is to explore what it means for a gene to be ER-differentially expressed in the context of covariates. ER focused differential expression analysis suggests many genes change expression depending on ER status. Keeping in mind that PR is highly correlated with ER, some ER-associated genes may be truly influenced by ER but some may be truly influenced by PR. The correlation between ER and PR makes these PR-influenced genes also appear ER-associated. Traditional differential expression does not discriminate between these situations, but covariate analysis can. For better understanding of HER2, ER, and PR influence of gene expression, differential expression analysis should be performed for all possible

combinations of HER2, ER, and PR as the factor of interest or as the covariate. For example, ER analysis with PR as the covariate should be compared with PR analysis with ER as the covariate, to determine which are truly ER or PR associated genes. With cross-validation, this can determine with confidence which genes are truly false positives or true positives for each condition.

Validation consistency of gene lists can be evaluated using other methods. In this thesis, we compared the overlap of significant genes between datasets as a metric for validation consistency. The next step would be to use the list of differentially expressed genes as predictors for the condition of interest. For example, a list of differentially expressed genes for HER2 status can be obtained from dataset A, then these genes could be used to predict HER2 status in dataset B. The prediction performance across different datasets would be a metric of validation consistency. In this way, smaller gene lists can be compared with larger gene lists without directly penalty to validation consistency – an issue we observed when using gene identity overlap as the performance metric.

So far, many biomarker studies are focused on genes and differential gene expression. In this thesis as well, we only examined covariate models within the context of single gene differential expression. Our list of genes could be further examined through gene clusters and networks and also at a protein level, looking at protein-protein interactions. At a higher level, these genes can be examined for relationships or patterns for survival annotation.

Outside of receptor statuses, comparisons can be made between healthy and tumour samples. Covariates are an even more important consideration with respect to diagnostic or survival biomarkers as those are involved in clinical decision making. Lastly, the next step after biomarker discovery is biological validation using cancer cell lines. Further collaborations would be required for this line of work.

From the results of this thesis, it is evident that the inclusion of covariates in differential expression is important for the detection of truly differentially expressed genes. Covariate models substantially change what appear to be relevant biomarkers, removing covariate-associated genes and revealing covariate-obscured genes. These changes are not limited to simply differences in the gene lists but also expand to downstream biological differences. Genes with strong biological implications were discovered only when accounting for covariates; such genes may have largely been ignored by traditional differential expression analysis. In this thesis, we show strong evidence that biomarker discovery has the potential to be improved, by considering clinical covariates and by using high-throughput data from large-scale studies. We believe optimized biomarker discovery can tackle some of the previous challenges that limited translational success.

6. REFERENCES

[1] K. Strimbu and J. Tavel, "What are biomarkers," *Current Opinion in HIV and AIDS,* vol. 5, no. 6, p. 463, 2010.

[2] I. Majewski and R. Bernards, "Taming the dragon: genomic biomarkers to individualize the treatment of cancer," *Nature medicine,* pp. 304-312, 2011.

[3] A. P. P. D. P. Antoniou, S. Narod, H. A. Risch, J. E. Eyfjord, J. L. Hopper and B. Pasini, "Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies.," *The American Journal of Human Genetics,* vol. 75, no. 5, pp. 1117-1130, 2003.

[4] G. Poste, "Bring on the biomarkers," *Nature,* vol. 469, no. 7329, pp. 156-157, 2011.

[5] M. Maemondo, A. Inoue, K. Kobayashi, S. Sugawara, S. Oizumi, H. Isobe and Y. Fujita, "Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR," *N Engl j Med,* vol. 362, pp. 2380-2388, 2010.

[6] G. Novelli, C. Ciccacci, P. Borgiani, M. Amati and E. Abadie, "Genetic tests and genomic biomarkers: regulation, quantification and validation," *Clinical cases in mineral and bone metabolism,* vol. 5, no. 2, p. 149, 2008.

[7] F. Rodriguez-Gonzalex, D. Mustafa, B. Mostert and A. Sieuwerts, "The challenge of gene expression profiling in heterogeneous clinical samples," *Methods,* vol. 59, no. 1, pp. 47-58, 2013.

