Transcriptomic Interrogation of the Drivers of High Grade Serous Ovarian Cancer Progression

by

John Alan Willis

BA, Princeton University, 2013

Submitted to the Graduate Faculty of
School of Medicine in partial fulfillment
of the requirements for the degree of
Master of Science

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

School of Medicine

This thesis was presented

by

John Alan Willis

It was defended on

April 27, 2020

and approved by

Neil Hukriede, Professor and Vice Chair,
Department of Developmental Biology

Steffi Oesterreich, Professor,
Department of Pharmacology and Chemical Biology

Thesis Advisor: Adrian Lee, Professor,
Department of Pharmacology and Chemical Biology

Transcriptomic Interrogation of the Drivers of Ovarian Cancer Progression

John Alan Willis, MS

University of Pittsburgh, 2020

Ovarian cancer is the deadliest gynecologic malignancy, particularly the High Grade Serous subtype (HGSOC). Unlike other subtypes of ovarian cancer, most initial HGSOC presentations respond well to chemotherapy, ususally a platinum-based agent and a taxane. Most HGSOC deaths follow initial round of treatment, when the recurrent disease fails to respond to standard therapy. A better understanding of how HGSOC progresses from a treatable disease to a chemoresistant one may assist in the development of more effective therapies. However, determining the drivers of this evolution has been stymied by a lack of longitudinal data.

Recent advances in sequencing technology and an increased understanding of the value of long-term follow-up in cancer patients present an opportunity to study the mechanisms of HGSOC evolution. Dissemination of sequencing results through public databases like the Sequence Read Archive offers researchers the opportunity to strengthen the conclusions drawn from their own sequencing studies by contextualizing their own results, or by pooling results for more powerful meta-analyses. I assembled eight RNA-Seq datasets (one generated in-house), four consisting of matched pairs of primary and recurrent HGSOC, two consisting of primary ovarian and metastatic samples from the same presentation, and two studies of ovarian cancer cell lines, to integrate in a study of HGSOC evolution. Combining gene expression and gene fusion analysis may suggests common and potentially actionable pathways of HGSOC evolution.

My analysis of 118 pairs of HGSOC samples suggests genes involved in tumor-microenvironmental interactions, immune response, epigenetic factors, and regulators of epithelial to mesenchymal transition (EMT) are altered with HGSOC progression. Transcript level analysis further reveals differential transcript use with progression, including differential transcript use of LPCAT2 in HGSOC tumor associated macrophages. The gene fusion profile of 36 pairs of HGSOC samples also reveals preserved expression of the potentially disease-relevant gene fusion CCDC6-ANK3 in multiple patients, also detected in cisplatin-resistant HGSOC lines. These results support the growing body of literature implicating altered tumor-TME interaction and epigenetic mechanisms in the evolution of HGSOC and calls for further longitudinal studies of HGSOC, and particularly the use of single cell sequencing to parse the contributions of multiple interacting cell.

Table of Contents

List of Tables

List of Figures

Preface

Acknowledgements

I'd like to that everyone who offered their expertise and valuable time to teach me the background, technical, and analytical skillset I used in carrying out this research. The mentorship, friendship, and sense of community I experienced from the Integrated Systems Biology Program, The Lee/Oesterriech lab, the Magee Women's Research Institute, the University of Pittsburgh School of Medicine and the broader University of Pittsburgh community were invaluable both in helping me perform this research, but in helping me grow as a researcher, caregiver, and member of the diverse and generous scientific community.

1.0 Background and Significance

1.1 High Grade Serous Ovarian Cancer (HGSOC) - Overview

Ovarian cancer(OVCA) is projected to kill over 14,000 women in 2020 according to the NIH Surveillance, Epidemiology, and End Results Program(SEER). While the umbrella term "ovarian cancer" includes epithelial and non-epithelial subtypes, the distribution of ovarian cancer fatalities is skewed towards the most common and most deadly subtype of ovarian cancer, high-grade serous ovarian carcinoma (HGSOC). This disease accounts for up to 80% of lives lost annually to ovarian cancer. No effective screening methods exist for the detection of HGSOC, so it is most often detected at a late stage via incidental imaging or in patients already presenting with genitourinary or gastrointestinal symptoms[1,2]. Initial management of ovarian cancer typically consists of a biopsy followed by surgical debulking[3]. Neoadjuvant chemotherapy has increasingly been used to reduce tumor volume prior to debulking, as residual tumor volume has been shown to be a major predictor of survival[4], but the contribution of neoadjuvant chemotherapy to the well documented evolution of HGSOC is under-studied. Treatment with the standard regimen of platinum and taxane-based chemotherapy typically leads to remission in HGSOC, but recurrence is nearly inevitable, causing a 5 year survival below 50%[2].

The limited set of drugs with proven efficacy further impedes management of recurrent, chemoresistant disease. The standard platinum and taxane therapy was developed for HGSOC over 30 years ago, and studies of alternative therapeutic approaches have only started to show promise in the past decade. PARP inhibition with olaparib[5,6], rucaparib, and niraparib[7], angiogenesis inhibition with bevacizumab[8], epigenetic modulation with HDAC inhibitors[9], and cell cycle regulation with CDK

inhibitors[10] are all actively being explored. Both PARP-Inhibition and VEGF inhibition have shown sufficient clinical benefit to be approved for treatment, but despite these successes, ovarian cancer survival remains low. These alternative treatments could all feasibly exploit known properties of HGSOC, but determining which patients benefit from specific therapies remains an unsolved challenge.

1.2 The Ovarian Cancer Genome

Researchers have recently turned to the ovarian cancer genome to answer the question of how to improve patient survival, with illuminating results. Cancers like ovarian clear cell carcinoma, endometrioid carcinoma, and low grade serous, classified as Type I OVCA, present as slow growing often chemoresistant tumors but patient survival exceeds that of HGSOC. Type II OVCA, which includes HGSOC, likely originates from fallopian tube epithelium and differs in its chemosensitivity, genomic profile, and overall patient survival. HGSOC accounts for over 70% of ovarian cancer cases overall[2]. It is the recurrence of this subtype of ovarian cancer that often presents with chemoresistance and causes patient death. To reduce the number of ovarian cancer deaths, many groups focus their research on HGSOC, due to both its prevalence and aggressive presentation.

The HGSOC genome is among the most unstable cancer genomes[11]. The disease is defined by near ubiquitous *TP53* mutations, frequently accompanied by DNA-repair defects or *CCNE1* amplifications. Subtypes have been described in HGSOC, with early studies suggesting up to seven distinct subtypes[12], but current advocates of HGSOC subtypes hold that 3 major subtypes exist: an immune active subtype, a mesenchymal subtype, and mixed[13]. The mesenchymal subtype,

associated with mutation and expression of stroma-associated genes has the worst prognosis, while the immune-active subtype has better odds of survival. It must be noted however that the validity of HGSOC subtypes is far less established than that of breast cancer.

Exactly what alterations beyond *TP53* mutations, HR defects, and *CCNE1* amplifications contribute to the onset and phenotype of HGSOC remains to be fully characterized. The disease has a relatively low burden of point mutations, while its DNA structural variation (SV) load is second only to breast cancer. HGSOC has long been associated with extrachromosomal DNA in the form of "double minute chromosomes" which can carry and express oncogenic material[14,15]. Additionally, the "Tandem Duplicator" phenotype, characterized in many cancers, is common in HGSOC[16]. Typically associated with the classic mutations of HGSOC(*TP53, BRCA1, CCNE1*), the tandem duplicator phenotype can be further reduced in HGSOC to several mutational signatures, the size and spacing of detected SVs corresponding to the n specific mutations constellations of mutations driving the phenotype[17].

While SV at the genome level is a potential mechanism of HGSOC evolution, additional regulatory mechanisms also exist. Studies of the epigenetics of ovarian cancer have shown epigenetic factors, including histone methylation and acetylation profiles and expression of regulators of histone markers, to be prognostic[9]. Histone deacetylase activity in OVCA has been implicated dysregulation of cyclins, cadherins, and secreted T-cell regulators. Elevated repressive hypermethylation in promoters of tumor suppressors including *BRCA1* have been found in ovarian tumors relative to normal tissue. The presence of both repressive(H3K27me3) and activating (H3K4me3) methylation in a "poised bivalent chromatin" appears to be a characteristic of genes downregulated with ovarian cancer[18,19]. In light of these findings, the use of drugs targeting epigenetic factors have been tested in small clinical trials, with promising results for the

methyltransferase inhibitor decitabine, but these studies have been complicated by the severe side effect profile of these drugs[9].

Looking beyond the genome has been informative in studies of HGSOC as well. Expressed gene fusions have been detected in the disease[20–23], most promisingly promiscuous ABCB1 fusions detected by Bowtell et al first in their landmark characterization of HGSOC progression in 2015, and later confirmed to be common in the AOCS cohort[24,25]. Characterization of the HGSOC proteome has primarily supported the findings from the genome and transcriptome, implicating JAK-STAT signaling, ECM signaling, and DNA repair pathways as key players in the disease[26].

A holistic view of the evolving -omic profile of HGSOC suggests that broad regulatory factors like TF and miRNA may be worthwhile targets in HGSOC[27]. While diverse mutations in a range of genes likely contribute to the presentation and evolution of HGSOC, many of these changes occur in common, well defined and targetable pathways[28]. Profiling HGSOC cases to determine which pathways are most frequently affected by potential driver mutations, and determining which proteins play major regulatory roles in the pathway offers an appealing path to improved HGSOC therapy.

Chemoresistance in cancer is thought to emerge through many mechanisms, including: reducing intracellular drug concentration, drug inactivation, and activation of DNA damage and repair pathways[29]. The problem of chemoresistance in HGSOC has been attacked from many angles. Characterizing DNA copy number variation (CNV) by sequencing and microarrays has implicated many CNVs in cancer associated genes[30], with *CCNE1* as the most distinct and well-studied amplification[25]. These analyses have produced experimental chemosensitivity signatures for HGSOC, predicting PARP-I response and cisplatin response based on copy number profile[31]. Transcriptomes have been similarly informative, demonstrating the frequent upregulation of genes

associated with the Mesenchymal subtype signature in chemoresistant HGSOC[12,32]. The response to neoadjuvant chemotherapy has also been interrogated, showing treatment-promoted upregulation of DNA repair pathways and activation of Wnt and TGFbeta pathways[33].

Attempts to characterize structural changes in the genome between primary and metastatic HGSOC have produced inconsistent results, some studies showing an enrichment in CNV in metastases, but others showing an equal or lower CNV burden in HGSOC metastases[25,34]. Single cell analysis of HGSOC suggests metastases are relatively depleted in HGSOC epithelial cells, with enrichment in immune and stromal components[35].

Genomic and transcriptomic data has been used to predict chemosensitivity, metastasis, and survival in HGSOC, with some clear prognostic factors found. HOX gene expression has been shown to predict recurrence in HGSOC[36], as has elevated expression of ECM proteins and epigenetic regulators. HGSOC stem cells, expressing ALDH and CD133, have been shown to be regulated by the wnt-bmp axis, with signaling between tumor cell and mesenchymal stem cells derived from normal stroma promoting stemness and chemoresistance[37,38].

1.3 **Fusion Transcripts in HGSOC**

Gene fusions have long been studied as drivers of cancer. Discovery of *BCR-ABL*, the Philadelphia Chromosome, in acute myelogenous leukemia (AML) provided oncologists with both a biomarker and a drug target[39]. In prostate cancer, *TMPRSS-ERG* fusions are prevalent[40], and the *WES-FLI1* fusion characteristic of Ewing's sarcoma is also frequent in that tumor[41]. *CCDC6-RET* is found in nearly 50% of papillary thyroid carcinomas[42]. More recently, *NTRK3* fusions have

successfully been targeted with kinase inhibitors in lung cancer[43], in a success for precision medicine.

The appeal of fusions as drivers of malignancy stems from the diverse mechanisms by which they might drive tumors. Domains gained and lost by fusion genes may remove or activate key regulatory elements, enabling pro-malignant signaling cascades. Fusions may also produce truncations of the 5` partner gene, again potentially creating a truncated protein with possible neo-function. Finally breakpoints at the 5` or 3` end of the fusions of interest may remove a gene from its appropriate regulatory context[44].

Gene fusions, which may ultimately produce chimeric fusion proteins, can emerge at the genomic or transcriptomic level. All SV are capable of producing gene fusions, and no single mechanism has been implicated. SVs are well documented drivers of atypical gene expression, and recent pan-cancer studies have shown their enrichment in relevant oncogenic pathways[28,45]. Frequently, fusions containing TSGs show reduced expression, while the opposite is true of oncogenic fusions. At the time of transcription, readthrough fusions may also be produced between adjacent genes when the appropriate stop codon is missed or skipped. Transcripts can also fuse via aberrant transsplicing, frequently producing fusions with a "2 and 2" pattern of the first two exons of the 5` gene and the last two exons of the 3` gene.

Kinase fusions are particularly common and appealing targets that can be activated by any of these mechanisms[44]. *CCDC6-RET*, a kinase fusion extremely common in papillary thyroid carcinoma, causes inappropriate RET signaling that drives proliferation. Thyroid cancer is one of the most fusion enriched cancers, though like prostate cancer it has several highly recurrent fusions[46–48]. Breast cancer and ovarian cancer differ in that they lack highly recurrent fusions, but the instability of their genomes leads to the accumulation of diverse fusions of unclear importance.

Despite this heterogeneity, some fusion events have been shown to promote clinically relevant phenotypes. *NTRK3* fusions in secretory breast carcinoma are potentially druggable alterations also seen in phenotypically similar salivary gland tumors[49,50]. *ESR1* fusions have been detected and validated as drivers of endocrine resistance[51,52]. One fusion that recurs primarily in the basal-like subtypes of breast cancer, *ESR1-CCDC170*, has been detected in both breast and ovarian cancer[51,53].

Methods that target both the transcriptome and genome have been developed for the detection of gene fusions, though the field now focuses primarily on expressed chimeric transcripts detection via RNA seq[54–57]. Any detection of SV with nucleotide resolution will allow possible fusions to be detected from WGS or exome sequencing, but in cancers like HGSOC, the high SV burden complicates separating pro-malignant "driver fusions" from inert "passenger fusions"[58]. By detecting chimeric transcripts via RNA-Seq, researchers can limit their search to fusions that are at least expressed in the transcriptome.

Numerous fusion callers have been developed, including FusionCatcher[55], STARFusion[59], TopHat-Fusion[60], Pizzly[61], and Ericscript[62]. The outputs of these fusion callers can be highly variable, motivating the development of integrated, multi-caller pipelines. Benchmarking these tools suggests FusionCatcher has a valuable mix of speed and accuracy[54], but studies focused on detecting clinically relevant, active fusions attempt to integrate the outputs of multiple fusion callers to reduce the high false positive rate of many fusion callers[63]. Once detected, bioinformatic validation of fusion expression requires mapping supporting reads to the breakpoint and confirming the specificity of the read support, which can be performed with IGV, SVViz[64]s, or ChimeraViz[65], an application designed for fusion visualization. Prioritization of possible driver fusions often involves mapping the reads contributing to the detected fusion against its wildtype

sequence to determine domains gained or lost. Oncofuse[66] and Pegasus[67] are programs designed to assign a driver likelihood score to fusions based on the similarity of their domain profiles to those of known malignancy promoting fusions. Following bioinformatic prioritization and biological detection via PCR of purported driver fusions, in vitro characterization is necessary to argue for a genuine effect of fusion expression.

The genomic instability of HGSOC has prompted many researchers to attempt to implicate fusions. Early work suggested that recurrent fusions including *BCAM-AKT2*[68], *CDKN2D-WDFY2*[69], and *ESRRA1-C11orf20*[23] might be valuable targets or HGSOC biomarkers, but the prevalence of those fusions was likely overestimated in all of those studies. Notably however, *ESR1—CCDC170*, a purported driver of basal-like breast cancer, has also been found in multiple studies of HGSOC[33].

The lack of highly recurrent fusions in HGSOC suggests that if fusions are effectors of its phenotype, their action is best understood at the population level, with diverse individual fusions acting through common oncogenic pathways. CNV are common in many oncogenically relevant pathways in HGSOC, and the same is true of fusions. A recent study of fusions in primary HGSOC is associated with dysregulation of the expression of nearby genes, and CNVs are well established drivers of altered gene expression[70]. These findings suggest that fusions in oncogenically relevant pathways dysregulate of those pathways, generating the phenotype in HGSOC. A pathway-based, systems biology approach to characterization of HGSOC fusions may be a powerful lens through which to view structural variation as a driver of HGSOC evolution.

A fusion-centric approach to HGSOC evolution also has many limitations. Searching for fusions in bulk RNA seq may be difficult if the driver fusion is only present in a subclonal population of cancer cells. Subclonal expansion has been postulated as a major mechanism of

HGSOC evolution[71], and in several studies the heterogeneity of metastatic tumors was actually reduced in comparison to the primary[35]. Additionally, many fusions have been detected in normal tissue, but it is unclear if they have oncogenic potential. Additional passenger fusions might be generated through splicing or readthrough mechanisms, but each of those mechanisms has also been implicated in generating pro-malignant fusions as well. Finally, fusion-centric study also offers limited perspective into the contributions of the tumor microenvironment to the disease.

1.4 **Study Objectives**

This work integrates and builds upon earlier transcriptomic studies of HGSOC progression with an emphasis on gene-set or pathway-level changes with progression. I further attempt to determine possible contributions of expressed chimeric transcripts to the observed systems-level changes in gene expression associated with HGSOC progression. My major objectives are as follows:

- Meta-analysis of HGSOC Progression via RNA-Seq
    - Characterize both differential gene expression between "progressed" and "unprogressed" individual HGSOC datasets and an appropriately normalized and batch-corrected integrated dataset
    - Contrast gene expression changes between ovarian primary tumors and early metastases with those detected between primary and recurrent HGSOC
    - Identify key pathways of and gene networks associated with progression
- Integration of HGSOC Fusion-Calling results with Gene Expression Results

- Develop a sensitive fusion detection pipeline to run on datasets with compatible sequencing parameters. Summarize those results with an interactive R-Shiny dashboard to facilitate exploration of fusion detection results
- Correlate the expression of any possible driver fusions with broader gene expression changes using the differential gene expression pipeline developed in 1.

2.0 Methods

2.1 Assembling the HGSOC pairs cohort through sequencing and download of public data

Searching the sequencing Read Archive(SRA[72,73]) using the terms “Ovarian cancer”, “pairs”, “recurrence”, and “metastasis” showed 5 datasets with usable sequencing data. The datasets varied in their preparation and final parameters. In addition to the publicly available data, we sequenced 19 pairs of samples in-house.

