**Distribution Agreement**

Emma Klein March 13, 2023

# Transcriptome Analysis of Primary Prostate Tumor Foci and Corresponding Lymph Node Metastases Identifies Pathways Associated with Metastatic Disease

By

Emma Klein

Dr. Carlos Moreno
Adviser

Biology

Dr. Carlos Moreno

Adviser

Dr. David Gorkin

Committee Member

Dr. Ymir Vigfusson

Committee Member

2023

# Transcriptome Analysis of Primary Prostate Tumor Foci and Corresponding Lymph Node Metastases Identifies Pathways Associated with Metastatic Disease

By

Emma Klein

Dr. Carlos Moreno

Adviser

An abstract of

a thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Science with Honors

Biology

2023

# Abstract

## Transcriptome Analysis of Primary Prostate Tumor Foci and Corresponding Lymph Node Metastases Identifies Pathways Associated with Metastatic Disease

By Emma Klein

Prostate cancer (PCa) is a highly heterogeneous disease, and mortality is mainly due to metastases. However, the molecular underpinnings that lead to the initial steps of metastasis have not been well characterized. We performed genomic analysis of primary prostate tumor foci (PTF) and corresponding lymph node metastases (LNM). PTF and LNM from patients with high-risk PCa were analyzed by RNAseq and Whole-Exome Sequencing (WES). Computational pipelines were developed with Linux, R, and Python scripting. Through the RNA-seq pipeline, differentially expressed genes between PTF, LNM, benign LNs, and normal prostate were identified. A median of 57 million paired-end reads were obtained per sample. Comparing PTF to LNM, 8110 transcripts were differentially expressed (p-adj < 0.01). PTF were enriched relative to LNM in gene sets associated with Notch signaling, TGFb signaling, hypoxia, and the epithelial to mesenchymal transition. Comparing PTF from metastatic patients to non-metastatic patients, 581 transcripts were differentially expressed (p-adj < 0.01). PTF from metastatic patients were enriched in cell cycle progression, MYC targets, ER stress, androgen response, and DNA repair. LNM gene sets were enriched in endoplasmic reticulum (ER) stress and oxidative phosphorylation. We also identified a set of 193 genes with significantly increased expression in primary tumors over benign LNs and in LNM over primary tumors. This gene set was significantly enriched in genes related to oxidative phosphorylation and included oncogenes such as *PIK3CB, NCOA2,* and *SCHLAP1*. The WES pipeline revealed genomic variant and tumoral heterogeneity information. The top mutated genes include *SPOP, EYA1,* and *NCOR2*. These somatic mutations may drive cancer proliferation via dysregulation of the AR signaling pathway. Through WES analysis, we identified mutations associated with PCa metastasis. The top mutated genes associated with metastasis are *SOGA1, LRRC4C, TP53, COL5A1, PCDHA13*, and *SLC16A14*. Our results are vital to the investigation of prostate cancer metastasis, as genomic changes drive oncogenic progression. By understanding the mechanism of metastasis, we may be able to improve clinical strategies to target PCa.

# Transcriptome Analysis of Primary Prostate Tumor Foci and Corresponding Lymph Node Metastases Identifies Pathways Associated with Metastatic Disease

By

Emma Klein

Dr. Carlos Moreno

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Science with Honors

Biology

2023

# Table of Contents

List of Figures

# 1. Introduction

## 1.1 Prostate Cancer Biology

Prostate cancer (PCa) is the most common cancer diagnosed in males and second highest cause of cancer-associated death in the United States (1). According to the American Cancer Society, 1 in 8 men are diagnosed with PCa and the mortality rate is 1 in 41. It is estimated that there will be approximately 288,300 new cases and 34,700 deaths from PCa in 2023 (2). Therefore, it is essential to investigate the mechanism of PCa progression in order to optimize patient treatment and understand disease pathogenesis.

Cancer is partly driven by somatic mutations – these are genomic alterations that emerge during the lifetime of an individual. A single abnormal cell with acquired genetic mutations can spark the creation of a neoplasm (3). When a single cell acquires mutations beneficial for tumor growth, it may encounter multiple rounds of selection and clonal expansion. As a result, cancerous cells proliferate and heterogeneous tumors are formed. Heterogeneous tumors include populations of cancerous cells with variable combinations of somatic mutations. This is a distinctive hallmark of PCa, as every tumor has a unique mutational signature.

Previous studies have aimed to create molecular and genetic profiles of primary prostate cancer. Multiple genomic alterations have been identified such as mutations, rearrangements, gene fusions, and DNA copy-number changes. Schultz et al. presented a primary PCa molecular analysis for the The Cancer Genome Atlas (TCGA). Based on 333 primary prostate carcinomas, they identified seven subtypes characterized by ETS fusions or mutations in *SPOP, FOXA1,* and *IDH1* (4). Other studies focus on metastatic, castration-resistant prostate cancer (mCRPC) through bone or soft tissue biopsies. Robinson et al. classified genomic aberrations in mCPRC pertaining to the *AR, PI3K, Wnt*, Cell-Cycle, and DNA Repair pathways. They found that nearly 90% of mCRPC harbor clinically-actionable molecular alterations (5). Despite the importance of genomic profiles in PCa, the molecular underpinnings that lead to the initial steps of metastasis have not been well characterized. Mortality is mainly due to PCa metastasis, so it is essential to understand the key drivers of tumor progression.

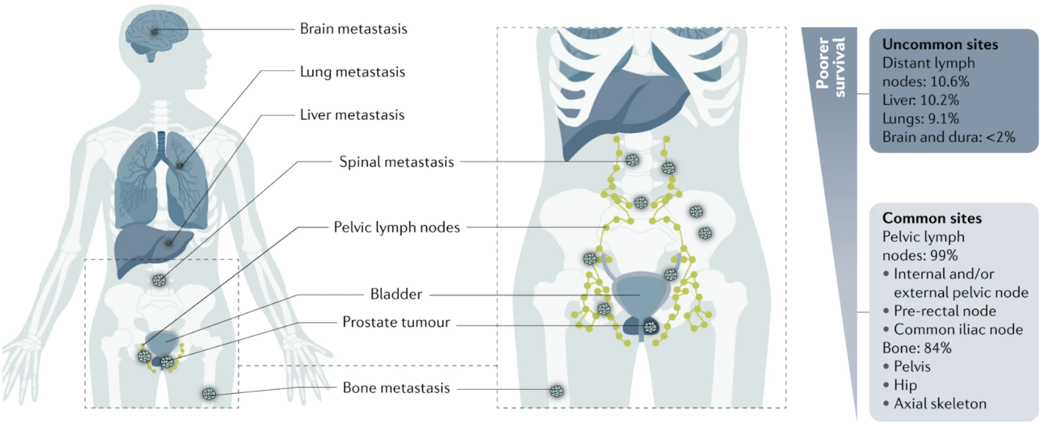
Prostate cancer cells spread through the bloodstream or lymphatic system, forming bone and lymph node (LN) metastases. PCa metastasis is a key component of diagnosis and defines a patient's stage of cancer. Staging is critical because it allows cancer care teams to categorize patient cases and optimize treatment plans. The most widely used staging system for PCa is the AJCC TNM staging system (6). The TNM system includes 5 parameters – the extent of the main primary tumor (PT), whether the cancer has spread to LNs, whether the cancer has metastasized to other parts of the body, prostate-specific antigen (PSA) level, and grade group.

Epithelial cells in the prostate express high levels of AR (androgen receptor), which drives hormone dependency in PCa. AR activates transcription of PSA, a serine protease, which is elevated in PCa patients. PSA blood tests are commonly used to screen for PCa in men without symptoms. High PSA levels indicate a higher likelihood of PCa. Furthermore, the grade group is determined based on the gleason score of a prostate biopsy. The Gleason System assigns grades based on tissue comparisons between cancer and normal prostate cells. Abnormal cells are assigned higher grades, while cells similar to normal prostate tissue are assigned low grades. Scores are generated by summing the grades of the two predominant lesions in the prostate. Grade groups include ranges of Gleason scores based on severity level.

Stages of PCa include Stage I, IIA, IIB, IIC, IIIA, IIIB, IIIC, IVA, and IVB. As stage number increases, symptoms and severity increase (7). According to the American Cancer

Society, the five-year survival rate is almost 100% for Stages I-IVA because PCa is still localized and regional. Earlier stages include cancer that has not spread outside of the prostate or cancerous cells that spread to nearby regions. However, the 5-year survival rate of Stage IVB patients is 31%. During Stage IVB, tumors infiltrate other parts of the body such as the bones, liver, or lungs (8). It is evident that most PCa mortality is driven by metastases (Fig. 1). The staging system facilitates the optimization of patient treatment and understanding of PCa progression.

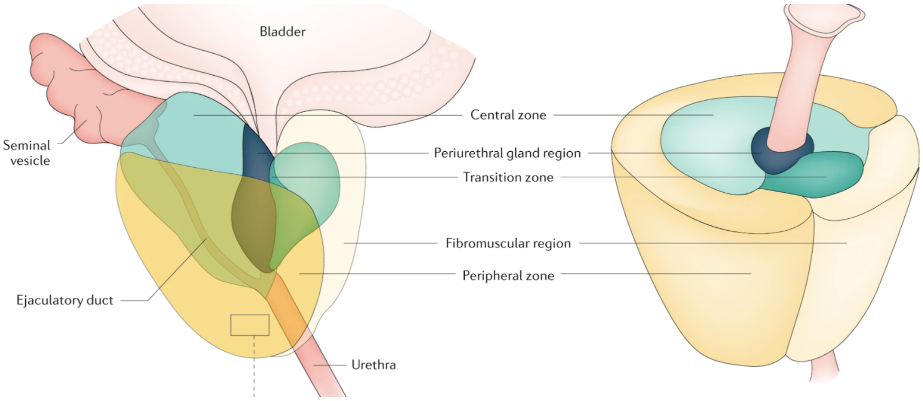
**Figure 1. Overview of PCa metastasis.** Most common sites of metastases are pelvic lymph nodes and bone. Adapted from Nature Reviews Disease Primers (Nat Rev Dis Primers) (13)



### 1.2 Symptoms & Treatment

The prostate is a component of the male reproductive system and located below the bladder (Fig. 2). It is a small organ, comparable to the size of a golf ball (9). Seminal fluid is created within the prostate and then secreted into the urethra. This fluid combines with sperm to form semen, which is released during ejaculation. Since the prostate is in front of the rectum, prostate health can be evaluated through a digital rectal exam (DRE). The American Cancer Society recommends prostate exams for men over 50 years old, every three to five years. Early detection is essential for the best prognosis and treatment strategies. Since early stages of PCa are not usually linked with symptoms, it is important to routinely schedule prostate exams. However, advanced PCa can be linked with a myriad of potential symptoms, mainly associated with the urinary tract. Specifically, advanced PCa symptoms include painful urination, weak urine streams, blood in semen, bone pain, fatigue, and leg swelling (10).

**Figure 2. Prostate Anatomy.** Five regions of the prostate include the central zone (green), periurethral region (dark blue), transition zone (dark green), peripheral zone (dark yellow), and fibromuscular region (light yellow). Adapted from Nature Reviews Disease Primers (Nat Rev Dis Primers) (11)



There are many treatments used for PCa, which may cure patients with localized cancer. Most patients with stage IV PCa cannot be cured with current treatments, but their symptoms can be mediated and metastasis slowed. Early detection is essential for the best prognosis and treatment strategies, as localized PCa is treated with the highest efficacy. Treatment options significantly vary between patients based on disease pathology and staging. If PCa is localized, surgery is a very common treatment option. Patients undergo surgical procedures to directly remove malignant regions of the prostate and some of the surrounding LNs. Radiation therapy is often used to target localized PCa; high energy rays destroy cancer cells through DNA damage. Furthermore, systemic therapies are also used to eliminate cancer cells. These treatments include medications, which can be administered intravenously or through oral consumption (swallowing a pill or capsule). There are several types of systemic treatments including androgen deprivation therapy, targeted therapy, chemotherapy, and immunotherapy (12).

### 1.3 Genomic Sequencing & Informatics

Sequencing technologies that identify the order of nucleic acid sequences in biological samples have revolutionized biomedical research. There are several types of genomic sequencing, such as Sanger Sequencing, Next Generation DNA Sequencing (NGS), and Long-Read Sequencing. Through the past few decades, sequencing technologies have become more efficient, more accurate, and cheaper. Whole-Genome Sequencing (WGS) uses an entire DNA sample in order to construct a donor's entire genome. WGS cuts the DNA into fragments using sonication, amplifies them with PCR, and then sequences all the fragments. Whole-Exome Sequencing (WES) follows a similar process but only includes the exons, protein-coding regions, of the genome.

Furthermore, RNA-seq is used to investigate the transcriptome, as it includes the sequencing of RNA transcripts. This method is helpful to understand genomic expression based on RNA transcript levels. Single-cell sequencing (scRNA-seq) conducts RNA sequencing on each individual cell in a sample. As a result, data can be clustered based on cell type, facilitating the analysis of individual cell gene expression. These sequencing technologies are commonly used in cancer genomics. Informatics tools are essential to the analysis of sequencing data. Through computational methods, sequencing data can be mapped to the human genome and transformed into interpretable genomic expression data. As a result, biological patterns and genomic variation can be uncovered.

# 2. Methodology

## 2.1 Main Objectives

Mortality is mainly due to metastases of highly heterogeneous tumors. However, the molecular underpinnings that lead to the initial steps of metastasis have not been well characterized. In order to optimize future treatments, it is essential to investigate the mechanism of PCa, as it proliferates and spreads throughout the human body. The main objective of this study is to identify potential key driver mutations and important genomic expression changes involved in PCa metastasis.

PCa is characterized by intratumoral heterogeneity (ITH) on multiple levels (13). Primary prostate cancers are known to be multifocal, which underlies the rationale for the Gleason scoring system. As aforementioned, Gleason scores sum the most predominant tumor grade with the second most predominant tumor grade in a radical prostatectomy specimen (14-17). Previous studies have found that over 70% of patients have multifocal disease representing multiple tumor grades (16). Furthermore, different primary tumor foci are composed of genetically distinct clones, suggesting independent carcinogenesis events within the prostate gland (21-24,25). ITH has been identified even within single primary tumor foci in the prostate (26), suggesting co-mingling of independent clones with branching and divergent evolution.

ITH is a poorly understood phenomenon that is critically important for understanding tumor progression and the development of drug resistance, as different sub-clones can respond differently to microenvironmental changes and selection pressure from therapies. Recent studies have indicated that independent foci and biopsies can have markedly different performance in commercially available biomarker panels (27). Thus, understanding ITH is essential for biomarker development and validation, prognosis, and therapeutic decision-making for precision medicine.

Mortality from prostate cancer is due to metastases, and thus, understanding the mechanisms of metastasis is also essential for improving patient outcomes. The heterogeneity of metastases is poorly understood, and there is some disagreement in the literature between opposing models of PCa metastasis. Some data using copy number analyses support a monoclonal model of metastasis, in which most metastatic lesions derive from a single clone despite the multiclonal nature of the primary tumor (28). Additional studies have shown limited genomic diversity in multiple metastases from the same men (29). However, conflicting studies support polyclonal seeding of metastases based on whole genome sequencing of 51 tumors from 10 patients (30). Some of the differences between these studies may be due to varying degrees of detail and granularity in the methods employed to molecularly characterize PCa metastases. In addition, it is not clear if the genomic variability observed in metastases in some studies are due to mutations that occur at the metastatic site or if polyclonal populations from different primary foci seeded those metastases. Most studies that have analyzed ITH have examined a limited number of patients and used only the index lesion of the primary tumor.

In this project, we aimed to address this gap in our understanding of heterogeneity in PCa metastasis and to leverage unique resources of many patients from a clinical trial that includes both radical prostatectomy and extensive dissection of all pelvic LNs. Since metastasis to surrounding LNs is one of the first steps in metastatic spread, understanding ITH in pelvic LNs will provide unique insights into the initial mechanisms and heterogeneity of PCa metastasis. Additionally, we aimed to analyze multiple primary foci beyond the index lesion, greatly increasing the richness of the dataset that we will generate in these studies.

Specifically, three main aims were identified to study PCa disease progression. Aim 1 is to perform RNA-seq analysis of multiple primary foci and corresponding LNs to identify gene expression changes and pathways associated with LN metastasis. RNA-seq analysis can be used to define tumor heterogeneity and enable identification of RNA gene expression patterns. Transcriptomics may help discriminate primary foci that can support metastasis from those that remain indolent and localized. Aim 2 is to perform whole exome sequencing (WES) of multiple primary foci and corresponding LN to identify somatic mutations associated with LN metastasis. WES analysis can also help define intratumoral heterogeneity in the earliest steps of the metastatic process. Lastly, Aim 3 is to determine RNA and DNA signatures associated with the presence and degree of uptake of [18F]-fluciclovine in LN metastases. Aim 3 is not included in the scope of this project.

