Integrated In Silico Analysis of NGS Prostate Cancer Data via High-Resolution RNA-Seq Analysis

Aubree Hoover[1], Sandeep Sanga[2], Antoaneta Vladimirova[2]

[1] CLC Bio LLC, Cambridge MA, [2] Ingenuity Systems, Redwood City, CA

Abstract

Prostate adenocarcinoma is the most frequent carcinoma in men and the second leading cause of death in the male population worldwide. The goal of our study was to get novel insights into the mechanisms of the disease by leveraging the rapidly growing next generation sequencing (NGS) data, and in particular, human transcriptome data through in silico data analysis and interpretation. The analysis of altered expression of genes and regulatory regions can pinpoint specific pathways and processes activated in growing cancer cells within tumors. Determining these activated pathways and networks can shed light on dysregulated processes, inform treatment options and highlight potential biomarkers with the ultimate goal to improve patient prognosis and treatment. High-resolution technologies, such as RNA-Seq, generate data that can be used to interrogate patient samples for expression changes and their patterns. Using short read RNA-Seq data from the NCBI SRA (Short Read Archive) public repository, gene expression changes from human prostate tumor and matched normal patient samples were assessed using CLC Genomics Workbench and CLC Genomics Server. To elucidate the underlying dysregulated biological processes, in silico pathway and mechanistic analysis was conducted in Ingenuity's IPA® software application by leveraging manually-curated biological information, canonical pathways and a variety of analytical tools. This poster highlights some of the results of this integrated in silico analysis and introduces a proposed workflow for the analysis and interpretation of RNA-Seq data.

Materials and Methods

A graphical overview of the analysis workflow is provided in Figure 1. Samples and sequencing reads were generated as described by Nacu, S & Wu, T . et.al.[1]. Publically available data from a total of three patients with matched normal and tumor prostate adenocarcinoma tissues were downloaded from the NCBI Short Read Archive, Geo Series GSE24283[2].

CLC Genomics Workbench version 4.5 and IPA 7.5 software platform were used to perform all data uploads and in silico analyses (see Figure 1). Fold change of 5 of tumor-to-normal experiments and p<0.05 cutoffs were used where applicable to analyze the data. Ingenuity® Knowledge Base was used as a reference set.
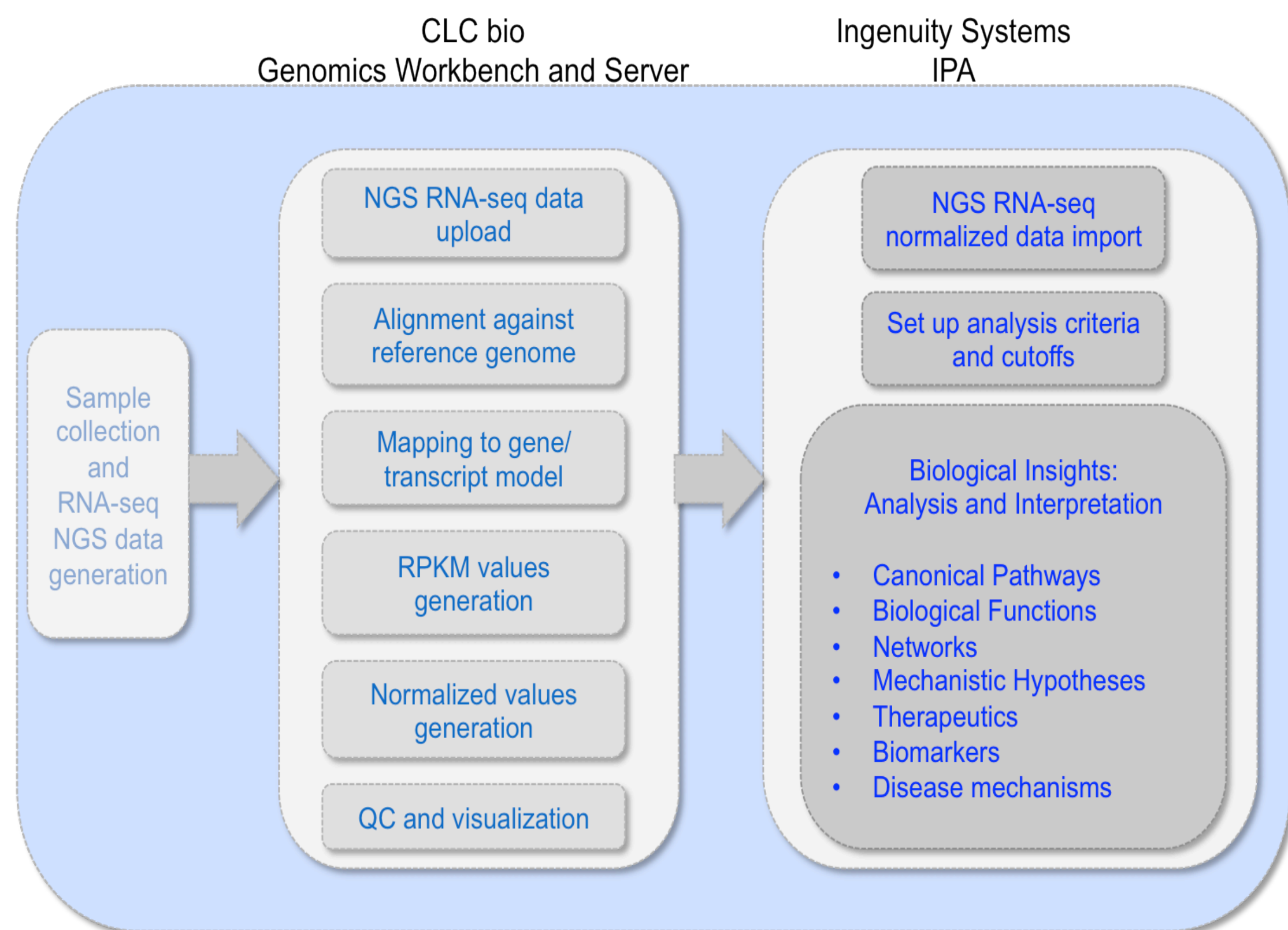