Sequencing of in-house HGSOC Pairs

Sample Acquisition

19 patient matched pairs of frozen tumor tissue were obtained from the Pitt Biospecimen Core (in collaboration with Dr Robert Edwards) and through collaboration with Roswell Park Cancer Institute(RPCI). RNA was extracted for sequencing with Quiagen’s RNAeasy RNA extraction kit according to standard protocols and extracted RNA quality was assessed using a NanoDrop spectrophotometer. Samples passing preliminary QC were provided to the University of Pittsburgh Genomics Research Core for further QC, library prep, and sequencing. RNA quality (fragment length distribution, RNA integrity, sample purity) was quantified by the Pitt Genomics Research Core and samples meeting the minimum thresholds for quality were processed for sequencing using the Illumina Tru-seq v2 paired-end stranded library preparation kit. Generated libraries were sequenced on an Illumina NextSeq 500 with a read length of 150bp, a target insert size of 400bp and a target read depth of 100x. The sequencing results were converted to fastq files that were transferred to long term and working storage on the computing cluster provided by the University of Pittsburgh Center for Research Computing.

Summary of other Datasets Employed:

Metastasis Datasets

The "Metastasis Datasets" were generated from 2 recent publications,[74,75] totaling 23 pairs. Both studies collected tissue from both the ovarian primary tumor, the "unprogressed" samples, and a metastasis, the "progressed" samples, during the initial debulking. Raw fastq files were downloaded from European Nucleotide Archive(ENA)[76] using curl in December 2019 . The integrity of downloaded data was verified by comparison of md5 checksums between downloaded files and the reported md5 checksums on ENA. The code used to download and verify download integrity is on github at https://github.com/johnalanwillis/ovarianCancerProject.

Recurrence Datasets

The "Recurrence Datasets" consisted of 3 datasets downloaded between 2016-2019 and the in-house dataset described above. Samples in this group consisted of HGSOC samples that had recurred, the "progressed" samples, and paired samples from a prior debulking, the "unprogressed" samples. The "Recurrence Datasets" varied more in their sequencing parameters and download methods than the "Metastasis Datasets". The AOCS and TCGA datasets were obtained by downloading the aligned bam files from the SRA using SRAtools and converting them to fastq files with the bam2fastq function provided by samtools[77]. The OCTIPS dataset was not available in raw form, but the gene-level counts matrix was downloaded from the ENA after obtaining permission from the original study authors.

MSC dataset

To investigate the contributions of the microenvironment to HGSOC evolution, additional raw RNA Seq data from a study comparing cancer associated mesenchymal stem cells(CA-MSCs) with MSCs from normal omentum. As with the Metastasis Datasets, the raw fastq files were

downloaded from the ENA (PRJNA384578, PRJNA564846) using curl in June 2019, and download integrity was verified by comparing md5 checksums.

CCLE dataset

To investigate the relationship between chimeric fusion transcript expression, overall gene expression profile, and bench-testable phenotypes, raw RNA-Seq data for ovarian cancer cell lines from the Cancer Cell Line Encyclopdia (CCLE)[78,79] was downloaded from the SRA using SRAtools in 2016 and the bam files were converted to fastq with samtools bam2fastq function.

2.2 Data Processing For Gene and Isoform Expression Analysis

QC of Raw and Processed Sequencing Data

Quality of raw sequencing reads was quantified using the java application FastQC(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)[80]fas. Consistency of reported parameters including fragment length and read depth with target values was checked prior to further analysis. Alignment quality was evaluated following alignment with STAR using dataset-specific alignment parameters using samtools(http://www.htslib.org/). Integrated read and alignment level QC was orthogonally verified with the java application Qorts(https://hartleys.github.io/QoRTs/)[81]. QoRts reported values were compared with samtools and fastqc reported values for consistency across methods. By-sample QC summaries were generated using the java application MultiQC(https://multiqc.info/)[82]. All QC operations were performed using BASH scripts that can be found on github at https://github.com/johnalanwillis/ovarianCancerProject

Data Processing For Expression Analysis

All expression analysis operations were performed using BASH scripts and within R markdown documents that can be found on github at https://github.com/johnalanwillis/ovarianCancerProject

Isoform Expression was quantified from raw reads using SALMON v0.14 (https://salmon.readthedocs.io/en/latest/salmon.html)[83,84] against an index generated from the gencode v29 transcriptome and genome[85]. Salmon employs a k-mer based pseudoalignment approach to isoform level expression quantification. Rather than mapping each reported read to a genome/transcriptome region with the highest sequence identity, Salmon breaks reads into "k-mers" of pre-defined length classifies each read as a likely member of a subset of "equivalence classes", or "k-mer" distribution profiles. This approach is both faster, and more accurate than alignment based read counting techniques. Isoform expression was studied using the R package IsoformSwitchAnalyzer v1.8[86]. The package reads transcript level quantification results as output by RSEM, cufflinks, SALMON, etc, and facilitates differential isoform use testing and annotation using a range of tools through a unified interface. The counts matrix generated by SALMON records transcript level counts, but most RNA seq Analyses are performed with the intention of uncovering genes of interest. Transcript counts matrices are transformed to gene counts by collapsing isoform counts to the gene level and correcting for gene-length effects. The resulting gene counts matrix derived from transcript quantification is more accurate than direct counting of reads mapping to genes without consideration of the contributions of different isoforms to the overall count.

Isoform abundances generated from Salmon's psuedomapping approach are collapsed into gene level counts with transcript length aware offsets using the txiMeta[87]t package when possible. txiMeta improves on the tximport approach by tying linking Salmon quantification results to their

target transcriptome by including a transcriptome signature in the quantification results. Each transcriptome, ie Gencode v.29, has a unique signature, and matching a quantification signature to a transcriptome signature simplifies the analysis process and prevents accidental mismatch of references between alignment and analysis.

Differential Expression Analysis

Gene expression level QC was examined using pcaExplorer[88], which wraps base R's princomp output in a range of visualization functions. Plots of samples along their principle components were generated using the function pcaPlot to determine if batch effects were present and if samples separated by covariates of interest along any of the principal component of the variance.

Differential gene expression analysis was performed with EdgeR[89], which employs a generalized linear model-based approach to differential expression testing. EdgeR, like DESeq2, assumes gene expression follows a negative binomial distribution, and uses an empirical bayes approach to derive sample and gene-level dispersion parameters for final modeling. To generate a counts matrix usable in visualizations or other statistical applications, the cpm() function from EdgeR was used. This generates a Trimmed Mean of M-values normalized matrix that corrects for differences in library size between samples. Gene level counts were prepared from the transcript level quantification matrix output from salmon using tximport, and differential expression testing was performed according to standard protocol using the EdgeR glm pipeline.

The limma[90] package offers another linear model-based approach to gene expression analysis which was originally developed for use with microarrays. For RNA-Seq experiments, it is often applied with the Voom[91] transformation, which transforms the gene counts

to to account for the measured mean-variance trend at both the sample and gene level. Differential expression analysis is performed on the transformed counts using a linear model-based approach.

The third and final differential expression technique employed was DESeq2[92]. DESeq2 normalizes between samples by calculating a by-sample size factor from the median by-gene geometric means. Scaled counts are then transformed using either the vst() or rlog() transformations to correct for the mean-variance trend, and a linear modeling approach is used to test for differential expression, based on the assumption of genes dispersed according to a negative binomial distribution.

Overrepresentation and Gene Set Enrichment Analysis

To identify patterns in gene expression between conditions and within clusters, overrepresentation analysis(ORA), was performed using the packages Clusterprofiler[93]. By testing for deviation in gene set membership for an experimentally uncovered gene set against a hypergeometric distribution, Clusterprofiler reveals large alterations in gene set activity.

Gene Set Enrichment Analysis improves on the ORA approach to gene set testing by comparing the distribution of gene expression in a gene set against a reference distribution. This method is more sensitive to subtle, coordinated changes in many genes. The package fGSEA[94] speeds up the typical gsea process by using preranked gene sets.

Gene Set Variation Analysis, GSVA[95], extends the gene set approach developed in GSVA to determine the relative enrichment of gene set activity in an individual sample against others in a cohort. By comparing the distribution of genes in a single sample against other genes in a cohort, and the relative distribution within the cohort to a predefined gene set, it generates a by-sample score for gene sets of interest. GSVA can transform a gene expression matrix into a gene-set

activation matrix, which can then be used in downstream applications, like for differential activation testing using limma.

Network analysis with VIPER and WGCNA

Virtual inference of Protein activity by Enriched Regulon Analysis(VIPER) and MAster Regulator INference Analysis (MARINA)[96] from the package viper facilitated the inference of "master regulator" gene activity from gene expression matrices using a mutual information approach. "Master regulators", or genes whose expression was highly correlated with the activity of several other genes, were inferred from HGSOC gene expression from the TCGA using the ARACNE algorithm[97]. These collections of mutually regulated genes and their "master regulators", termed "regulons" were accessed from the package TCGARegulons[98] and used to uncover regulons strongly associated with progression using the marina() function. Gene expression matrices were also transformed into regulon activity matrices using the viper() function, and the transformed matrices were used with limma to test for differential regulon activation associated with progression, using the standard limma linear modeling approach.

Correlated Expression network analysis with WGCNA

Identification of correlated gene expression modules using Weighted Gene Correlation Network Analysis(WGCNA)[99,100] was also performed. In WGCNA, a scaled correlation matrix is generated and the goal of hierarchical clustering by gene. The modules are identified by the hierarchically clustered gene-correlation, and modules are collapsed together based on the similarity of module eigengenes. These modules can then be correlated to covariates of interest by transforming the by-sample gene expression matrices into module activation matrices.

Methods for Exploring and summarizing gene expression results

Findings from HGSOC paired data analysis were explored in HGSOC TCGA cases as an external validation set. TCGABiolinks[101] provides a standardized interface for processing TCGA data. I used this package to obtain gene counts and sample metadata for TCGA HGSOC cases. Batch effects resulting from combining heterogenous datasets were corrected using the removeBatchEffects function from limma, which employs a linear modelling approach to regress batch effects out from a gene expression matrix. Success of batch effect correction was confirmed by PCA with the PCAExplorer package.

The ComplexHeatmap[102] package provides a powerful and flexible interface to visualize data with attractive and detailed annotations. In addition to the heatmap and annotation functions, the package also generates visualizations of set relationships with the UpsetPlot function. This package was used to generate heatmaps summarizing differential expression results and Upset plots to visualize the intersections between differential gene expression tools.

The statistical package cluster[103] provides numerous supervised and unsupervised clustering methods and cluster quality analysis functions silhouette width and elbow plots in R. Cluster stability was further characterized by clustering of bootstrapped subsamples of the data using the package ConsensusClusterPlus[104]

Survival analysis was performed by a Cox proportional-hazards model fitting patient survival by quantized gene expression or regulon cluster membership with the package Survival. Survival plots were then generated with survminer[105], which plots survival analysis results using the ggplot2 framework, rather than base R graphics.

Methods for Fusion Detection and Exploration

The contribution of expressed fusion genes to the progression of HGSOC was investigated through the use of fusion detection algorithms when sequencing parameters were appropriate. Overall 36 pairs of samples were sequenced with paired-end libraries and long read length, in the WCRC-RPCI, AOCS, and TCGA cohorts. To improve the sensitivity of the screen, three fusion detection and annotation tools were employed, and their combined results were then filtered using a custom R Shiny dashboard.

Expressed gene fusions were detected using FusionCatcher[55]. FusionCatcher detected expressed gene fusions by aligning raw sequencing reads against a transcriptome, using a battery of aligners including BWA, STAR, and TOPHAT. Reads that appear split across the transcriptome or pairs that lack an adjacent partner are collected and their mapping against putative fusions is then quantified. The generated fusion list is subsequently compared to an internal database of known false positives.

Expressed gene fusions were detected using FusionZoom. FusionZoom detected expressed gene fusions by aligning raw sequencing reads against a transcriptome, using Tophat to perform alignment. Putative fusions are scored for their driver-likelihood to generate a list of likely drivers from the tophat output.

Expressed gene fusions were detected using STARFusion[59]. STARFusion detected expressed gene fusions by aligning raw sequencing reads against a transcriptome, using STAR to perform alignment.

The tabular output of the fusion calling battery was parsed, cleaned, and harmonized using custom R scripts. The cleaned fusion detection results were then explored using an R ShinyDashboard[106], named FusionExplorer, as described in the Results.

Methods for Bench Validation of Detected Fusions in human tissue

To confirm the expression of bioinformatically detected fusions in-vivo, PCR primers were designed against the putative novel exon-exon junctions in the fusions of interest. Primers were designed using Primer3[107] with standard parameters. The desired primers were ordered from Genewiz, and stored in 20uM aliquots at -20C.

RNA was extracted from frozen sections of HGSOC tumor tissue using the Quiagen RNeasy FFPE kit according to standard manufacturers protocols. Prior to final elution, an on-column DNA digest was performed with the provided DNAase. The concentration and quality of eluted RNA was quantified by UV spectroscopy on the Nanodrop 2000, and samples failing to meet quality cutoffs were re-extracted from additional sections. cDNA was generated from the eluted RNA using Takara Primescript and the RNA was stored at -80C.

Patient Samples were screened for novel fusion junctions via RT-PCR with BioRad's SYBRGreen Master Mix. The reaction was prepared on ice with sterile, nuclease-free H20 and the previously obtained primers and cDNA. Optimal reaction parameters were determined by performing test reactions at a range of concentrations and cycler temperature was determined with gradient PCR. Final reactions were run with a fusion-negative control sample, typically either OVCAR3 or MCF7 cDNA, and primers for B2M, GAPDH, or ACTB were used as positive controls. 1% polyacrylamide gels were run for initial PCR optimization, through gels with concentrations 0.5%-1% were used for gel extraction and sequencing.

To confirm the identity amplified products, bands were cut from PCR gels using sterile scalpels under UV light and collected in sterile, nuclease-free PCR Tubes. Excess polyacrylamide was trimmed from the bands and amplified sequences were purified using the Monarch gel extraction kit**.** The quality of extracted products was determined by UV spectroscopy on the

nanodrop and samples meeting minimum quality and concentration requirements were sent for Sanger Sequencing by GeneWiz**.** Sequencing results from gel extracted fusion junction bands were downloaded and the sequences from high confidence results were aligned against putative fusion junctions reported by the fusion calling algorithms employed in the in-silico screen.

2.3 **Methods for Bench Validation of Detected Fusions in Cell lines**

Culture of NIH_OVCAR3, SKOV3, BT474 Cells

Cell lines were obtained from ATCC and stored at -80C in LN2 until needed. Periodic mycoplasma testing was performed using a MycoAlert Mycoplasma Detection. Once thawed for use, cells were maintained at 37°C in 5% CO_2. All culture media and media supplements were obtained from Life Technologies. NIH_OVCAR3, SKOV3, BT474 cells were cultured in Dulbecco's modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS).

2.4 Methods for characterizing the contribution of fusion expression to phenotype

siRNA Knockdown of CCDC6--ANK3

To assess the contribution of CCDC6—ANK3 expression in NIH_OVCAR3 cells, siRNA were designed targeting the CCDC6—ANK3 exon-exon junction expressed in NIH_OVCAR3 cells using dharmacon's siRNA Design tool and were ordered from Dharmacon, along with non-targeting control pool siRNA. Once obtained, siRNA stocks were aliquoted and stored at -20C. siRNA kd of fusion Expression using lipofectamine RNAiMax liposomal vectors

CCDC6—ANK3 Expression in NIH_OVCAR3 cells was knocked down via liposomal transfection with lipofectamine RNAiMax via reverse transfection. siRNA-liposome complexes in OptiMem media were prepared according to the manufacturers protocol at in 6 well plates. Cells were subsequently added to the growth media to achieve a target count of ~125K cells per well. After 24hrs of kd, siRNA kd media was replaced with standard growth media.

confirmation of fusion kd via qPCR

Fusion knockdown (kd) was confirmed by qPCR, using Primers designed against the exon-exon junction targeted by the siRNA. RNA was harvested from siRNA kd cells, untreated and vehicle-treated, and non-targeting siRNA treated cells 72 hrs after reverse transfection using the Quiagen RNAeasy kit according to standard protocols, and cDNA was generated using Takara Primescript. The qPCR reaction mixture was prepared on ice in 96-well plates from SsoAdvanced Green Supermix,and run on the CFX96 thermocycler (Bio-Rad) according to the manufacturer's protocol. Samples were normalized to a battery of housekeeping controls and CCDC6—ANK3 kd

was assessed using the $2^{-\Delta\Delta C_T}$ method against Vehicle treated control.

Determining Growth Effects of fusion kd with Prestoblue

Once the efficacy of siRNA kd of CCDC6—ANK3 was demonstrated, the effects of anti CCDC6—ANK3 siRNA treatment on growth in NIH OVCAR3 cells was assessed using Prestoblue and FluoReporter. In both cases, growth assays were performed by reverse transfecting NIH_OVCAR3 cells with siRNA or control for 24 hrs and seeding cells in normal growth media at 10K cells per well in 96 well plates. 5 replicates per condition were plated and samples were collected at 24, 48, 72, 96, and 120hrs after plating. For FluoReporter Quantification, plates were collected at each timepoint, growth media was dumped, and the plates were stored at -80C until scoring. Fluoreporter quantification was performed by lysing stored cells by freeze-thaw cycling, adding DNA intercalating fluorescent dye, and scoring fluorescence of siRNA kd relative to control. PrestoBlue quantification was performed at the time of collection for each timepoint by adding prestoblue dye in 1:1 ratio to the growth media, and scoring fluorescence at 5, 15, 30min, 1hr, 2hr, 4hr, and 8hrs of incubation. The fluorescence by incubation time was plotted and the fluorescence score detected at the midpoint of the exponential growth phase of the saturation curve was taken as the true value. Final scoring compared siRNA kd with vehicle.

3.0 Results and Discussion

3.1 Characterization of available HGSOC Sequencing Data

Aggregating all available paired sets of RNA-Seq Data from the "Metastasis Datasets" and "Recurrence Datasets" yielded a collection of 118 pairs of "progressed" and "un-progressed" HGSOC. While the focus of these analyses is on the contrasting gene expression between HGSOC samples over time and space in the body, I also interrogated the contributions of changing isoform and gene fusion expression. Overall 8 datasets were employed in these analyses, 7/8 of which had raw fastq files available. The sequencing parameters varied between datasets, with the Metastasis datasets consisting only of short (<75bp) single end reads suitable for gene expression and transcript expression analysis. The Recurrence datasets were generated using higher read depth, longer read length, and with a paired-end modality suited to fusion detection in addition to gene/transcript expression analysis. Only the OCTIPS dataset (n=128) lacked accessible sequencing reads, but the gene counts matrix provided was still useful for gene expression analyses.

The details of the sequencing parameters can be found at the SRA sites for each project. **Figure 1A** summarizes the raw or processed data available by cohort, illustrating the gaps in the Metastasis and OCTIPS datasets.

Dataset	Group	Available Format	Publication Date	PMID
WCRC/RPCI (In-house Sequenced)	Recurrence	fastq	N/A	N/A
EGAD00001000877 (AOCS)	Recurrence	fastq	May 2015	26017449
EGAD00001000877 (TCGA)	Recurrence	fastq	Nov 2011	21720365
EGAS00001002660 (OCTIPS)	Recurrence	Gene-counts	Dec 2017	28972047
PRJNA384578	Metastasis	fastq	Nov 2019	31744494
PRJNA564846	Metastasis	fastq	Oct 2019	31600962

Table 1: Studies Explored in HGSOC Progression Meta-Analysis.