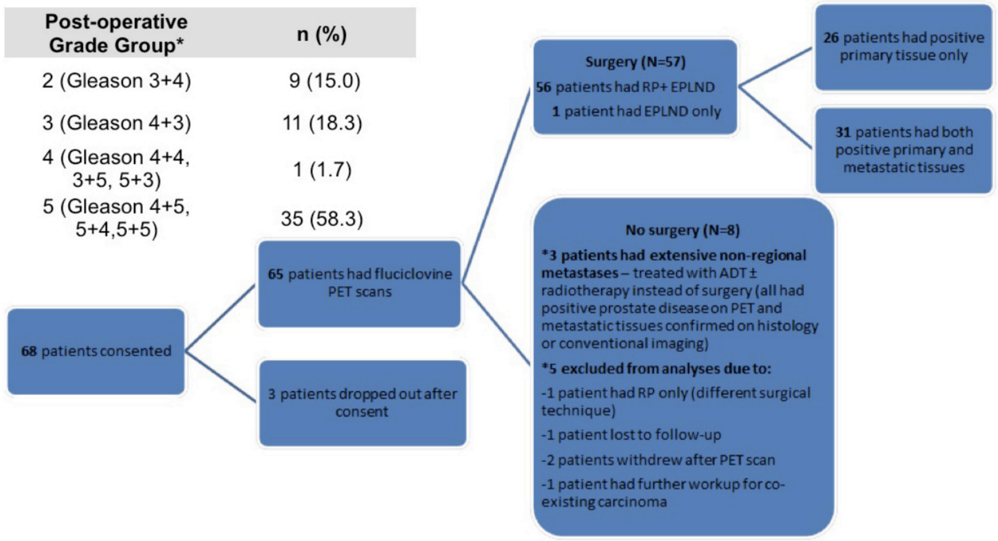
It is essential to investigate PCa disease progression in order to optimize treatment and patient prognosis. This study uses an integrated functional and clinical genomics approach to reveal genes driving aggressive metastatic PCa.

### 2.2 Experimental Design

The Emory, Harvard, & University of Washington Prostate Cancer Biomarker Center, led by Dr. Sanda, conducted a clinical trial (NCT01808222) to determine if [18F]-fluciclovine PET imaging can detect significant occult metastatic disease in patients with high risk PCa (Fig. 3). These patients had negative or equivocal conventional imaging such as CT, MR, and bone scan. In this study, 56 PCa patients within high or very high risk groups (T3a, Gleason score 8-10, or PSA greater than 20 ng/ml) were selected based on criteria that correspond to a 50-80% PSA failure rate in the first 5 years after prostatectomy.

Patients underwent radical prostatectomies and extended pelvic lymph node dissections. The surgical plan generally involves nodal dissection of left and right obturator, external iliac, and internal iliac nodes for the high-risk patients. Each group of nodes is removed as a packet, labeled separately, and sent to pathology for routine histopathologic examination. If either conventional imaging or [18F]-fluciclovine PET imaging demonstrates potential other pelvic or extrapelvic nodal disease, the surgeon may choose to extend the nodal dissection to other sites. When this is complete, the additional nodes are labeled to indicate the site and they are also submitted for pathology review. In this trial, 56 patients underwent radical prostatectomies. Of these, 30/56 (54%) of patients had metastases. Of those, 7 had only one positive lymph node, and 23 had two or more positive lymph nodes. A total of 92 positive LNs were identified out of a total of 2480 excised LNs of which 58 positive LNs were > 4mm in diameter.

**Figure 3. Summary of EDRN Fluciclovine Trial**



| Post-operative Grade Group* | n (%) |
|---|---|
| 2 (Gleason 3+4) | 9 (15.0) |
| 3 (Gleason 4+3) | 11 (18.3) |
| 4 (Gleason 4+4, 3+5, 5+3) | 1 (1.7) |
| 5 (Gleason 4+5, 5+4,5+5) | 35 (58.3) |

68 patients consented
65 patients had fluciclovine PET scans
3 patients dropped out after consent
Surgery (N=57)
56 patients had RP+ EPLND
1 patient had EPLND only
26 patients had positive primary tissue only
31 patients had both positive primary and metastatic tissues
No surgery (N=8)
*3 patients had extensive non-regional metastases – treated with ADT ± radiotherapy instead of surgery (all had positive prostate disease on PET and metastatic tissues confirmed on histology or conventional imaging)
*5 excluded from analyses due to:
-1 patient had RP only (different surgical technique)
-1 patient lost to follow-up
-2 patients withdrew after PET scan
-1 patient had further workup for co-existing carcinoma

This project uses tissue from the aforementioned clinical trial, in which patients with aggressive prostate cancer underwent radical prostatectomies and pelvic LN dissections. Multiple samples were collected from each patient, so it would be possible to track changes in genomic expression as the cancer metastasizes from the prostate to the LNs. Patient samples include PT, normal LN, and metastatic LN. This experimental design is advantageous so that comparisons can be made between and within patients, using their own genomes as a baseline.

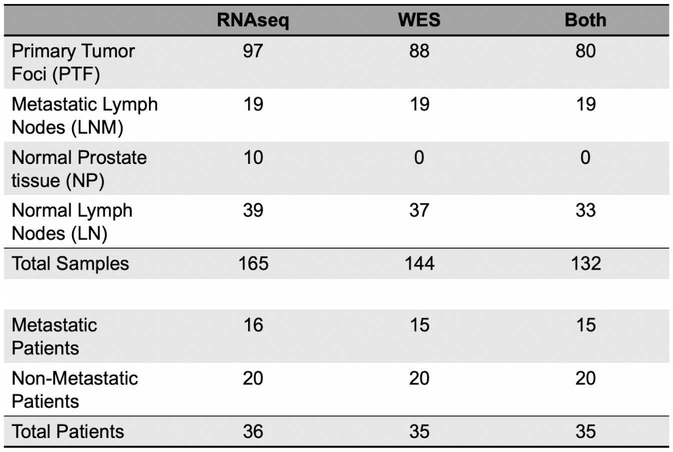
Emory University Hospital (EUH) pathology surgical services has Formalin-Fixed Paraffin-Embedded (FFPE) tissue samples from the previous clinical trial, consisting of prostate and lymph node dissections from 56 cases. For 30 (54%) of cases, the cancer had metastasized to the surrounding lymph nodes. Of these 56 cases, 7 did not give consent (4 metastatic and 3 non-metastatic) and were removed from the study. For the patients in which the cancer did not metastasize (n=23), we identified at least 3 FFPE prostate tissue samples and 1 LN (benign) for examination. For the cases in which the cancer did metastasize (n=26), we identified at least 4 prostate samples, 1 benign LN and 1 metastatic LN. All available FFPE blocks were collected from EUH pathology services and coded to remove any PHI. For 3 of the cases, the met LN blocks were not available, and were removed from the study.

All collected FFPE blocks were sent to Winship Cancer Tissue and Pathology Shared Resource Core (CTPSR) for sectioning. Seven sections, at 5 µm each, were made from each FFPE block. One section from each block was stained with Hematoxylin and Eosin (H&E) for cancer cell identification. The H&E stained sections were analyzed by a GU pathologist to identify regions of interest for macrodissection and sequencing. For the non-metastatic patients, two prostate samples and one benign lymph node sample were selected for sequencing. For metastatic patients, three prostate samples, one met lymph node, and one benign lymph node were selected. Six metastatic patients were omitted from the study, as all corresponding lymph node metastases were under 4mm.

The pathologist also used the H&E stained sections to identify and mark the cancerous regions within the tissue samples. This served as a template to identify and mark the same region on the six unstained sections from the same block. The identified regions from all 6 sections were macrodissected, and the tissue collected into a shearing microtube provided in the Covaris truXTRAC FFPE total NA kit. The FFPE tissue samples were then sent to Emory Integrated Genomics Core (EIGC) for RNA and DNA isolation using the aforementioned Covaris kit. QC analysis was also performed by EIGC on all RNA and DNA extractions. For any sample whose RNA was below a concentration of 5 ng/ul, and/or DNA under 500 ng, more sections of the tissue block were made. Macrodissection of the identified area was performed again, for isolation of more RNA and DNA. If the necessary amount could not be isolated, another FFPE block from the same patient was selected and used for RNA/DNA extraction. Once the total RNA and DNA collected were above the given threshold, the samples underwent sequencing.

The final patient count included 36 individuals with high risk PCa. A total of 165 tissue samples (51 Met PT, 46 Non-Met PT, 19 LN Met, 39 Normal LN, 10 Normal Prostate) were sequenced via RNA-seq and and 144 tissue samples were sequenced via WES (Fig. 4). RNAseq and WES were performed at HudsonAlpha, part of Discovery Life Sciences.

**Figure 4.** Sample distribution across 40 PCa patients

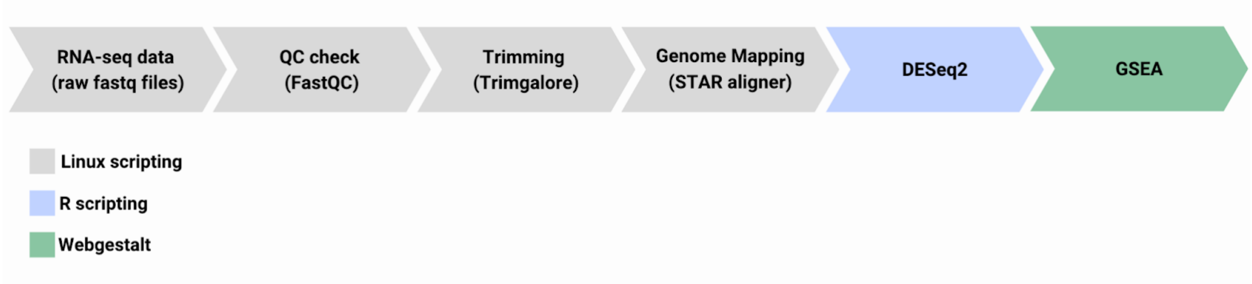


| | RNAseq | WES | Both |
|---|---|---|---|
| Primary Tumor Foci (PTF) | 97 | 88 | 80 |
| Metastatic Lymph Nodes (LNM) | 19 | 19 | 19 |
| Normal Prostate tissue (NP) | 10 | 0 | 0 |
| Normal Lymph Nodes (LN) | 39 | 37 | 33 |
| Total Samples | 165 | 144 | 132 |
| | | | |
| Metastatic Patients | 16 | 15 | 15 |
| Non-Metastatic Patients | 20 | 20 | 20 |
| Total Patients | 36 | 35 | 35 |

### 2.3 RNA-seq Pipeline

The RNA-seq pipeline was constructed with Linux shell programming, R scripting, and Gene Set Enrichment Analyses (GSEA) (Fig. 5). Raw FASTQ files served as input to analyze RNA-seq data from patient samples. First, FastQC was used for quality control checks of the high throughput data to eliminate poor quality reads (31). Next, the reads were trimmed with TrimGalore to remove adapter sequences and poor quality reads. Genome mapping with the trimmed reads was conducted with STAR mapper (32). Trimmed reads were mapped to the human transcriptome based on the GRCh38 reference. STAR Aligner determines locations in the human genome associated with read data. This alignment strategy is highly accurate and outperforms other aligners in mapping speed.

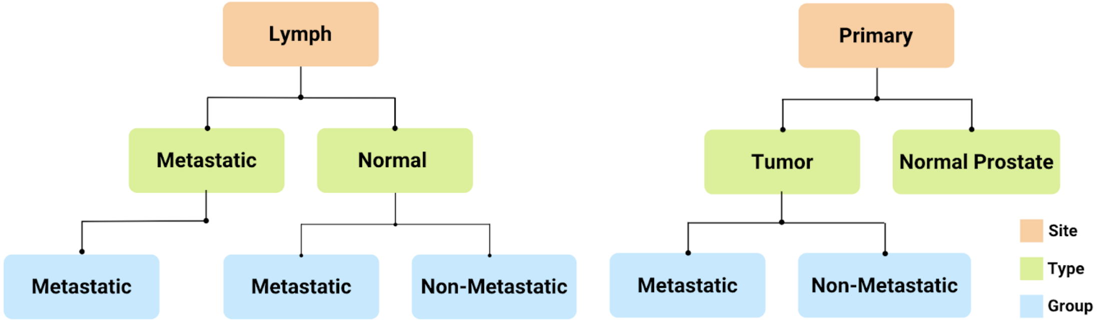
**Figure 5.** Full flowchart of the constructed RNA-seq pipeline. Linux scripting (gray), R scripting (blue), and Webgestalt (green) were used as computational tools.



The STAR alignment algorithm includes two main steps: (1) Seed searching and (2) Clustering, stitching, and scoring. In seed searching, STAR aligns reads with the longest sequence that matches one or more locations on the reference genome. Seeds are different parts of a particular read that are mapped separately to different genomic locations. This alignment method is sequential – STAR continues to search for unmapped sections of each read that matches the reference genome. STAR uses an uncompressed suffix array to search for the longest matches. Separate seeds are combined to create a full read by clustering, stitching, and scoring. The output of STAR aligner is read counts per gene.

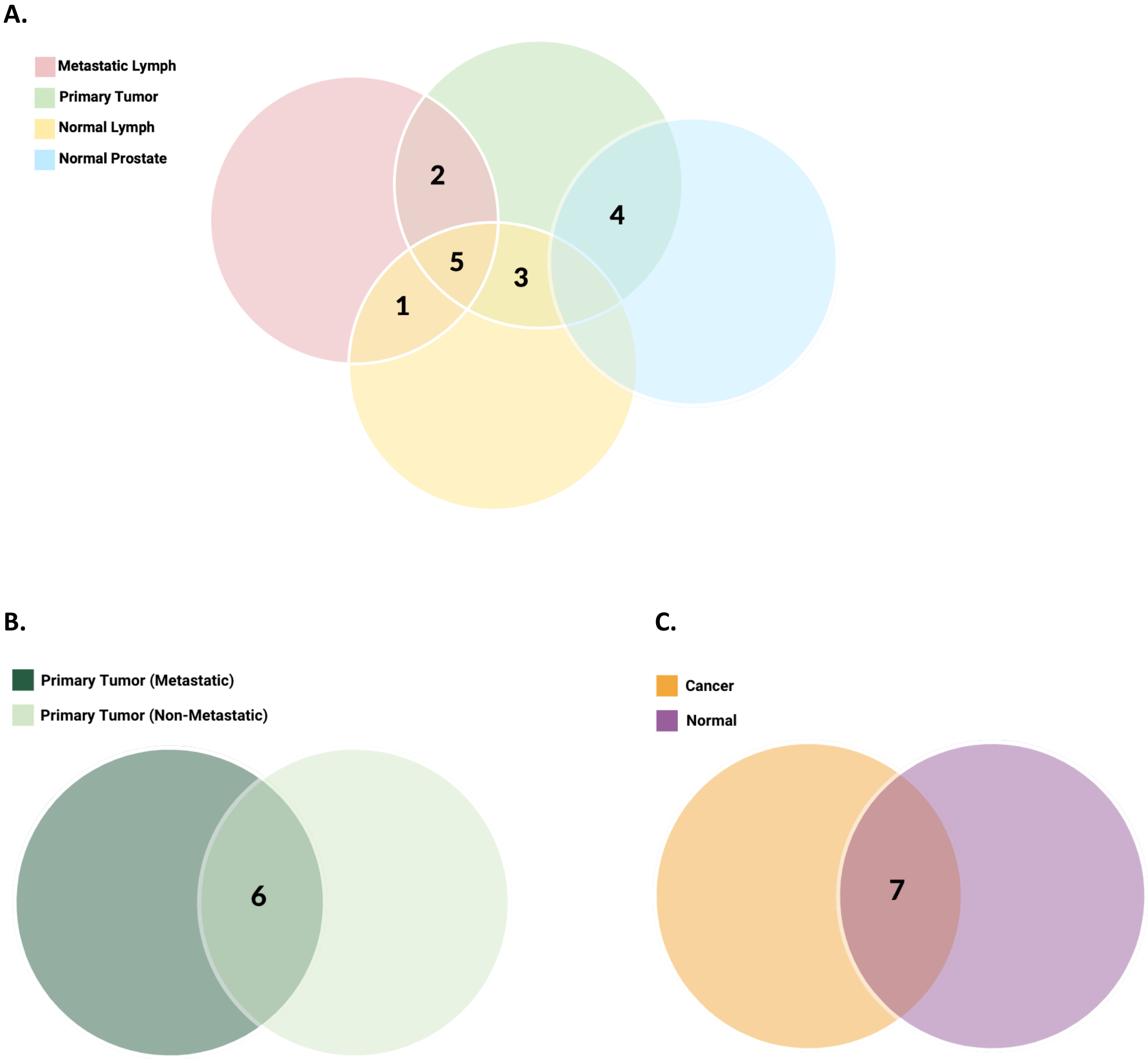
Through R scripting, DESeq2 was used to determine differences in RNA expression among the samples (33). The DESeq2 package includes negative binomial generalized linear models and read count normalization. Through DESeq2, we aimed to identify differentially expressed genes between sample groups (Fig. 6). Each patient sample is associated with a specific site, type, and group.