Figure 1. Combined CLC Bio & Ingenuity Workflow
Data were loaded in CLC Bio's Genomics workbench for all data preparation steps: alignment, mapping, expression value generation and QC. Resulting datasets were then loaded into IPA, filtered and then analyzed against the Ingenuity Knowledge Base for biological interpretation.

References and Acknowledgements

1. "Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples", Nacu, S. et.al. BMC Medical Genomics 2011, 4:11.
2. National Center for Biotechnology Information: Short Read Archive http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24284 Accessed January 11, 2011.
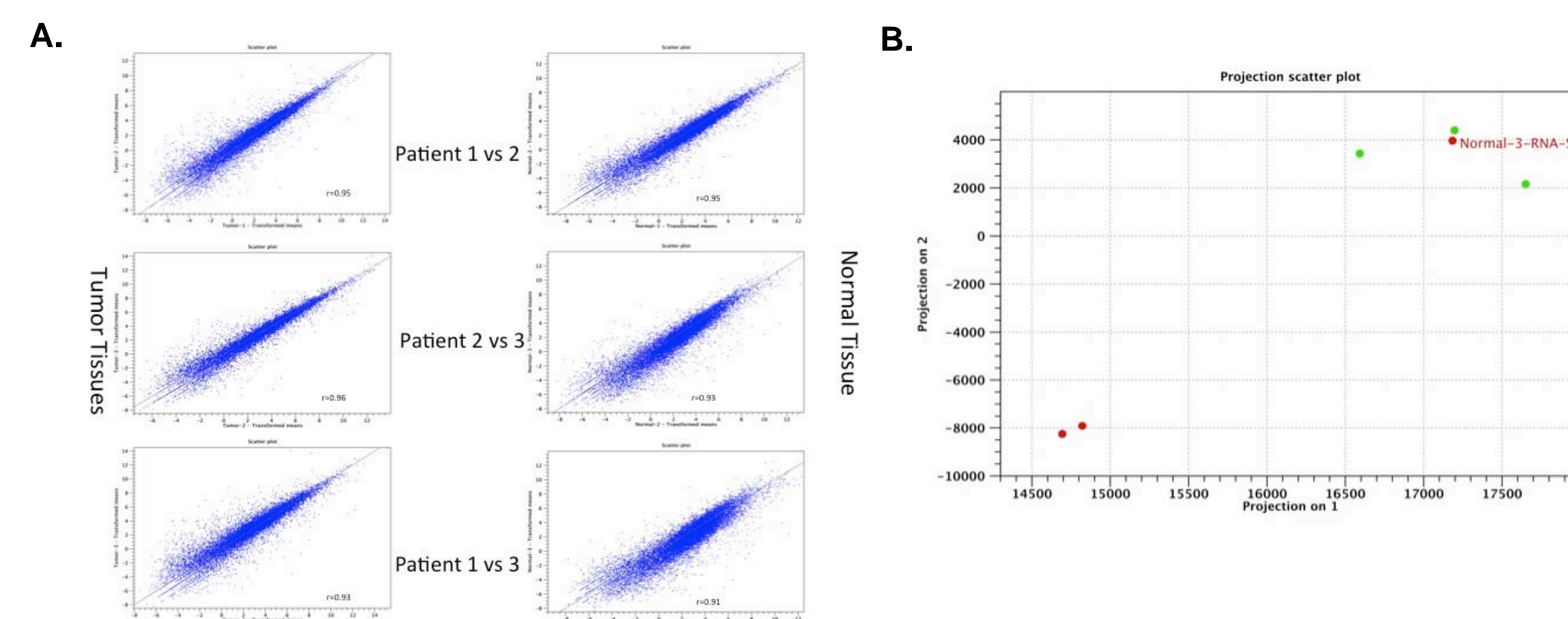
Results