Depth and quality of sample metadata varied widely between datasets. While some datasets included potentially valuable information like chemosensitivity and tumor cellularity, most datasets lacked sufficient annotation to perform rigorous analysis based on those covariates. The contributions of those factors to expression patterns was explored in the single dataset analyses, but the focus of the aggregated analyses ("Metastasis", "Recurrence", "Progression") on the within-patient paired contrasts.

The metadata tables processed into the final metadata table and the final table used in analysis is available in **SI8-17. Figure 1B** summarizes the metadata available by cohort, illustrating the gaps in the relevant covariates including chemoresponse and tumor cellularity. **Table 2** further describes the characteristics of the datasets for annotated covariates of interest.

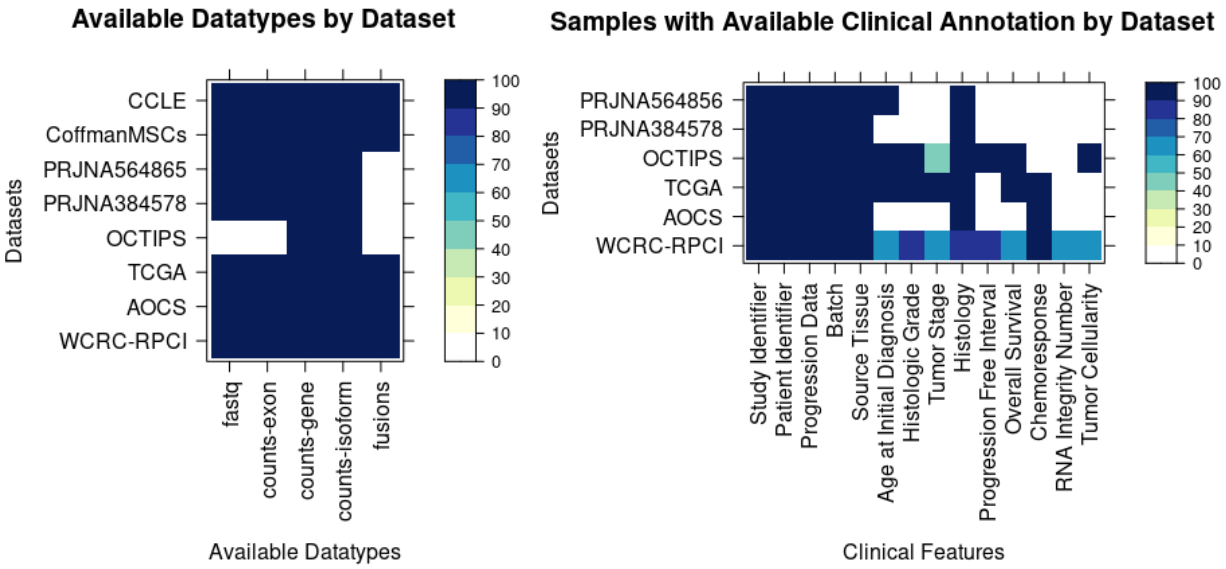


Figure 1: Summary of Available Datasets and Metadata by Dataset. Available Data by dataset for the 8 studies investigated B: Metadata available for the 6 datasets of sequenced tumors.

HGSOC Pairs Sample Characteristics

	AOCS	FROZEN	OCTIPS	PRJNA384578	PRJNA564846	TCGA
n	27	43	132	20	22	19
TIMEPOINT = late (%)	15 (55.6)	21 (48.8)	66 (50.0)	10 (50.0)	11 (50.0)	5 (26.3)
grade (%)						
2	0 (NaN)	2 (5.6)	14 (10.6)	0 (NaN)	0 (NaN)	11 (57.9)
3	0 (NaN)	28 (77.8)	118 (89.4)	0 (NaN)	0 (NaN)	8 (42.1)
4	0 (NaN)	6 (16.7)	0 (0.0)	0 (NaN)	0 (NaN)	0 (0.0)
histology (%)						
mucinous	0 (0.0)	2 (5.6)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
other	0 (0.0)	6 (16.7)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
pap serous	0 (0.0)	6 (16.7)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
serous	27 (100.0)	22 (61.1)	132 (100.0)	20 (100.0)	22 (100.0)	19 (100.0)
response (%)						
NA	0 (0.0)	5 (11.6)	0 (NaN)	0 (NaN)	0 (NaN)	5 (26.3)
resistant	13 (48.1)	2 (4.7)	0 (NaN)	0 (NaN)	0 (NaN)	0 (0.0)
sensitive	14 (51.9)	36 (83.7)	0 (NaN)	0 (NaN)	0 (NaN)	14 (73.7)

Table 2: Characteristics of samples studied by Dataset. Summary of tumor characteristic for sequenced tumor samples. Raw numbers are reported next to percentages. Datasets lacking annotation of the characteristic of interest report 0 counts and NaN percents

3.2 **Gene Expression Changes associated with Metastasis, Recurrence, and Progression**

As all datasets employed in this analysis could be simplified to a gene-level counts matrix as described in the methods, the most detailed analyses were performed at the gene expression level. Fortunately, numerous gene expression analysis methods have been developed for RNA-Seq data. Recent studies of DEA have suggested that using multiple gene expression analysis tools may offer improved sensitivity and specificity, so I used the three most popular tools available through the Bioconductor framework. DESeq2, EdgeR, and Limma–Voom are frequently used to explore differential gene expression results and offer flexible means for testing for differential gene expression. The major differences between DEA methods are the underlying model used to describe the relationship of covariates to expression value, the hypothesized underlying distribution of the reported gene counts, the means of correcting for the tendency towards heteroscedasticity in the counts data, and the tests used to define significance of differential expression. DESeq2 and EdgeR both use a generalized linear model based approach with an underlying count distribution drawn from a negative binomial distribution, which assumes that most genes are not differentially expressed. These techniques differ in their methods of controlling for the mean variance trend in the data and the tests used for determining the significance of differentially expressed genes. EdgeR is generally regarded to be a less sensitive tool, and our results bear this out. The limmaVoom pipeline differs in that it assumes a poisson distribution for counts and controls for variance using the “voom” transformation, which employs an empirical bayes estimation-based method to model the distribution of counts based on the observed counts within and between different genes and conditions

In addition to basic DEA, I used both overrepresentation analysis and genes set enrichment analysis(GSEA) approaches to extract biologically meaningful summaries from noisy gene expression data.

I also used two co-expression network-based approaches to detect patterns of gene expression associated with conditions of interest. MaRina/Aracne-AP attempt to infer the activity of known biologically relevant hub genes and their targets, termed “regulons”, by determining which genes provide significant information about the expression of other genes within the same sample. It then tests for differential activation of regulons across conditions by bootstrap subsampling to define a null distribution of regulon activity in a dataset and testing the actual data against that null. Weighted Gene Correlation Network Analysis(WGCNA) similarly focuses on networks of related genes and their hubs by generating a correlation matrix of all genes with each other, transforming that matrix to achieve a “scale free” topology, and clustering genes by their scaled correlation. Closely related genes are then integrated into modules, and the correlation of covariates with module eigengene expression by sample is tested. **Table 3** summarizes the major techniques employed in our gene expression analysis and the underlying models they employ.

Application	Model	Underlying Distribution	Variance-Stabilization	Test
DESeq2	GLM	Negative Binomial	VST, RLOG	OR, Log-likelihood
EdgeR	GLM	Negative Binomial	Log	anova
Limma-Voom	GLM	Poisson	Voom	fishers
MaRina/Aracne-AP	N/A	Mutual Information network	VST	Bootstrap t-testing
WGCNA	N/A	Scale-Free Correlation Network from Biweight Midcorrelation	VST	Fisher's transformed rho and t-test

Table 3: Analysis Methods Employed in Gene Expression Meta-Analysis of HGSOC Progression

The sensitivity of sequencing results to subtle factors like date of extraction, technician, or other "batch effects" is well documented. While many tools designed explicitly to perform differential gene expression analysis allow batch to be specified and corrected for within the analysis, many other functions, including clustering and coexpression network analyses, require a batch-corrected gene expression matrix as an input. The limma package contains a function that uses a generalized linear model approach to regress out the contributions of user defined batches to overall expression results.

To explore the contribution of batch effects to the sequencing data under investigation, I used PCAExploreR to investigate the relationship between the covariates of interest and the major sources of variance in the data. To assess the contribution of batch effect to the variance, I analyzed the data both with and without batch correction. As shown in **Figure 2**, BATCH is a dominant

contributor to the variance in the data, and the removeBatchEffects() function effectively removes the contributions of batch to the variance of the resulting, batch-effect-corrected gene expression matrix. **Figure 2C** illustrates the pearson correlation coefficient and significance between the eigengene of each principle component and the covariates PATIENT_ID, TISSUE, BATCH, and TIMEPOINT. Notably, regressing out the batch effect improves the sensitivity to detect correlations between principle component eigengenes and the remaining covariates, including revealing significant correlations between TIMEPOINT and principle components 2 and 8.

Differential Gene Expression analysis was performed using DESeq2, EdgeR, and LimmaVoom. All 3 applications use a model formula interface to specify the relationship between covariates and expression, so the same formula, Expression ~ BATCH+PATIENT_ID+TIMEPOINT, was used in each case. TISSUE was not included in our model due to a lack of coverage across our datasets. Genes differentially expressed with alpha < 0.1 were counted as significant, and the similarity between DGEA results across libraries was compared as shown in the upset plot in **Figure 4**. In all contrasts, DESeq2 was the most sensitive, and the DESeq2 results were used for all contrasts in downstream analyses including clustering and GSEA.

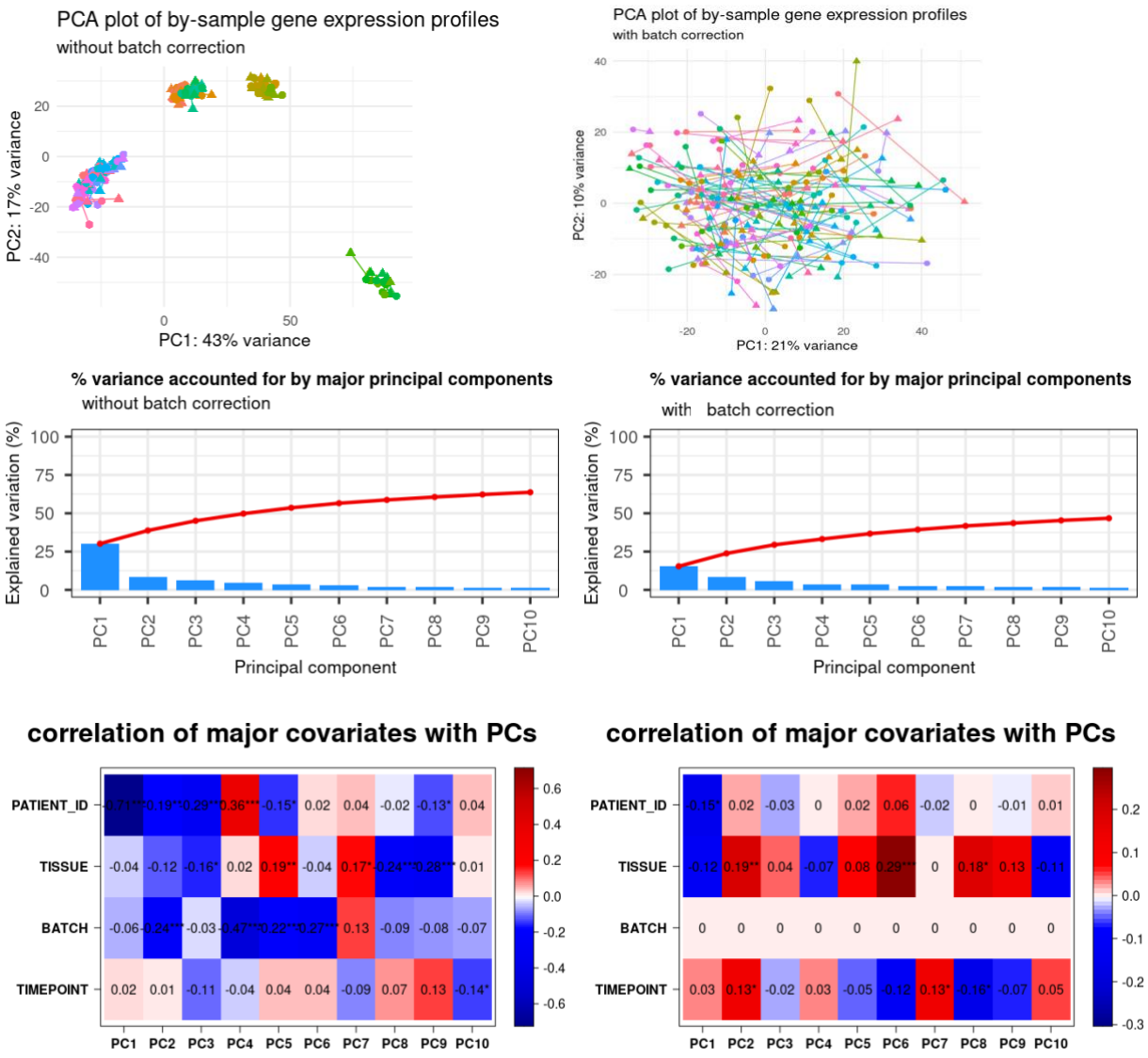


Figure 2: Batch Effects and their Correction using GLMs. (A): PCA plot of sample pairs along the two dominant principal components – Circles represent "un-progressed" samples, triangles represent "progressed" samples, and a line connects pairs of samples from the same patient. Batch is indicated by color. Batch effects are clearly illustrated on the left, and the residuals following regression for batch show an elimination of by-batch clustering along the dominant principal components (B) SCREE Plot of principal components of the data shows a reduction in batch associated variance – The magnitude of the principal components is indicated by bar height, with cumulative variance indicated by the red line. (C) Principal Component Eigencorrelation Table: The degree of correlation(pearson) between the first 10 principal components and the major known covariates with correlation significance

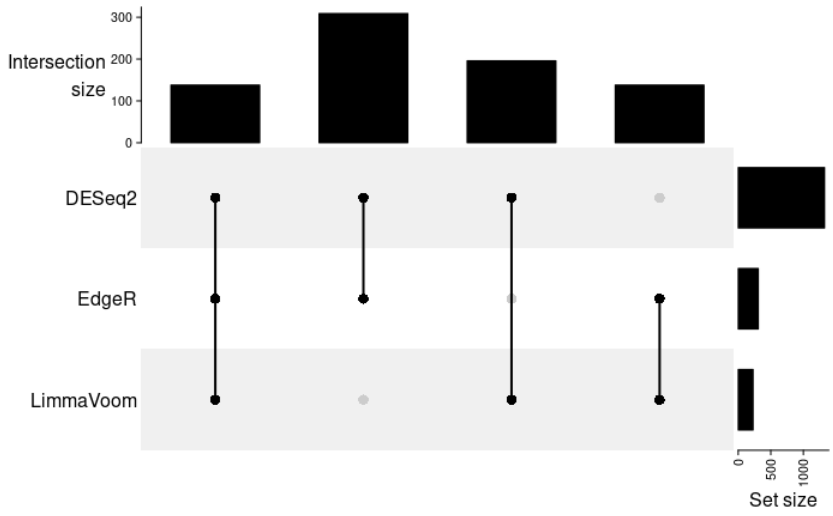


Figure 3: Comparison of Differential Gene Expression Results Between Major DEA Tools
Overlap in progression-associated differentially expressed genes

The top 25 and 1,500 differentially expressed genes were then studied to identify possible clusters associated with progression as shown in **Figure 4**. Optimal K for K-means clustering was selected by clustering of bootstrapped subsamples with ConsensusClusterPlus. K-means clustering with k=3 the full cohort by the top 25 DE genes showed one cluster strongly associated with the unprogressed samples, with relative overexpression of genes including TSPAN8 and GATA4 and underexpression of FAPB4 and SFRP2. Notably this clustering shows no clear correlation with TISSUE or BATCH. This clustering of samples was then applied to the top 1500 DE genes, which were subsequently k-means clustered with k=4 to less obvious success.

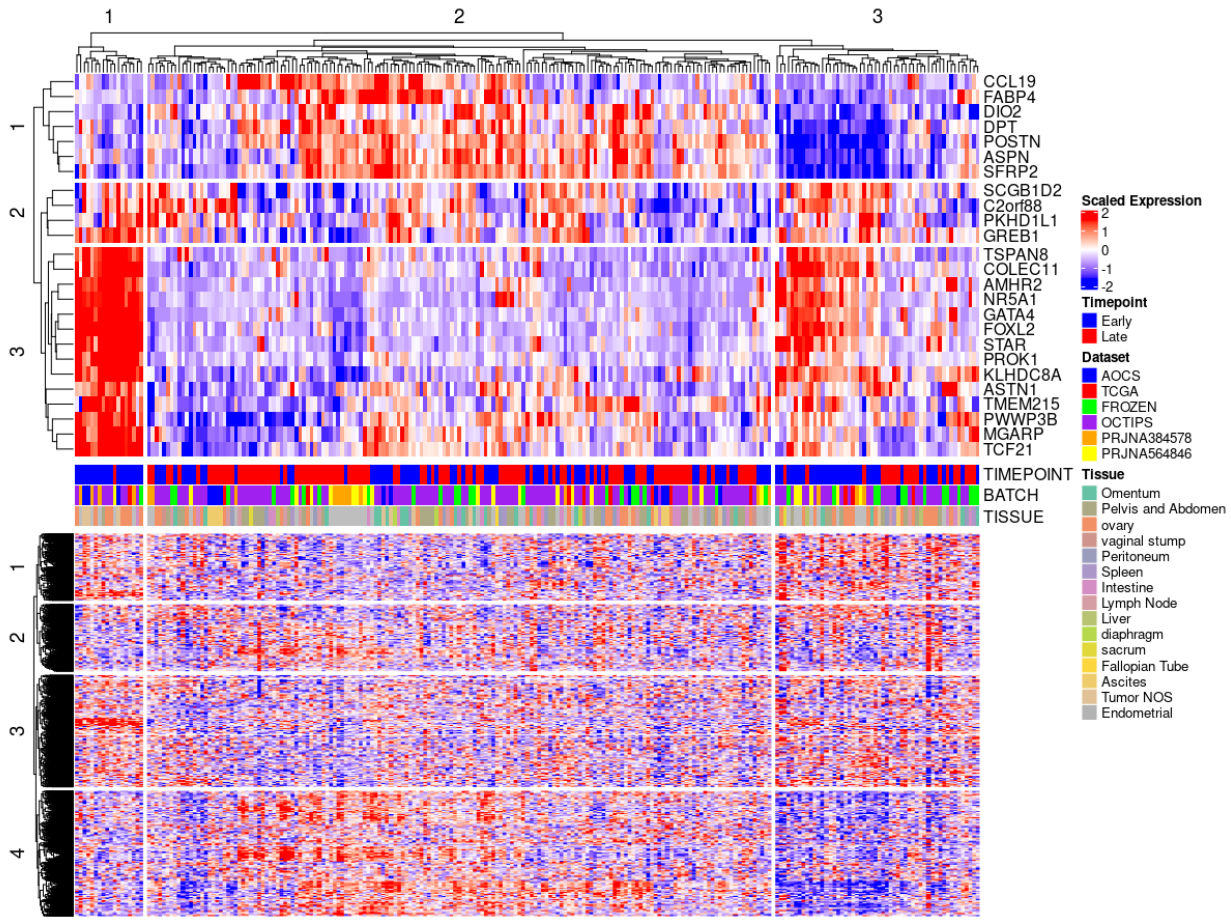


Figure 4: Clustered Heatmap of the Most Significantly Differentially Expressed Genes with Progression by Test Statistic. The top 25 differentially expressed genes as ranked by test statistic calculated by the DESeq2 method are shown in the top heatmap, biclustered using hierarchical clustering. The bottom graph shows the top 1500 differentially expressed genes ranked by test statistic, with the same column order as determined by the upper clustering and the rows hierarchically clustered

These differential expression results were used to generate a list of DE genes ranked by test statistic for use in Gene Set Enrichment Analysis (**GSEA)**. GSEA abstracts noisy gene expression data to the pathway or gene-set level, and unlike overrepresentation analysis, detects subtle coordinated changes in gene expression. I compared the gene expression results to the C6 Oncogenic and Hallmark gene sets from MSigDB, uncovering differentially regulated gene sets associated with progression in both, as shown in **Figure 5**. Hallmark gene set results showed

upregulation of EMT genes and immune-associated gene sets, and decreased activation of Myc target genes with progression. The C6 gene set also showed significant dysregulation of genes dysregulated by shRNA kd of BMI1, a polycomb group gene associated with cancer stem cell maintenance.

