Integrative bioinformatics in the age of massive throughput sequencing: from the transcriptome to the proteome in prostate cancer

by

Lee T. Sam

**A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2014**

Doctoral Committee:

- **Professor Arul M. Chinnaiyan, Co-Chair**
- **Associate Professor Alexey I. Nesvizhskii, Co-Chair**
- **Assistant Professor Jun Z. Li**
- **Professor Gilbert S. Omenn**
- **Assistant Professor Maureen A. Sartor**

DEDICATION

This work is dedicated to my family.

ACKNOWLEDGEMENTS

This work would not have been possible without the guidance of my co-chairs Professors Chinnaiyan and Nesvizhskii, who have both contributed tremendous amounts of constructive advice, criticism, and context. My other committee members, Professors Jun Li, Gil Omenn, and Maureen Sartor were also critical in guiding and supporting this research. A number of people from the Chinnaiyan and Nesvizhskii labs and the Michigan Center for Translational Pathology have contributed to this work. Prof. Chandan Kumar, Dr. Anastasia Youcum, Dr. Christopher Maher, and Dr. Damian Fermin lent their substantial statistical, experimental, laboratory, and bioinformatics expertise in developing and evaluating the data and methods in this work. Prof. Dan Robinson, Xuhong Cao, Bob Lonigro, and Shanker Kalyana-Sundaram also contributed substantial advice guidance, and constructive criticism to my work. The faculty of the Department of Computational Medicine and Bioinformatics, including Professors George Michailidis, Brian Athey, Margit Burmeister, and Dan Burns, contributed an incredible amount of guidance and support. Many of my fellow bioinformatics staff and students also shared their specialized experience and support in addressing some of the challenges in this research; this includes Dr. Jeffrey DeWet, Douglas Jacobsen, Chee Lee, Avinash Shanmugam, and Dattatreya Mellacheruvu. The support of the advising and administrative staff across the departments and labs was substantial; Julia Eussen, Christine Betts, Karen Giles, Jyoti Athanikar, and Alex Terzian were central in managing and reviewing the progression of this work.

Table of Contents:

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF TABLES vii

LIST OF FIGURES ix

LIST OF APPENDICES xvii

ABSTRACT xviii

Chapter 1: Introduction 1

NGS and RNA-seq for interrogating the transcriptome 2

Disease and the dysregulation of biological networks 3

The transcriptome, proteome, and the challenges of data integration 9

The transcript-protein relationship and its role in disease and cancer 12

Application of methods to other studies 14

List of contributions to publications 15

Chapter 2: A Comparison of Single Molecule and Amplification Based Sequencing of Cancer Transcriptomes 17

Introduction 17

Results 18

Assessment of SMS RNA-Seq through transcript profiling 18

Global properties of AS and SMS results 19

Coverage bias in amplification-based sequencing 20

Increased SMS sensitivity results from high coverage of low-abundance transcripts 22

Uniquely detected genes in SMS 22

Consistent over-representation of high-expression genes in amplification-based sequencing 24

Impact of duplicated reads in amplification-based sequencing 27

Gene Ontology analysis of the set of 393 recurrently over-expressed genes ... 30

Re-discovery of known gene fusions using single-molecule sequencing ... 31

Discussion ... 32

Methods ... 36

Chapter 3: A Framework for integrating transcriptome and proteome data ... 43

Introduction ... 43

Background ... 43

Results and Discussion ... 45

RNA-seq and Proteomics Results ... 45

Calculating FDR in transcriptome and proteome data ... 49

Analyzing the impact of data processing methodology on correlation ... 50

Conclusion ... 55

Materials and Methods ... 55

RNA-seq expression data ... 57

Comparison of RNA-seq data with spectral counts ... 59

Chapter 4: The transcript-protein relationship in human prostate cancer ... 62

Introduction ... 62

Results and discussion ... 63

Transcript and Protein abundance in each of the cell line models ... 63

Comparison of VCaP and RWPE cell lines ... 69

Conclusion ... 77

Methods ... 79

Chapter 5: Conclusion ... 81

APPENDICES ... 85

APPENDIX A: Chapter 1 Supplementary Methods, Figures, and Tables ... 86

Methods ... 86

APPENDIX B: Chapter 2 Supplementary Figures and Tables..104

APPENDIX C: Chapter 3 and 4 Supplementary Figures and Tables..122

References: 238

LIST OF TABLES

Table 1: Unsupervised k-means clustering illustrates the poor separability of the data, with 1631 (55.4%) instances incorrectly clustered 6

Table 2: Classification of CD, SGD, and GO classes using all variables. While the complex disease (CD) subnetworks and derived from the Gene Ontology (GO) demonstrated relatively good classification performance, the subnetworks associated with single gene dieases (SGD) were very poorly separable. 8

Table S 1: Complete results of unsupervised k-means clustering of the data. 94

Table S 2: Classification results from each of nine classification attempts using complete GO set.. 96

Table S 3: Ranked Features By Parameter Type. A. Biological Parameters Only B. Topological Parameters Only C. Combined Parameterization 103

Table S 4: Sample statistics in A. amplification-based and B. single-molecule sequencing technologies.. 117

Table S 5: Recurrently over-represented genes in amplification-based sequencing in ten or more samples. Of the 393 genes are recurrently within the top 500 over-represented genes by total read count in five (40%) or more samples, these 59 are seen most often, occurring in at least 10 samples.. 118

Table S 6: Sum of normalized expression values per quartile by sample in AS and SMS. We observe that the number of reads aligning to transcripts seen in the third and fourth quartiles is consistently greater in A. SMS than B. AS across the sample set... 119

Table S 7: Gene-level read coverage of observed transcripts. A. and B. illustrate the number of genes with coverage values at various depths in single molecule and amplification-based sequencing, respectively. 120

Table S 8: Primers used for validating transcripts seen only by SMS. All experiments were performed in duplicate using two primer pairs per candidate gene when possible. 121

Table S 9: Sum of reads and spectra for top and bottom 50 in RWPE and VCaP by Gene Ontology class, ordered by VCaP relative enrichment 140

Table S 10: Sum of reads and spectra for top and bottom 50 in RWPE and VCaP by Gene Ontology class with ribosomal genes removed, ordered by VCaP relative enrichment 144

Table S 11: Top 150 Highest Transcript-protein correlation by Gene Otology class in VCaP 152

Table S 12: Top 150 Highest Transcript-protein correlation by Gene Otology class in RWPE 161