Figure 2. Sample Quality Control . A.) Scatter plots of biological replicates for all tissue and Pearson Coefficient calculation (r). Patient 3 Normal shows the most variation from other Normal samples with a lower r value in plots. **B.)** Bi-plot of Principal Component Analysis of all samples. All normal samples in red, all tumor samples in green. Patient-Normal 3 (labeled) appears to group with it's tumor counterpart rather than other normal samples.
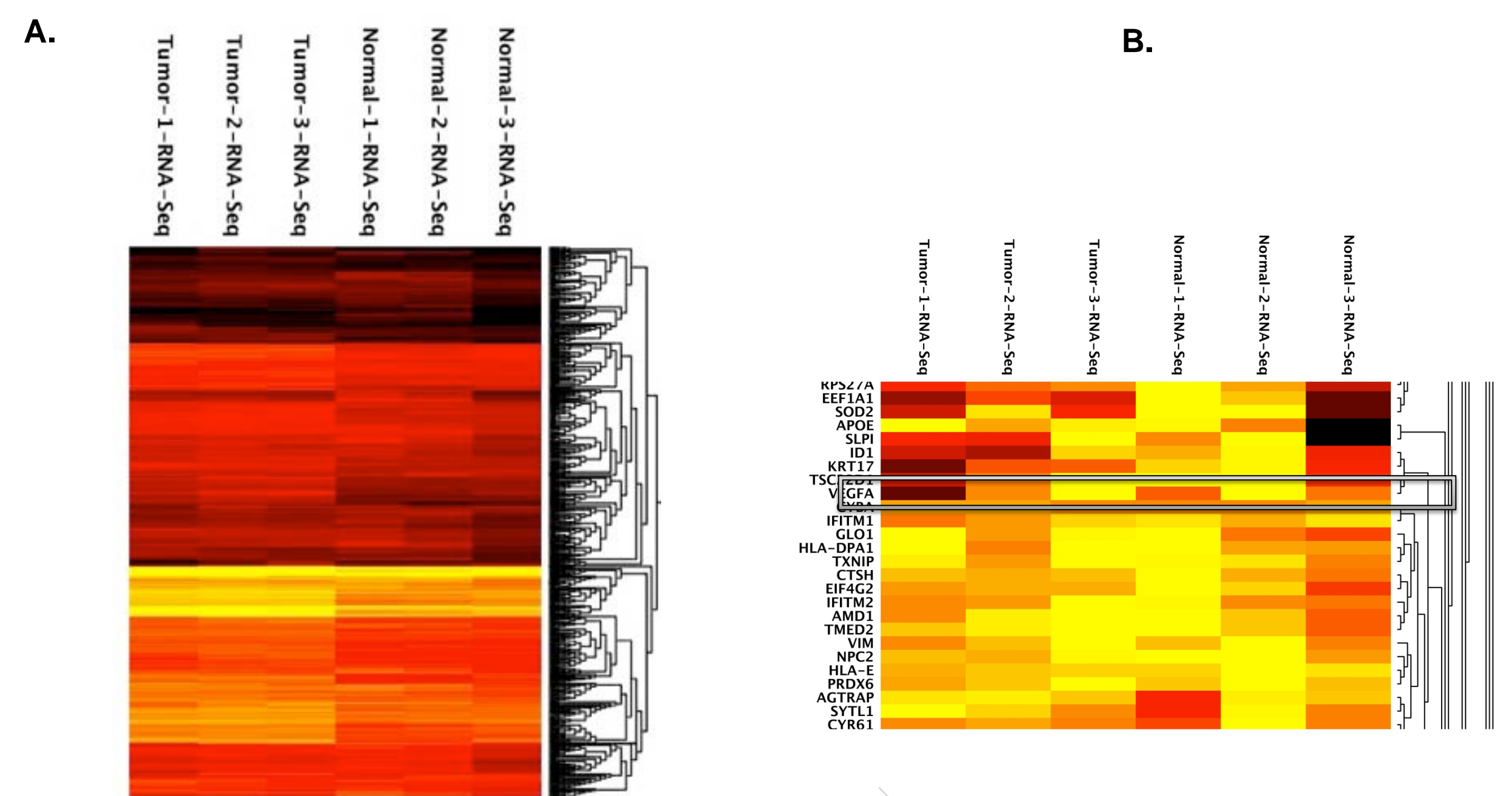