To determine potentially differentially active genes and pathways associated with recurrence, additional analyses were performed on the metastasis datasets and the recurrence datasets only. Performing DGE on only the recurrence datasets revealed many of the same genes differentially expressed at alpha <= 0.1, including TSPAN8 and COLEC11. Integration of those results to gene sets shows downregulation of MYC targets and alteration of BM1 activity, as detected in the integrated datasets as seen in **Figure 6**.

This BM1 activity is not detected in the Metastasis datasets, which failed to detect any C6 gene sets with differential activity, but did detect significant (alpha <0.05) differential activation of Hallmark gene sets associated with lipid processing, cell cycle progression, and EMT, as shown in **Figure 7**

An orthogonal approach to GSEA to identify differentially active gene sets across contrasts is the use of coexpression networks. Unlike GSEA, which tests the similarity of the input dataset to published gene sets, these approaches derive their own sets of coordinated genes from the data itself. I used two different coexpression analysis techniques. VIPER, or Virtual Inference of Protein activity by Enriched Regulon analysis, and WGCNA, or Weighted Gene Correlation Network analysis. While the two techniques can both explore gene networks and associate those networks with covariates of interest, they differ in how they derive their networks.

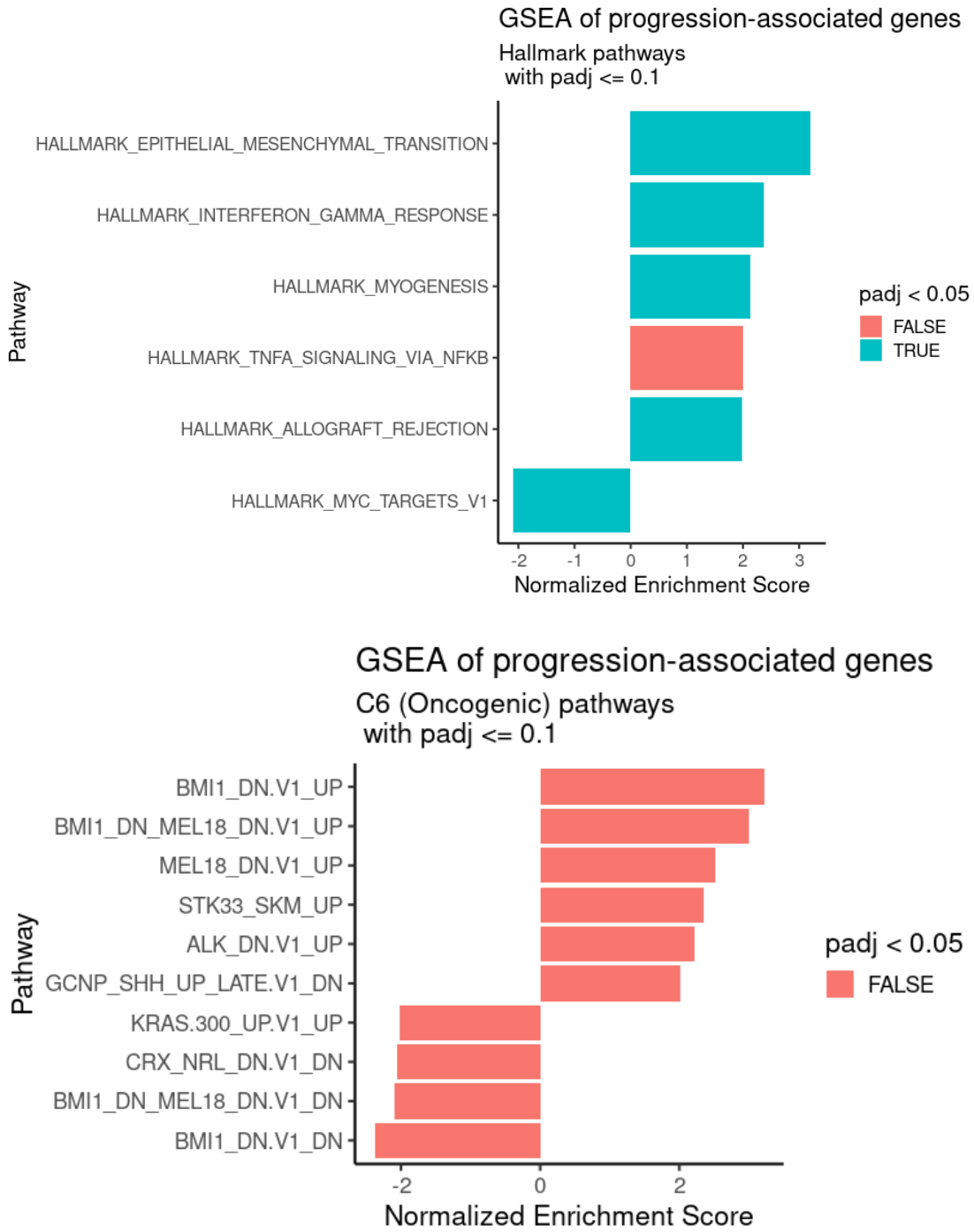


Figure 5: Gene Set Enrichment Analysis of Progression Associated Genes Against Hallmark and C6(Oncogenic) Gene Sets from MSigDB. (A) Significantly altered Hallmark gene sets associated with progression – Genes dysregulated with progression are involved in EMT, immune response, and MYC signaling, consistent with prior results (B) Significantly altered C6 (Oncogenic) sets associated with progression – Genes dysregulated with progression are involved in Epigenetic and kinase signaling processes

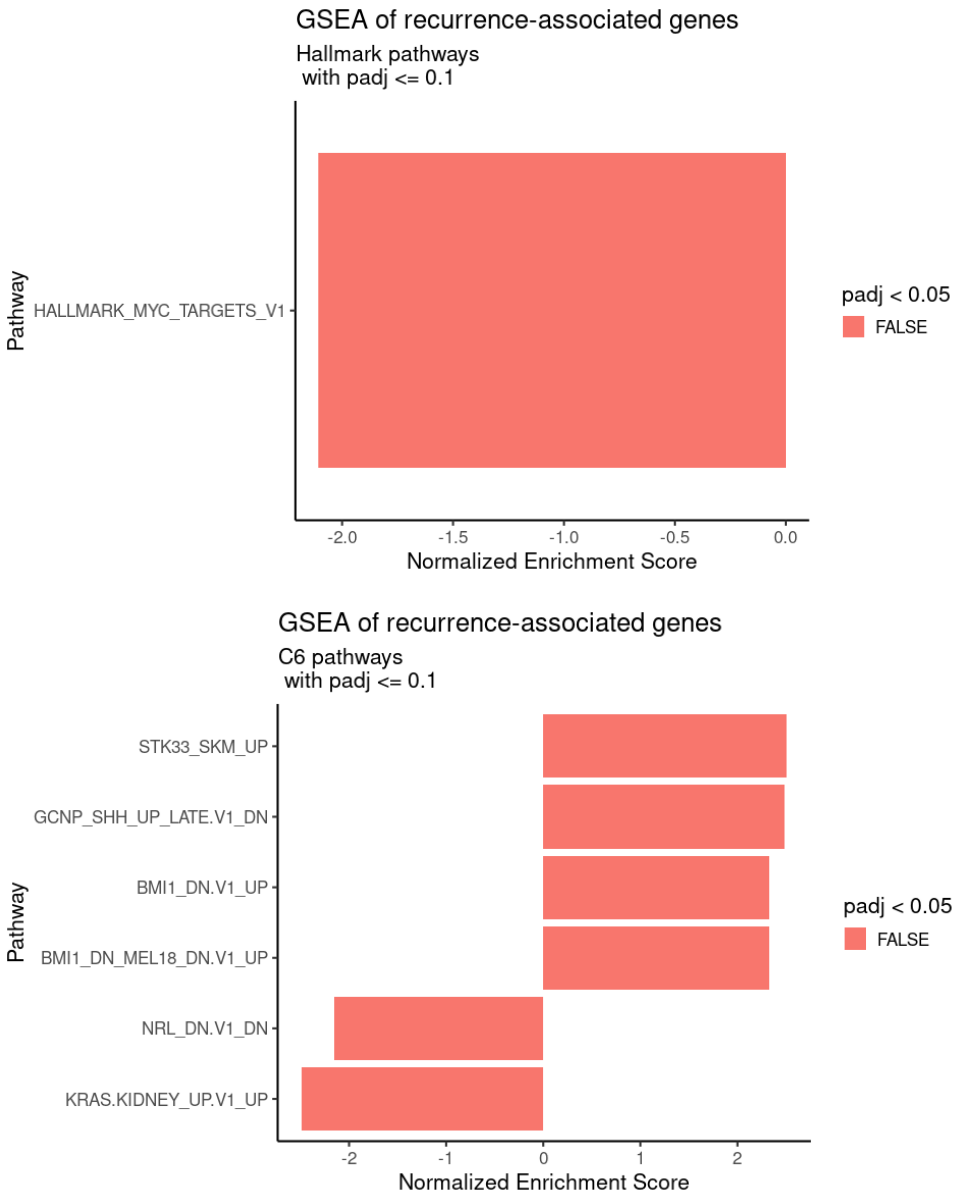


Figure 6: Gene Set Enrichment Analysis of Recurrence Associated Genes Against Hallmark and C6(Oncogenic) Gene Sets from MSigDB
A: Significantly altered Hallmark gene sets associated with progression – Genes dysregulated with progression are involved in EMT, immune response, and MYC signaling, consistent with prior results **B: Significantly altered C6 (Oncogenic) sets associated with progression** – Genes dysregulated with progression are involved in epigenetic and kinase signaling processes

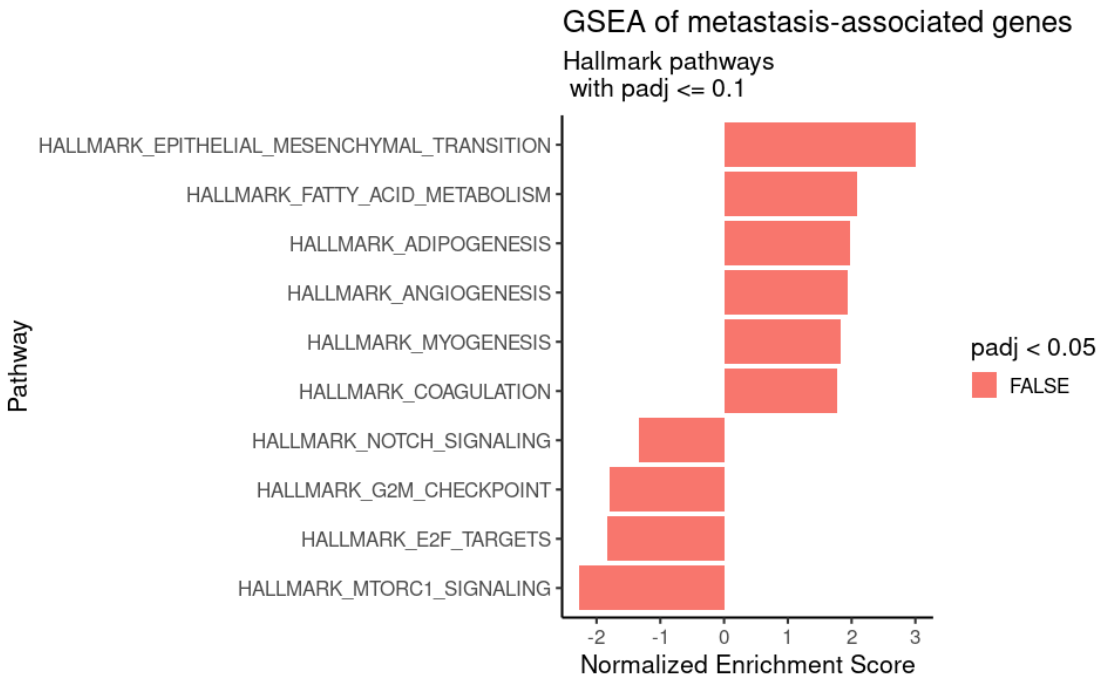
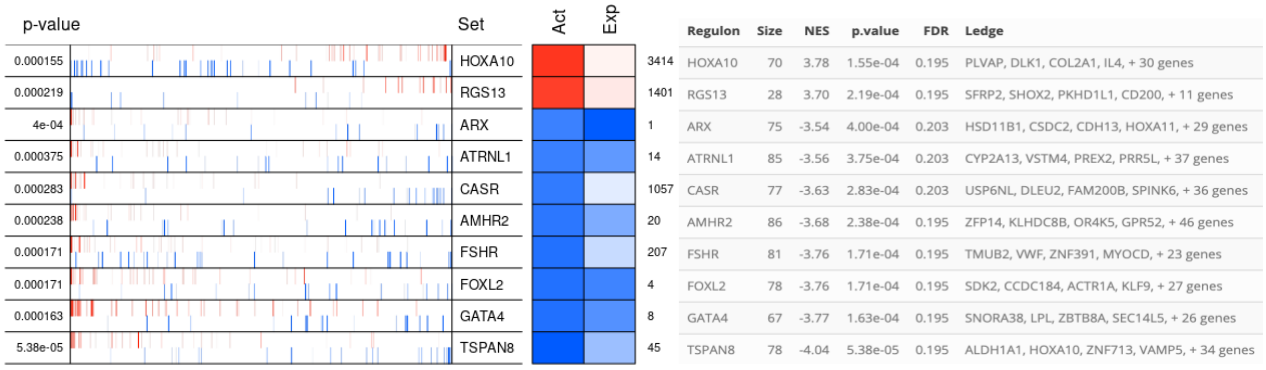


Figure 7: Gene Set Enrichment Analysis of Metastasis Associated Genes Against Hallmark and C6(Oncogenic) Gene Sets from MSigDB: Significantly altered Hallmark gene sets associated with metastasis – Diverse Hallmark gene sets were significantly associated with metastatsis only, including EMT and cell-cycle associated sets

VIPER, which uses a mutual-information based approach called ARACNE to derive a gene-correlation matrix, compares the expression profile of samples of interest to gene sets, "Regulons", associated with the activity of "Master Regulator" genes that appears to strongly correlate with many related genes. In this analysis, I used a set of regulons derived from TCGA ovarian cancer RNA-Seq data from the TCGARegulons package to determine how regulon activity changes with progression. VIPER tests the measured contrast in regulon activity against a bootstrapped null distribution of regulon activities to determine what regulons are differentially active under some condition. The advantage of this approach is that it can detect changes in regulon activity in settings where regulon expression may not be significantly changed. This approach is similar to the use of GSEA with a set of transcription factor associated genes, but the use of

regulons derived directly from HGSOC RNA-Seq data simplifies the interpretation of the results specifically in the context of HGSOC.