Table S 13: DAVID GO clustering analysis results in VCaP 176

Table S 14: Discordance-based LRPath Gene Ontology and pathway analysis results with FDR <= 0.05........ 214

Table S 15: Concordance-based LRPath Gene Ontology and pathway analysis results with FDR <= 0.05 237

LIST OF FIGURES

Figure 1: Derivation of an example subnetwork composed of five annotated genes and three intermediate genes 5

Figure 2: Variables ranked by importance in classification based on Gini coefficient. The most informative variables in classification were a mix of both biological and topological parameters. 7

Figure 3: Observed bias in amplification-based sequencing. A. Single-best mapping method-based quantile-quantile plot demonstrates evidence of over-representation of highly expressed transcripts in amplification-based sequencing compared to single-molecule methods. B. Distribution of reads across genes by transcript concentration shows decreased SMS coverage of the most highly expressed genes, with those reads going to mid- and low-level expressors. C. Differences in the distribution of reads lead to increased sensitivity of low-expressing transcripts. D. Nine of the candidate genes seen above the 0.3 RPKM noise level demonstrated any amplification by RT-PCR, although only HIST1H4C showed high abundance. 21

Figure 4: High-concentration transcript bias leads to differences in gene coverage in amplification-based sequencing. Coverage maps from amplification-based and single molecule sequencing demonstrate significantly greater coverage of A. RPLP0, B. RPL31, and C. SPINT2. Removal of reads with the same start positions, strictly suppressing amplification of specific mRNA fragments, significantly reduces the “spikiness” seen in these cases. D. Duplicate reads, defined as reads in excess of one per start locus and read length, are relatively evenly distributed along the length of all observed transcripts across all samples in our evaluation set. 26

Figure 5: Global representation of Gene Ontology classes in Amplification-based sequencing. GO analysis of the 392 most over-represented genes found using our recurrence analysis in the Molecular Function (MF) and Biological Process (BP) subtrees demonstrates that translational processes and components of the ribosome are over-represented across samples in amplification-based sequencing. 30

Figure 6: Single molecule sequencing “re-discovers” known gene fusions. Schematic of the intra-chromosomal rearrangement on chromosome 21 fusing TMPRSS2 (yellow) to ERG (purple). 32

Figure 7: Data Processing and Integration Pipeline. Three replicates of each sample were generated using MS/MS and RNA-seq and quantified against a common reference library of mRNA and protein sequences. Tandem mass spectrometry data were processed with the TPP and post-processed with Abacus to yield spectral count data. RNA-seq reads were aligned to the common reference using Bowtie and post-processed with in-house Perl scripts to yield RPKM (reads per Kilobase million) quantification. 46

Figure 8: The data filtering and integration statistics producing the core and extended datasets. Data is merged to the gene level before filtering by FDR and integrated using 1% and 5% FDR thresholds resulting in the core and extended datasets, respectively. 47

Figure 9: Analysis of transcript and protein datasets. A. The distribution of real and decoy gene values. The mean abundance of all decoy genes is 22.52 RPKM while decoys have a mean abundance of 0.46 RPKM. B and C. True positive detections across FDR values in protein and transcript data. D. Spearman correlation coefficient values for each alignment method and protein data reduction step in extended dataset. E. Protein detection at increasing transcript abundance levels. F. Distribution of RNA-seq reads and tandem MS spectra across all genes detected in the extended dataset for VCaP and RWPE. G. Most over-represented (as a fraction of all reads) Gene Ontology classes in transcript and protein data. 51

Figure 10: Analysis of the transcript-protein relationship in VCaP and RWPE. A. Division of genes by relative protein-transcript relationship with zero values. B. Plot of cancer-related GO class genes of interest. C. Relationship between GO class size and transcript-protein Spearman correlation coefficient. D. Relationship between transcript and protein abundance per Gene Ontology class. E. Distribution of transcript stability used to segment extended dataset into high and low stability sets. F and G. Correlation of transcript and protein levels for low and high stability genes in VCaP. 66

Figure 11: Detecting Dysregulation of transcript-protein relationships in prostate cancer. A. z-transformed fold changes of transcript and protein observed between VCaP and RWPE and major Gene Ontology class clusters in red and orange classes, corresponding to enrichment in protein and transcript abundance, respectively. Representative GO classes for each annotation cluster were chosen by examining overlap between DAVID clusters and LRPath results with FDR ≤ 0.05, in order of observed genes B. Number of GO classes resulting from LRPath analysis before and after removal of genes overlapping between concordance and discordance classes. 70

Figure 12: The dysregulated networks surrounding Akt. A. Network closely linked to Akt formed by genes constituting the GTPase regulator activity, polyol metabolic process, guanyl-nucleotide exchange factor activity, and regulation of immune response classes. B and C. Distribution of genes in the GTPase regulator activity and activation of immune response Gene Ontology classes highlighted in red in the context of all dataset genes. 72

Figure 13: Insights from joint transcriptome-proteome analysis. Heatmap of log-odds ratios for KEGG and Biocarta pathways with p-values ≤ 0.05 in each of four categories – results derived from protein only, transcript only, discordance index, and concordance index data. ... 74

Figure S 1: Principal Components Analysis demonstrates the poor separability of the data. A principal components analysis of the combined sets using all the parameters, suggests that the difference between disease-related subnetworks and the GO baseline subnetworks are subtle and not easily derived. When the PCA is done over just the CD and SGD sets, we see a similar pattern where there is no clear separation. However the non-continuous nature of the features may be a confounding factor when applying the PCA approach. With that in mind, a simple k-means clustering approach was taken where k = 3 to represent the three source types. A. Principal components analysis of all sets using all parameters. 95% of data points fall within the ellipse. B. Principal components analysis of SGD and CD sets using all parameters. 95% of data points fall within the ellipse. 92

Figure S 2: A. Size Distribution of SGD Subnetworks B. Size Distribution of CD Subnetworks C. Size Distribution of GO Subnetworks... 95

Figure S 3: Single-molecule mRNA-sequencing. mRNAs are purified using poly-A selection and then fragmented. 1st-strand cDNA is synthesized from the fragmented mRNA, and then poly-A tailed using terminal transferase. Polyadenylated cDNA fragments are hybridized to poly-T oligomers bound to a glass substrate, excess A bases are “filled ,“ and then “locked” with an A, C, or G base attached to a virtual terminator. The sequencing process then occurs with repeated cycles of virtual terminator cleavage, bases addition, and image readout. 104