[8] A. Potti, H. Dressman, A. Bild, R. Riedel, G. Chan, R. Sayer and D. Harpole, "Genomic signatures to guide the use of chemotherapeutics," *Nature medicine,* vol. 12, no. 11, pp. 1294-1300, 2006.

[9] D. Ransohoff and M. Gourlay, "Sources of bias in specimens for research about molecular markers for cancer," *Journal of clinical oncology,* vol. 28, no. 4, pp. 698-704, 2009.

[10] C. Smigal, A. Jemal, E. Ward, V. Cokkinides, R. Smith, H. Howe and M. Thun, "Trends in breast cancer by race and ethnicity: update 2006," *CA: a cancer journal for clinicians,* vol. 56, no. 3, pp. 168-183, 2006.

[11] B. Kwabi-Addo, W. Chung, L. Shen, M. Ittmann, T. Wheeler and J. Issa, "Age-related DNA methylation changes in normal human prostate tissues," *Clinical cancer research,* vol. 13, no. 13, pp. 3796-3802, 2007.

[12] E. Ward, A. Jemal, V. Cokkinides, G. Singh, C. Cardinez, A. Ghafoor and M. Thun, "Cancer Disparities by Race/Ethnicity and Socioeconomic Status," *CA: a cancer journal for clinicians,* vol. 54, no. 2, pp. 78-93, 2004.

[13] N. Christenfeld, R. Sloan, D. Carroll and S. Greenland, "Risk factors, confounding, and the illusion of control," *Psychosomatic medicine,* vol. 66, no. 6, pp. 868-875, 2004.

[14] M. Pourhoseingholi, A. Baghestani and M. Vahedi, "How to control confounding effects by statistical analysis," *Gastroenterology and Heptaology from bed to bench,* vol. 5, no. 2, p. 79, 2012.

[15] N. Zaitlen, S. Lindström, B. Pasaniuc, M. Cornelis, G. Genovese, S. Pollack and B. I. Freedman, "Informed conditioning on clinical covariates increases power in case-control association studies," *PLoS genetics,* vol. 8, no. 11, p. e1003032, 2012.

[16] R. De Angelis, R. Capocaccia, T. Hakulinen, B. Soderman and A. Verdecchia, "Mixture models for cancer survival analysis: application to population-based data with covariates," *Statistics in medicine,* vol. 18, no. 4, pp. 441-454, 1999.

[17] R. K. Rogers, G. J. Stoddard, T. Greene, A. D. Michaels, G. Fernandez, A. Freeman and J. Stehlik, "Usefulness of adjusting for clinical covariates to improve the ability of B-type natriuretic peptide to distinguish cardiac from noncardiac dyspnea," *The American journal of cardiology,* vol. 104, no. 5, pp. 689-694, 2009.

[18] W. D. Shannon, M. A. Watson, A. Perry and K. Rich, "Mantel statistics to correlate gene expression levels from microarrays with clinical covariates," *Genetic Epidemiology,* vol. 23, no. 1, pp. 87-96, 2002.

[19] T. Sing, A. J. Low, N. Beerenwinkel, O. Sander, P. K. Cheung, F. S. Domingues and P. R. Harrigan, "Predicting HIV coreceptor usage on the basis of genetic and clinical covariates," *Antiviral therapy,* vol. 12, no. 7, p. 1097, 2007.

[20] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of RNA-seq data," *BMC bioinformatics,* vol. 14, no. 1, p. 91, 2013.

[21] W. B. Ershler and E. T. Keller, "Age-associated increased interleukin-6 gene expression, late-life diseases, and frailty," *Annual review of medicine,* vol. 51, no. 1, pp. 245-270, 2000.

[22] B. Jovov, F. Araujo-Perez, C. S. Sigel, J. K. Stratford, A. N. McCoy, J. J. Yeh and T. Keku, "Differential gene expression between African American and European American colorectal cancer patients," *PloS one,* vol. 7, no. 1, p. e30168, 2012.

[23] P. A. Stewart, J. Luks, M. D. Roycik, Q. X. A. Sang and J. Zhang, "Differentially expressed transcripts and dysregulated signaling pathways and networks in African American breast cancer," *PloS one,* vol. 8, no. 12, p. e82460, 2013.