Figure 3. Global Expression Patterns A.) hierarchical cluster of all significantly expressed genes in tumor and normal samples. Patient Normal 3 appears to have an independent expression pattern. **B.)** Closer view of differentially expressed gene shows variation in specific genes, notably, VEGF whose role has been previously described in prostate and other cancers.
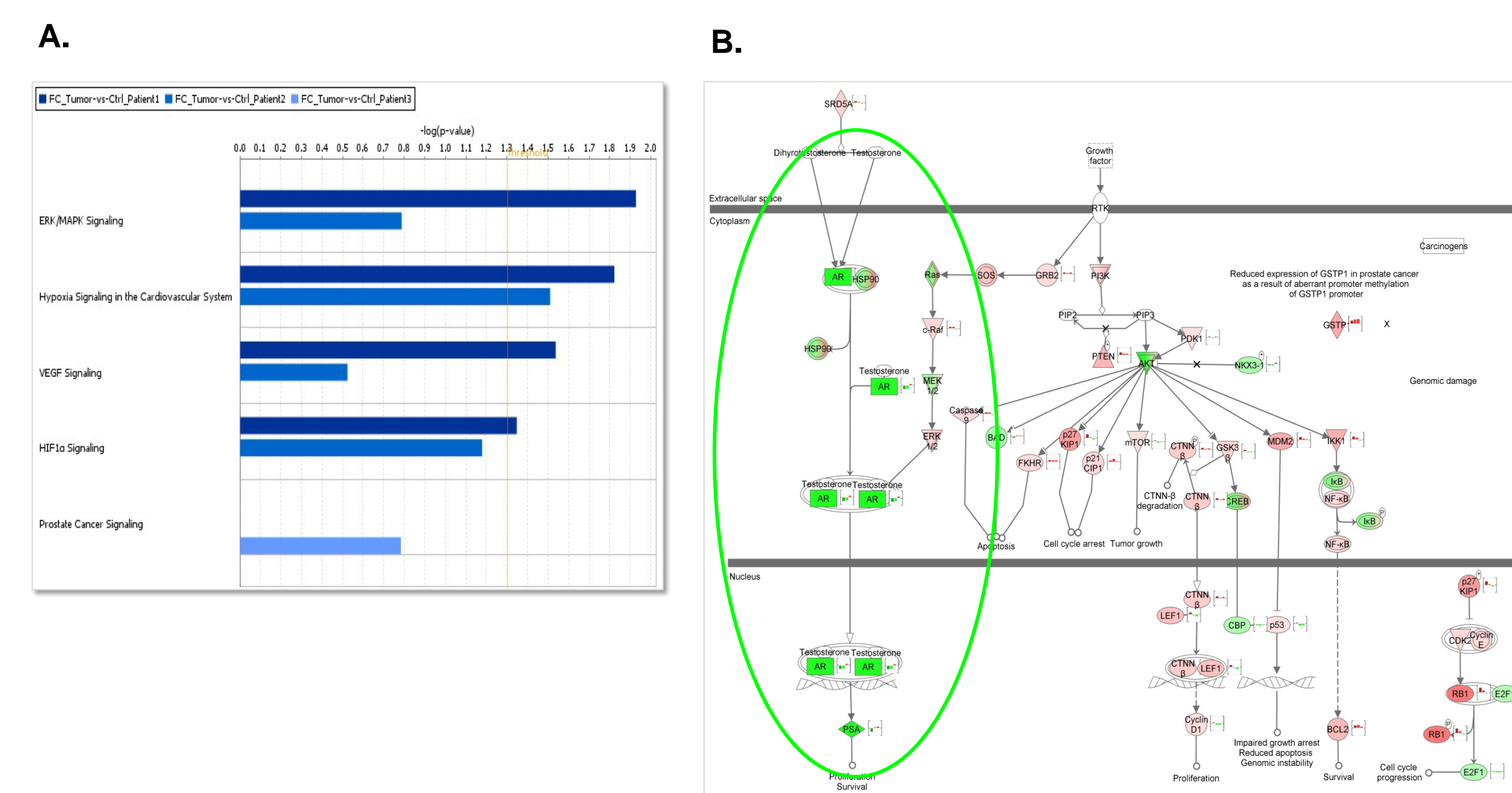