Figure S 4: Read alignment with Bowtie and IndexDP. Bowtie was used for amplification-based sequencing read alignment and IndexDP for single molecule read alignment. While different in their parameters, the effective alignments and specificity between the aligners are similar, although Bowtie has a slightly higher cutoff 105

Figure S 5: Length distribution of aligned SMS reads. Aligned SMS read lengths varied between 24bp to 57bp in our first set of samples and 25bp to 63bp in our second set. The majority of reads are between 25bp and 45bp in length. 105

Figure S 6: Sample Profiling Reproducibility in SMS and AS. Bowtie was used for amplification-based sequencing read alignment and IndexDP for single molecule read alignment. Pearson correlation for log2-transformed, normalized tag counts is r=0.98 for both SMS and AS. 106

Figure S 7: Log2 correlation between amplification-based and single-molecule sequencing. Log2 correlation between single-molecule and amplification-based RNA-Seq single-best read mappings in these samples show that in broad terms the two sequencing methods yield similar results, suggesting the observed bias is not due to sample differences. 107

Figure S 8: Correlation between IndexDP and Bowtie alignment of amplification-based sequencing reads. The correlation between Bowtie and IndexDP within the subset of samples was relatively high, with Pearson correlation values above r=0.95 in all samples. 108

Figure S 9: IndexDP realignment of amplification-based sequencing reads. Alignment of amplification-based sequencing reads using the IndexDP alignment tool used to align single-molecule reads shows persistence of the observed bias in amplification-based technology. This provides evidence that the alignment method is not responsible for this bias towrds high-concentration transcripts. 109

Figure S 10: Unique gene detection in AS and SMS across threshold values, by transcript length. The pattern of increased sensitivity in SMS is uniform as the baseline noise level is varied from 0.1 to 3.0 RPKM. Low representation by short transcripts show that this effect is not due to the lack of a size-selection step in SMS. 110

Figure S 11: Expression values of validation candidate genes showing amplification. Out of the set of genes chosen for RT-PCR validation for their detection over the 0.3 RPKM noise threshold by only SMS, diffuse read alignment pattern, and the presence of long reads aligned to their transcripts, these ten genes showed detectable amplification. 111

Figure S 12: RPLP0 coverage in other samples. Coverage plots of the over-represented gene RPLP0 in the LNCaP-24h, LNCaP-48h, VCaP-24h, VCaP-48h, and PrCa-Met samples show that this gene is often more highly sequenced using the amplification-based method. 112

Figure S 13: Quantile-quantile plot of AS and SMS reads with duplicates removed. Reads in excess of a single read per aligned locus were removed from both AS and SMS data sets. The result of this procedure was inconsistent across the data set; some samples saw reduced representation of high expressing genes while the high-concentration bias remained in others 113

Figure S 14: Effect of duplicate removal in AS. Reads in excess of a single read per aligned locus were removed from both AS and SMS data sets, resulting in (A) a median 47% drop in the number of usable reads across the 12 samples in the evaluation set and (B) the loss of dynamic range for genes in with high coverage levels. 114

Figure S 15: Gene Fusion Discovery Using SMS Reads. All possible reads were aligned against the transcriptome and genome using IndexDP. The set of non-mapping reads (some of which harbor chimeras) were subsequently aligned against the transcriptome, returning reads that had a partial alignment of at least 18 nucleotides. All reads having the same partial alignments, suggesting a common breakpoint, were clustered. All clusters were then compared to determine if the non-aligning “overhang” portion of the read from one breakpoint region had similarity to the overhang of an independent breakpoint, thereby reconstructing the fusion junction. Finally, all remaining non-mapping reads were aligned against the candidate novel fusion junctions. 115

Figure S 16: Alternate mappings for genes detected by SMS only in DU145. We analyzed alternate mappings for the reads attributable to each of the nine genes we observed to be detectable only by SMS in DU145 using reads from both replicates. In all nine cases, reads mapped most strongly to the genes of interest, suggesting that the detection of these genes is not an artifact of mis-mapping. The top 20 alternate mappings, ordered by mapping read count, are shown in the graph. 116

Figure S 17: Reproducibility between replicates. A. RWPE RNA-seq, B. VCaP RNA-seq, C. RWPE tandem MS, and D. VCaP tandem MS. Data are derived from the extended dataset. 122

Figure S 18: Correlation between VCaP and RWPE by RNA-seq and tandem MS. Within both the transcriptome and proteome data, both cell lines showed relatively high similarity in abundance profile. This is expected, owing to their common prostate tissue origin. 123

Figure S 19: False Discovery Rate estimation in RNA-seq data. We measured the FDR at increasing RPKM cutoffs to determine the 1% and 5% FDR levels in each cell line in our RNA-seq data 124

Figure S 20: False Discovery Rate estimation in protein data. We used the output of TPP and Abacus to determine appropriate parameter values for controlling FDR in our protein data. Three parameters were considered to control FDR at 1% and 5% FDR levels; peptide probability, protein group probability, and protein probability. 125

Figure S 21: RNA-seq False Discovery Rate estimation methodology. We used a methodology similar to that of Ramskold, et al. to estimate FDR in our RNA-seq data. Corresponding decoy sequences were sampled without replacement from the intergenic regions in hg19 for each representative transcript in our database, for a total of 34,728 decoys. These decoy sequences were of equal length as the real transcripts. We aligned reads to the merged total set of these decoy and real mRNA transcripts. Abundance data was summarized at the gene level using the same transcript-gene mappings for both the real and decoy transcript set. FDR was calculated as the number of decoy genes detected divided by the number of non-decoy genes detected at each threshold value. 126

Figure S 22: Comparison of mapping methodologies. The number of reads assigned to this hypothetical gene, reflecting its abundance, is highly variable based on both the mapping and counting parameters can constraints. 127

Figure S 23: Coverage of the transcriptome by observed RNA-seq reads. 128

Figure S 24: Coverage of the proteome by observed peptides. 129

Figure S 25: Distribution of reads and spectra among observed genes in VCaP and RWPE. A and B. Distribution of reads and spectra, respectively, across extended dataset. C and D. Distribution of reads and spectra, respectively, across extended dataset after removal of top 100 most abundantly observed genes. E and F. Distribution of reads and spectra, respectively, across extended dataset after removal of ribosome-associated genes. 130

Figure S 26: Segregation of RWPE genes into broad categories by transcript-protein relationship 131

Figure S 27: Correlation of protein and transcript in high and low stability groups chosen using transcript half-life 132

Figure S 28: Correlation of protein and transcript in high and low stability groups chosen using protein half-life 133

Figure S 29: PI3K/Akt Signaling Pathway colored by VCaP/RWPE transcript fold change, protein fold change, discordance index, and concrodance index data. 136

LIST OF APPENDICES

APPENDIX A: Chapter 1 Supplementary Methods, Figures, and Tables 86

APPENDIX B: Chapter 2 Supplementary Figures and Tables 104

APPENDIX C: Chapter 3 and 4 Supplementary Figures and Tables 122

ABSTRACT

The proliferation of massively parallel nucleotide sequencing and increases in the throughput of mass spectrometry has produced an unprecedented volume of highly specific, highly accurate data elucidating the transcriptome and proteome. This data explosion has facilitated a tremendous number of novel discoveries in both disease and basic biology. It has also presented a number of challenges due to the characteristics of these cutting-edge technologies. Across these studies, we focus on the context of human cancer where these technologies are increasingly being used to characterize and target molecular aberrations for treatment tailored to individuals' cancer biology.