[24] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne and C. Tyler-Smith, "Relative impact of nucleotide and copy number variation on gene expression phenotypes.," *Science,* vol. 315, no. 5813, pp. 848-853, 2007.

[25] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen and T. Thorsen, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences,* vol. 98, no. 19, pp. 10869-10874, 2001.

[26] K. S. Wilson, H. Roberts, R. Leek, A. L. Harris and J. Geradts, "Differential gene expression patterns in HER2/neu-positive and-negative breast cancer cell lines and tissues.," *The American journal of pathology,* vol. 161, no. 4, pp. 1171-1185, 2002.

[27] H. Y. Chang, D. S. Nuyten, J. B. Sneddon, T. Hastie, R. Tibshirani, T. Sørlie and M. van de Rijn, "Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, no. 10, pp. 3738-3743, 2005.

[28] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin and W. Hiller, "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer," *New England Journal of Medicine,* vol. 351, no. 27, pp. 2817-2826, 2004.

[29] M. I. Love, W. Huber and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome biology,* vol. 15, no. 12, p. 550, 2014.

[30] M. D. Robinson, D. J. McCarthy and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics,* vol. 26, no. 1, pp. 139-140, 2010.

[31] X. Li, Y. Shi, Z. Yin, X. Xue and B. Zhou, "An eight mi-RNA signature as a potential biomarker for predicting survival in lung adenocarcinoma," *Journal of translational medicine,* vol. 12, no. 1, p. 1, 2014.

[32] K. Tomczak, P. Czerwińska and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemporary oncology,* vol. 19, no. 1A, p. A68, 2015.

[33] C. W. Brennan, R. G. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S. R. Salama and R. Beroukhim, "The somatic genomic landscape of glioblastoma," *Cell,* vol. 155, no. 2, pp. 462-477, 2013.

[34] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature,* vol. 490, no. 7418, pp. 61-70, 2012.

[35] C. DeSantis, J. Ma, L. Bryan and A. Jemal, "Breast cancer statistics, 2013," *CA: a cancer journal for clinicians,* vol. 64, no. 1, pp. 52-62, 2014.

[36] Breastcancer.org, "Non-Invasive or Invasive Breast Cancer," 26 January 2017. [Online]. Available: http://www.breastcancer.org/symptoms/diagnosis/invasive. [Accessed 9 December 2017].

[37] American Cancer Society, "Breast Cancer," 25 September 2017. [Online]. Available: https://www.cancer.org/cancer/breast-cancer.html. [Accessed 9 December 2017].

[38] D. J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde and J. Baselga, "Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2," *New England Journal of Medicine,* vol. 344, no. 11, pp. 783-792, 2001.

[39] M. E. H. Hammond, D. F. Hayes, M. Dowsett, D. C. Allred, K. L. Hagerty, S. Badve and D. G. Hicks, "American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version)," *Archives of pathology & laboratory medicine,* vol. 134, no. 7, pp. e48-e72, 2010.

[40] Early Breast Cancer Trialists' Collaborative Group, "Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials," *The Lancet,* vol. 365, no. 9472, pp. 1687-1717, 2005.

[41] P. M. Ravdin, S. Green, T. M. Dorr, W. L. McGuire, C. Fabian, R. P. Pugh and R. J. Belt, "Prognostic significance of progesterone receptor levels in estrogen receptor-positive patients with metastatic breast cancer treated with tamoxifen: results of a prospective Southwest Oncology Group study," *Journal of clinical oncology,* vol. 10, no. 8, pp. 1284-1291, 1992.

[42] I. Smith and M. Dowsett, "Aromatase inhibitors in breast cancer," *New England Journal of Medicine,* vol. 348, no. 24, pp. 2431-2442, 2003.

[43] K. B. HORWITZ, Y. KOSEKI and W. L. McGUIRE, "Estrogen control of progesterone receptor in human breast cancer: role of estradiol and antiestrogen," *Endocrinology,* vol. 103, no. 5, pp. 1742-1751, 1978.

[44] G. M. Clark, W. L. McGuire, C. A. Hubay, O. H. Pearson and A. C. Carter, "The importance of estrogen and progesterone receptor in primary breast cancer," *Progress in clinical and biological research,* vol. 132, pp. 183-190, 1982.