**Figure 6.** Group divisions of PCa patient samples. Site (orange) is the sample location, Type (green) represents type of sample, and Group (blue) is the sample group based on patient status.



Several comparisons were analyzed based on group, site, and type of tumor. DESeq2 objects were created for each comparison based on specific subsets of the data. There are 7 total comparisons across each of the main 4 categories: primary tumor foci (PTF), normal lymph nodes (LN), metastatic lymph nodes (LNM), and normal prostate (NP). Specifically, the comparisons 1-5 are based on overlap between the 4 aforementioned categories (Fig. 7a). Comparison 6 is based on overlap within PTF samples – between metastatic and non-metastatic patients (Fig. 7b). Lastly, comparison 7 compares all normal patient samples with all cancer samples. In the normal category, patient samples include NP and LN, while all tumors and metastases are in the cancer category (Fig. 7c).

**Figure 7. Comparisons of PCa samples. (A)** Comparisons (1-4) between the four categories, including LNM (pink), PTF (green), LN (yellow), and NP (blue). A three-way comparison (5) was conducted between LNM, PTF, and LN. **(B)** Comparison 6 compares samples within the PTF group between metastatic (dark green) and non-metastatic (light green) patients. **(C)** Comparison 7 compares all cancer (orange) samples to normal (purple) samples. Cancer samples include all tumors and metastases, while normal includes LN and NP.



The output includes variation between group read counts for each gene: baseMean, log2FoldChange, log fold change standard error, stat, p-value, and adjusted p-value (p-adj). DESeq2 detects outliers and removes those genes from the analysis, based on a normalized count threshold value. The DESeq2 stat value is the Wald statistic for significance testing. This value is helpful to identify genes that are differentially expressed between the two compared groups. Genes are considered differentially expressed genes (DEG) if there is significant variation in RNA transcript level or normalized read count data. DEGs can reveal genomic variation and particular expression patterns in each comparison.

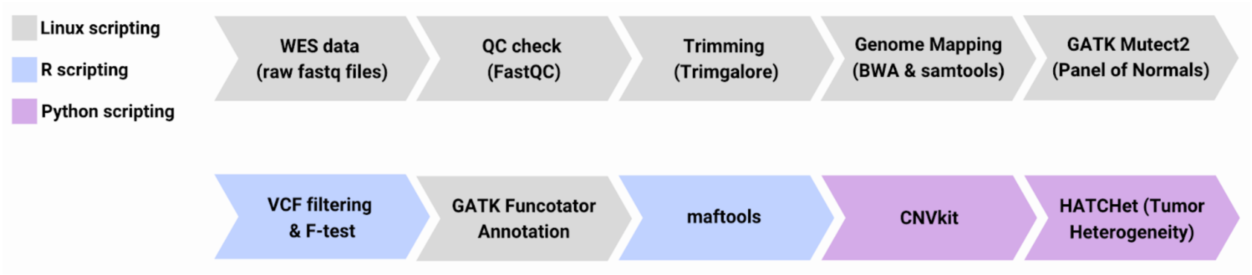
Gene set enrichment analysis (GSEA) is a computational method to identify DEGs. In this case, GSEA input files were generated from the DESeq2 results. There is one RNK file for each comparison with two columns – ensembl gene ID and DESeq2 stat. WebGestalt (WEB-based Gene SeT AnaLysis Toolkit) was used to conduct several GSEA for each comparison (34). WebGestalt is a functional analysis webtool, which utilizes gene rank values to explore genomic enrichment patterns. Four main GSEA methods were applied to each of the seven comparisons. These databases include geneontology (GO), KEGG, Hallmark50, and Wikipathway cancer. The GO database includes a wide array of gene ID with associated biological functions (35). KEGG PATHWAY database consists of pathway maps across various categories such as metabolism, genetic information processing, and human diseases (36). The community-contributed Hallmark50 database includes 50 gene sets related to cancer progression. Similarly, the Wikipathway cancer database includes various biological pathways linked with cancer. Each database provides additional parameters to reveal genomic variation across the comparisons.

### 2.4 WES Pipeline

The WES pipeline was constructed with Linux, R, and Python scripting (Fig. 8). The initial input is raw FASTQ files from the WES output. Similar to the RNA-seq pipeline, FastQC was used for quality control checks of the high throughput data and eliminated poor quality reads. Next, the reads were trimmed with TrimGalore to remove adapter sequences. Genome mapping was conducted with the Burrows-Wheeler Alignment (BWA) algorithm from the GATK package (37). BWA is a software package for mapping WES data, which aligns short reads with a reference genome. To be consistent with the RNA-seq pipeline, the GRCh38 human genome reference was used for the mapping process. The output of BWA is SAM files (Sequence Alignment/Map), which are text files that contain alignment information. Next, Samtools was used to convert SAM files into Binary Alignment Map (BAM) files (38). BAM files are essentially compressed SAM files that can serve as the input for Picard.

Genome Analysis Toolkit (GATK) and Picard Tools were used to calibrate read group qualities and mark duplicates (39). Afterwards, the Mutect2 algorithm in the GATK package was used to call mutations to detect SNVs and indels in the sample data. First, a panel of normals was run through Mutect2 to identify all the germline and technical variation. This step filters out polymorphisms and sequencing errors based on normal tissue, and calls somatic mutations in tumor samples. Mutect2 produces Variant Call Format (VCF) files, which includes genomic variant information. VCF files were filtered with R scripts in order to exclude insignificant variants. Resulting filtered VCF files had sufficient read depth, coverage, allele frequency values, and high significance levels based on F-tests. Funcotator is a functional annotator that marks genomic variants based on read depth, allele frequency, and F-tests. This tool uses VCF files as input and outputs mutation annotation format (MAF) output files. These MAF files were analyzed with the R package, maftools, for variant analysis (40). Maftools identified recurrent mutations and patterns to build mutational profiles. BAM files were also analyzed using a python copy number calling pipeline (CNVkit) (41). This script outputs copy number (CN) changes with CN ratios and CN segments. Through CNVkit, CN variant files can serve as input for HATCHet, an algorithm to compute tumor heterogeneity (42). As a result, copy number aberrations (CNAs) and whole-genome duplications (WGDs) can be used to investigate tumor evolution and metastatic seeding patterns (43).

**Figure 8. Full flowchart of the constructed WES pipeline.** Linux scripting (gray), R scripting (blue), and Python scripting (purple) were used as computational tools to build the pipeline.


Linux scripting
R scripting
Python scripting
WES data (raw fastq files)
QC check (FastQC)
Trimming (Trimgalore)
Genome Mapping (BWA & samtools)
GATK Mutect2 (Panel of Normals)
VCF filtering & F-test
GATK Funcotator Annotation
maftools
CNVkit
HATCHet (Tumor Heterogeneity)

# 3. Results

## 3.1 RNA-seq Analysis

GSEA results include enrichment plots and genomic expression tables. Four databases were used for GSEA, including GO, KEGG, Hallmark50, and Wikipathway Cancer. As a result, there are four GSEA results for each comparison (Fig. 9-15). Furthermore, ORA results include similar bar plots. Darker colors represent a false discovery rate (FDR) less than or equal to 0.05, which indicates significant differential expression. Various plots were generated through R scripting to further investigate the comparisons. Principal Component Analysis (PCA) plots cluster the patient samples based on their similarity (Fig. 16). Heatmaps were created to visualize differential gene expression across different samples (Fig. 17).

**Figure 9. GSEA Results for C1.** Enrichment plots for LNM vs. LN for 4 different pathways. Blue bars signify enrichment in LNM. Orange bars signify enrichment in LN. (A) KEGG Pathway (b) Wikipathway Cancer (C) Hallmark50 (D) Geneontology (GO)

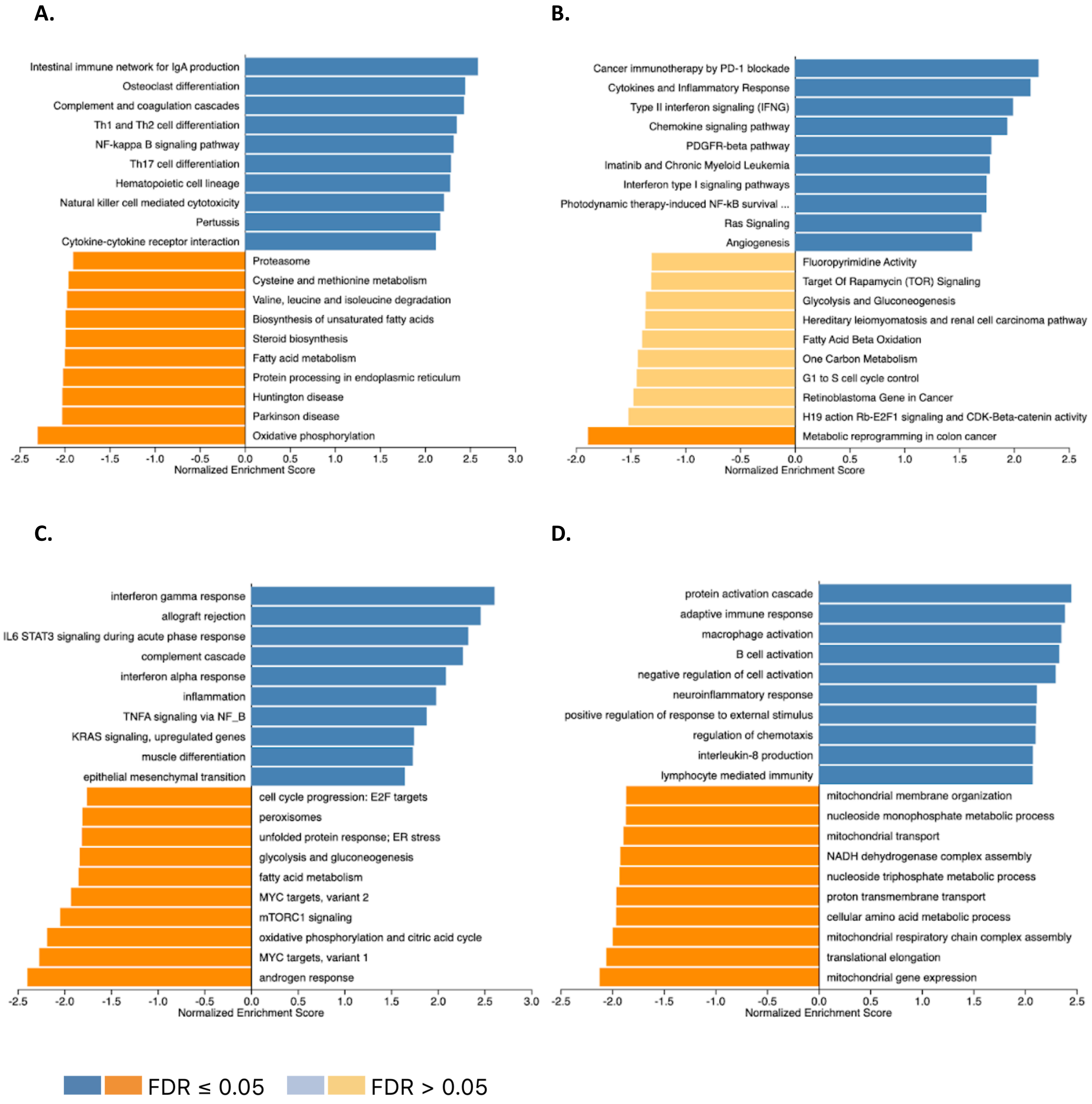


**Figure 10. GSEA Results for C2.** Enrichment plots for LNM vs. PTF for 4 different pathways. Blue bars signify enrichment in PTF. Orange bars signify enrichment in LNM. (A) KEGG Pathway (b) Wikipathway Cancer (C) Hallmark50 (D) Geneontology (GO)

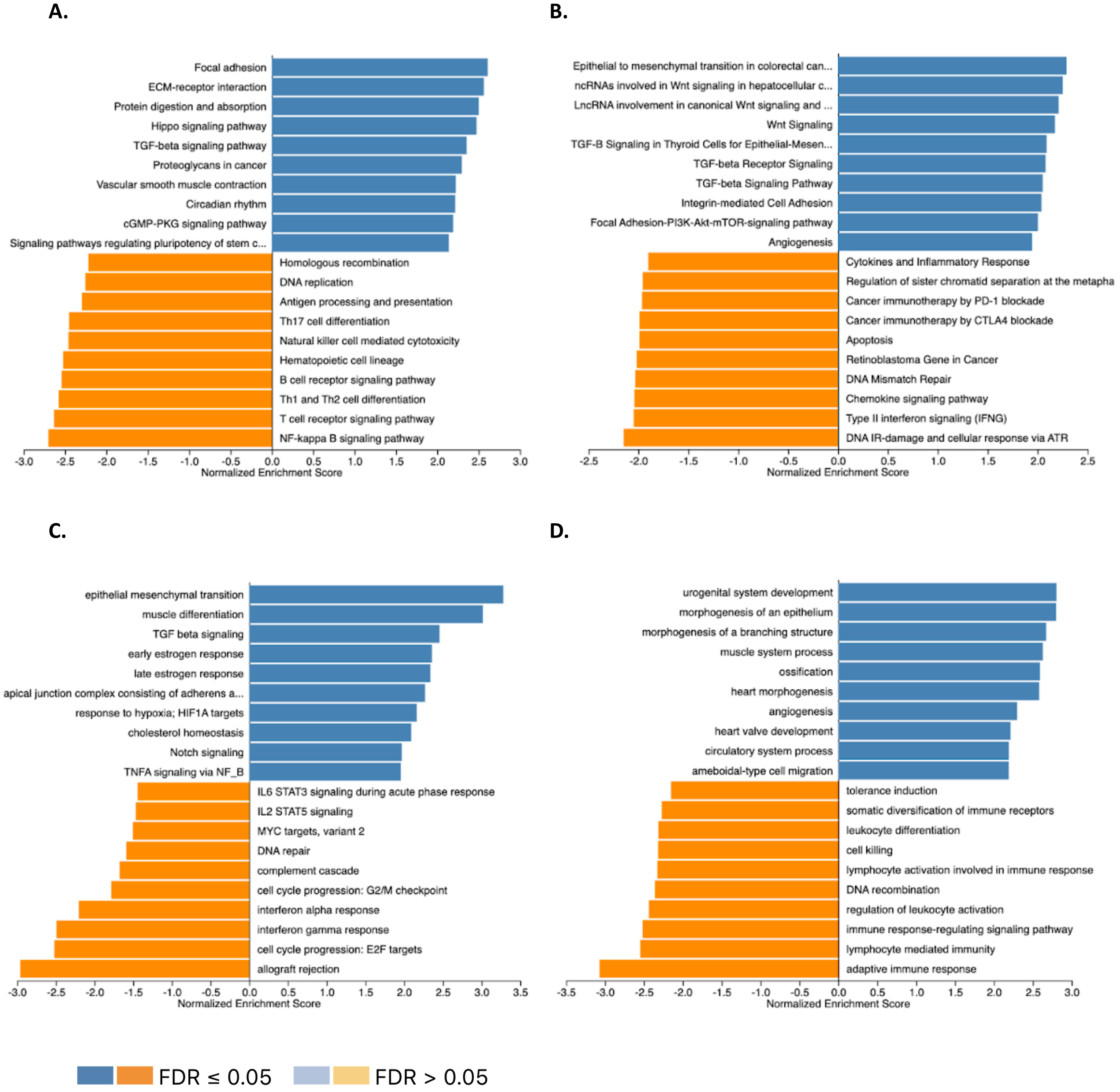


**Figure 11. GSEA Results for C3.** Enrichment plots for LN vs. PTF for 4 different pathways. Blue bars signify enrichment in PTF. Orange bars signify enrichment in LN. (A) KEGG Pathway (b) Wikipathway Cancer (C) Hallmark50 (D) Geneontology (GO)