This analysis showed HOXA10 and RGS13 as HGSOC-associated regulons significantly activated with progression and a decrease in the activity of EMT associated regulons GATA4, FOXL2, and TSPAN8 and the hormone signaling related regulons FSHR, ARX, and CASR, as shown in **Figure 8A**



Regulon	Size	NES	p.value	FDR	Ledge
HOXA10	70	3.78	1.55e-04	0.195	PLVAP, DLK1, COL2A1, IL4, + 30 genes
RGS13	28	3.70	2.19e-04	0.195	SFRP2, SHOX2, PKHD1L1, CD200, + 11 genes
ARX	75	-3.54	4.00e-04	0.203	HSD11B1, CSDC2, CDH13, HOXA11, + 29 genes
ATRNL1	85	-3.56	3.75e-04	0.203	CYP2A13, VSTM4, PREX2, PRR5L, + 37 genes
CASR	77	-3.63	2.83e-04	0.203	USP6NL, DLEU2, FAM200B, SPINK6, + 36 genes
AMHR2	86	-3.68	2.38e-04	0.195	ZFP14, KLHDC8B, OR4K5, GPR52, + 46 genes
FSHR	81	-3.76	1.71e-04	0.195	TMUB2, VWF, ZNF391, MYOCD, + 23 genes
FOXL2	78	-3.76	1.71e-04	0.195	SDK2, CCDC184, ACTR1A, KLF9, + 27 genes
GATA4	67	-3.77	1.63e-04	0.195	SNORA38, LPL, ZBTB8A, SEC14L5, + 26 genes
TSPAN8	78	-4.04	5.38e-05	0.195	ALDH1A1, HOXA10, ZNF713, VAMP5, + 34 genes

Figure 8: Differentially Activated Regulons as Detected By VipeR Implicate Increased HOXA10 Signaling in HGSOC Progression. (A) Regulon activation plot for the top 10 regulons associated with progression: the top 10 differentially activated regulons show a clear activation of inflammatory and kinase signaling associated regulons but on overall decline in the activation of hormone signaling and EMT-associated regulons (B) Significance and major contributing genes for the top differentially activated regulons

VIPER provides a function to transform an input gene expression matrix to a protein activity matrix, offering the opportunity to validate these results generated by the MaRina approach in VIPER. Using the transformed protein activity matrix as an input to limma's empirical Bayesian DE pipeline, I found many of the same regulons significantly differentially activated, as shown in **Figure 9.**

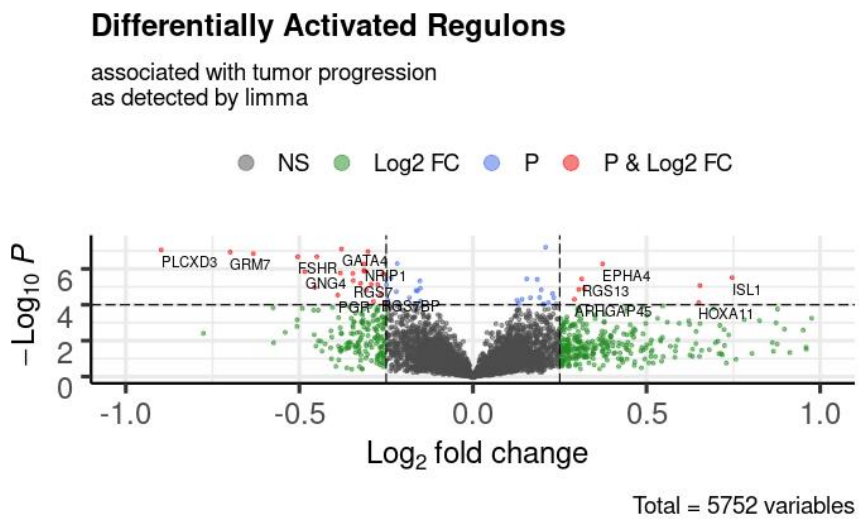


Figure 9: Differentially Activated Regulons as Detected By limma Confirm VIPER Regulon Analysis Results. Magnitude and significance of differentially activated regulons as detected by limma

WGCNA derives its own gene networks from a transformed gene expression correlation matrix and tests the correlation of eigengenes for the networks with covariates. First a correlation matrix is constructed from the gene expression data, expressing each gene in terms of its similarity to others in terms of expression. In my analysis, the "midweight bicorrelation statistic" was chosen as the correlation statistic for its robustness to outliers. To improve the interpretability of the data by reducing noise, the correlation matrix is raised to a power with the goal of scaling the values in the correlation matrix so that most genes have low scaled correlation value with most other genes, often below a cutoff of 0.2, but a few highly connected network "hub" genes can be found in a "Scale Free Toplogy". The ideal exponent is typically selected by hand to yield a manageable number of hub genes. Plotting the network characteristics "mean connectivity" and "scale independence" as a function of soft threshold, a "scale free" network can be constructed. The genes are then clustered by their scaled coexpression to identify "network modules".

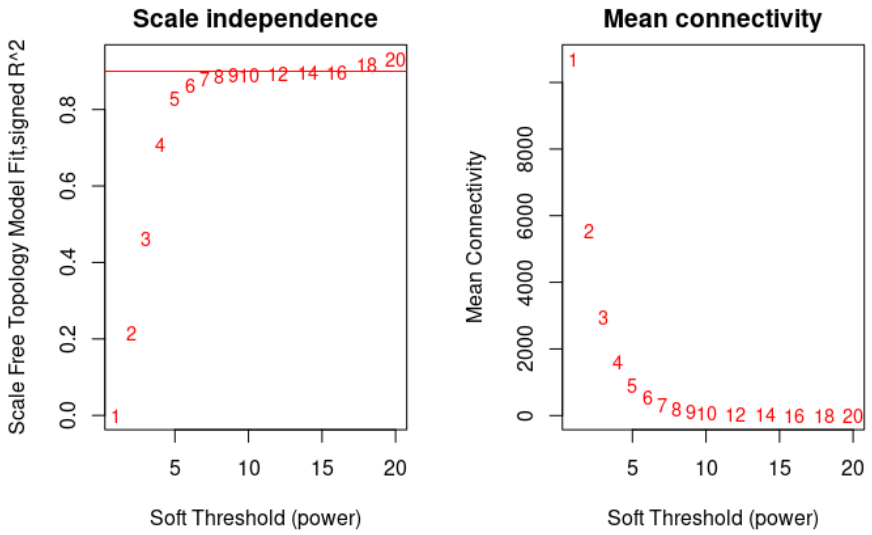


Figure 10: Weighted Gene Correlation Network Analysis (WGCNA) of HGSOC Progression - Selection of "Soft Threshold" for correlation matrix transformation. A soft threshold of 8 reduces mean connectivity to achieve a "scale-free topology", in which most genes are connected only to one other highly connected "hub" nodes

In constructing a scale free network, a soft threshold of 8 was chosen, as shown in **Figure 10**. The module eigengenes of the resulting modules were then compared, and highly similar modules by eigengene similarity were collapsed together, as shown in **Figure 11**. I found 25 modules with the parameters employed. As a simple negative control, a nonsense covariate, FILELOC, was included in the analysis, and showed the lowest correlation with ME expression of all covariates measured, consistent with expectations. Factors like RNA Integrity and tumor cellularity were strongly associated with several modules, but most notably one module, steelBlue, was significantly associated with TIMEPOINT only, shown in **Figure 12**.

Examining specifically the steelBlue module showed a strong correlation between a gene's module membership statistic and its correlation to TIMEPOINT, as **Figure 13A** shows.

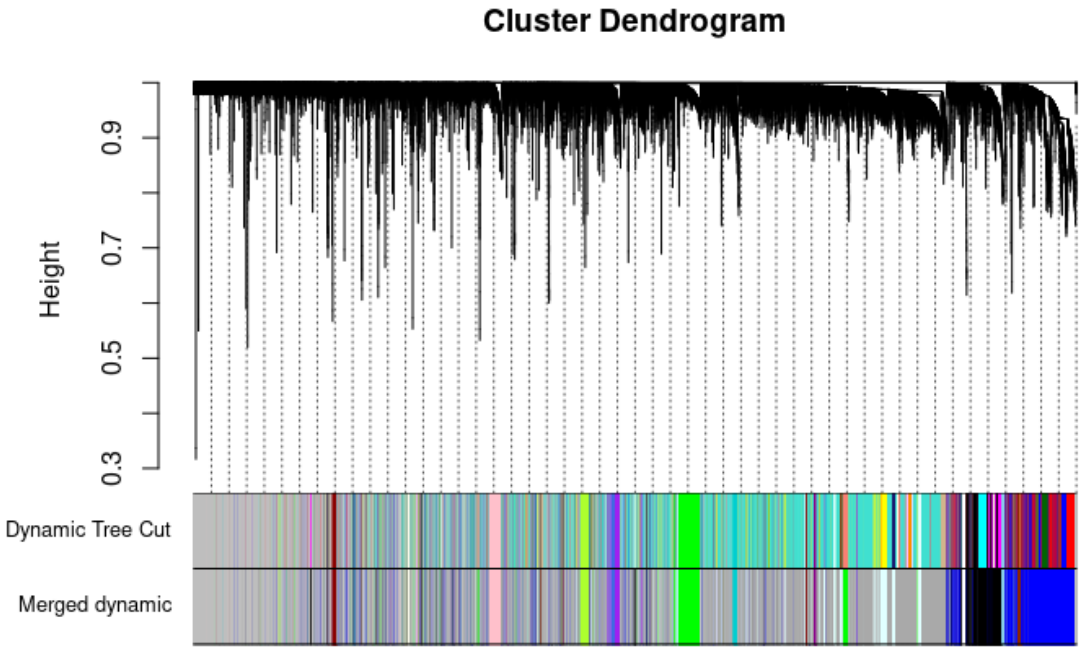


Figure 11: Weighted Gene Correlation Network Analysis (WGCNA) of HGSOC Progression - Initial clustering of genes by scaled correlation and final clustering following collapsing of similar modules. Many initially detected modules were collapsed due to high eigengene similarity

These genes include many of the regulons detected by VIPER as significantly associated with TIMEPOINT, including NR5AX, GATA4, and STAR, and several genes differentially expressed between TIMEPOINT as detected by standard DGEA. The steelBlue network is shown in **Figure 13B**. Overrepresentation analysis testing the steelBlue module genes against a background of all genes in the final WGCNA input showed significant enrichment in embryonic development, cell-microenvironmental interactions, and intracellular signaling.

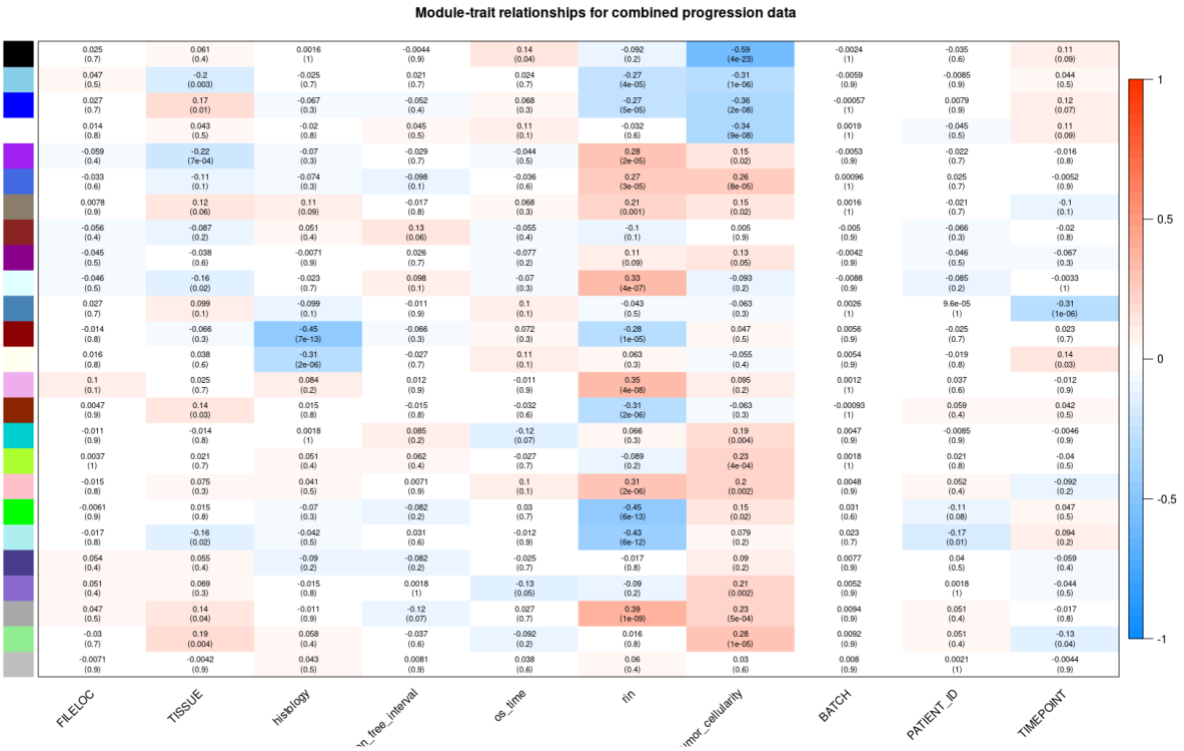


Figure 12: Weighted Gene Correlation Network Analysis (WGCNA) of HGSOC Progression - Correlation of identified coexpression modules with sample metadata. The correlation of each module's eigengene expression with traits was quantified. Tumor cellularity has the most significantly coexpressed modules. The steelblue module is significantly correlated with progression, but no other recorded traits

The consistency of both coexpression-based analyses motivated deeper analysis of network activity in the datasets. By VIPER transforming the gene expression matrix, as was performed for the limma validation of MARINA results, generates a matrix similar to a gene expression matrix. Selecting the top 15 differentially activated regulons associated with progression and clustering samples by their VIPER scores. An elbow plot generated using consensusClusterPlus suggested 4 to be the most stable number of clusters by K means clustering(**Figure 14A)**, and so samples were clustered with k=4 using the cluster package. A heatmap of the resulting clusters is presented in **Figure 14B.** To search for patterns in cluster transition, I generated a Sankey plot of cluster

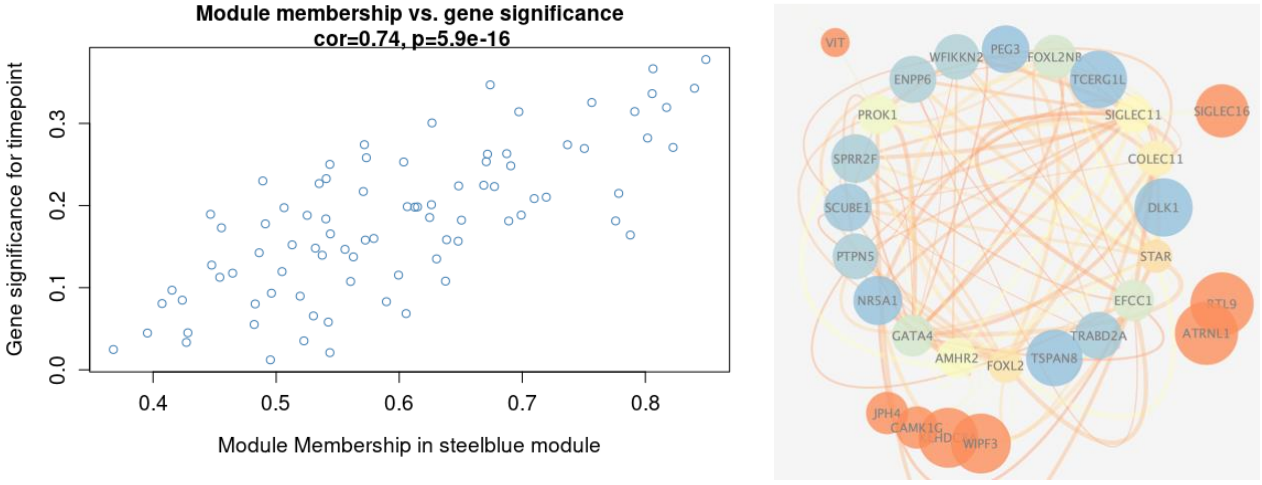


Figure 13: Weighted Gene Correlation Network Analysis (WGCNA) of HGSOC Progression - Characterization of the Progression-Associated Gene Correlation Module "steelblue"
(A) Correlation of module membership significance and progression – a strong correlation between module membership and significance of correlation with progression is shown (B) Dominant contributors to the "steelblue" module – many genes detected by WGCNA were also detected to be progression-associated by MaRina, including NR5A1, GATA4, TSPAN8, and many genes included within their regulons

membership transitions, separated by Batch. While samples are somewhat evenly distributed across clusters in the unprogressed samples, with progression, clusters 1 and 3 become dominant shown in **Figure 15**. Clusters 1 and 3 show increased HOX gene activity and decreased hormone signaling and EMT associated regulons.

Clustering TCGA gene expression data using the same regulons produced similar clusters, but those clusters showed correlation with overall patient survival. These findings suggest that while there are modules of coexpressed genes, many associated with "master regulators" that are frequently active in HGSOC progression, further work needs to be done to uncover a regulon progression signature with clinical utility.

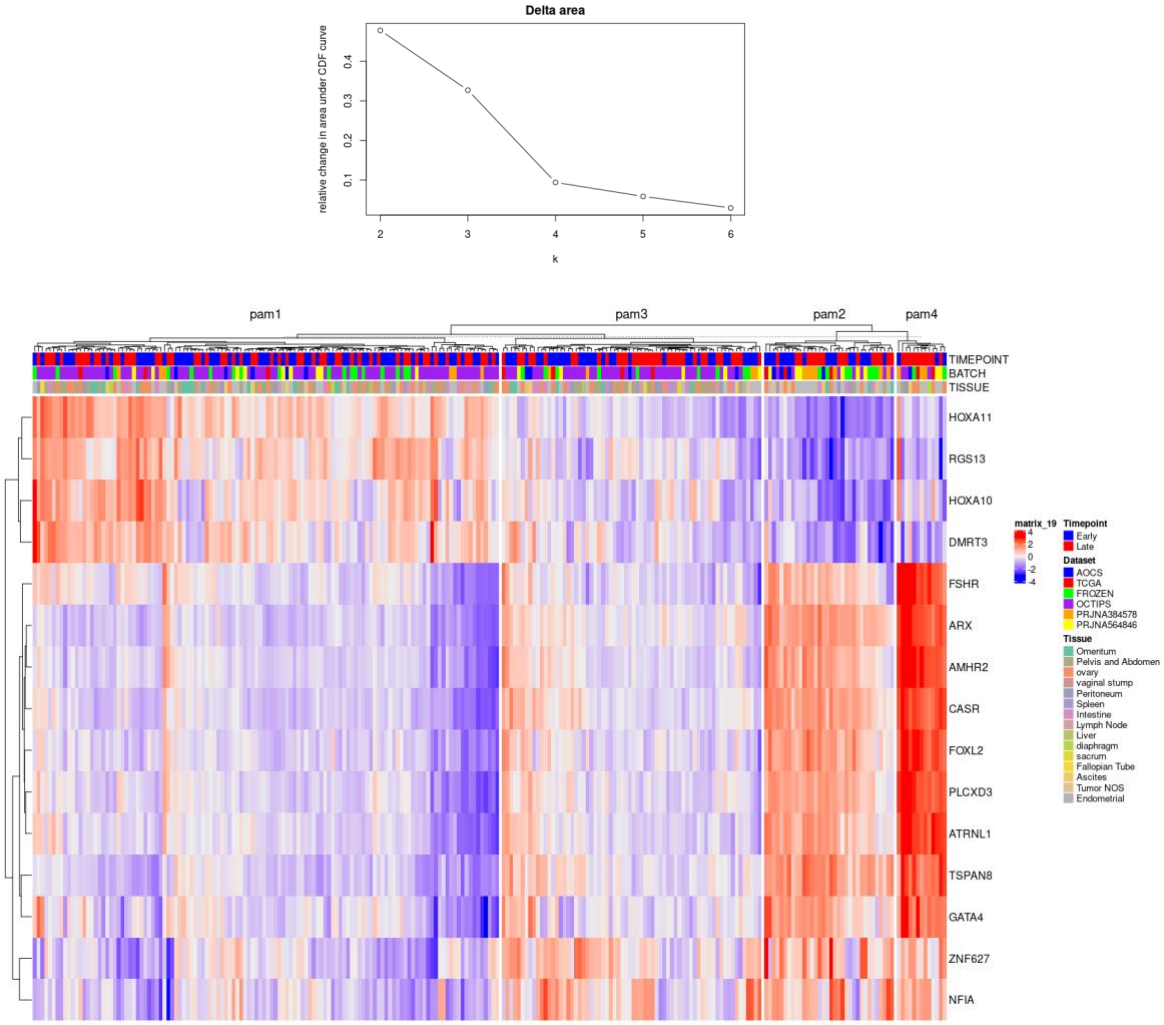


Figure 14: Consensus Clustering of Viper Regulon Activation Scores for Top Progression-Associated Regulons Yields 4 Stable Clusters. (A) Elbow plot of clustering CDF suggests 4 clusters (B) Clustered Heatmap of VIPER transformed regulon activation matrix for the top differentially activated regulons shows 4 clusters