Figure 4. Top Impacted Pathways by Genes Passing the Cutoff Criteria (RPKM>5 and FC tumor/control >5) and Selected Pathway Depictions. A.) Top Prostate Carcinoma pathways include ERK/MAPK Signaling, Hypoxia- and VEGF-related Signaling Pathways. Blue bars indicate significance. Different shades of blue correspond to different patients data. Orange line indicates p value cutoff of 0.5. **B.)** Depiction of the Prostate Cancer Signaling pathway with gene expression data overlay from Patient 1. The bar charts next to each gene represent the expression values in Patient 1, 2 and 3, in that order (no cutoff on expression data applied). **C.)** Depiction of the Hypoxia Signaling in the Cardiovascular System pathway with gene expression data overlay. from Patient 1. The bar charts next to each gene represent the expression values in Patient 1, 2 and 3, in that order (FC>5 cutoff applied).
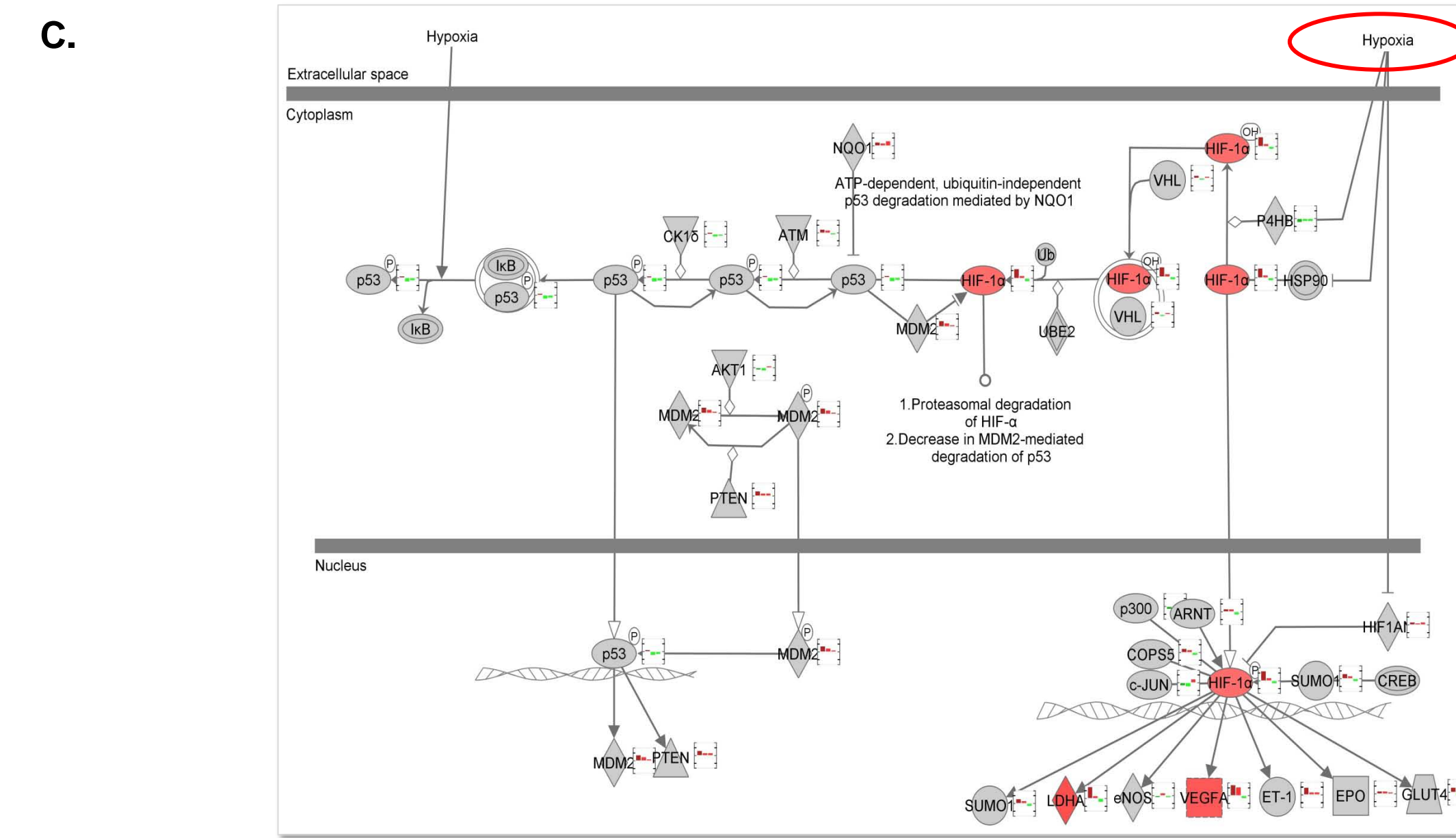

Figure 5. Top Impacted Prostate-related Functions Based on Expression Data from the Three Patients RPKM>5, FC>5). Genes impacted in the corresponding functions and patients are displayed along with the significance p-value of that function per patient.