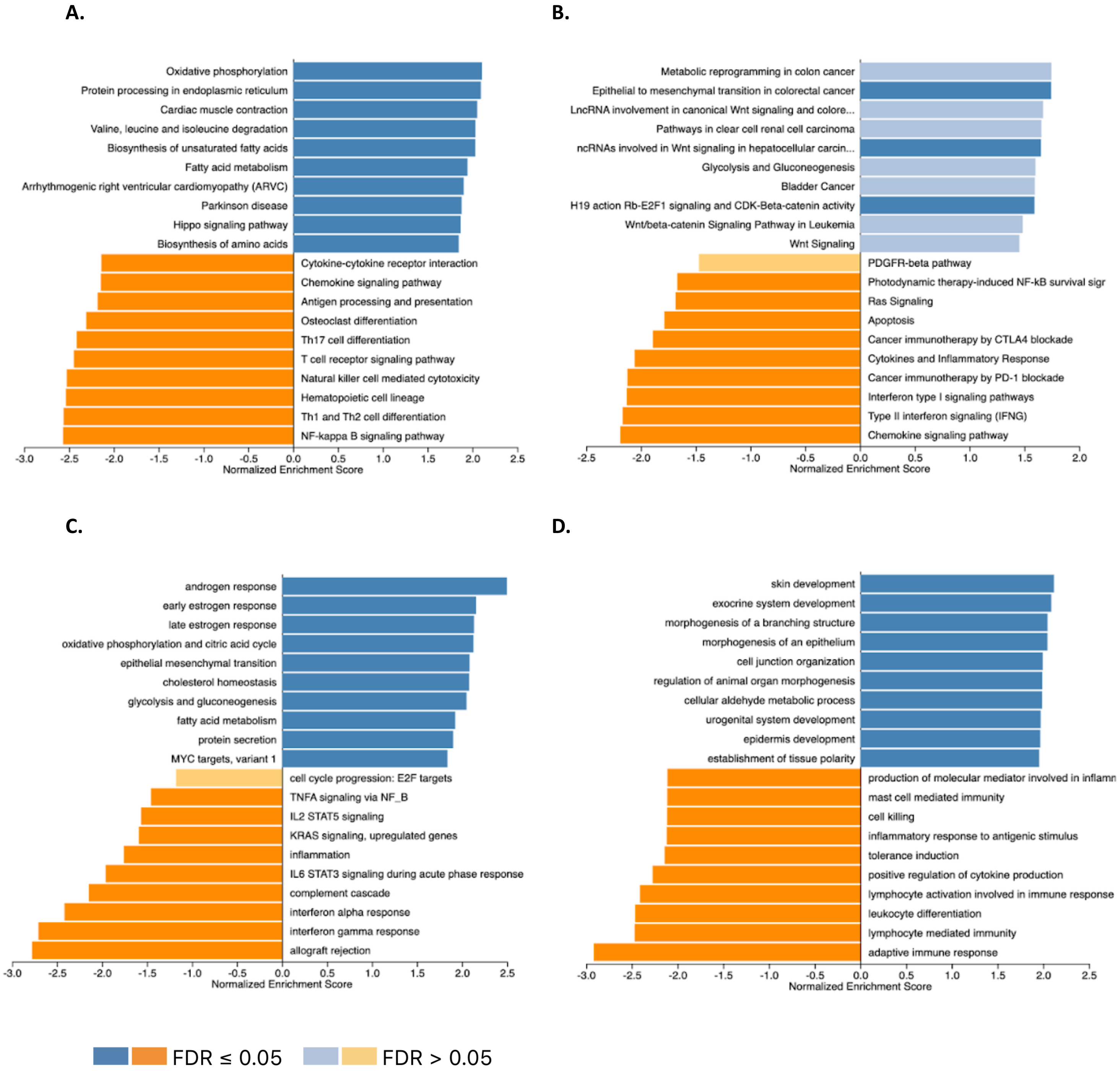


**Figure 12. GSEA Results for C4.** Enrichment plots for NP vs. PTF for 4 different pathways. Blue bars signify enrichment in PTF. Orange bars signify enrichment in NP. (A) KEGG Pathway (b) Wikipathway Cancer (C) Hallmark50 (D) Geneontology (GO)

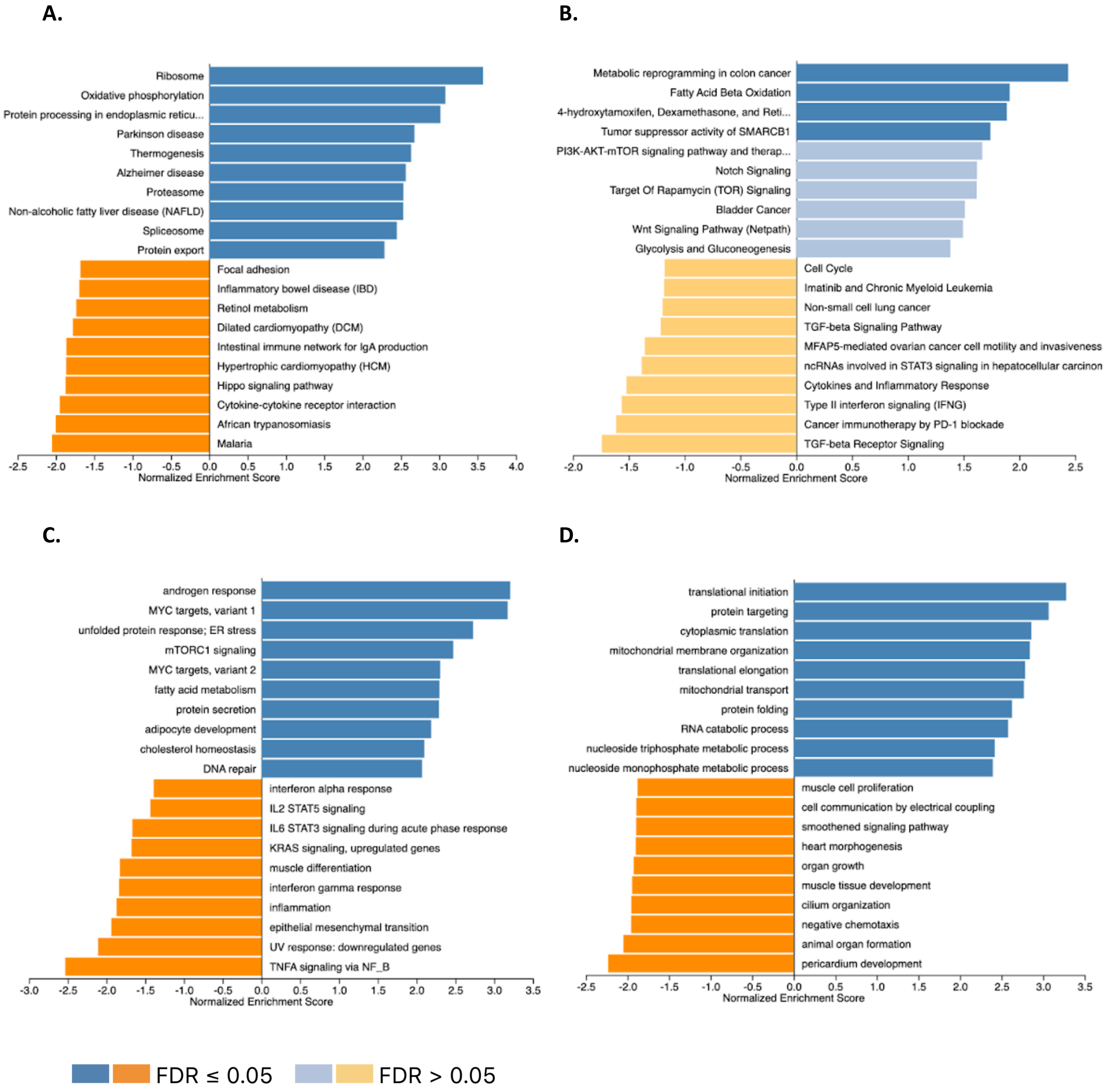


**Figure 13. GSEA Results for C5.** Enrichment plots for the three-way comparison (LN vs. LNM vs. PTF) for 3 different pathways. Wikipathway Cancer is excluded because it did not provide significant enrichment patterns. (FDR > 0.05) (A) KEGG Pathway (B) Hallmark50 (C) Geneontology (GO)

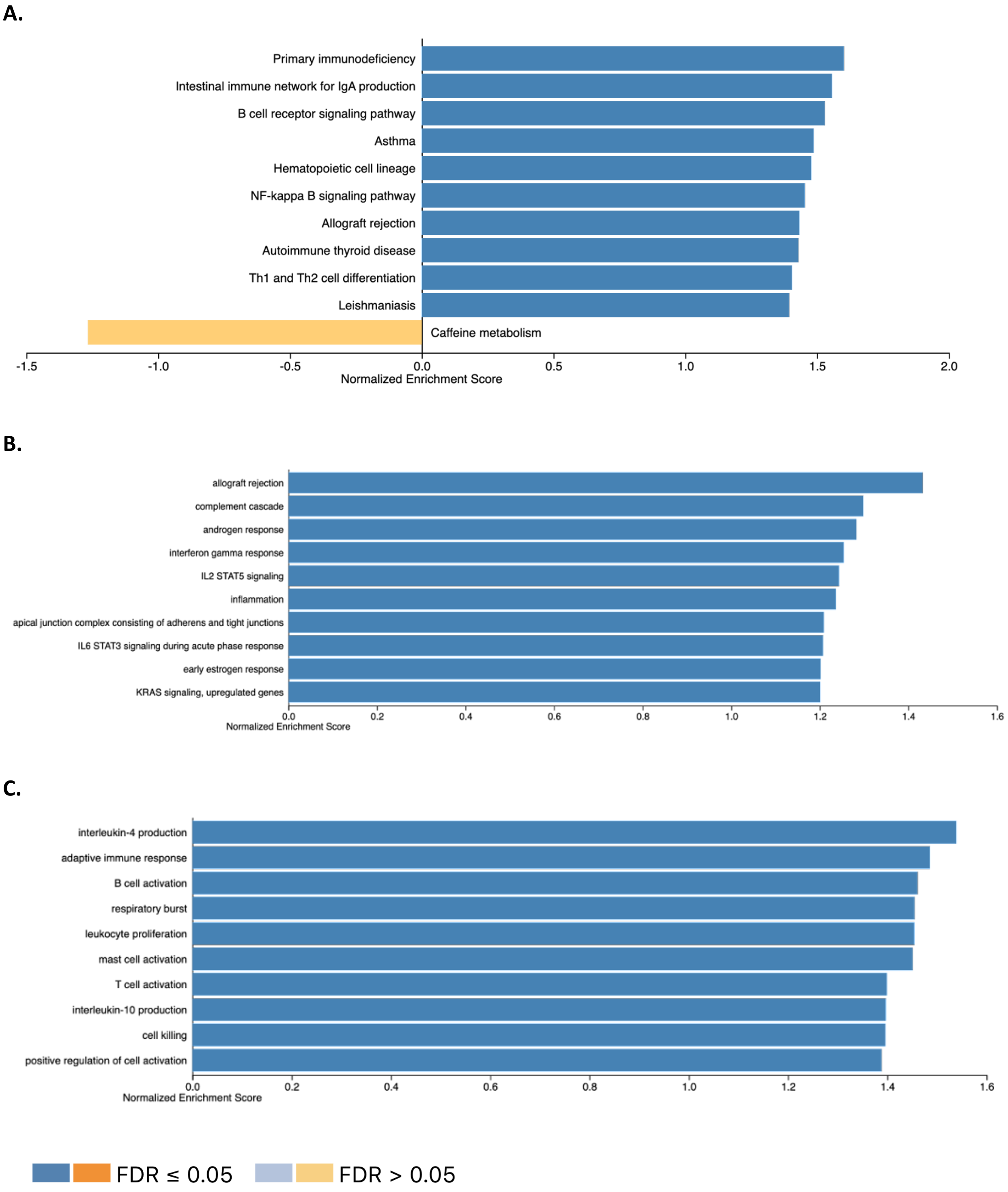


**Figure 14. GSEA Results for C6.** Enrichment plots for the PTF comparison between metastatic and non-metastatic patients for 4 different pathways.  Blue bars signify enrichment in non-metastatic PTF, while orange bars signify enrichment in metastatic PTF. (A) KEGG Pathway (b) Wikipathway Cancer (C) Hallmark50 (D) Geneontology (GO)

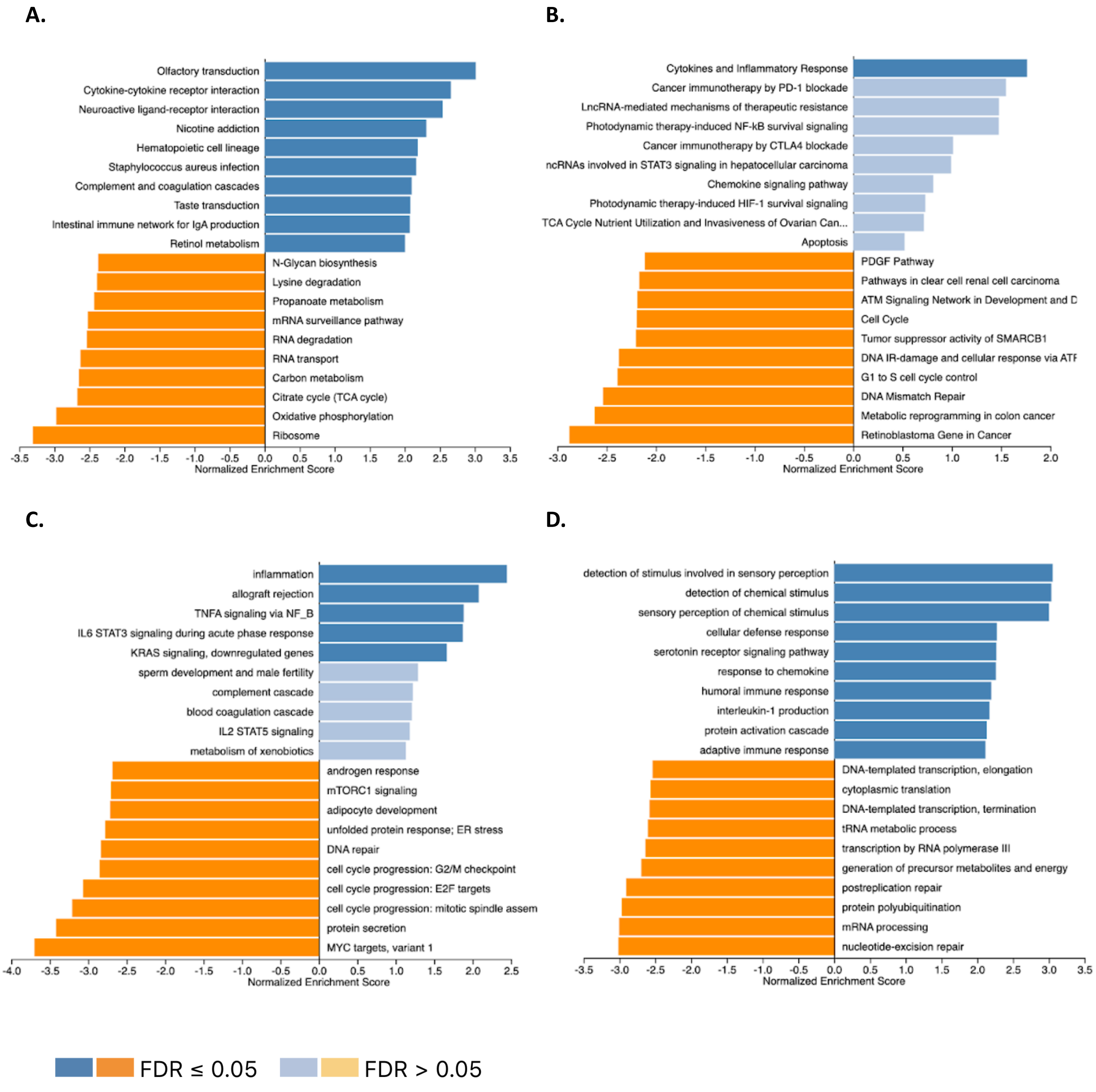


**Figure 15. GSEA Results for C7.** Enrichment plots for the comparison between all normal vs. all cancer for 4 different pathways. The normal category includes NP and LN, while all tumors and metastases are in the cancer category. Blue bars signify enrichment in cancer, while orange bars signify enrichment in normal. (A) KEGG Pathway (b) Wikipathway Cancer (C) Hallmark50 (D) Geneontology (GO)