First, we evaluate the emerging technology of single-molecule sequencing (SMS), which may provide a clearer picture of the biological activity in the cell by avoiding the sample amplification steps that may introduce biases in the data. We compare transcriptome data from both SMS and a method employing amplification, noting the effects that the differences in sample preparation may have on the resulting data in terms of dynamic range and coverage bias. In particular, we find that SMS has greater dynamic range, providing more resolution for low abundance transcripts while avoiding coverage peaks which may result from the amplification process.

We then turn to the challenge of integrating NGS-derived transcriptome data with tandem mass spectrometry data quantifying the proteome. The relationship between the transcriptome and the proteome is broadly defined by the central dogma of molecular biology. However, previous attempts at integrating data from the transcriptome and the proteome have seen large variation in the correlation between transcript and protein. To address this, we developed a framework for integrating data from these two realms using a novel common reference employing corresponding transcript and protein sequences. We apply this framework to integrate data derived from the RWPE and VCaP prostate cell lines and show how a number of methodological factors and sources of error can impact the correlation between transcript and protein.

Finally, we analyze the results of our data integration pipeline with a focus on the transcript-protein relationship. We classify the genes in our dataset into broad categories, and show how their biological roles as well as experimental characteristics impact the relationship we observe between transcript and protein. To compare the cell lines in terms of their genes' transcript-protein relationship with the goal of uncovering the uncoupling of this relationship in prostate cancer, we apply a novel concordance and discordance index to the genes in the dataset. Using these indices, we show how variations in protein abundance drive many of the differences between the cell lines and how stability has substantial impacts on the transcript-protein relationship.

The results and methods derived from this work can be used by researchers in the future to better understand the characteristics of emerging NGS technologies and integrate this data across scales of biology to better understand the molecular underpinnings of disease.

Chapter 1: Introduction

The advent of massively parallel “next-generation” sequencing (NGS) technologies has made for an unprecedented explosion in both the volume and depth of data derived from biological experiments. While opening new avenues of research and enabling the most precise view of the cell’s molecular machinery to date, the massive volume and characteristics of NGS has created many new questions in the analysis, interpretation, and integration of the experimental results. This research focuses on its impact in characterizing the transcriptome, the most frequently assessed metric of molecular activity in cells, and extending those results by coupling them with results from tandem mass spectrometry to gain a multi-scale picture of cellular activity in a cancer context. First, we characterize one of the emerging single-molecule sequencing technologies. Without using an amplification step in sample preparation, we assess the advantages and disadvantages of these methods in assessing the abundance distribution and aberrations in the transcriptome. This is contextualized by our study of the biological characteristics and topological structure of biological networks, subsets of which are often seen dysregulated in human disease. Then, we turn our attention to the proteome, and the particular technical challenge of integrating the rapidly increasing yield of tandem mass spectrometry experiments with the tremendous output of massively parallel mRNA sequencing, or RNA-Seq. Finally, we focus on the results from our computational framework to analyze the transcript-protein relationship and how it is dysregulated in our VCaP prostate cancer model. We pay particular attention to the effect these derangements have on important pathways and networks which may confer growth, survival, and apoptotic escape advantages in our cancer model.

NGS and RNA-seq for interrogating the transcriptome

Next generation sequencing is typically used to describe the massively parallel methods for sequencing nucleic acids that do not employ the Sanger sequencing chemistry that the first generation of sequencing machinery relied upon. The 454 pyrosequencing methodology was the first of the NGS technologies, making its debut in 2005 [1]. Unlike the Sanger chemistry-based sequencing method which produces reads up to 800bp in length, most of the NGS approaches produce short reads of ranging from 50-150bp.

The chemistry underlying each of the methods is highly varied; most methods use a sequencing-by-synthesis approach, such as those from Illumina [2], Ion Torrent [3], and Pacific Biosciences [4, 5]*,* although more exotic approaches, such as the ABI SOLiD sequencing-by-ligation approach exist. Most methods generally involve the construction of a library involving the attachment of adapter molecules to the ends of sheared DNA or RNA and an amplification step [6]. A subset of these sequencing methods are considered “single-molecule” approaches, which do not involve amplification steps that may affect sensitivity and bias sequencing results, ultimately producing a clearer picture of the experimental sample.

In this work, I focus on the use of NGS methods to sequence mRNA, an application commonly called RNA-seq. This is an application where DNA microarrays had previously been the standard for global transcriptome characterization, followed by more exotic methods such as the Serial Analysis of Gene Expression (SAGE) [7] and Massively Parallel Signature Sequencing (MPSS) [8], which were much more rarely used. RNA-seq offered a number of clear advantages over microarrays; not requiring *a priori* knowledge of the transcripts under study (observing only the transcripts for which there is a probe), the production of sequence information about the transcripts under study (due to the direct sequencing of the transcripts in the sample), a much lower background signal, and much higher potential dynamic range [9-11]. Compared to SAGE and MPSS, it has the advantage of higher sensitivity as a result of higher throughput.

The characteristics and possible advantages of the single-molecule approach (in comparison to competitive approaches utilizing an amplification step) had not been well characterized in a transcriptome context, and are the first topic of study. The characteristics of single-molecule based methods are discussed in **Chapter 2** using data from the Helicos Heliscope. While the technology had been proven in sequencing the human genome [12], its application to sequence the transcriptome had not yet been explored in depth.

To date, NGS technology has yielded an unprecedented amount of data – never has so much data been produced. The simultaneous factors of constantly plummeting costs, improving quality, and increases in experimental yield guarantee that these technologies will become widespread in the future, thus making the characterization of experimental results and development of methods to leverage their data invaluable.