[45] G. M. Clark, W. L. McGuire, C. A. Hubay, O. H. Pearson and J. S. Marshall, "Progesterone receptors as a prognostic factor in stage II breast cancer," *New England Journal of Medicine,* vol. 309, no. 22, pp. 1343-1347, 1983.

[46] G. Arpino, H. Weiss, A. V. Lee, R. Schiff, S. De Placido, C. K. Osborne and R. M. Elledge, "Estrogen Receptor–Positive, Progesterone Receptor–Negative Breast Cancer: Association With Growth Factor Receptor Expression and Tamoxifen Resistance," *Journal of the National Cancer Institute,* vol. 97, no. 17, pp. 1254-1261, 2005.

[47] Z. Wang, M. Gerstein and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature reviews genetics,* vol. 10, no. 1, pp. 57-63, 2009.

[48] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson and A. Mortazavi, "A survey of best practices for RNA-seq data analysis," *Genome biology*, vol. 17, no. 1, p. 13, 2016.

[49] S. Goodwin, J. D. McPherson and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333-351, 2016.

[50] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature methods*, vol. 9, no. 4, pp. 357-359, 2012.

[51] C. Trapnell, L. Pachter and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105-1111, 2009.

[52] Genomic Data Commons, "mRNA Analysis Pipeline," Genomic Data Commons, 24 October 2017. [Online]. Available: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#rna-seq-alignment-workflow. [Accessed 12 December 2017].

[53] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15-21, 2013.

[54] S. Anders, P. T. Pyl and W. Huber, "HTSeq—a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166-169, 2015.

[55] N. J. Schurch, P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh and M. Blaxter, "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?," *Rna*, vol. 22, no. 6, pp. 839-851, 2016.

[56] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome biology*, vol. 11, no. 10, p. R106, 2010.

[57] M. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome biology*, vol. 11, no. 3, p. R25, 2010.

[58] P. E. McKnight and J. Najab, "Mann-Whitney U Test," *Corsini Encyclopedia of Psychology*, 2010.

[59] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang and P. D. Thomas, "PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements," *Nucleic acids research*, vol. 45, no. D1, pp. D183-D189, 2017.

[60] A. Lánczky, Á. Nagy, G. Bottai, G. Munkácsy, A. Szabó, L. Santarpia and B. Győrffy, "miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients," *Breast cancer research and treatment*, vol. 160, no. 3, pp. 439-446, 2016.

[61] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini and M. Ceccarelli, "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data," *Nucleic acids research,* vol. 44, no. 8, pp. e71-e71, 2016.

[62] S. Anders, D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber and M. D. Robinson, " Count-based differential expression analysis of RNA sequencing data using R and Bioconductor," *Nature protocols,* vol. 8, no. 9, pp. 1765-1786, 2013.

[63] R. Fisher, Statisical Methods for Research Workers, 4th Edition ed., Edinburgh: Oliver and Boyd, 1932.

[64] M. Whitlock, "Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach," *Journal of evolutionary biology,* vol. 18, no. 5, pp. 1368-1373, 2005.

[65] D. V. Zaykin, "Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis," *Journal of evolutionary biology,* vol. 24, no. 8, pp. 1836-1841, 2011.

[66] F. K. Bedford, A. Ashworth, T. Enver and L. M. Wiedemann, "HEX: a novel homeobox gene expressed during haematopoiesis and conserved between mouse and human," *Nucleic acids research,* vol. 21, no. 5, pp. 1245-1249, 1993.

[67] C. Tomasetto, C. Regnier, C. Moog-Lutz, M. G. Mattei, M. P. Chenard, R. Lidereau, P. Basset and M. C. Rio, "Identification of four novel human genes amplified and overexpressed in breast carcinoma and localized to the q11-q21. 3 region of chromosome 17," *Genomics,* vol. 28, no. 3, pp. 367-376, 1995.

[68] B. Győrffy, A. Lanczky, A. C. Eklund, C. Denkert, J. Budczies, Q. Li and Z. Szallasi, "An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients," *Breast cancer research and treatment,* vol. 123, no. 3, pp. 725-731, 2010.

[69] C. Sawyers, "The cancer biomarker problem," *Nature,* vol. 452, no. 7187, pp. 548-552, 2008.