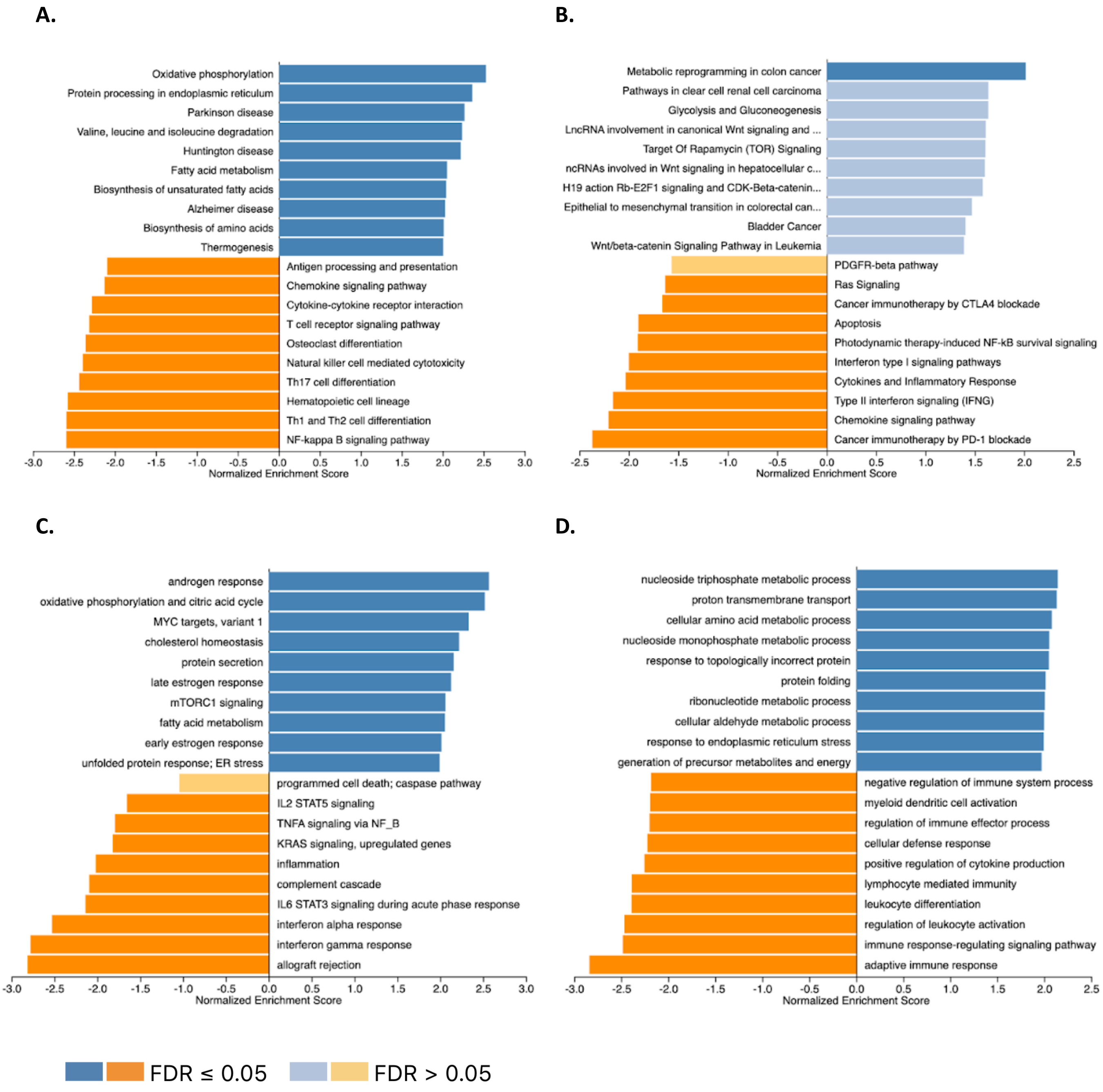
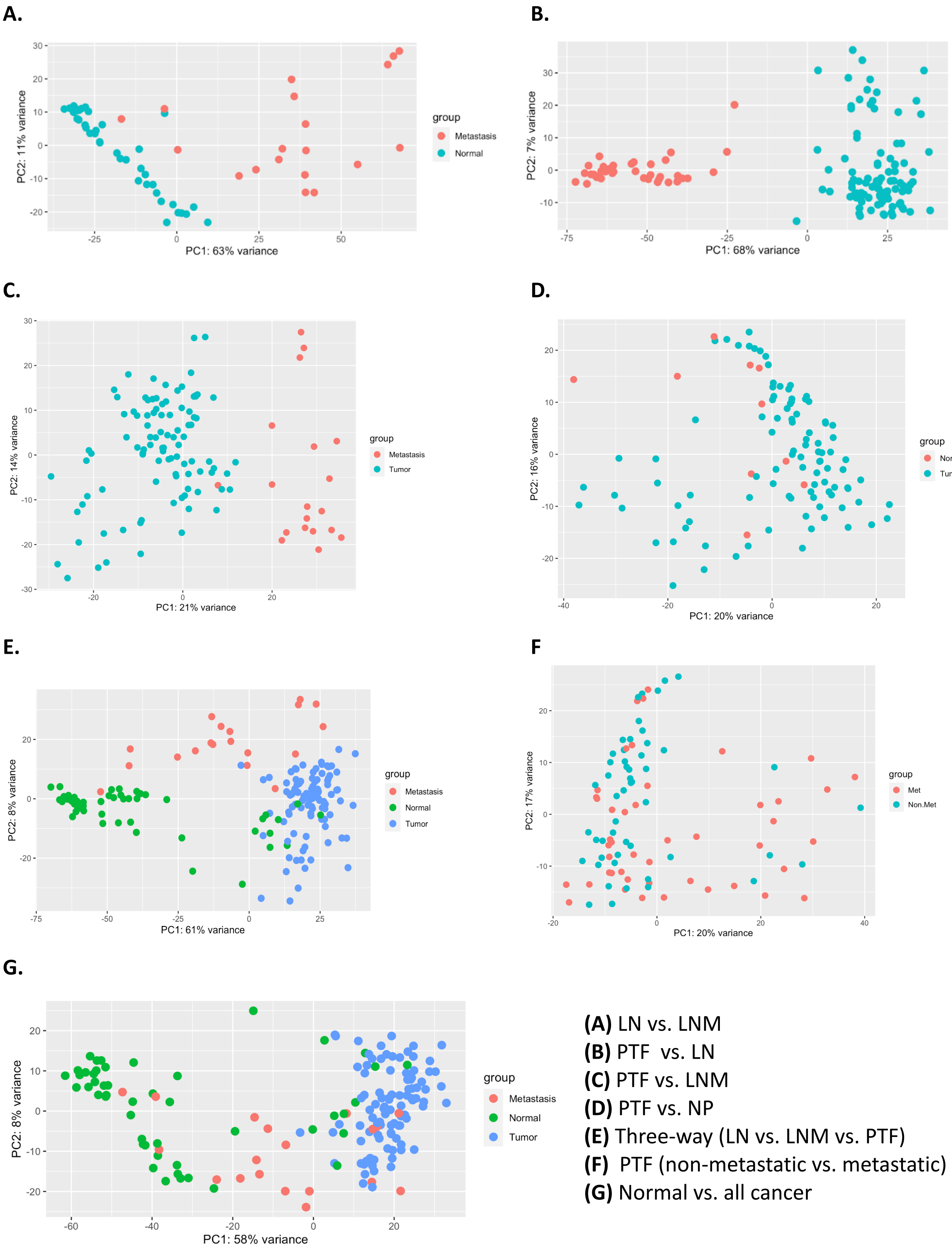


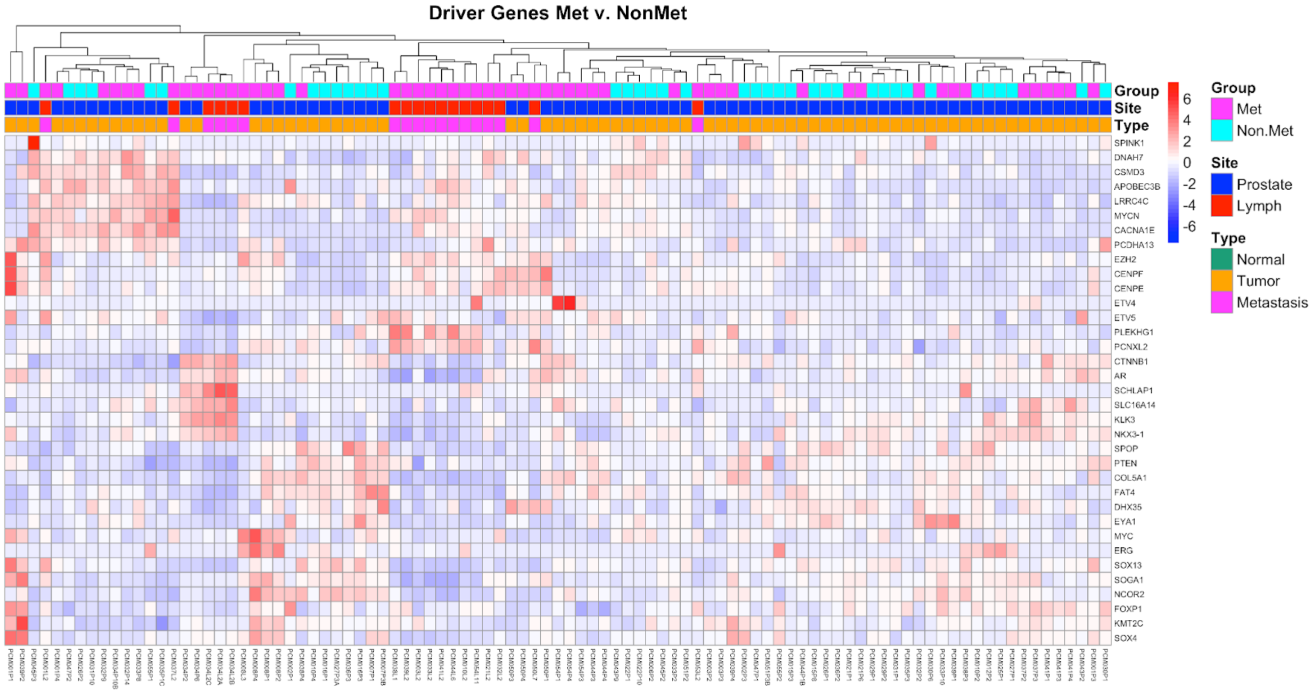
**Figure 16. Principal Component Analysis (PCA) plots.** PCA plots cluster samples based on gene expression similarity.



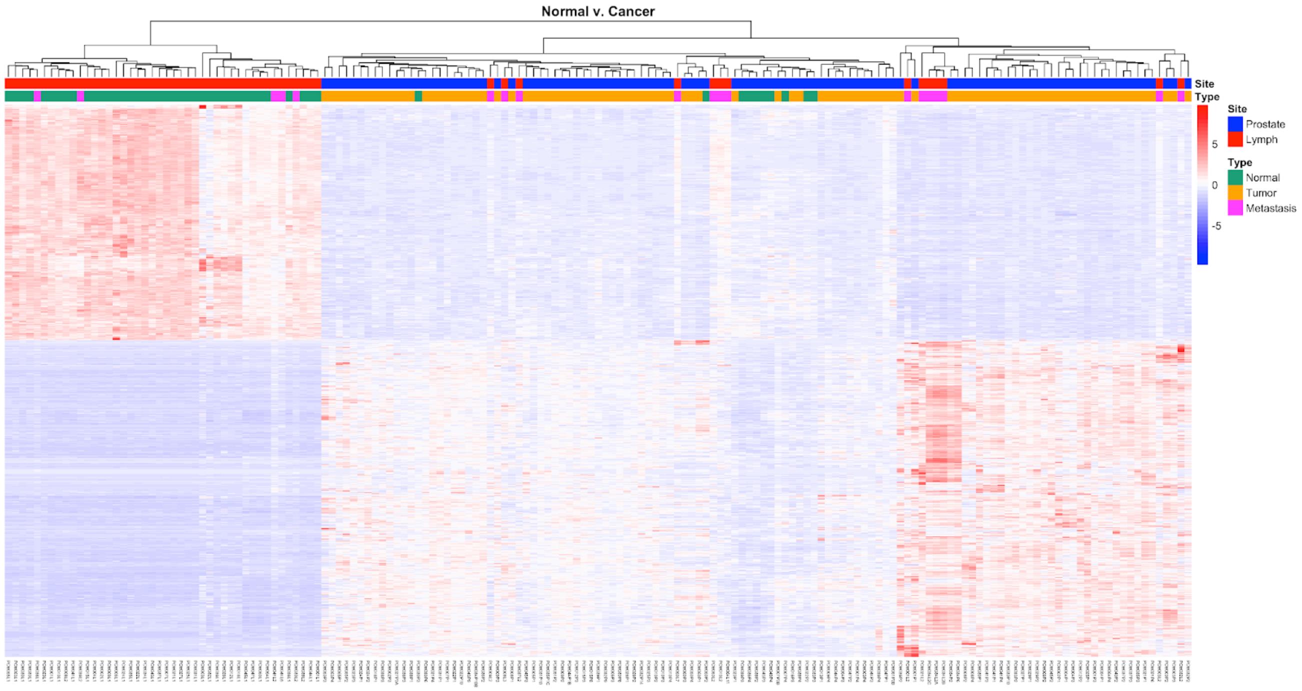
**(A)** LN vs. LNM
**(B)** PTF vs. LN
**(C)** PTF vs. LNM
**(D)** PTF vs. NP
**(E)** Three-way (LN vs. LNM vs. PTF)
**(F)** PTF (non-metastatic vs. metastatic)
**(G)** Normal vs. all cancer

**Figure 17. Hierarchical Clustering.** Heatmaps of differential gene expression. Only includes gene expression with padj<0.01. (A) PTF metastatic vs. non-metastatic (B) Normal vs. all cancer

**A.**



**B.**



### 3.2 WES Analysis

Maftools determines mutational patterns to build genomic profiles. Significant mutations were identified based on an allele frequency (AF) difference > 0.05 and false discovery rate (FDR) < 0.1. This filtering identified 1927 mutations in 91 patient samples within the aforementioned threshold values. Mutations affected several biological pathways such as RTK-RAS, WNT, and PI3K (Fig. 18). Variant information is highlighted, including variant classification, type, SNV class, and variants per sample (Fig. 19). Most variants are missense SNVs, and there is a high frequency of C>T mutations. The top mutated genes are also indicated, based on the number of samples across each site (Fig. 20). The top three mutated genes include *SPOP*, *EYA1*, and *NCOR2* (Fig. 21). Missense mutations are most common in the MATH domain of *SPOP,* which is involved in receptor binding and oligomerization.