Disease and the dysregulation of biological networks

High-throughput techniques for determining molecular interactions have opened the door to genome scale evaluation of the molecular interactome of many species due to the quickly growing pool of data. A number of databases have been developed in order to integrate protein interaction data from high throughput experiments such as DIP, BIND, HPRD, and several others. Studies looking at this data across a number of organisms have indicated that these networks are organized into functional biomodules that function at multiple scales [13-15].

Analysis of disease gene knowledge coupled with data from large-scale protein interaction networks to form a phenome-interactome network have revealed that a significant portion of disease-associated genes form small sub-networks. The networks formed by the interactions of known disease genes have been used to relate phenotypically similar inherited diseases together [16]. Similarly, subnetworks that represent protein complexes have been used to relate diseases with similar phenotypes and provide novel disease gene candidates when melded to association data [17]. The disease-associated genes themselves also seem to possess a number of characteristics within the interactome. Compared to the mean degree values of all proteins, many disease related proteins display relatively elevated degree and tend to interact with other disease-related proteins [18, 19]. This property has been used to propose likely candidate genes for disease association [20]. Taken together, it suggests that the intermediate nodes in the interactome play a contributory factor. In addition to the importance of highly interconnected “hub” proteins [21, 22], certain topological features were found to be associated with essentiality/lethality [23]. Additional research has suggested that genes expressing proteins of similar importance also share topological characteristics in the interaction network [24]. These topological characteristics have been used to explain variable disease outcome [25], making an argument for their role in the progression of disease.

To study this phenome-interactome network in human disease, we integrated data from several protein interaction networks with gene-disease relationships to create a set of sub-networks that form functional biomodules for over 4,300 diseases in single-gene disease (SGD) and complex disease (CD) categories, as well as over 6,600 functional sub-networks derived from Gene Ontology (GO) classes. The diseases in the SGD category were primarily caused by aberrations in one of several individual genes in the derived sub-network, in contrast to those in the CD category where several genes in the derived sub-network often influenced the resultant disease phenotype.

The subnetworks associated to human diseases and biological processes were built by the determination of all shortest pairs paths between all distinct associated genes found in the protein interaction network for each disease or biological process. Shortest paths in the interaction subnetwork are determined using Dijkstra's shortest paths algorithm [26]. For example, **Figure 1** illustrates a hypothetical disease of interest associated to UMLS concept 'UMLS:000000', associated with genes A, B, C, D, and E. The shortest path between pairs {A,B}, {A,C}, {A,D}, {A,E}, {B,C}, {B,D}, {B,E}, {C, D}, {C, E}, and {D, E} would be analyzed, noting the identities of the original nodes, the original node also found in the protein interaction network (as many nodes are not represented within the network), the intermediate connecting nodes, and the respective counts of each class. This process discovers intermediate nodes X, Y, and Z in the process of deriving the subnetwork and associates these nodes.

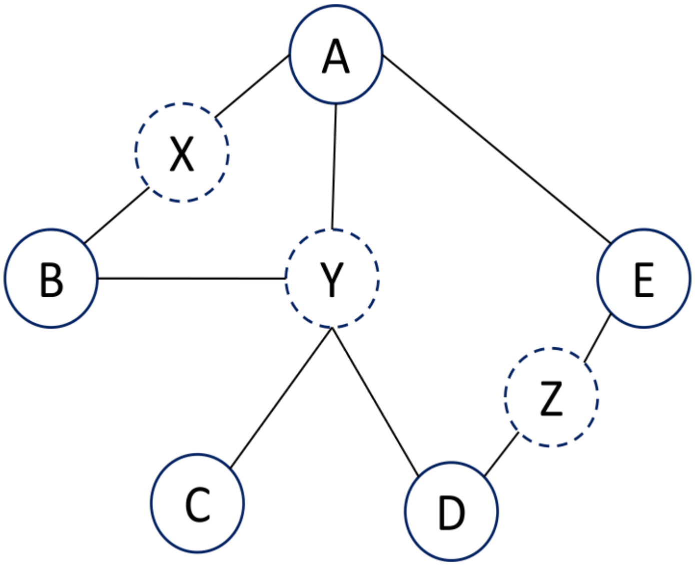


Figure 1: Derivation of an example subnetwork composed of five annotated genes and three intermediate genes

We analyzed the structure and characteristics of these functional and disease networks using network analysis tools and unsupervised machine learning techniques described in detail in .

Subnetwork Characteristics

As expected, the OMIM-derived SGD set demonstrated a smaller range in size in terms of total gene count from 3 to 32 genes with a median of five genes, while the complex disease set was composed of networks of much more varied size, ranging from 3 to 127 genes, with a median of eight genes. The Gene Ontology derived background set had the largest range from 3 to 968 genes. As shown in **Figure S 2a-c**, most subnetworks tended to remain small, generally involving between three and nine genes. The GO background set exhibits a long-tailed distribution with most networks remaining under seventeen genes in size.

Classification accuracy

Unsupervised Principal Components Analysis and k-means clustering methods were first attempted in order to assess the separability of the three classes of subnetworks. As shown in **Table 1** and **Figure S 1a and b**, clustering mirrored the results of the PCA with high misclassification levels (misclassifying ~55% of the data), further demonstrating the poor separability of the data.

		Assigned to Cluster		
		GO	SGD	CD
Source	GO	59	4	16
	SGD	1220	435	932
	CD	158	31	89

Table 1: Unsupervised k-means clustering illustrates the poor separability of the data, with 1631 (55.4%) instances incorrectly clustered

As a result, machine learning techniques must be applied to derive the subtle differences between the CD, SGD, and GO sets. As shown in **Table S 2a-i**, the overall misclassification error rate remains relatively low across several subsets of the subnetwork parameter data, never exceeding 5%. Other measures – precision, recall, f-measure- exhibit very satisfactory performance. However, a close inspection of the results for the three class problems (SGD, GO, CD) reveals that the results for the SGD class are not satisfactory. Confusion matrices from these analyses show the classifier tends to assign those subnetworks to the GO class, an issue likely due to the single-point driver nature of single-gene disease. Further analysis of the data by breaking down the features into biological and topological characteristics further revealed the similarities between the SGD and GO set, further detailed in **APPENDIX A**. The separability of the SGD and CD sets as shown in **Table S 2j** demonstrates the differences in subnetwork characteristics between those primary involved with single-gene disorders and those associated with multigenic, complex disorders. A reclassification of all the study data was also done using a GO dataset that included only the “Biological Process” entries, with similar results. The complete results of the classifications as well as additional methods and analyses are available in **APPENDIX A**.