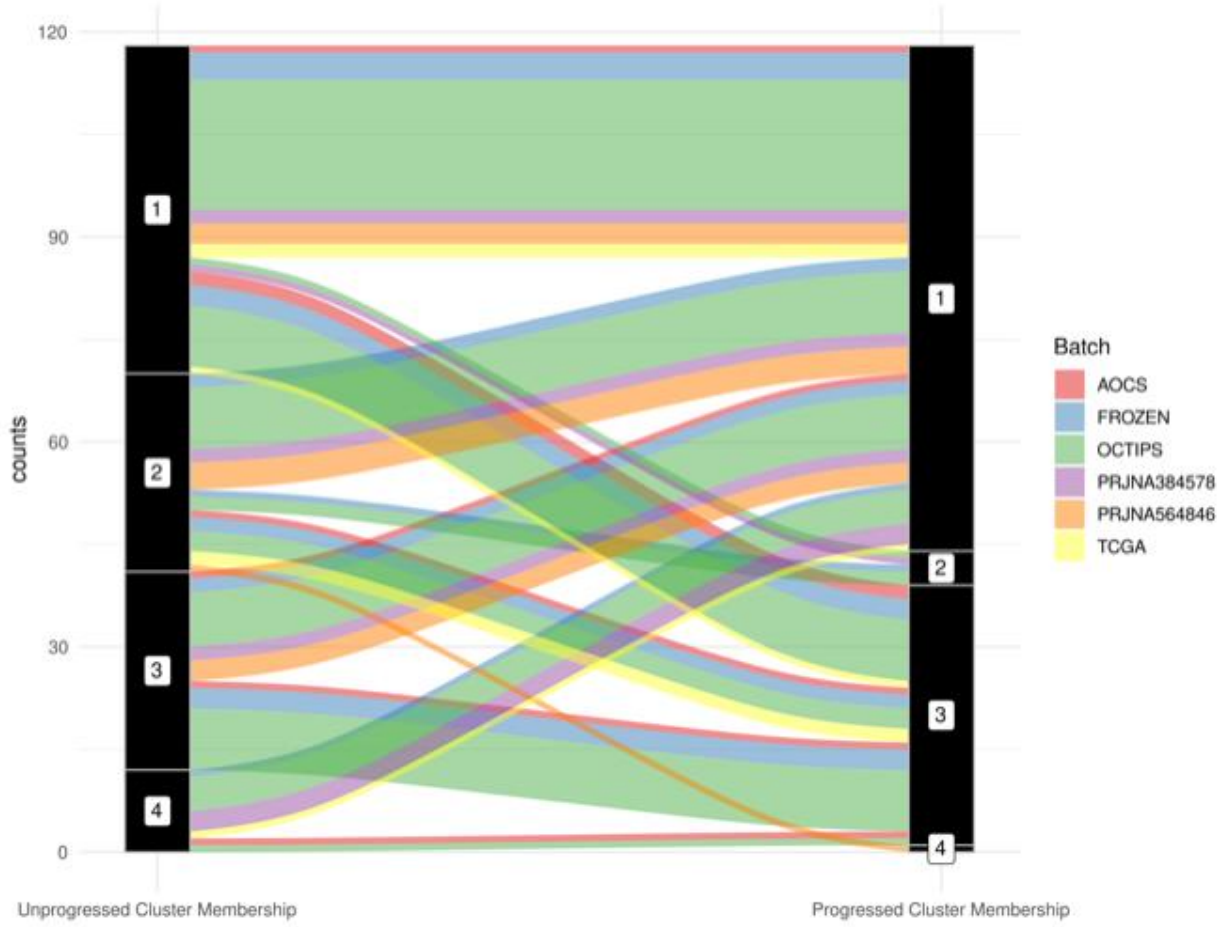


Figure 15: Master Regulator Cluster Transitions with Progression in HGSOC. Clusters 1 and 3 dominate with progression in HGSOC

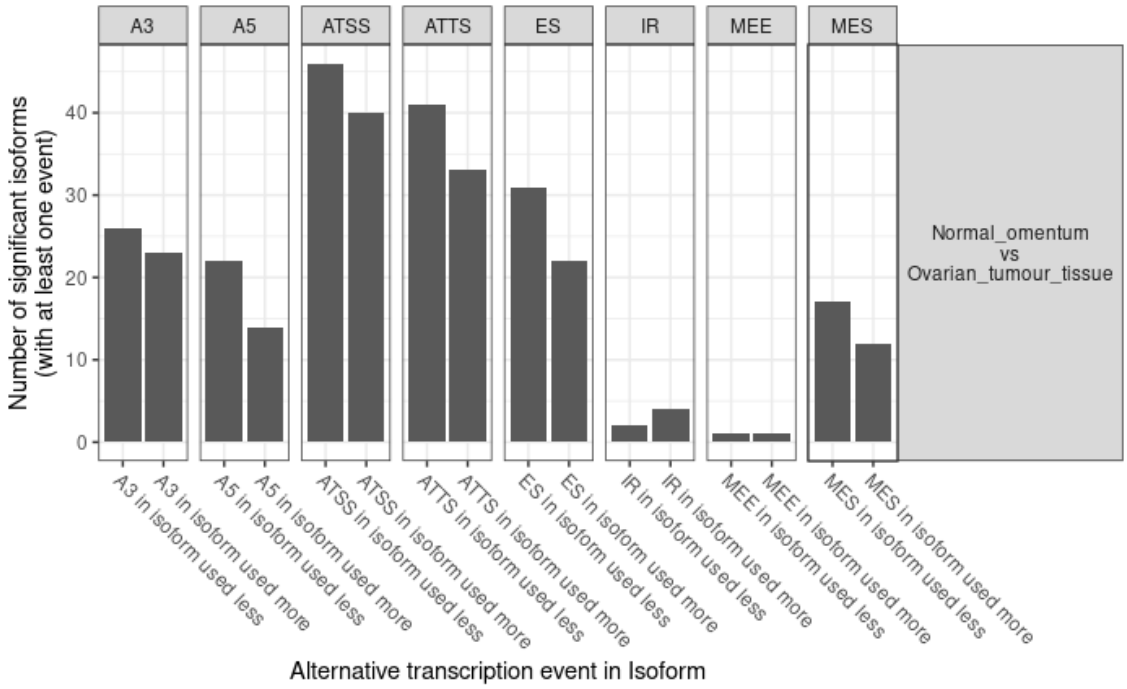


Figure 16: HGSOC Derived Cancer-Associated Mesenchymal Stem Cell Lines Show Differential Transcript Use Patterns Compared to Wild Type Omental Mesenchymal Stem Cells

3.3 Gene Isoform Expression Analysis and the Contribution of CA-MSCs to HGSOC Progression

While gene-level expression analysis and its abstraction to pathway and regulon-level trends offers insight into the mechanics of HGSOC progression, a growing body of literature suggests that differential isoform expression may play a major role in the progression of a range of diseases, including cancer. Salmon performs isoform level quantification by default, and the Bioconductor ecosystem has several established tools for investigating patterns in this isoform level data. I applied differential isoform analysis methods to all cohorts for which the raw fastq

files were available, and found changes in several potentially disease relevant genes that would were not detected by gene-level analysis.

Using the R package IsoformSwitchAnalyzer, I uncovered many instances of differential isoform use associated with progression, including exon skipping, intron retention, and the use of alternative transcriptional start and end sites in HGSOC progression, as shown in **Figure 16.**

In analysis of patient derived cancer-associated mesenchymal stem cells (CA-MSCs) and mesenchymal stem cells derived from normal omentum(MSCs), the gene LPCAT2, a lipid droplet associated protein, was not differentially expressed between conditions, but showed differential isoform use with potential functional consequence. As **Figure 17** shows, a truncated version of LPACT2 is preferentially expressed in unprogressed HGSOC and in HGSOC CA-MSCs. In CA-MSCs, a full length version of the gene is expressed.

The expression of this gene has been implicated in chemoresistance in colorectal cancer, and lipid metabolism is currently an active area of research in the study of chemoresistace. This finding suggests that in addition to changes in gene expression, changes in isoform expression can play a role in the progression of HGSOC.

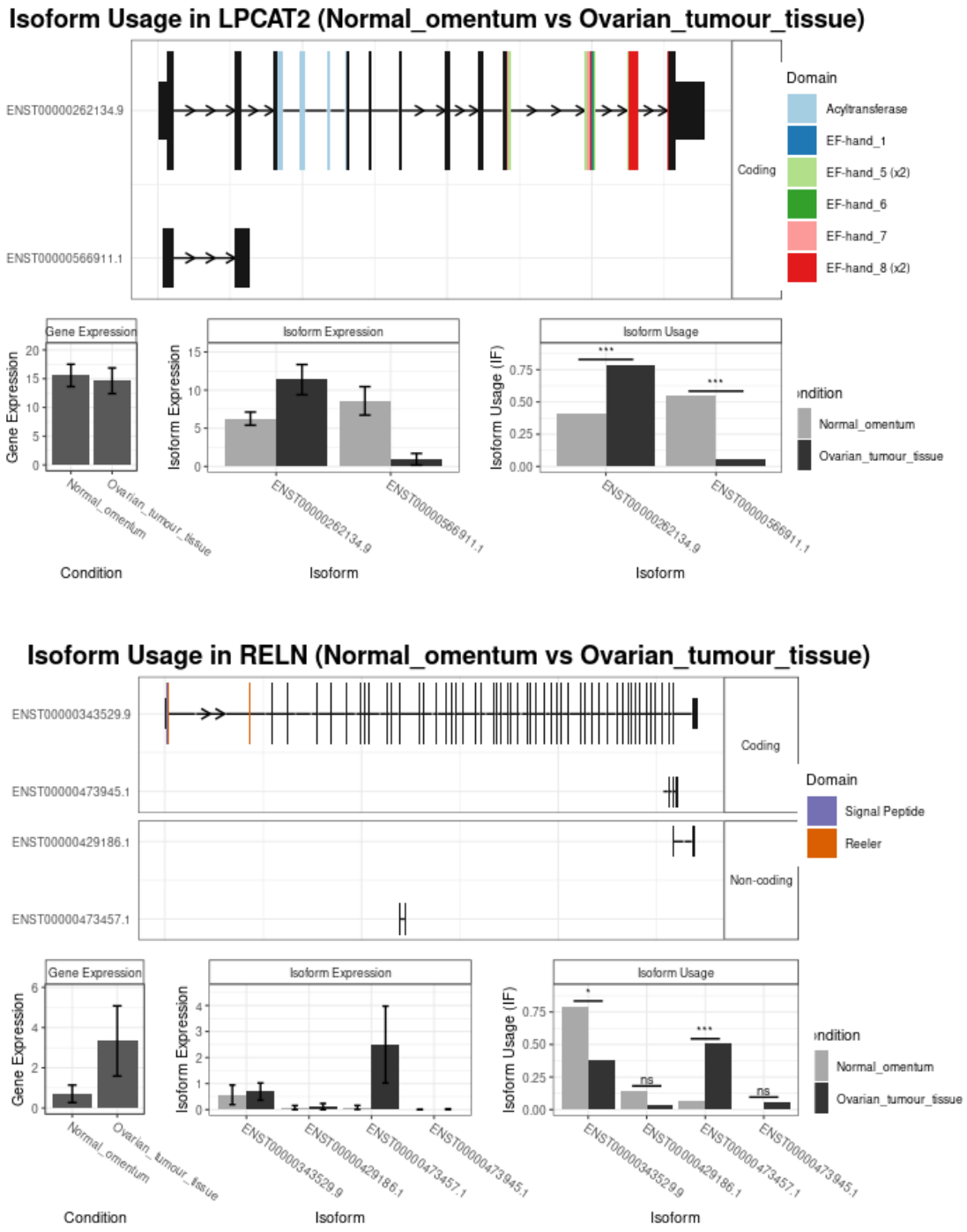


Figure 17: Differential Transcript Use in CA-MSCs Alters the Domain Structure of Cancer-Relevant Proteins. Schematic of exon isoform usage in the gene RELN. Top panel represents isoforms from the ensemble database. Graphs below how gene expression, isoform expression and usage.

3.4 **Detection of Expressed Gene Fusions form RNA-Seq and the development of FusionExplorer, an R Shiny appliction for exploration of cohort fusion profiles**

The ability to detect fusions and differential splicing is a major advantage of interrogating HGSOC evolution via sequencing rather than microarray. By searching for structural variation at the transcriptome level, this approach also improves the signal/noise ratio, a particular problem in SV prone cancers like HGSOC and breast cancer. While only 3 of our datasets (AOCS, TCGA, WCRC/RPCI) were sequenced with parameters adequate for fusion detection, this still revealed the expressed fusion profile of 36 pairs of HGSOC samples.

Consistent with standard practice, I screened the sequencing-amenable cohorts for interesting fusions by applying 3 fusion detection algorithms, FusionCatcher, STARFusion, and FusionZoom. While the sensitivity varied greatly between applications and datasets, most samples were found to have expressed fusions. Although tools exist for visualizing the basic fusion profile of single samples, no tools are explicitly designed to facilitate quick exploration of large collections of fusions across groups are available on Bioconductor. As fusion calling results are typically rectangular delimitted text files, these outputs could be parsed and basic analysis could be performed in excel by a lay biologist, but an R ShinyDashboard approach could simplify fusion exploration further by providing a graphical user interface and more robust bioinformatic tools.

To simplify exploration of multiple fusion datasets, I developed FusionExplorer, an R package consisting of parsing and analysis functions and a ShinyDashboard gui for interaction and the display of results. The basic components of FusionExplorer is detailed in **Figure 18**

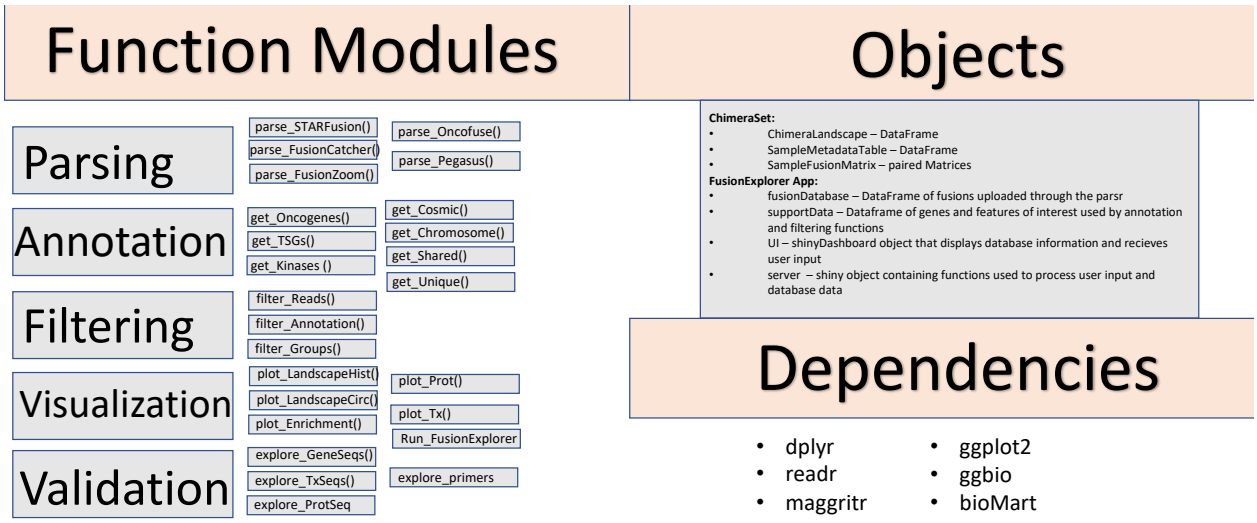


Figure 18: Key Features of FusionExplorer, an R Shiny Dashboard Interface for Cohort-Level Fusion Exploration.

The structure of FusionExplorer is further detailed in **Figure 19.** As with any Shiny object, FusionExplorer is best understood as a User Interface(UI) and Server. In this case, the Server performs all analysis functions, generates all visual summaries of data, and hosts all the necessary data. The UI primarily serves to transmit data from bidirectionally between the User and Server, through a user input pane and two data display panes. Screenshots of the application demonstrating the UI in action is included in the SI and the code is available on github at https://github.com/johnalanwillis/ovarianCancerProject. In this initial design of FusionExplorer, the full parsed fusion cohort must either be explicitly generated and loaded onto the server or generated with custom functions within the server. Generating a fusion set usable by the application is simple, requiring only a metadata table containing the location of the fusion caller results file and any relevant metadata. A worthwhile improvement would be to move the parsing functions to the User Interface Pane, for a more conceptually consistent design.

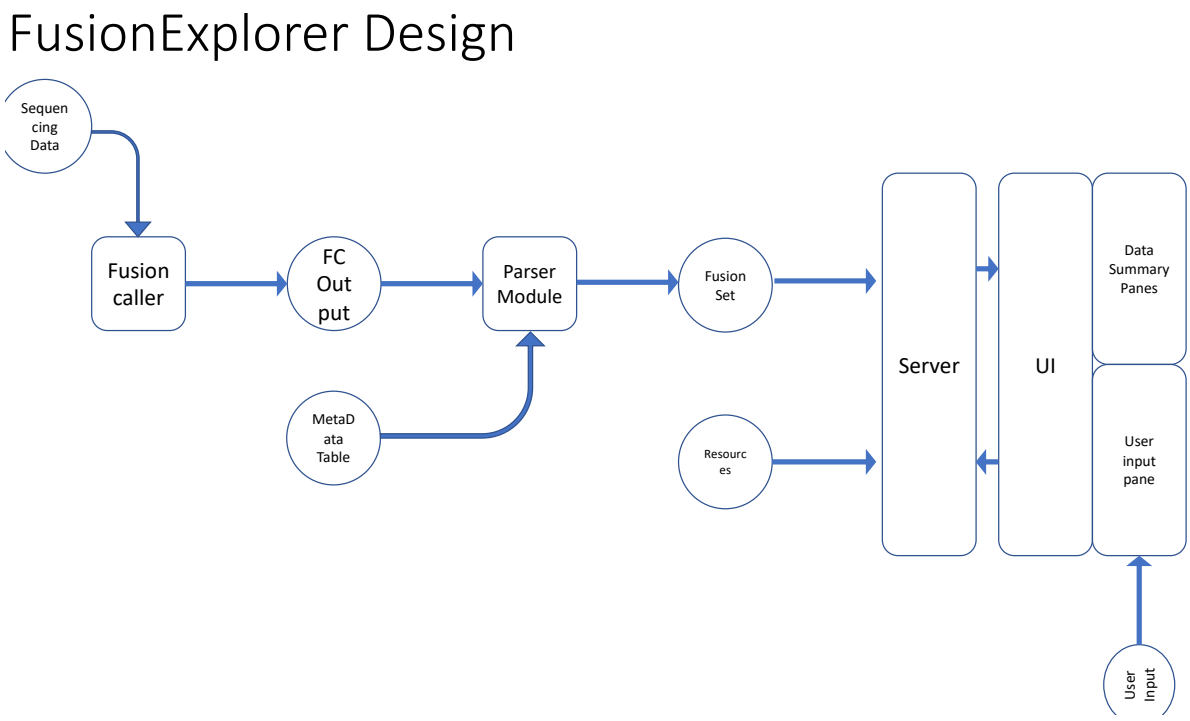


Figure 19: Structure of FusionExplorer, an R Shiny Dashboard Interface for Cohort-Level Fusion Exploration

In developing FusionExplorer, I collected 10 fusion datasets, from my HGSOC investigations, from other Lee-Oesterreich lab investigations of breast tumor and cell line fusions, and from publicly available fusion datasets. Subsetting the data through the User Input Pane to only include HGSOC fusions involving in-frame, cosmic genes with greater than 5 supporting reads, I was able to identify several recurrent fusions, typically preserved between samples.