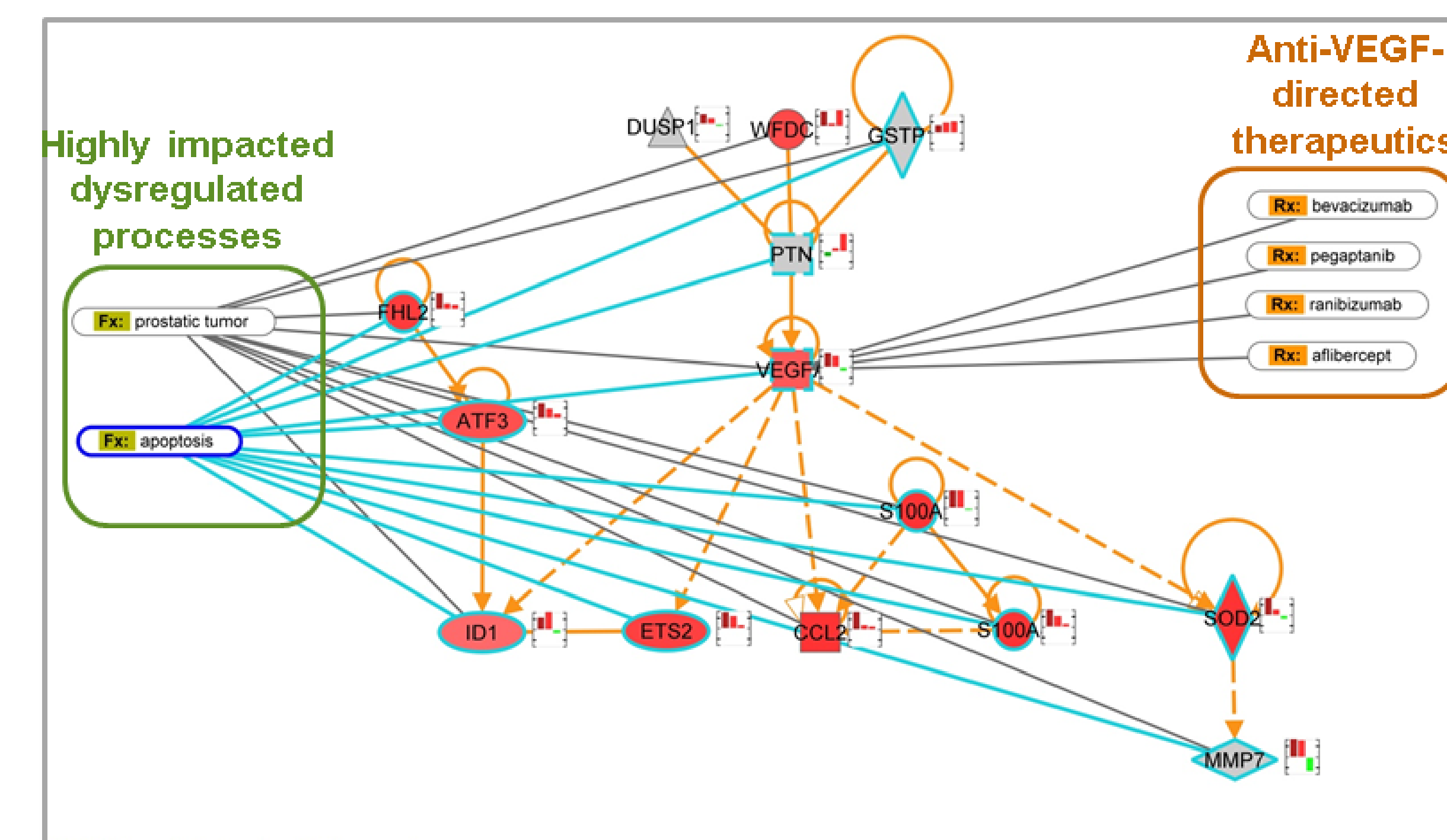| Function Annotation	p-value	Molecules	# Molecules
prostatic tumor	4.10E-11	ASS1, B2M, CCL2, CFB, CYP1B1, DUSP1, FHL2, ID1, KLK4, KRT7, MSMB, PSCA (includes	19
prostatic carcinoma	4.35E-11	ASS1, B2M, CCL2, CFB, CYP1B1, DUSP1, FHL2, ID1, KLK4, KRT7, MSMB, PSCA (includes	18
prostate cancer	5.81E-09	ASS1, B2M, CCL2, CFB, CHRNA2, CYP1B1, DUSP1, FHL2, ID1, KLK4, KRT7, LYZ, MSMB,	21
prostatic intraepithelial tumor	2.73E-07	ASS1, CFB, FHL2, KRT7, S100A8, S100A9, TFF3, WFDC2	8
developmental process of prostate cancer cell lines	4.05E-04	ETS2, ID1, KLK4, LCP1, PMEPA1, YBX1	6
growth of prostate cancer cell lines	2.26E-03	ETS2, KLK4, LCP1, PMEPA1, YBX1	5
apoptosis of prostate cancer cells	4.01E-03	DPP4, FHL2	2
invasion of prostate cancer cell lines	6.69E-03	ATF3, ETS2, LCP1	3
arrest in cell cycle progression of prostate cancer cells	7.40E-03	DPP4	1
prostatic tumor	2.23E-08	AMACR, B2M, EIF4EBP1, GDF15, ID1, MME, MMP7, MSMB, RARRES1, S100A9, SGK1	12
prostatic carcinoma	6.25E-08	AMACR, B2M, EIF4EBP1, GDF15, ID1, MME, MMP7, MSMB, RARRES1, S100A9, SGK1	11