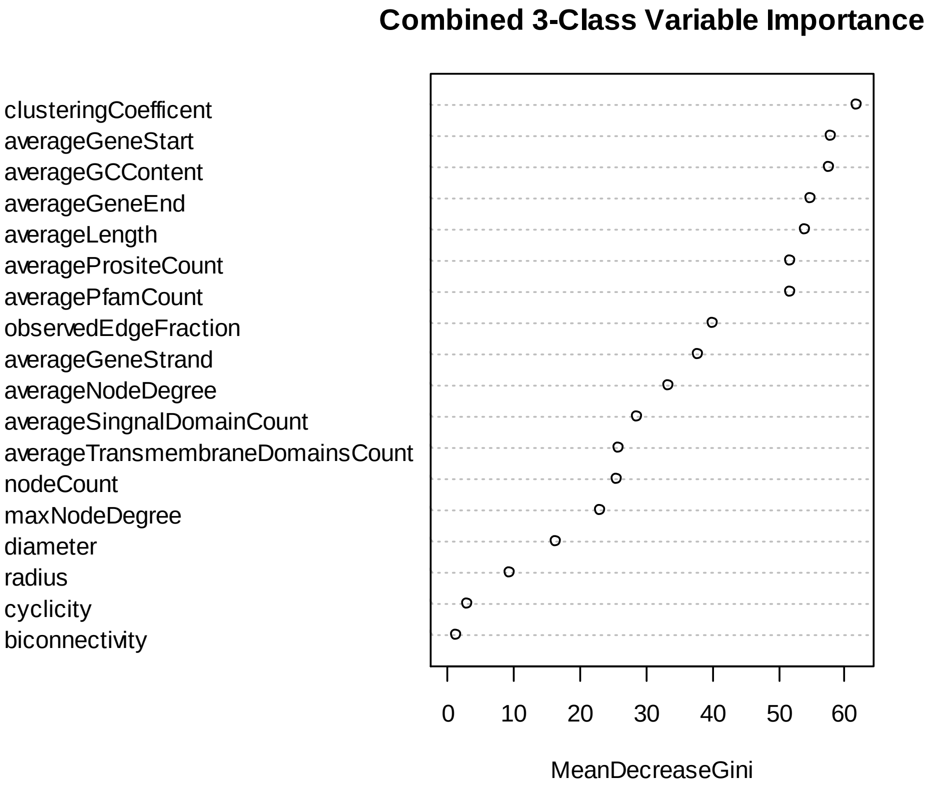


Figure 2: Variables ranked by importance in classification based on Gini coefficient. The most informative variables in classification were a mix of both biological and topological parameters.

The most important variables in the classification of subnetworks to their individual classes is illustrated in **Figure 2** as derived using the reduction in Gini index, a measure of the reduction in misclassification when a particular variable is used.

Lessons from the analysis of disease- and function- associated subnetworks

The relative paucity of data describing disease-associated subnetworks continues to present a serious challenge in the analysis of the functional biomodules underlying human disease. While the classification of complex disease-associated subnetworks appears to achieve reasonable results, the underlying heterogeneity of human disease, as evidenced by the SGD set in **Table 2**, will always present a problem in classification.

Correctly Classified Instances		2795		94.94 %	
Incorrectly Classified Instances		149		5.06%	
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.101	0.003	0.5	0.101	0.168	SGD
0.997	0.387	0.949	0.997	0.972	GO
0.752	0.001	0.986	0.752	0.853	CD

Table 2: Classification of CD, SGD, and GO classes using all variables. While the complex disease (CD) subnetworks and derived from the Gene Ontology (GO) demonstrated relatively good classification performance, the subnetworks associated with single gene dieases (SGD) were very poorly separable.

It is notable that the variables with the highest influence are a mix of both topological and biological factors, confirming previous findings that characteristics from both categories play an important role in the susceptibility to biological disruption and resulting disease. The relative importance of clustering coefficients confirms recent results examining the differences between disease-associated genes and essential genes [27]. The inclusion of mean gene start locus and GC content confirm the relative importance of genomic localization and transcriptional propensity [28]. While the examination of individual factors increases confidence in the findings through recapitualation of established study results, the random forest is able to capture the interaction between these variables. These inter-variable interactions are a prime target for continued study.

It is not completely surprising that the SGD subnetworks appear to bear a strong resemblance to the GO background considering the pathogenesis of diseases that arise from anomalies in a single gene. In many cases, the GO-derived subnetworks can be considered functional biomodules of the interactome. The disruption of certain genes in these functional biomodules is likely to manifest in the form of disease phenotypes if they are not serious enough to result in lethality. This can result in failures of protein complex assembly and complementation such as in Xeroderma Pigmentosum, a single gene disease that can arise from any one of the seven known genes in the XPA-XPG complementation group associated with nucleotide excision repair [29]. As such, these two classes are relatively poorly separable even in a supervised machine learning context.

As we expected, the differences between the networks formed by sets of genes associated with biological processes and those associated with human disease are subtle and not easily derived as they are, by definition, intimately linked. The similarity between the single gene disease-associated subnetworks and those derived from the Gene Ontology demonstrates the multiscale behavior of a single disruption in a functional biomodule, and its ability to cause debilitating effects. The need for additional data and high specificity data is made abundantly clear in this study, as demonstrated by the propensity for misclassification of complex disease-associated subnetworks as well as the limited number of subnetworks derived from the data due to lack of representation in the interaction network. The limited availability of interaction propensity or data quality measures associated with individual interactions in the particular version of the interaction database we employed led us to treat all interactions as equally probable and equally correct. This may be a source of error in the process that may be ameliorated in the future with additional data and quantitative measures associated with the interactions. As more gene-disease association data becomes available, the effectiveness of this method should be re-evaluated.

The transcriptome, proteome, and the challenges of data integration

The transcriptome is a common metric for assessing the biological activity in cells, used with the implicit assumption that the activity of the proteome follows. The relationship between mRNA transcripts and proteins is described by the central dogma of molecular biology; the information in DNA is transcribed into mRNA, which is subsequently translated into protein products [30]. However, this transfer of information from DNA to protein is mediated by a large number of intermediary factors such as ribosome stalling during translation [31], nonsense-mediated decay of transcripts [32], transcript degradation by small [33] and other noncoding RNAs, protein decay, and a large number of post-translational modifications [34].