**Figure 18. Affected Pathways.** Maftools output displays proportions of affected biological pathways.

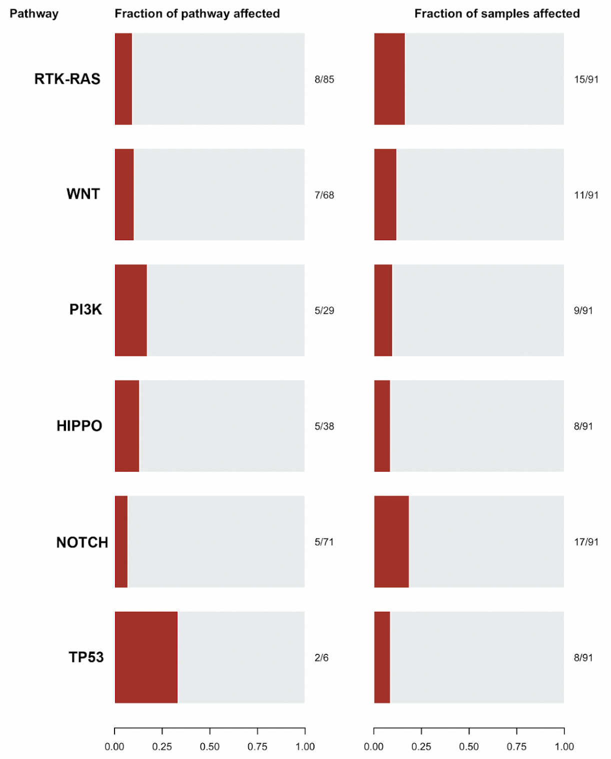


**Figure 19. WES Variant Summary.** Maftools output displays variant information across patient samples.

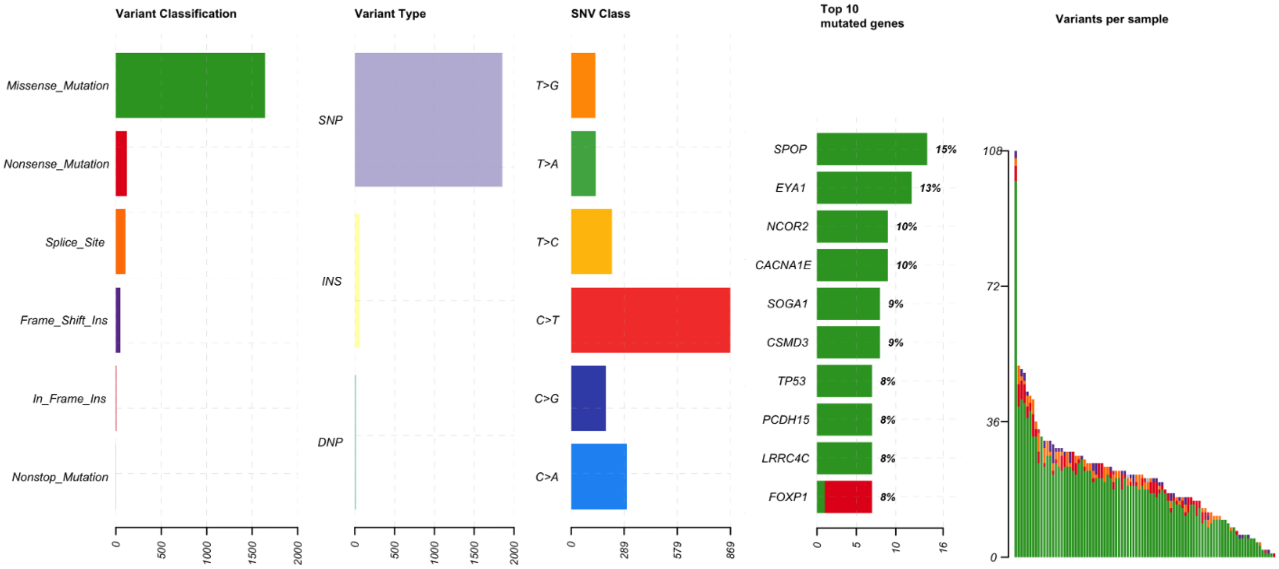


**Figure 20. Top Mutated Genes.** Maftools output displays mutations across patient samples.

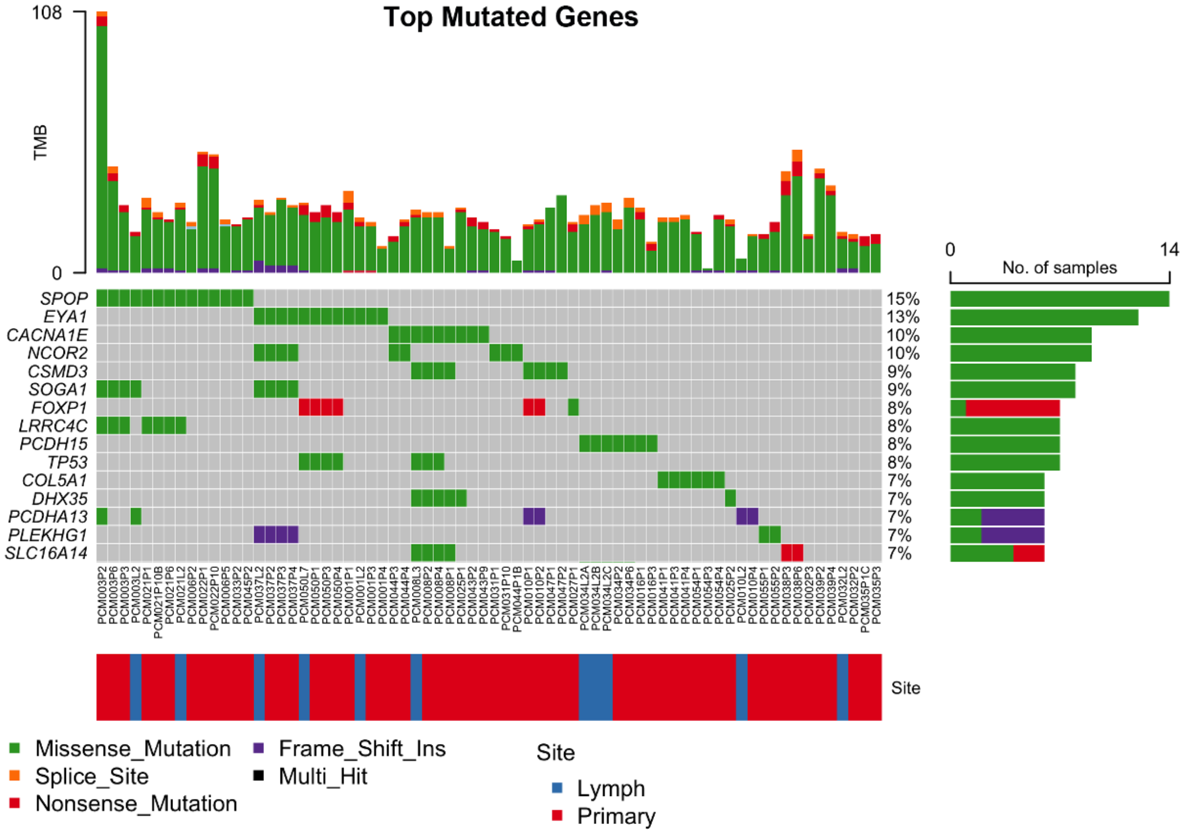
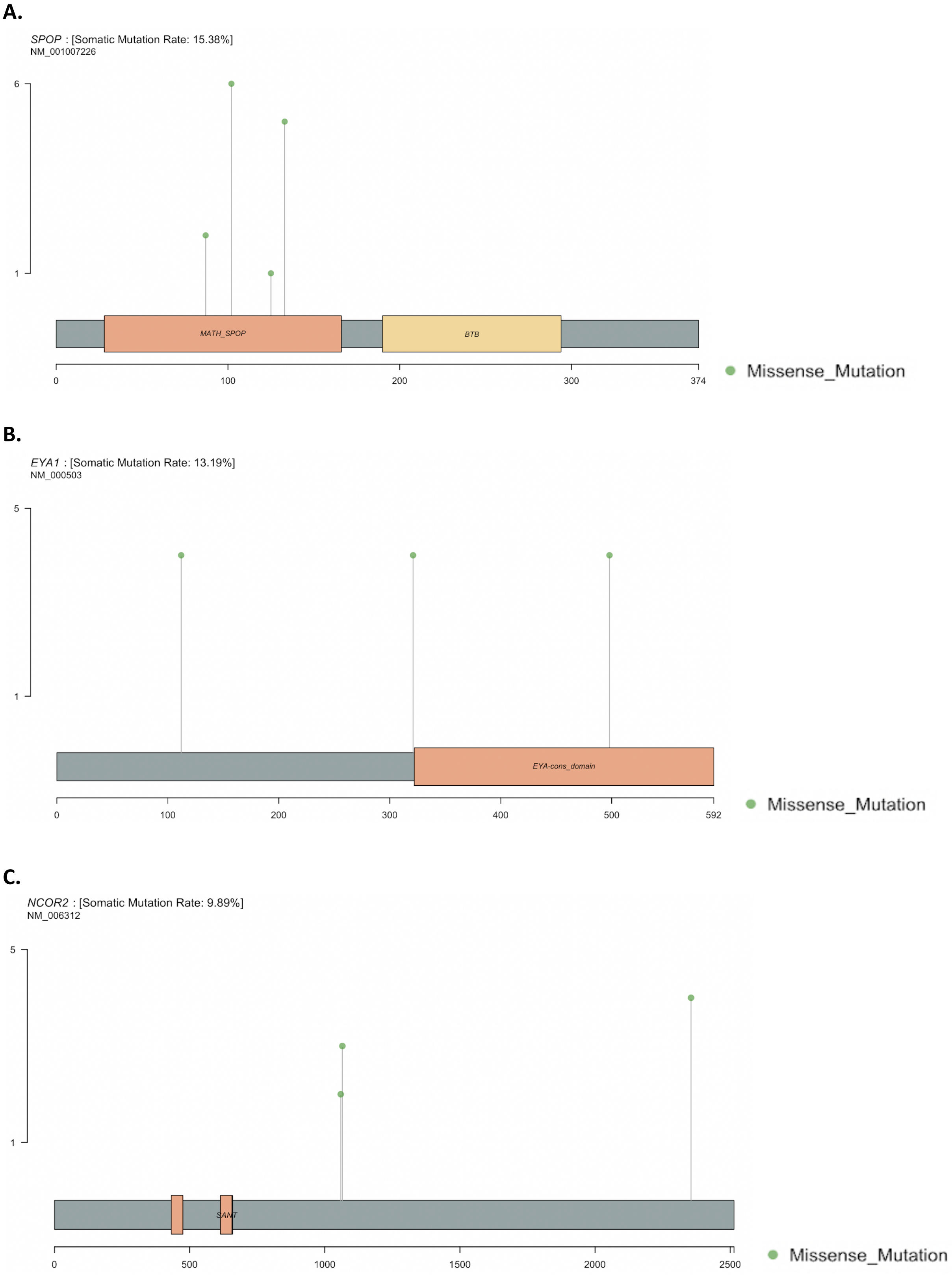
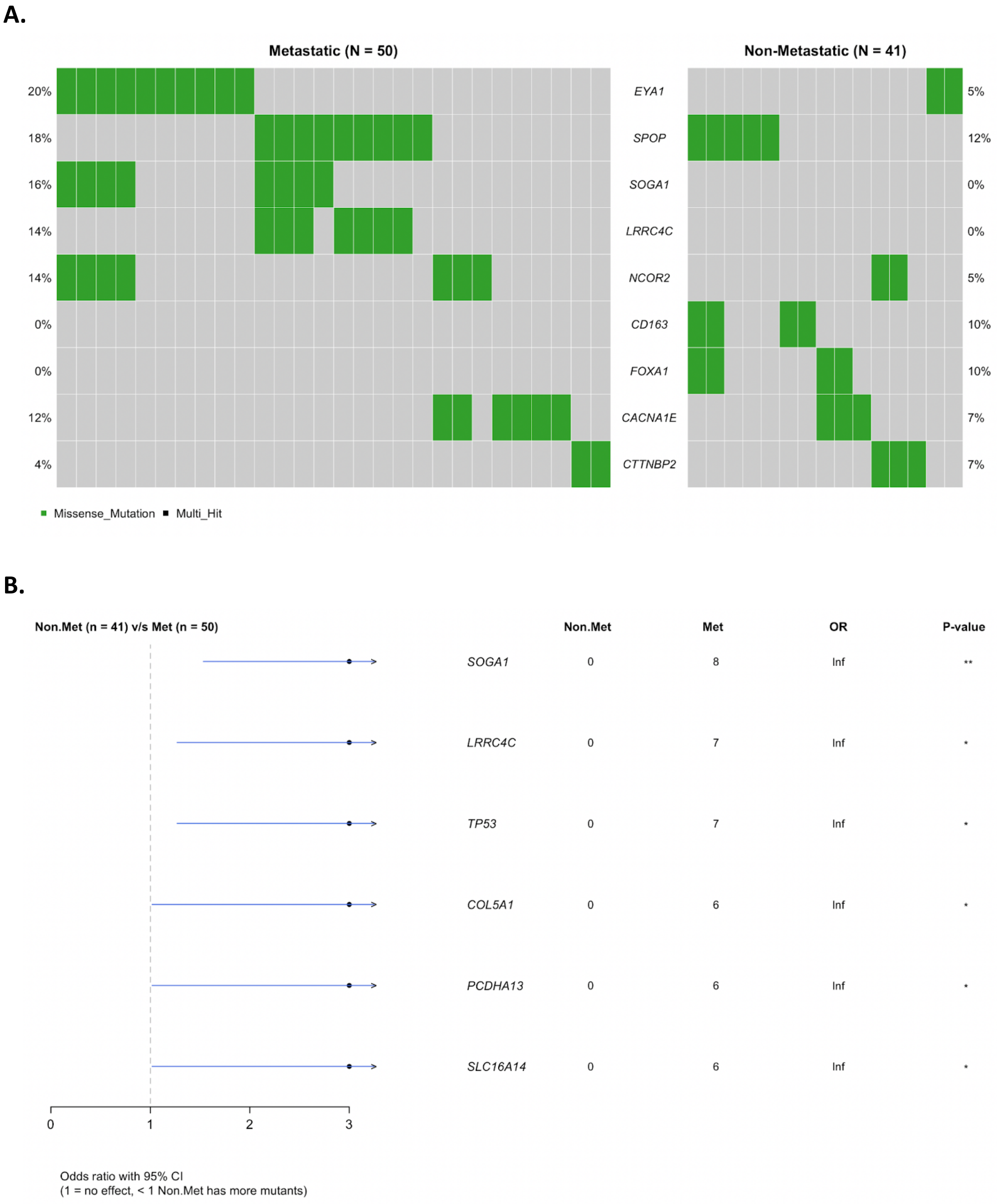


**Figure 21. Mutation Loci.** (A) *SPOP* mutation sites (B) *EYA1* mutation sites (C) *NCOR2* mutation sites



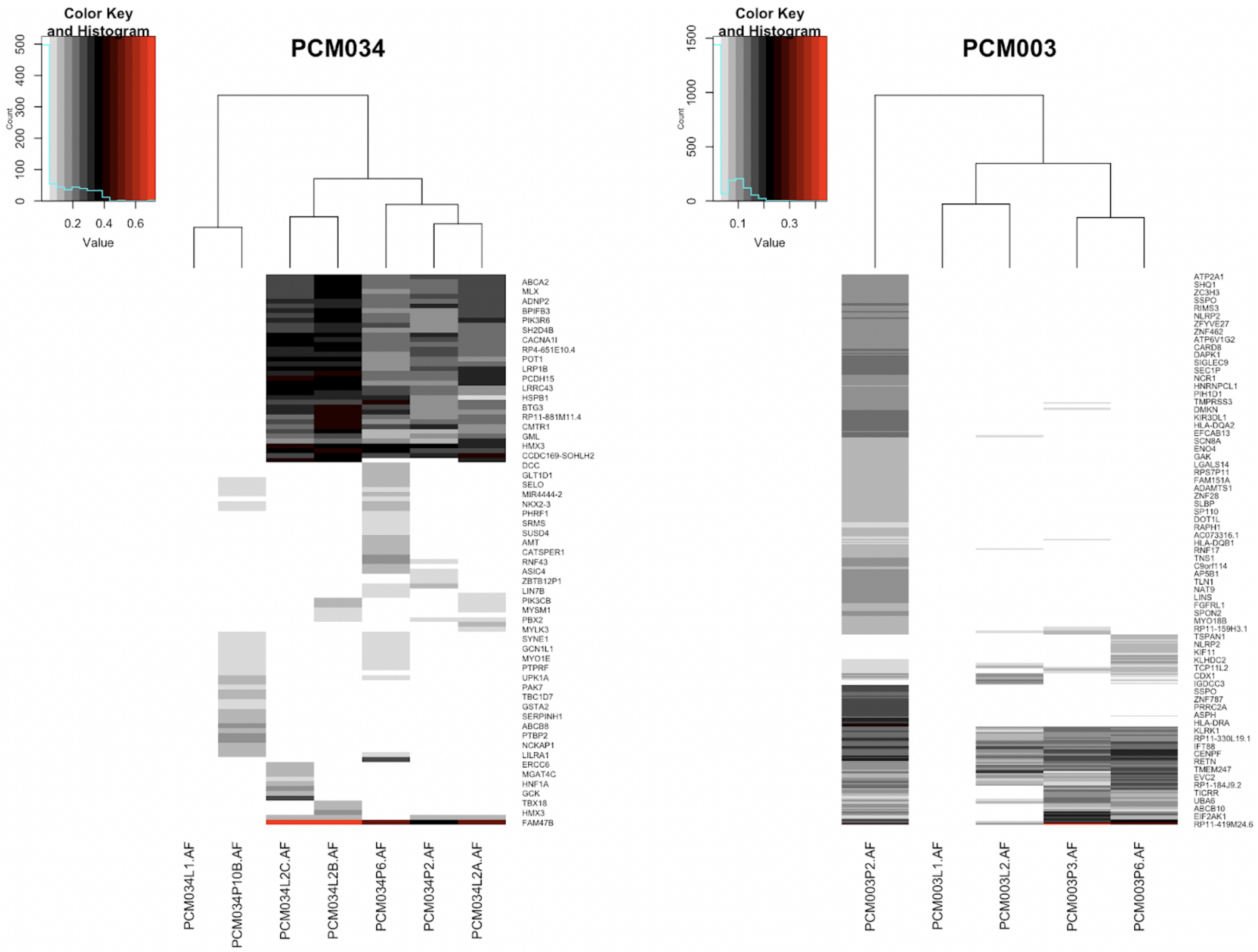
Furthermore, metastatic samples have a higher frequency of missense mutations compared to non-metastatic samples. Specific mutations are enriched in metastatic samples; the top three mutated genes in metastases include *SOGA1*, *LRRC4C*, and *TP53* (Fig. 22).