The properties of the HGSOC fusion cohort are summarized in **Figure 20**. Unsurprisingly, sequencing depth plays a significant role in the sensitivity of fusion detection, with the deeply sequenced AOCS samples exhibiting the highest counts of detected fusions. Consistent with the prior result reported by the AOCS consortium, progressed AOCS samples appear to have more fusions detected than the early samples, though differences in the sample acquisition parameters could also play a role. The AOCS cohort compares solid tumor samples to tumor cells obtained

from ascites in recurrent HGSOC. The pattern of increased detectable fusions does not hold for the WCRC-RPCI or TCGA cohorts however, as **Figure 20A** shows.

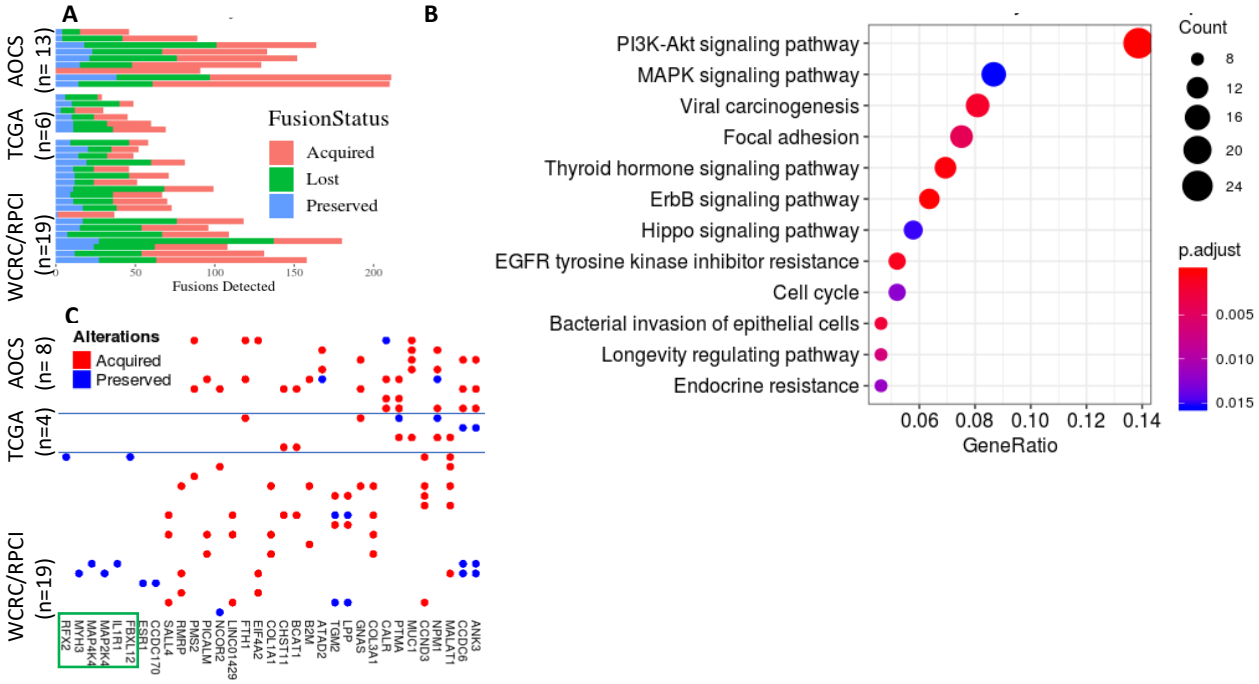


Figure 20: HGSOC Cases Preserve and Acquire Fusions in Oncogenically Relevant Pathways (A) Counts of fusions passing read filtering by sample and cohort from the union of Fusioncatcher, STARFusion, and FusionZoom (B) KEGG Pathways overrepresented in the list of genes involved in acquired or preserved fusions (C) Distribution of highly recurrent genes involved in Preserved or Acquired fusions across datasets

The fusions detected that had supporting reads above a predefined cutoff of 5 supporting reads are enriched in oncogenically relevant pathways. In particular, kinase signaling pathway genes appear to frequently contribute to fusions detected in our datasets. One challenge to interpreting this result is the possibility that this enrichment in detected genes is simply a function of gene length or baseline expression levels. Highly expressed genes are more likely to be detectable by sequencing and therefore fusions involving those same genes may also be more likely to be detectable by sequencing, assuming that the fusion does not alter the regulation of the gene. This is most likely to play a role for genes detected at the 5' end of the fusions, which should

have their relationships to upstream regulatory sequences preserved. Gene length could also bias the detection of patterns in expressed fusion genes, because these genes occupy a greater fraction of the genome and are therefore more likely to be disrupted by acquisition of structural variation by simple geometric probability.

Focusing in on the genes frequently detected as fused in our cohorts, a clearer view of the expressed fusion landscape develops. CCDC6—ANK3 recurs in all datasets, acquired in the AOCS cohort, but preserved in the WCRC-RPCI and TCGA dataset. Fusions with the lncRNA MALAT1 are also frequently acquired in the WCRC-RPCI and TCGA cohorts. Within the WCRC-RPCI cohort, we also uncovered a number of ”Preserved” fusions, shared between both early and late timepoints within a single patient. While their status as preserved fusions limits the likelihood that they are drivers of the evolution of these cases, they are excellent test cases for the bench validation of the bioinformatic fusion detection results. These preserved fusions, indicated by the green box, were selected for bench validation in patient samples by RT-PCR.

3.5 **Bioinformatic and Benchtop investigation of the contributions of fusion expression to HGSOC progression**

Recurrent Fusions detected in the in-silico screen of HGSOC pairs yielded many potential targets. The highly supported fusion FBXL12—RFX2 was among the first on the target list to validate. It was detected in both unprogressed and progressed samples in one patient from the WCRC/RPCI cohort, OVCA_14. In **Figure 21-22**, I summarize my bioinformatic characterization and benchtop validation of FBXL12—RFX2 as a "preserved" fusion in HGSOC.

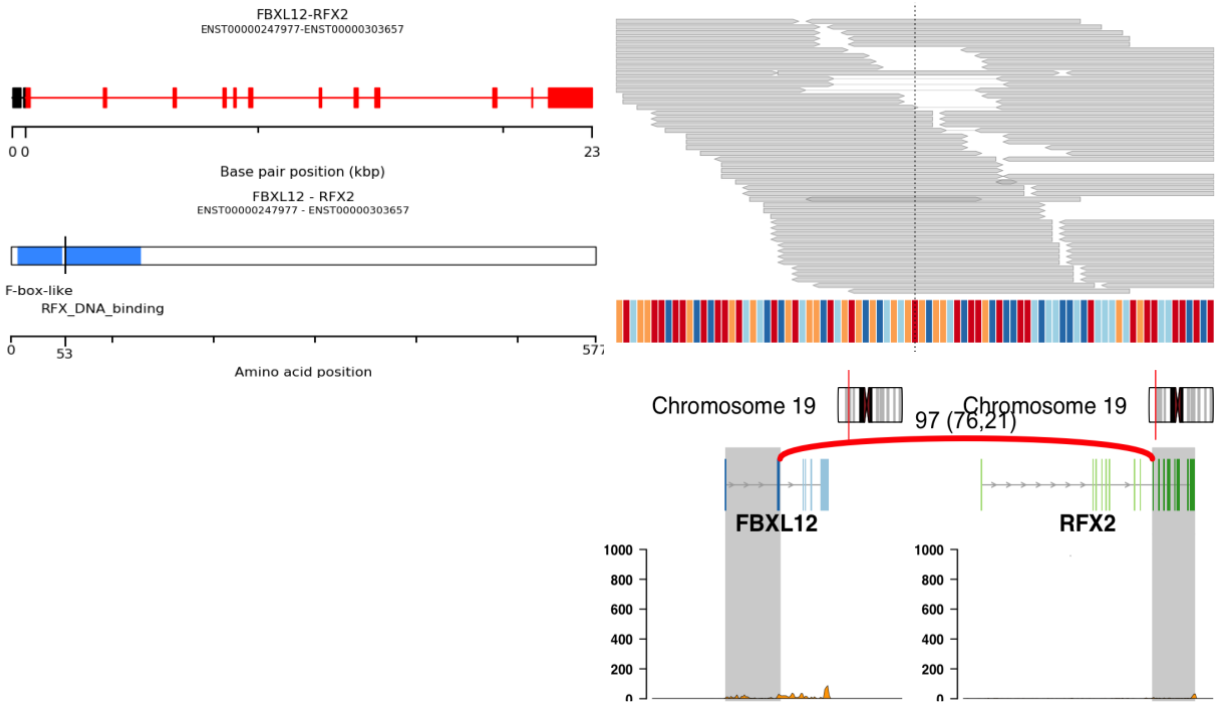


Figure 21: FBXL12—RFX2 is a Preserved fusion with high oncogenic potential detected in one HGSOC patient. Mapping fusion breakpoints to WT genes reveals a fusion that preserves F-box-like and RFX-DNA-binding domains

Detected by FusionCatcher in 2 samples in the WCRC-RPCI cohort, FBXL12-–RFX2 is an example of a bioinformatically well supported and bench-validated expressed gene fusion. In this figure, we see the bioinformatic evidence for the fusion –the grey bars represent RNA-sequencing reads that map to the new fusion junction. The high number of reads spanning the putative fusion junction, indicated by the dotted line, suggest that this detected fusion is not simply a bioinformatic or sequencing artefact.

The exon structure of the fusion is diagrammed in A, showing the earliest exons of FBXL12 fused with the later exons of RFX2. Below that diagram, we see a map of the known protein domains associated with the new structure. An Fbox-like domain and an DNA binding domain from RFX2 are preserved by the fusion. This fusion had a high Oncofuse "Driver Score", which attempts to score the similarity of fusions to fusions known to have oncogenic activity. This fusion derives oncogenic potential from both its partners.

FBXL12 may fail to perform its normal regulation of CDK4, contributing to proliferation, while aberrant RFX2 activity driven by FBXL12 expression may alter collagen expression patterns, possibly contributing to chemoresistance or invasion.

To confirm the expression of this fusion in patient samples, primers were designed against the component genes of the fusion, oriented to span the putative fusion junction. RT-PCR was used to amplify a fragment of the fusion containing the fusion junction, tested against a fusion-negative control, in this case RNA from the HGSOC-like cell line NIH_OVCAR3. The strong band present in the fusion positive sample and absent in the control cell line was extracted with a clean scalpel and the amplified cDNA was sent for sequencing, to confirm the presence of the expected fusion junction sequence, shown in **Figure 22**. Multiple sequence alignment performed with DECIPER showed inconsistency between multiple sequencing runs of the same sample, but

the resulting E score still allows us to confidently conclude that the fusion junction was indeed the amplified fragment visualized on the gel.

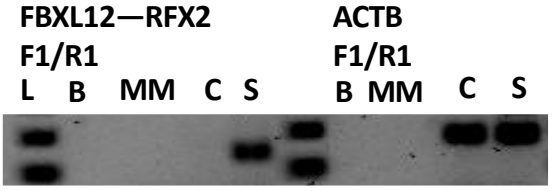


Figure 22: PCR Confirmation of FBXL12—RFX2 expression in one HGSOC patient. RT-PCR with fusion-junction targeting primers shows the fusion junction of FBXL12—RFX2 is uniquely expressed in one patient. L: 1kb+ ladder, W: water control, MM: master mix and primers without cDNA, C: fusion negative control NIHOVCAR3 cDNA, S: Fusion positive sample OvCa1-L.

In addition to FBXL12—RFX2, the HGSOC-associated fusions AGFG—MFF, IL1R1—MAP2K4—MYH3, and CCDC6—ANK3 were detected using the same PCR strategy, as shown in **Figure 23A**. As the most common in-frame, cosmic-gene-associated, and not "banned" by fusionCatcher expressed fusion in HGSOC, CCDC6—ANK3 was chosen for more detailed characterization. Fortunately, CCDC6—ANK3 is also expressed in the HGSOC-derived cell lines NIH_OVCAR3 and ONCO-DG1, so NIH_OVCAR3 was chosen as the initial platform for phenotypic studies. The exon structure of the recurrently fused genes CCD6, ANK3 and the structure of the fusion shows that CCDC6 only contributes one exon, without any complete functional domains to the abbreviated ANK3, shown in **Figure 23B**. This suggests that rather than creating a functional protein of combined domains or a hyperactive version due to truncation, CCDC6—ANK3 may function by putting the expression of a truncated ANK3 under the control of a promoter usually driving the expression of CCDC6, a gene with roles in DNA Damage Repair and Cell Cycle progression.

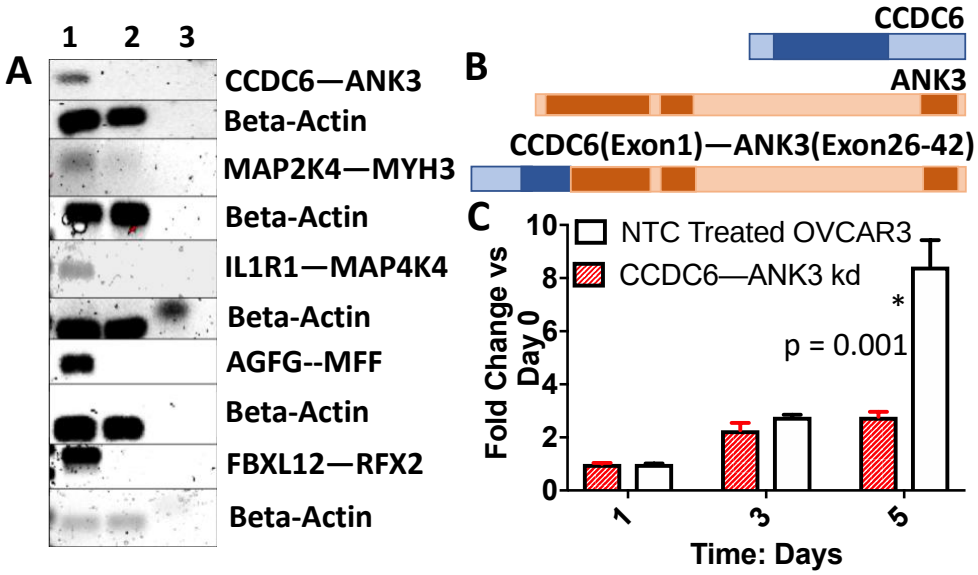


Figure 23: Investigation of Recurrent Fusions Uncovered in the WCRC-RPCI Cohort
(A) PCR Confirmation of Recurrent Fusions Detected by RNA Seq (B) Structure of the Highly Recurrent Fusion CCDC6—ANK3 C: CCDC6—ANK3 expression contribution to growth

To determine if CCDC6—ANK3 expression plays a role in NIH_OVCAR3 cells, siRNA targeting the novel exon-exon junction between the two partner genes were designed using Dharmacon's siRNA design tool. Comparing the growth rates of NIH_OVCAR3 cells treated with a non-targeting siRNA control library to that of NIH_OVCAR3 cells treated with fusion-targeting siRNA shows a clear difference in growth at 5 days, as shown in **Figure 23C.**

While liposomal transfection of NIH_OVCAR3 cells with siRNA against the CCDC6—ANK3 fusion junction appears to alter growth, I attempted to confirm the specific effect of CCDC6—ANK3 siRNA treatment on the expression of the fusion and its component genes via qPCR. As shown in **Figure 24A**, another set of CCDC6—ANK3 siRNA kds in NIH_OVCAR3 cells showed one siRNA pair, siRNA2, has a clear effect on growth compared to non-targeting control and another set of siRNA targeting the same fusion junction. Puzzlingly, both sets of siRNA appeared to alter fusion expression compared to NTC, but siRNA2, the one with a clear growth effect, did have the mose significant and dose dependent decrease. Attempts to characterize

the effect of CCDC6—ANK3 expression and kd on the expression of wild-type genes was also confounded by irreconcilable results.

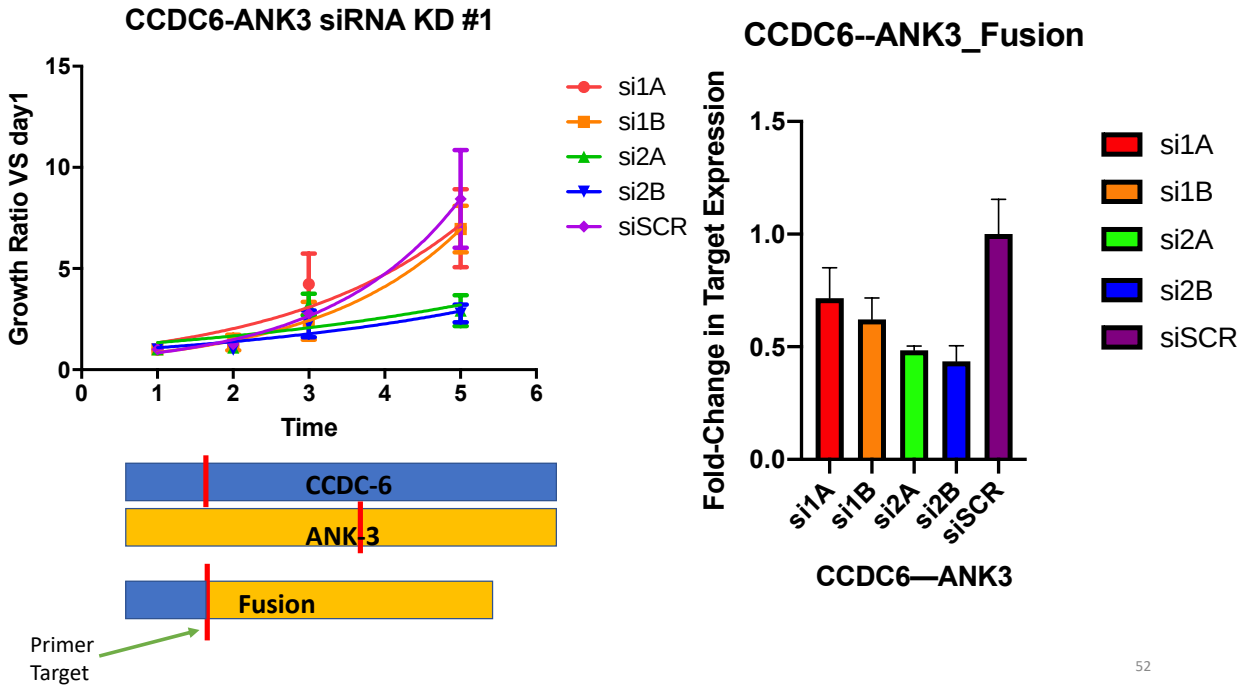


Figure 24: siRNA Knockdown of CCDC6—ANK3 Alters Growth in NIH_OVCAR3 Cells
A: Growth effect of liposomal transfection of anti-CCDC6—ANK3 siRNA alters growth in NIH_OVCAR3 cells. Two different siRNA sequences were tested(si1, si2) and two concentrations (A/B) were tested against a non targeting control pool **B:** Design strategy for CCDC6—ANK3 siRNA knockdown **C:** Treatment of NIH_OVCAR3 cells with anti CCDC6—ANK3 siRNA reduces expression of CCDC6—ANK3 as quantified by qPCR

4.0 Conclusion

HGSOC continues to present a challenge to clinicians and researchers, as its unstable genome promotes its evolution into treatment resistant forms. While recent trials of PARP inhibitors and cell cycle inhibitors have shown promise in treatment of HGSOC, the disease lacks the clear subtypes or molecular markers to guide treatment decisions. This analysis uncovered diversity in molecular changes associated with HGSOC progression, but also revealed that many of these changes ultimately act through common mechanisms.