Understanding the role each of these distinct regulatory mechanisms plays in affecting the resulting abundance of protein is a central motivator for studying the transcript-protein abundance relationship.

The task of correlating the transcriptome to the proteome has historically been subject to a number of challenges. In the past, the most restrictive of these has been the limited ability to sample both the proteome and the transcriptome. For example, early studies investigating this relationship often employed gel-based methods, which limited them to small subsets of genes due to these experimental limitations [35, 36]. As a consequence, many of their results were inconclusive and were difficult to generalize to the broader set of genes.

While not matching the beyond-exponential growth rate of nucleotide sequencing, the throughput of proteome profiling techniques has grown significantly. A number of methods have been applied for the quantitative profiling of proteins using both radio-labeled and label-free methods, reviewed in [37-39]. While monitoring post-transcriptional modifications remain a challenge, these advances have enabled the profiling of nearly the entire proteome, estimated at 10-12,000 proteins [40].

These capacity increases from the development of mass spectrometry (MS) methods and the advent of DNA microarrays enabled large increases in the number of genes that could be studied simultaneously. These technologies were the first to enable the comparison of complex transcriptomes to equally complex proteomes. However, experimental results remained highly variable. For example, a study of transcript and protein levels across 98 genes in 78 lung adenocarcinomas using microarrays and a MALDI-MS method found highly varying correlation levels between r = -0.4 and 0.4 for each gene in their set, and a global correlation of r = -0.025 [41]. A slightly larger study by Cox, et al using microarray and LC-MS measured a correlation of r = 0.63 in approximately 900 genes in developing mouse lung tissue [42]. A more recent study by Gry, et al. utilizing microarrays looking at 1066 genes across 23 cell lines found a correlation of r = 0.52 [43]. A similar study across the NCI-60 set of cell lines by Shankavaram, et al. observed correlations ranging from r = 0.48 to r = 0.58 in a set of 162 feature set of assayed proteins, across four related microarray platforms [44]. An analysis looking at protein and transcript levels by Ghazalpour, et al. in 97 strains of mice found a correlation of only r = 0.27 [45].

The application of RNA-seq to this task, with its much greater dynamic range and specificity, refined the biological picture. A study combining previously published SILAC-labeled protein abundance with separate RNA sequencing data in three cancer cell lines, A431, U251MG, and U2OS found transcript-protein relationships correlated at levels from *r* = 0.55-0.61 [46]. A similar study focusing on deeply profiling both the transcriptome and proteome in the HeLa cell line found a correlation of *r* = 0.6 [40].

These studies used a diverse set of techniques for data integration – though all attempt to address two fundamental challenges in this process, noted in [47-49]; First, how to match transcripts to proteins to ensure that the same entities were being compared. Different gene annotation, definition, and naming schemes must be harmonized in order to ensure a fair comparison, a process complicated by the multiplicity of transcript isoforms and incomplete transcript and protein databases. The second major challenge is that of comparing the transcript and protein abundance values. Derived from two different technologies, the transcript and protein abundance values have very different ranges of sensitivity, distributions, and error profiles. The very computation of these abundance values is an important factor, but is beyond the scope of this research and is reviewed in [50-52] and [53, 54] for RNA-seq and label-free proteomics, respectively. RNA-seq is often quantified using the TopHat and Cufflinks suite of tools [55] using related sources of annotation. Quantification of protein abundance from MS experiments is more varied and largely dictated by the choice of experimental procedure and processing pipelines.

The issue of integrating transcriptome data from RNA-seq and protein data from tandem mass spectrometry methods is an active area of research, and the focus of **Chapter 3**. While a number of studies have focused on examining the relationship between transcript abundance and the resulting protein products, methodologies and resultant correlation relationships vary significantly. In particular, few studies analyze the impact of their abundance measurement and data integration methods on the resulting relationship. Consequently, it is unclear how these differences in bioinformatics methodology affect the final correlation. To address the issues of data integration, we construct a common reference database from the RefSeq database composed of corresponding transcript and protein sequences against which we applied the Trans-Proteomic Pipeline [56] and Abacus [57] for protein abundance quantification and an in-house pipeline for RNA-seq quantification. This unique approach allowed for a one-to-one comparison of transcripts and protein products. Using this method to ensure proper comparison of genes, we normalize the abundance values derived from each of the experiments in the transcriptome and proteome.With the data derived with this methodology, we explored how technical factors, namely identification and counting methods in both transcriptome and proteome data, contribute to uncertainty in correlating transcript and protein abundances. In particular, we show how read mapping for transcriptome data and multiple assignment of MS spectra lead to variation in the correlation coefficient of abundance.

The transcript-protein relationship and its role in disease and cancer

One of the primary motivations for studying the transcript-protein abundance relationship and the factors that affect it is the desire to dissect the regulatory mechanisms of the cell [58]. To address this question, several studies have examined the influence of a number of regulatory factors on the correlation observed between protein and transcript abundance. For example, Vogel, et al. examine the role of sequence features that may affect transcription, degradation, or translation, such as 5’ and 3’ UTR lengths, local secondary structure, and the number of miRNA target sites on the transcript in the Daoy medulloblastoma cell line [59]. In a similar study, Schwanhäusser, et al. derive a quantitative model of protein abundance, noting that both transcript and protein half life have significant impacts on the transcript-protein abundance relationship [60]. In both of these studies, the authors note that sequence features play a role in the abundance of protein products in addition to transcript abundance, although a significant amount of this variability is still not accounted for. The biological role of the transcripts and proteins play a part in this – for example, several studies have observed that highly stable structural proteins have higher correlation with their cognate transcripts [43].

Although the dysregulation of the transcript-protein relationship can play a role in disease, it has not been well studied. While a number of the studies analyzing the transcript-protein relationship are focused on cancer, most of these studies used microarray or older techniques, and suffer from the issues of small gene sets and limited dynamic range discussed earlier. On the other hand, the studies that were aimed at a thorough analysis of both the transcriptome and the proteome generally did so outside of a disease context.

In the landmark “Hallmarks Of Cancer,” Hanahan and Weinberg broadly classify the set of biological characteristics commonly acquired by cancers into a set of six traits. These acquired traits, enabled by genome instability, are self-sufficiency in growth signals, insensitivity to growth-inhibitory signals, the evasion of apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis [61]. To these, the subsequent “next generation” of cancer hallmarks added the deregulation of cellular energetics and avoidance of immune destruction. In addition, it recognized the role of inflammation in promoting tumorgenesis and progression [62].