**Figure 22. Enrichment in Metastases.** (A) Mutations in *EYA1* and *SPOP* are enriched in metastatic samples (B) Mutations associated with metastasis. Top 3 genes include *SOGA1, LRRC4C,* and *TP53.*



Mutational allele frequency clustering identified clonally independent tumor foci. PCM034 and PCM003 display distinctive primary clones, which appear to be independent from the metastases (Fig. 23). However, most primary tumor foci seem to be clonally related to the metastases (Fig. 24). The CNVkit pipeline was executed with sample BAM files in order to identify copy number changes. This tool allows for detection and visualization of copy number variation (CNV). The output includes .cns files and scatter plots, which displays the genome-wide copy ratio (Fig. 25). This analysis is currently ongoing and will continue with HATCHet as the next step.

**Figure 23. Mutational Allele Frequency Clustering identifies clonally independent tumor foci.**



### Figure 24. Most Primary Tumor Foci are Clonally related

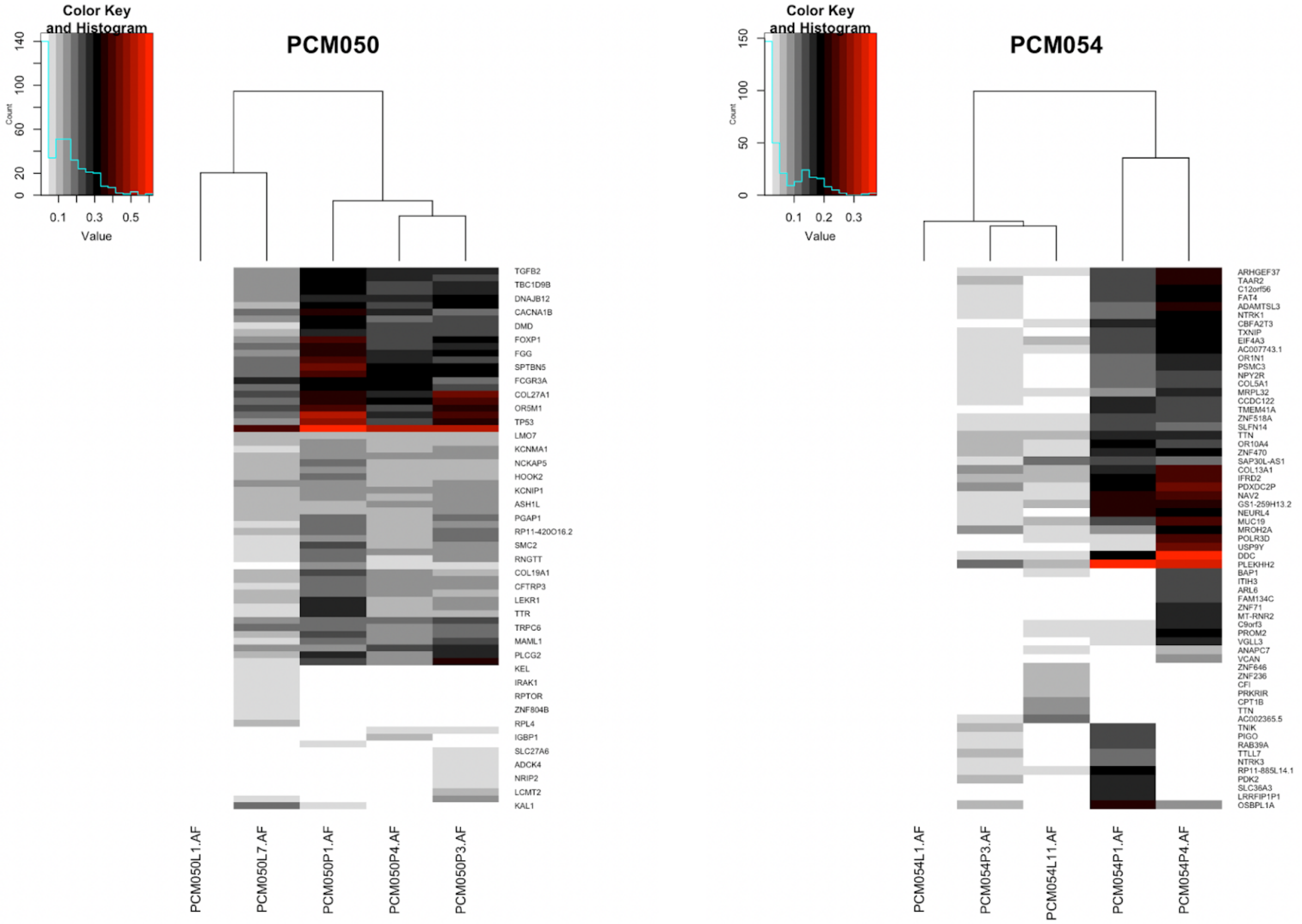
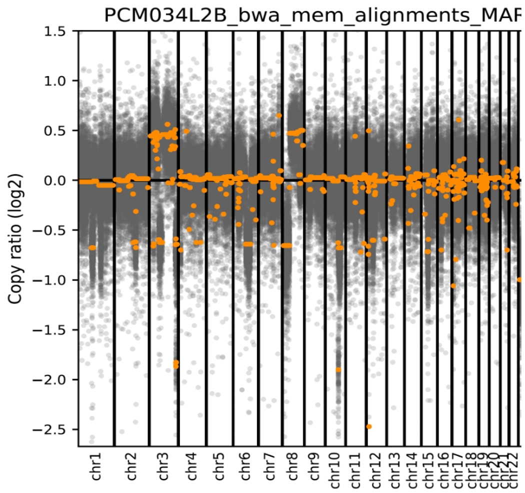


**Figure 25. CNV Pipeline Scatter Plot.** Genome-wide copy ratios across PCM034L2B.



### 3.2 Integrative DNA/RNA Analysis

Based on the GSEA results from the RNA-seq pipeline, it is clear that some pathways are enriched in metastatic clones compared to non-metastatic clones. The Hallmark50 Gene Set analysis reveals that metastatic clones were enriched in the G2M checkpoint (Fig. 26). This checkpoint stops the proliferation of damaged cells – it allows DNA repair to occur before moving to the mitotic stage of the cell cycle. However, non-metastatic clones were enriched in epithelial mesenchymal transition (EMT). EMT is a process in which epithelial cells gain mesenchymal features, such as migratory capabilities (44). These cells may become invasive and contribute to cancer metastasis. Hierarchical clustering of genes associated with G2M and EMT display the aforementioned patterns (Fig. 27). G2M is upregulated in metastatic clones, while EMT is downregulated.

**Figure 26. Hallmark50 Metastatic Enrichment Patterns.** Barplots displays enriched G2M checkpoint in metastatic clones, and enriched EMT in non-metastatic clones.

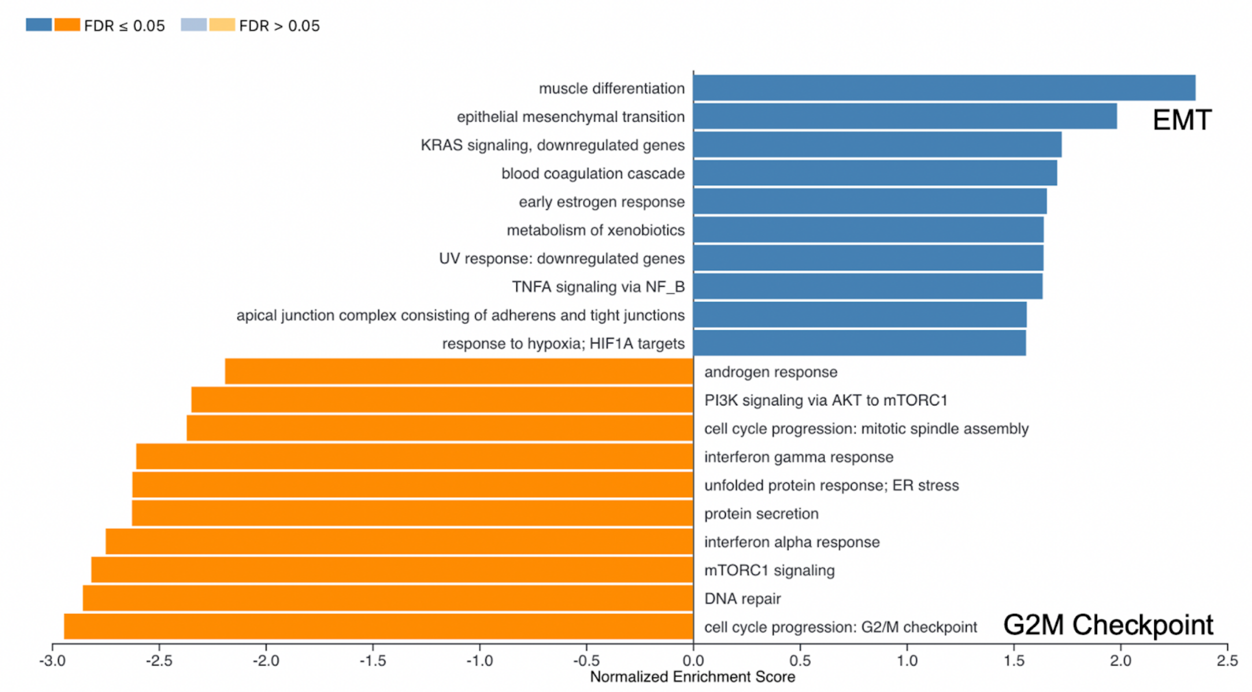
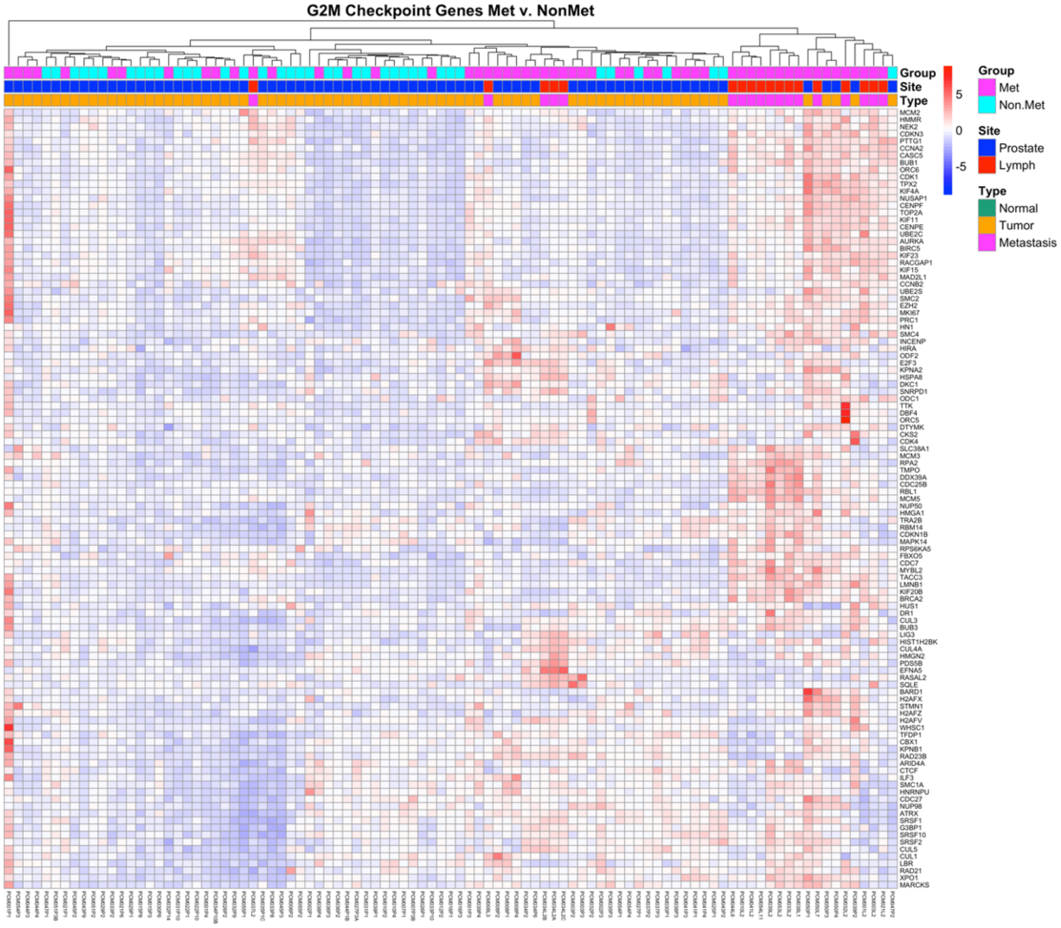
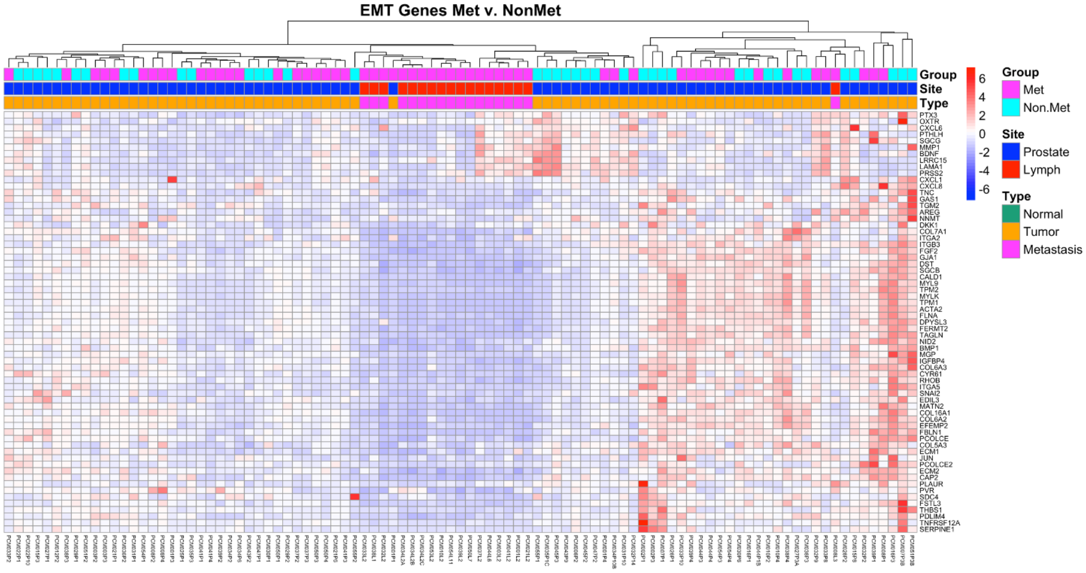


**Figure 27. Hierarchical Clustering of EMT and G2M Genes.** (A) G2M genes are upregulated in metastatic clones (B) EMT genes are downregulated in metastatic clones

**A.**



**B.**



# 4. Discussion

## 4.1 RNA-seq Pipeline

A median of 57 million paired-end reads were obtained per sample, with a median of 10 million total read counts per sample across the transcriptome, and 44,289 transcripts were detected in at least 5% of samples. Comparing PTF to LNM, 8110 transcripts were differentially expressed (p-adj < 0.01). PTF were enriched relative to LNM in gene sets associated with Notch signaling, hormone signaling, TGFb signaling, hypoxia, and the epithelial to mesenchymal transition. Comparing PTF from metastatic patients to non-metastatic patients, 581 transcripts were differentially expressed (p-adj < 0.01). PTF from metastatic patients were enriched in cell cycle progression, MYC targets, ER stress, androgen response, and DNA repair. LNM gene sets were enriched in endoplasmic reticulum (ER) stress and oxidative phosphorylation.

The top 500 upregulated genes in malignant tissues were significantly enriched in genes related to androgen and estrogen signaling as expected. We also identified a set of 193 genes whose expression was significantly increased in primary tumor over benign LNs and in LNM over primary tumors. This gene set was significantly enriched in genes related to oxidative phosphorylation and included oncogenes such as *PIK3CB, NCOA2*, and *SCHLAP1.* Based on the PCA plots, the primary tumor samples are closer together in a tighter cluster compared to the metastatic samples. This result was expected since there is more variation in the metastatic samples due to tumor heterogeneity. The heatmaps facilitate the visualization of differential gene expression, in which there is distinct clustering of metastatic samples. In the Hallmark50 GSEA, it is clear that genes associated with the G2M checkpoint are upregulated in metastatic clones, while genes associated with EMT are downregulated.

### 4.2 WES Pipeline

Most identified variants are missense mutations, which includes SNPs across different SNV classes. Frequently mutated genes were determined; the top three mutated genes are *SPOP, EYA1,* and *NCOR1.* The *SPOP* gene encodes for Speckle-type POZ Protein (SPOP), which is essential for ubiquitination and subsequent proteasomal degradation (45). One of the many substrates of SPOP is the androgen receptor (AR). Therefore, functional SPOP is necessary for the degradation of AR. Mutated SPOP may fail to ubiquitinate AR and allow for an increase of AR in the cell (46). As a result, AR signaling increases and encourages PCa proliferation. Results from the WES pipeline support this mechanism by identifying mutations in *SPOP.* We discovered mutations in the MATH binding domain of *SPOP,* which may alter protein-protein interactions with AR. These mutations could facilitate the survival and metastasis of cancer cells. Furthermore, past literature reveals important interactions between *SPOP* and c-JUN (47). Overexpression of the c-JUN protein leads to accelerated cell proliferation and induced gene expression. Mutated SPOP can bind to c-JUN, which stabilizes the complex and may further inhibit AR degradation.