migration of prostate cancer cell lines	1.61E-06	ID1, LGALS3, MME, RARRES1, VEGFA	5
prostate cancer	4.93E-06	AMACR, B2M, EIF4EBP1, GDF15, ID1, MME, MMP7, MSMB, RARRES1, S100A9, SGK1	12
arrest in growth of prostate cancer cell lines	3.15E-04	AMACR, GDF15	2
proliferation of prostate cancer cell lines	6.38E-04	AMACR, GDF15, LGALS3, MSMB	4
invasion of prostate cancer cell lines	1.11E-03	ETS2, LGALS3, VIM	3
developmental process of prostate cancer cell lines	1.69E-03	AMACR, ETS2, GDF15, ID1	4
apoptosis of prostate cancer cell lines	2.40E-03	ETS2, GDF15, ID1, MME	4
prostatic intraepithelial tumor	3.04E-03	AMACR, LGALS3	2
prostatic intraepithelial tumor	5.00E-03	AMACR, MMP7, S100A9	3
epithelial-mesenchymal transition of prostate cancer cell lines	7.85E-03	ID1	1
prostatic carcinoma	6.60E-04	AMACR, GSTP1, MSMB, TFF3, WFDC2	5
prostatic intraepithelial tumor	8.54E-04	AMACR, TFF3, WFDC2	3
proliferation of prostate cancer cell lines	1.24E-03	AMACR, MSMB, PTN	3
prostate cancer	1.90E-03	ALDH1A1, AMACR, GSTP1, MSMB, TFF3, WFDC2	6
arrest in G2/M phase of prostate cancer cell lines	6.38E-03	AMACR	1
arrest in growth of prostate cancer cell lines	1.48E-02	AMACR	1