Previous studies of HGSOC progression have implicated EMT and tumor microenvironmental interactions in the emergence of chemoresistance, and these results are consistent with that finding. There is clear evidence from histopathology, omics studies, and single cell sequencing studies that demonstrates HGSOC progression involves an increase in the relative mass of tumor epithelium to stroma in progressed disease. This combined study of HGSOC progression implicated genes associated with EMT and immune response in HGSOC progression, supporting the current emphasis on therapies that exploit the immune system to overcome the inherent heterogeneity of the disease.

Our study failed to replicate the results of prior fusion detection studies in HGSOC, but did implicate one highly recurrent fusion, CCDC6—ANK3, in the presentation of some HGSOC cases. While The exact nature of its contribution to the presentation of HGSOC remains unclear, this fusion merits further investigation. Additionally, numerous other potentially druggable fusions were detected in the course of our analysis, supporting the further investigation of fusion detection as a means to target therapies for HGSOC.

Bibliography

1. Peres, L. C. *et al.* Invasive epithelial ovarian cancer survival by histotype and disease stage. *J. Natl. Cancer Inst.* **111**, (2019).
2. Bowtell, D. D. Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. *Nat Rev Cancer* **15**, 668–679 (2016).
3. Raja, F. A., Chopra, N. & Ledermann, J. A. Optimal first-line treatment in ovarian cancer. *Ann. Oncol.* **23**, (2012).
4. Wright, A. A. *et al.* Neoadjuvant chemotherapy for newly diagnosed, advanced ovarian cancer: Society of Gynecologic Oncology and American Society of Clinical Oncology Clinical Practice Guideline. *Gynecol. Oncol.* (2016) doi:10.1016/j.ygyno.2016.05.022.
5. Kaufman, B. *et al.* Olaparib monotherapy in patients with advanced cancer and a germline BRCA1/2 mutation. *J. Clin. Oncol.* (2015) doi:10.1200/JCO.2014.56.2728.
6. Audeh, M. W. *et al.* Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: A proof-of-concept trial. *Lancet* (2010) doi:10.1016/S0140-6736(10)60893-8.
7. Cortez, A. J., Tudrej, P., Kujawa, K. A. & Lisowska, K. M. Advances in ovarian cancer therapy. *Cancer Chemother. Pharmacol.* **81**, 17–38 (2018).
8. Wang, H., Xu, T., Zheng, L. & Li, G. Angiogenesis Inhibitors for the Treatment of Ovarian Cancer: An Updated Systematic Review and Meta-analysis of Randomized Controlled Trials. *Int. J. Gynecol. Cancer* **28**, 903–914 (2018).
9. Moufarrij, S. *et al.* Epigenetic therapy for ovarian cancer: Promise and progress. *Clin. Epigenetics* **11**, 1–11 (2019).

10. Jung-Min Lee, M. *et al.* Prexasertib, a cell cycle checkpoint kinase 1 and 2 inhibitor, in BRCA wild-type recurrent high-grade serous ovarian cancer: a first-in-class proof-of-concept phase 2 study. *Lancet Oncol.* **19**, 207–215 (2018).

11. Mittal, V. K. & McDonald, J. F. Integrated sequence and expression analysis of ovarian cancer structural variants underscores the importance of gene fusion regulation. *BMC Med. Genomics* **8**, 1–12 (2015).

12. Konecny, G. E. *et al.* Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. *J. Natl. Cancer Inst.* (2014) doi:10.1093/jnci/dju249.

13. Chen, G. M. *et al.* Consensus on molecular subtypes of high-grade serous ovarian carcinoma. *Clin. Cancer Res.* (2018) doi:10.1158/1078-0432.CCR-18-0784.

14. McGill, J. R. *et al.* Double minutes are frequently found in ovarian carcinomas. *Cancer Genet. Cytogenet.* **71**, 125–131 (1993).

15. Yu, L. *et al.* Gemcitabine Eliminates Double Minute Chromosomes from Human Ovarian Cancer Cells. *PLoS One* **8**, (2013).

16. Alaei-Mahabadi, B., Bhadury, J., Karlsson, J. W., Nilsson, J. A. & Larsson, E. Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers. *Proc. Natl. Acad. Sci.* (2016) doi:10.1073/pnas.1606220113.

17. Menghi, F. *et al.* The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. *Cancer Cell* (2018) doi:10.1016/j.ccell.2018.06.008.

18. Chen, Y., Breeze, C. E., Zhen, S., Beck, S. & Teschendorff, A. E. Tissue-independent and tissue-specific patterns of DNA methylation alteration in cancer. *Epigenetics and Chromatin* **9**, 1–11 (2016).

19. Tan, J. *et al.* Integrative epigenome analysis identifies a polycomb-targeted differentiation program as a tumor-suppressor event epigenetically inactivated in colorectal cancer. *Cell Death Dis.* **5**, 1–11 (2014).

20. Smebye, M. L. *et al.* Involvement of DPP9 in gene fusions in serous ovarian carcinoma. *BMC Cancer* **17**, 1–10 (2017).

21. Galagoda, A., Kannan, K., Yen, L., Kordestani, G. & Coarfa, C. Aberrant MUC1-TRIM46-KRTCAP2 Chimeric RNAs in High-Grade Serous Ovarian Carcinoma. *Cancers (Basel).* **7**, 2083–2093 (2015).

22. Agostini, A. *et al.* Identification of novel cyclin gene fusion transcripts in endometrioid ovarian carcinomas. *Int. J. Cancer* **143**, 1379–1387 (2018).

23. Salzman, J. *et al.* ESRRA-C11orf20 is a recurrent gene fusion in serous ovarian carcinoma. *PLoS Biol.* (2011) doi:10.1371/journal.pbio.1001156.

24. Christie, E. L. *et al.* Multiple ABCB1 transcriptional fusions in drug resistant high-grade serous ovarian and breast cancer. *Nat. Commun.* **10**, 5–14 (2019).

25. Patch, A. M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489–494 (2015).

26. Coscia, F. *et al.* Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status. *Nat. Commun.* **7**, 1–14 (2016).

27. Yang, D. *et al.* Article Integrated Analyses Identify a Master MicroRNA Regulatory Network for the Mesenchymal Subtype in Serous Ovarian Cancer. *Cancer Cell* **23**, 186–199 (2013).

28. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* (2018) doi:10.1016/j.cell.2018.03.035.

29. Kim, C. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* 1–15 (2018) doi:10.1016/j.cell.2018.03.041.

30. Vaidyanathan, A. *et al.* ABCB1 (MDR1) induction defines a common resistance mechanism in paclitaxel- and olaparib-resistant ovarian cancer cells. *Br. J. Cancer* (2016) doi:10.1038/bjc.2016.203.

31. Pujade-Lauraine, E. *et al.* Olaparib tablets as maintenance therapy in patients with platinum-sensitive, relapsed ovarian cancer and a BRCA1/2 mutation (SOLO2/ENGOT-Ov21): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Oncol.* (2017) doi:10.1016/S1470-2045(17)30469-2.

32. Zhang, Z. *et al.* Molecular Subtyping of Serous Ovarian Cancer Based on Multi-omics Data. *Sci. Rep.* **6**, 1–10 (2016).

33. Arend, R. C. *et al.* Molecular Response to Neoadjuvant Chemotherapy in High-Grade Serous Ovarian Carcinoma. *Mol. Cancer Res.* (2018) doi:10.1158/1541-7786.mcr-17-0594.

34. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).

35. Shih, A. J. *et al.* Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. *PLoS One* (2018) doi:10.1371/journal.pone.0206785.

36. Özeş, A. R. *et al.* NF-κB-HOTAIR axis links DNA damage response, chemoresistance and cellular senescence in ovarian cancer. *Oncogene* **35**, 5350–5361 (2016).

37. Virant-Klun, I., Kenda-Suster, N. & Smrkolj, S. Small putative NANOG, SOX2, and SSEA-4-positive stem cells resembling very small embryonic-like stem cells in sections of ovarian tissue in patients with ovarian cancer. *J. Ovarian Res.* **9**, 1–15 (2016).

38. Coffman, L. G. *et al.* Human carcinoma-associated mesenchymal stem cells promote ovarian cancer chemotherapy resistance via a BMP4/HH signaling loop. *Oncotarget* (2016) doi:10.18632/oncotarget.6870.

39. Kurzrock, R., Kantarjian, H. M., Druker, B. J. & Talpaz, M. Philadelphia Chromosome-Positive Leukemias: From Basic Mechanisms to Molecular Therapeutics. *Annals of Internal Medicine* (2003) doi:10.7326/0003-4819-138-10-200305200-00010.

40. Tomlins, S. A. *et al.* Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia* **10**, 177–88 (2008).

41. Riggi, N. *et al.* EWS-FLI-1 expression triggers a ewing's sarcoma initiation program in primary human mesenchymal stem cells. *Cancer Res.* (2008) doi:10.1158/0008-5472.CAN-07-1761.

42. Luise, C. *et al.* Identification of Sumoylation Sites in CCDC6, the First Identified RET Partner Gene in Papillary Thyroid Carcinoma, Uncovers a Mode of Regulating CCDC6 Function on CREB1 Transcriptional Activity. *PLoS One* (2012) doi:10.1371/journal.pone.0049298.

43. Tognon, C. *et al.* Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. *Cancer Cell* **2**, 367–376 (2002).

44. Stransky, N., Cerami, E., Schalm, S., Kim, J. L. & Lengauer, C. The landscape of kinase fusions in cancer. *Nat. Commun.* **5**, 1–10 (2014).

45. Zhang, Y. *et al.* A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases. *Cell Rep.* (2018) doi:10.1016/j.celrep.2018.06.025.

46. Prescott, J. D. & Zeiger, M. A. The RET oncogene in papillary thyroid carcinoma. *Cancer* (2015) doi:10.1002/cncr.29044.

47. Vu-Phan, D. & Koenig, R. J. Genetics and epigenetics of sporadic thyroid cancer. *Molecular and Cellular Endocrinology* (2014) doi:10.1016/j.mce.2013.07.030.

48. Yoshihara, K. *et al.* The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* (2015) doi:10.1038/onc.2014.406.

49. Gatalica, Z., Xiu, J., Swensen, J. & Vranic, S. Molecular characterization of cancers with NTRK gene fusions. *Mod. Pathol.* (2019) doi:10.1038/s41379-018-0118-3.

50. Märkl, B., Hirschbühl, K. & Dhillon, C. NTRK-Fusions – A new kid on the block. *Pathology Research and Practice* (2019) doi:10.1016/j.prp.2019.152572.

51. Veeraraghavan, J. *et al.* Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers. *Nat. Commun.* (2014) doi:10.1038/ncomms5577.

52. Hartmaier, R. J. *et al.* Recurrent hyperactive ESR1 fusion proteins in endocrine therapy-resistant breast cancer. *Ann. Oncol.* **29**, 872–880 (2018).

53. Jiang, P. *et al.* The Protein Encoded by the CCDC170 Breast Cancer Gene Functions to Organize the Golgi-Microtubule Network. *EBioMedicine* (2017) doi:10.1016/j.ebiom.2017.06.024.

54. Liu, S. *et al.* Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Res.* **44**, 1–15 (2015).

55. Nicorici, D. *et al.* FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv* 011650 (2014) doi:10.1101/011650.

56. McPherson, A. *et al.* Defuse: An algorithm for gene fusion discovery in tumor rna-seq data. *PLoS Comput. Biol.* **7**, (2011).

57. Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, e178 (2010).

58. Latysheva, N. S. & Babu, M. M. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. **44**, 4487–4503 (2018).

59. Haas, B. J. *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv* 120295 (2017) doi:10.1101/120295.

60. Kim, D. & Salzberg, S. L. TopHat-Fusion: An algorithm for discovery of novel fusion transcripts. *Genome Biol.* (2011) doi:10.1186/gb-2011-12-8-r72.

61. Melsted, P. *et al.* Fusion detection and quantification by pseudoalignment. *bioRxiv* (2017) doi:10.1101/166322.

62. Benelli, M. *et al.* Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics* (2012) doi:10.1093/bioinformatics/bts617.

63. Stransky, N., Cerami, E., Schalm, S., Kim, J. L. & Lengauer, C. The landscape of kinase fusions in cancer. *Nat. Commun.* (2014) doi:10.1038/ncomms5846.

64. Spies, N., Zook, J. M., Salit, M. & Sidow, A. Svviz: A read viewer for validating structural variants. *Bioinformatics* **31**, 3994–3996 (2015).

65. Lågstad, S. *et al.* Chimeraviz: A tool for visualizing chimeric RNA. *Bioinformatics* **33**, 2954–2956 (2017).

66. Shugay, M., De Mendíbil, I. O., Vizmanos, J. L. & Novo, F. J. Oncofuse: A computational framework for the prediction of the oncogenic potential of gene fusions. *Bioinformatics* **29**, 2539–2546 (2013).

67. Abate, F. *et al.* Pegasus: A comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Syst. Biol.* **8**, 1–14 (2014).

68. Kannan, K. *et al.* Recurrent *BCAM-AKT2* fusion gene leads to a constitutively activated AKT2 fusion kinase in high-grade serous ovarian carcinoma. *Proc. Natl. Acad. Sci.* **112**, E1272–E1277 (2015).

69. Kannan, K. *et al.* CDKN2D-WDFY2 Is a Cancer-Specific Fusion Gene Recurrent in High-Grade Serous Ovarian Carcinoma. *PLoS Genet.* **10**, (2014).

70. Krzyzanowski, P. M. *et al.* Regional perturbation of gene transcription is associated with intrachromosomal rearrangements and gene fusion transcripts in high grade ovarian cancer. *Sci. Rep.* (2019) doi:10.1038/s41598-019-39878-9.

71. Kurman, R. J. & Shih, I. M. The dualistic model of ovarian carcinogenesis revisited, revised, and expanded. *Am. J. Pathol.* **186**, 733–747 (2016).

72. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* (2011) doi:10.1093/nar/gkq1019.

73. Kodama, Y., Shumway, M. & Leinonen, R. The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Res.* (2012) doi:10.1093/nar/gkr854.

74. Mitra, S. *et al.* Transcriptome profiling reveals matrisome alteration as a key feature of ovarian cancer progression. *Cancers (Basel).* (2019) doi:10.3390/cancers11101513.

75. Sallinen, H. *et al.* Comparative transcriptome analysis of matched primary and distant metastatic ovarian carcinoma. *BMC Cancer* (2019) doi:10.1186/s12885-019-6339-0.

76. Leinonen, R. *et al.* The European nucleotide archive. *Nucleic Acids Res.* (2011) doi:10.1093/nar/gkq967.

77. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp352.

78. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* (2012) doi:10.1038/nature11003.

79. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* (2019) doi:10.1038/s41586-019-1186-3.

80. Andrews, S. FASTQC A Quality Control tool for High Throughput Sequence Data. *Babraham Inst.* (2015).

81. Hartley, S. W. & Mullikin, J. C. The QoRTs Analysis Pipeline Example Walkthrough. *Nucleic Acids Res.* (2016) doi:10.1093/nar/gkw501.

82. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btw354.

83. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* (2017) doi:10.1038/nmeth.4197.

84. Patro, R. *et al.* Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. **14**, 417–419 (2017).

85. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* (2012) doi:10.1101/gr.135350.111.

86. Vitting-Seerup, K. & Sandelin, A. IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz247.

87. Love, M. I. *et al.* Tximeta: Reference sequence checksums for provenance identification in RNA-seq. *PLoS Comput. Biol.* (2020) doi:10.1101/777888.

88. Marini, F. & Binder, H. PcaExplorer: An R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinformatics* (2019) doi:10.1186/s12859-019-2879-1.

89. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp616.

90. Smyth, G. K. limma: Linear Models for Microarray Data. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (2005). doi:10.1007/0-387-29362-0_23.

91. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom : precision weights unlock linear model analysis tools for RNA-seq read counts. 1–17 (2014).

92. Love, M. I., Anders, S. & Huber, W. *Differential analysis of count data - the DESeq2 package*. *Genome Biology* (2014). doi:110.1186/s13059-014-0550-8.

93. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omi. A J. Integr. Biol.* (2012) doi:10.1089/omi.2011.0118.

94. Sergushichev, A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv* (2016) doi:10.1101/060012.

95. Sonja, H., Castelo, R. & Guinney, J. GSVA : The Gene Set Variation Analysis package for microarray and RNA-seq data. *Bioconductor.org* 1–20 (2014).

96. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* (2016) doi:10.1038/ng.3593.

97. Margolin, A. A. *et al.* ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* (2006) doi:10.1186/1471-2105-7-S1-S7.

98. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* (2019) doi:10.1101/gr.240663.118.

99. Li, J. *et al.* Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design. *Sci. Rep.* (2018) doi:10.1038/s41598-017-18705-z.

100. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* (2008) doi:10.1186/1471-2105-9-559.

101. Colaprico, A. *et al.* TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* (2016) doi:10.1093/nar/gkv1507.

102. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btw313.

103. Maechler, M. *et al.* Package 'cluster': Cluster Analysis Basics and Extensions. *R Top. Doc.* (2015) doi:ISBN 0-387-95457-0.

104. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics* (2010) doi:10.1093/bioinformatics/btq170.

105. Kassambara, A., Kosinski, M., Biecek, P. & Fabian, S. survminer: Drawing Survival Curves using 'ggplot2' (R package). *version 0.4.3* (2018).

106. Ishimaru, S. *et al.* Shiny. in (2014). doi:10.1145/2638728.2638798.

107. Rozen, S. & Skaletsky, H. J. Primer3. *Bioinformatics Methods and Protocols Methods in Molecular Biology* (1998).