Furthermore, *EYA1* encodes a transcription factor (TF) that interacts with SIX1. The EYA1-SIX1 complex plays a critical role in cell proliferation and gene regulation. Mutations in *EYA1* may cause dysregulation and act as a tumor promoter with SIX1 (48). Previous literature demonstrates that the EYA1-SIX1 complex activates STAT3 signaling. STAT (Signal transducer and activator of transcription) proteins serve as TFs and influence various biological processes including cell proliferation, apoptosis, mitosis, and differentiation (49). Consequently, STAT proteins are highly regulated in normal cells in order to prevent overexpression of genes. Elevated STAT-3 activity has been observed in many different cancers and is often associated with tumor progression. Since the EYA1-SIX1 complex activates STAT3 signaling, mutations in *EYA1* may promote the proliferation of PCa through increased STAT3 activity. Our WES pipeline further validates this hypothesis, as we located significant mutations in *EYA1.* Additionally, most mutations in *SPOP* and *EYA1* were found in metastatic PCa patients.

The *NCOR2* gene encodes a nuclear co-repressor (NCOR2) that mediates gene silencing. NCOR2 interacts with nuclear receptors, such as AR, to promote gene repression. Specifically, it binds to histone deacetylases to alter histone modifications. Mutations in *NCOR2* could inhibit the ability of NCOR2 to regulate and maintain the epigenome. This may change AR genomic interactions and support PCa proliferation. When AR genomic binding is redirected via mutated NCOR2, alternative malignant pathways may be upregulated. Previous literature investigates the role of NCOR2 in androgen deprivation therapy (ADT). Long MD, et al. found that reduced NCOR2 expression accelerates ADT failure in PCa (50). This phenomenon aligns with our findings that mutated NCOR2 alters the AR signaling pathway and drives cancer progression.

Through WES analysis, we identified mutations associated with PCa metastasis. The top six mutated genes are *SOGA1, LRRC4C, TP53, COL5A1, PCDHA13*, and *SLC16A14*. These mutations will be further analyzed using external, independent datasets. We aim to identify similar mutational patterns in additional samples in order to increase the validity of our results. This next step will be conducted with the Stand Up To Cancer dataset, which includes metastases and primary cancers. Additionally, the WES pipeline is under construction, as we are currently investigating intratumoral heterogeneity with HATCHet. Next steps will include additional variant analyses with the CNVkit output. HATCHet can determine CNAs and WGDs for tumor clones within patient samples.

## 5. Conclusions

Through building RNA-seq and WES pipelines, raw sequencing data was transformed into interpretable biological information. Mutational signatures were identified based on transcriptional enrichment patterns across seven main comparisons. Signaling pathways associated with ER stress, oxidative phosphorylation, metabolism, and cell cycle progression are prominent in LNM of aggressive PCa. Furthermore, expression of *PIK3CB, NCOA2,* and *SCHLAP1* are significantly increased in LNM. These results are vital to the investigation of prostate cancer metastasis. We have identified enhanced signaling pathways and overexpression of particular oncogenes. Based on the WES pipeline, cancer cell proliferation may be sparked by mutations in *SPOP, EYA1,* and *NCOR2.* Alterations in these genes may lead to AR upregulation or misregulation. As a result, these genomic changes may drive oncogenic progression. Furthermore, we identified several genes that are significantly enriched in metastases compared to non-metastatic samples. These genes include *SOGA1, LRRC4C, TP53, COL5A1, PCDHA13,* and *SLC16A14*. By understanding the mechanism of metastasis, we may be able to improve clinical strategies to target PCa.

## 6. Acronym Appendix

| **Abbreviation** | **Description** |
|---|---|
| AF | Allele Frequency |
| AJCC | American Joint Committee on Cancer |
| AR | Androgen Receptor |
| BAM | Binary Alignment/Map |
| BWA | Burrows-Wheeler Alignment |
| CN | Copy Number |
| CNA | Copy Number Aberrations |
| CNVkit | Copy Number Calling Pipeline |
| DEGs | Differentially Expressed Genes |
| DRE | Digital Rectal Exam |
| EIGC | Emory Integrated Genomics Core |
| EMT | Epithelial Mesenchymal Transition |
| EUH | Emory University Hospital |
| FDR | False Discovery Rate |
| FFPE | Formalin-Fixed Paraffin-Embedded |
| Funcotator | Functional Annotator |
| GATK | Genome Analysis Toolkit |
| GO | Geneontology |
| GSEA | Gene Set Enrichment Analysis |
| H&E | Hematoxylin and Eosin |
| HATCHet | Holistic Allele-specific Tumor Copy-number Heterogeneity |
| ITH | Intratumoral Heterogeneity |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| LN | Lymph Node |
| LNM | Lymph Node Metastasis |
| MAF | Mutation Annotation Format |
| mCRPC | Metastatic Castration-resistant Prostate Cancer |

| | |
|---|---|
| NGS | Next Generation Sequencing |
| NP | Normal Prostate |
| ORA | Over-representation Analysis |
| PCa | Prostate Cancer |
| PCA | Principal Component Analysis |
| PCR | Polymerase Chain Reaction |
| PET imaging | Positron Emission Tomography imaging |
| PHI | Protected Health Information |
| PSA | Prostate-Specific Antigen |
| PTF | Primary Tumor Foci |
| QC | Quality Control |
| SAM | Sequence Alignment/Map |
| scRNA-seq | Single cell RNA sequencing |
| SNV | Single Nucleotide Variant |
| STAR | Spliced Transcripts Alignment to a Reference |
| TCGA | The Cancer Genome Atlas |
| TNM Staging System | Tumor Nodes Metastasis Staging System |
| VCF | Variant Call Format |
| WebGestalt | WEB-based Gene SeT AnaLysis Toolkit |
| WES | Whole-Exome Sequencing |
| WGD | Whole-Genome Duplications |
| WGS | Whole-Genome Sequencing |

## 7. References

(1) "An Update on Cancer Deaths in the United States." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 28 Feb. 2022, https://www.cdc.gov/cancer/dcpc/research/update-on-cancer-deaths/index.htm.

(2) "Key Statistics for Prostate Cancer." *Prostate Cancer Facts*, American Cancer Society,https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html#:~:text=Other%20than%20skin%20cancer%2C%20prostate,34%2C500%20deaths%20from%20prostate%20cancer

**(3)** Nowell PC. The clonal evolution of tumor cell populations. Science. 1976 Oct 1;194(4260):23-8. doi: 10.1126/science.959840. PMID: 959840.

(4) Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015 Nov 5;163(4):1011-25. doi: 10.1016/j.cell.2015.10.025. PMID: 26544944; PMCID: PMC4695400.

(5) Robinson, Dan et al. "Integrative clinical genomics of advanced prostate cancer." *Cell* vol. 161,5 (2015): 1215-1228. doi:10.1016/j.cell.2015.05.001

(6) "Cancer Staging Systems." *ACS*, https://www.facs.org/quality-programs/cancer-programs/american-joint-committee-on-cancer/cancer-staging-systems/.

(7) American Cancer Society. Cancer Statistics Center. http://cancerstatisticscenter.cancer.org. 02/21/2023

(8) "Survival Rates for Prostate Cancer." *American Cancer Society*, https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/survival-rates.html.

(9) Division of Cancer Prevention and Control, Centers for Disease Control and Prevention, 02/22/2023, https://www.cdc.gov/cancer/prostate/basic_info/what-is-prostate-cancer.htm

(10) American Cancer Society. Cancer Statistics Center. https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/signs-symptoms.html. 02/21/2023

(11) Rebello, R.J., Oing, C., Knudsen, K.E. *et al.* Prostate cancer. *Nat Rev Dis Primers* 7, 9 (2021). https://doi.org/10.1038/s41572-020-00243-0

(12) American Cancer Society. Cancer Statistics Center, by Cancer.Net Editorial Board, https://www.cancer.net/cancer-types/prostate-cancer/types-treatment#advanced. 02/23/2023

(13) 9. Yadav SS, Stockert JA, Hackert V, Yadav KK, Tewari AK. Intratumor heterogeneity in prostate cancer. Urol Oncol. 2018;36(8):349-60. Epub 2018/06/12. doi: 10.1016/j.urolonc.2018.05.008. PubMed PMID: 29887240.

(14) 10. Aihara M, Wheeler TM, Ohori M, Scardino PT. Heterogeneity of prostate cancer in radical prostatectomy specimens. Urology. 1994;43(1):60-6; discussion 6-7. Epub 1994/01/01. doi: 10.1016/s0090- 4295(94)80264-5. PubMed PMID: 8284886.

(15) 11. Gleason DF. Classification of prostatic carcinomas. Cancer Chemother Rep. 1966;50(3):125-8. PubMed PMID: 5948714.

(16) 12. Gleason DF. Histologic grading of prostate cancer: a perspective. Hum Pathol. 1992;23(3):273-9. PubMed PMID: 1555838.

(17) 13. Gleason DF, Mellinger GT. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urol. 1974;111(1):58-64. PubMed PMID: 4813554.

(18) 14. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA, Grading C. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. Am J Surg Pathol. 2016;40(2):244-52. Epub 2015/10/23. doi: 10.1097/PAS.0000000000000530. PubMed PMID: 26492179.

(19) 15. Osunkoya AO. Update on prostate pathology. Pathology. 2012;44(5):391-406. Epub 2012/07/10. doi: 10.1097/PAT.0b013e32835657cf. PubMed PMID: 22772344.

(20) 16. Ruijter ET, van de Kaa CA, Schalken JA, Debruyne FM, Ruiter DJ. Histological grade heterogeneity in multifocal prostate cancer. Biological and clinical implications. J Pathol. 1996;180(3):295-9. Epub 1996/11/01. doi: 10.1002/(SICI)1096-9896(199611)180:3<295::AID-PATH663>3.0.CO;2-W. PubMed PMID: 8958808.

(21) 17. Cheng L, Song SY, Pretlow TG, Abdul-Karim FW, Kung HJ, Dawson DV, Park WS, Moon YW, Tsai ML, Linehan WM, Emmert-Buck MR, Liotta LA, Zhuang Z.

Evidence of independent origin of multiple tumors from patients with prostate cancer. J Natl Cancer Inst. 1998;90(3):233-7. Epub 1998/02/14. doi: 10.1093/jnci/90.3.233. PubMed PMID: 9462681.

(22) 18. Mehra R, Han B, Tomlins SA, Wang L, Menon A, Wasco MJ, Shen R, Montie JE, Chinnaiyan AM, Shah RB. Heterogeneity of TMPRSS2 gene rearrangements in multifocal prostate adenocarcinoma: molecular evidence for an independent group of diseases. Cancer Res. 2007;67(17):7991-5. Epub 2007/09/07. doi: 10.1158/0008-5472.CAN-07-2043. PubMed PMID: 17804708.

(23) 19. Boutros PC, Fraser M, Harding NJ, de Borja R, et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. Nat Genet. 2015;47(7):736-45. Epub 2015/05/26. doi: 10.1038/ng.3315. PubMed PMID: 26005866.

(24) 20. Lovf M, Zhao S, Axcrona U, Johannessen B, Bakken AC, Carm KT, Hoff AM, Myklebost O, MezaZepeda LA, Lie AK, Axcrona K, Lothe RA, Skotheim RI. Multifocal Primary Prostate Cancer Exhibits High Degree of Genomic Heterogeneity. Eur Urol. 2019;75(3):498-505. Epub 2018/09/06. doi: 10.1016/j.eururo.2018.08.009. PubMed PMID: 30181068.

(25) 21. Lindberg J, Klevebring D, Liu W, Neiman M, Xu J, Wiklund P, Wiklund F, Mills IG, Egevad L, Grönberg H. Exome Sequencing of Prostate Cancer Supports the Hypothesis of Independent Tumour Origins. European Urology. 2013;63(2):347-53. doi: https://doi.org/10.1016/j.eururo.2012.03.050.

(26) 22. Cooper CS, Eeles R, Wedge DC, et al. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nature Genetics. 2015;47(4):367-72. doi: 10.1038/ng.3221.

(27) 23. Wei L, Wang J, Lampert E, Schlanger S, et al. Intratumoral and Intertumoral Genomic Heterogeneity of Multifocal Localized Prostate Cancer Impacts Molecular Classifications and Genomic Prognosticators. Eur Urol. 2017;71(2):183-92. Epub 2016/07/28. doi: 10.1016/j.eururo.2016.07.008. PubMed PMID: 27451135; PMCID: PMC5906059.

(28) 24. Liu W, Laitinen S, Khan S, Vihinen M, Kowalski J, Yu G, Chen L, Ewing CM, Eisenberger MA, Carducci MA, Nelson WG, Yegnasubramanian S, Luo J, Wang Y, Xu J, Isaacs WB, Visakorpi T, Bova GS. Copy number analysis indicates monoclonal origin of

lethal metastatic prostate cancer. Nature Medicine. 2009;15(5):559-65. doi: 10.1038/nm.1944.

(29) 25. Kumar A, Coleman I, Morrissey C, Zhang X, et al. Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. Nat Med. 2016;22(4):369-78. Epub 2016/03/02. doi: 10.1038/nm.4053. PubMed PMID: 26928463; PMCID: PMC5045679.

(30) 26. Gundem G, Van Loo P, Kremeyer B, et al. The evolutionary history of lethal metastatic prostate cancer. Nature. 2015;520(7547):353-7. doi: 10.1038/nature14347.

(31) Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

(32) Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras, STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, Volume 29, Issue 1, January 2013, Pages 15–21, https://doi.org/10.1093/bioinformatics/bts635

(33) Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8

(34) Wang, J., Vasaikar, S., Shi, Z., Greer, M., & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Research.

(35) Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. Jan 2019;47(D1):D419-D426.

(36) Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000). [pubmed] [doi]

(37) Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18. PMID: 19451168; PMCID: PMC2705234.

(38) Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, Volume 25, Issue 16, August 2009, Pages 2078–2079, https://doi.org/10.1093/bioinformatics/btp352

(39) McKenna, Aaron et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." *Genome research* vol. 20,9 (2010): 1297-303. doi:10.1101/gr.107524.110

(40) *Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. 2018. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Resarch. PMID: 30341162*

(41) Talevich, E., Shain, A.H., Botton, T., & Bastian, B.C. (2014). CNVkit: Genome-wide copy number detection and visualization from targeted sequencing. *PLOS Computational Biology* 12(4):e1004873

(42) Simone Zaccaria and Benjamin J. Raphael, Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. bioRxiv (Dec. 17, 2018) doi.org/10.1101/496174

(43) Zaccaria, Simone, and Benjamin J Raphael. "Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data." Nature communications vol. 11,1 4301. 2 Sep. 2020, doi:10.1038/s41467-020-17967-y

(44) Yang, J., Antin, P., Berx, G. et al. Guidelines and definitions for research on epithelial–mesenchymal transition. Nat Rev Mol Cell Biol 21, 341–352 (2020). https://doi.org/10.1038/s41580-020-0237-9

(45) Cuneo, Matthew J, and Tanja Mittag. "The ubiquitin ligase adaptor SPOP in cancer." The FEBS journal vol. 286,20 (2019): 3946-3958. doi:10.1111/febs.15056

(46) Barbieri CE, Baca SC, Lawrence MS, et al. Nat Genet. 2012;44(6):685-689. An J, Wang C, Deng Y, Yu L, Huang H. Cell Rep. 2014;6(4):657-669.

(47) Mo X, et al. Cell. 2022;185(11):1974-1985.e12.

(48) Kong, Deguang et al. “SIX1 Activates STAT3 Signaling to Promote the Proliferation of Thyroid Carcinoma via EYA1.” Frontiers in oncology vol. 9 1450. 20 Dec. 2019, doi:10.3389/fonc.2019.01450

(49) Gu, Yuchen et al. “Overview of the STAT-3 signaling pathway in cancer and the development of specific inhibitors.” Oncology letters vol. 19,4 (2020): 2585-2594. doi:10.3892/ol.2020.11394

(50) Mark D. Long, Justine J. Jacobi, et al. “Reduced NCOR2 expression accelerates androgen deprivation therapy failure in prostate cancer”, Cell Reports, Volume 37, Issue 11, 2021, 110109, ISSN 2211-1247, https://doi.org/10.1016/j.celrep.2021.110109.