Figure 6. Mechanistic Hypothesis with Patient Gene Expression, Impacted Functions and Potential Therapeutics. Displayed are the genes and their relationships that may help explain the mechanism of disease pathogenesis and progression. Genes with highest average expression values across all three patients and passing the RPKM>5 and FC>5 cutoff values were interconnected using the Ingenuity Knowledge Base in IPA to help understand mechanism. Patient 1 expression data is overlaid on the diagram genes. The expression values in the bar chart next to each gene correspond to patients 1, 2 and 3, respectively. Red color indicates up-regulation and green indicates down-regulation. Link of genes to prostate tumor and apoptosis functions are depicted. Therapeutics for which VEGFA is a known target are displayed.

Figure 7. Table of Putative Prostate Carcinoma Biomarkers and their Characteristics. Genes induced in all three prostate cancer patients were selected (RPKM>5). The table includes the fold increase in tumor vs. control for each patient, average value across all three patients, gene symbol, gene name, gene product location, whether the gene has been proposed in the literature to be a prostate carcinoma biomarker, and whether the gene is considered to be a biomarker for any other application such as efficacy, prognosis, safety or diagnosis. In bold are the genes with more favorable drug-target or biomarker characteristics.
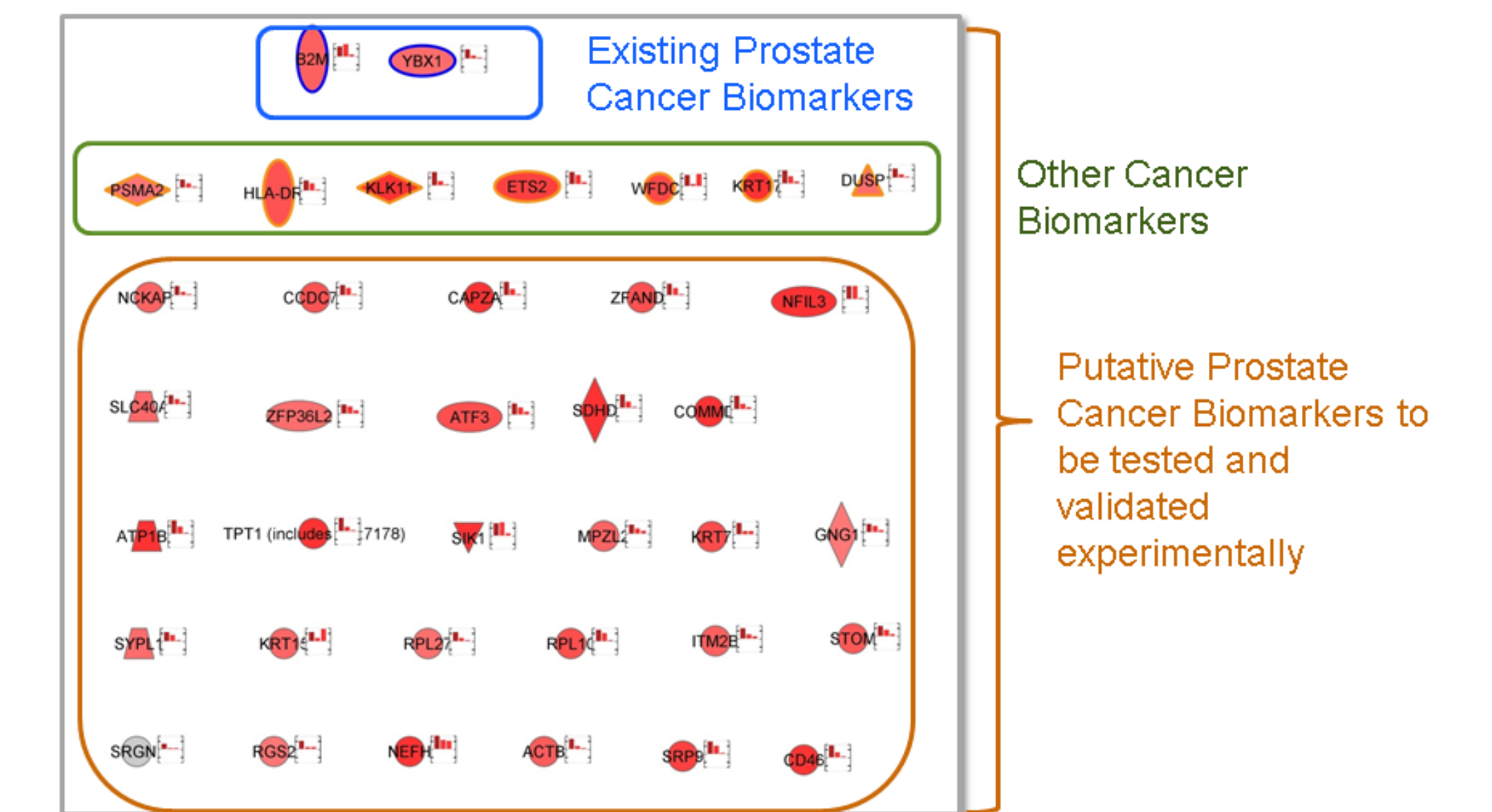

Figure 8. Putative Prostate Cancer Biomarkers Categories. Depiction of the putative prostate carcinoma biomarkers with their expression values across the three patients and across the different biomarker categories: Existing prostate cancer biomarker, Other cancer biomarker and Putative Prostate Cancer Biomarker to be tested and validated experimentally.

Conclusions

Analysis results:

- Identified impacted prostate cancer-related functions and the genes that drive them (to be investigated further in more detail)
- Identified preliminary mechanistic hypothesis involving dysregulated apoptosis and hypoxia likely promoting tumor angiogenesis
- Highlighted potential therapeutic target with existing marketed drug (clinical trials undergoing)
- Identified preliminary prostate cancer biomarkers for experimental validation and confirmed some already published biomarkers

Combined CLC bio and IPA joint workflow functionalities:

- Upload and normalization of raw data
- Visualization and QC data
- Calculation of RPKM and mapping results into gene/transcript model
- Creation of ratios and option to set up cutoffs
- Upload of individual sample files or pre-calculated metrics (fold changes, ratios, cutoffs) on a gene or transcript level into IPA
- Setting up analysis criteria and cutoffs
- Analysis of data similarly to the microarray data analysis to investigate pathways, networks, functions, toxicities, generate hypotheses about mechanism, therapeutic targets, biomarkers etc.