Changes in the transcript protein relationship can have important functional consequences due to alterations in the regulatory structure of the cell, giving rise to the hallmark characteristics of cancer. In breast cancer, it has been observed that the stabilization of DNA Methyltransferase 1 (*DNMT1*) causes its dysregulation leading to aberrant genomic hypermethylation [63]. This is primarily seen as an increase in protein levels without an increase in the cognate transcript. There is also significant evidence that stabilization of transcription factors that regulate transcript abundance is a common mechanism of gene regulation, with several examples in human cancers. *HIF-1a* is a commonly studied transcription factors due to its widespread effects on cell survival and angiogenesis as well as its activation under the hypoxic stress often seen in tumors [64]. It has been shown to be stabilized by interaction with another transcription factor, YY1, itself implicated in tumorgenesis [65]. This stabilization was also noted as a change in protein abundance without an effect on mRNA level [66]. Another example is observed in the interaction of the *LMO2* and *SCL* transcription factors, both of which have been implicated in hematopoietic cancers [67, 68]. The protein product of the *LMO2* gene, a central component of several transcription factor complexes, has been shown to be stabilized by interaction with *SCL*, which prevents its degradation allowing for more widespread transcription factor complex assembly [69].

To address this paucity of cancer-focused research encompassing a comprehensive set of genes from the transcriptome and proteome, the work in **Chapter 4** focuses on the characterization of the transcript-protein relationship in the RWPE prostate epithelial cell line and the VCaP prostate cancer cell line and compares the two. We focus on the transcript-protein relationship within various biological functional classes, and how that relationship is altered in a cancer context. In particular, we use a novel transcript-protein discordance index to assess the level of transcript-protein dysregulation in VCaP when compared to RWPE.

Application of methods to other studies

The methods and knowledge developed in the course of these studies for profiling the landscape of the cellular transcriptome using RNA-seq were also applied to a number of other studies. The derived expression measurements were often used in tandem with other types of profiling to analyze multiple facets of biological activity. In Maher, et al. [70], an early version of the quantification methodology was used to contextualize the abundance of known and novel gene fusion transcripts in prostate cell lines in terms of the broader transcriptome. This was one of the first studies to utilize RNA-seq data to infer gene expression values from read mapping counts, and showed the relative abundance of several gene fusion transcripts such as the well-characterized *TMPRSS2-ERG* and the novel *USP10-ZDHHC7* fusion transcripts compared to some of the most highly expressed genes in the transcriptome. This expression calculation technique was applied in Kim, et al. [71], where transcriptome sequencing derived expression values were integrated with microarray and NGS-derived DNA promoter methylation data to analyze the impact of differential methylation patterns in prostate cancers.

The lessons in read mapping, detection, and dynamic range were also applied in contributions to studies using other sequencing-based methods to profile the genome and exome. Accurate read mapping and artifact filtering techniques derived from the development of the RNA-seq quantification methodology were applied to an evaluation of variants in the transcribed portion of the genome, or exome, in Grasso CS, et al. [72] focused on castration-resistant prostate cancer. These efforts led to increased detection of rare single-nucleotide variants and short insertions and deletions in the exome.

Similarly, methods to filter out artifacts resulting from ambiguously mapping reads and aberrant characteristics of the sequencing process were applied to develop a pipeline for characterizing the structure of the cancer genome in a personalized oncology context in Roychowdhury S, et al. [73]. This method derived both copy number and aberrant mapping data across the sampled genomes and integrated them to increase confidence in the derived candidates. Consequently, this methodology was able to recover genome-scale rearrangements from low-depth genomic sequencing of a cancer sample, in absence of a corresponding normal sample, in a subset of the total four patients with advanced or refractory cancer enrolled in the study. These genomic rearrangements demonstrated the genomic basis for the RNA fusion transcripts observed in this study and several subsequent patients enrolled in the MI-ONCOSEQ personalized oncology program.

List of contributions to publications

1. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, Quist MJ, Jing X, Lonigro RJ, Brenner JC, Asangani IA, Ateeq B, Chun SY, Siddiqui J, **Sam L**, Anstett M, Mehra R, Prensner JR, Palanisamy N, Ryslik GA, Vandin F, Raphael BJ, Kunju LP, Rhodes DR, Pienta KJ, Chinnaiyan AM, Tomlins SA. The mutational landscape of lethal castration-resistant prostate cancer. *Nature*. 2012 Jul 12;487(7406):239-43.

   In this study, I developed a pipeline using the BWA short read aligner and samtools suite to discover single-nucleotide polymorphisms and short insertions and deletions (indels) specific to cancer across samples in our cohort.

2. Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, Cao X, Kalyana-Sundaram S, **Sam L**, Balbin OA, Quist MJ, Barrette T, Everett J, Siddiqui J, Kunju LP, Navone N, Araujo JC, Troncoso P, Logothetis CJ, Innis JW, Smith DC, Lao CD, Kim SY, Roberts JS, Gruber SB,

Pienta KJ, Talpaz M, Chinnaiyan AM. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med*. 2011 Nov 30;3(111):111ra121.

In this study, I contributed structural variation and copy number results derived from a single channel of low-depth genomic sequencing of the cancer sample without a matched normal using the Breakdancer and ReadDepth tools. These results were derived from a comprehensive profiling pipeline I developed for the characterization of large scale variation in the genome.

3. Kim JH, Dhanasekaran SM, Prensner JR, Cao X, Robinson D, Kalyana-Sundaram S, Huang C, Shankar S, Jing X, Iyer M, Hu M, **Sam L**, Grasso C, Maher CA, Palanisamy N, Mehra R, Kominsky HD, Siddiqui J, Yu J, Qin ZS, Chinnaiyan AM. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. *Genome Res*. 2011 Jul;21(7):1028-41.

   I contributed gene expression profiling for RNA-seq data, used in conjunction with methylome (Methyl-seq) profiling data, to survey the effects of dysregulated methylation in prostate cancer. I also contributed assistance in optimizing the alignment of Methyl-seq derived reads.

4. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, **Sam L**, Barrette T, Palanisamy N, Chinnaiyan AM. Transcriptome sequencing to detect gene fusions in cancer. *Nature*. 2009 Mar 5;458(7234):97-101.

   Using an early version of the transcriptome quantitative profiling pipeline I developed, I contributed RNA-seq based gene expression to this study. This provided context for the relative abundance of aberrant gene fusion products to the rest of the transcriptome.

Chapter 2: A Comparison of Single Molecule and Amplification Based Sequencing of Cancer Transcriptomes

Introduction

Sequencing samples at single-molecule resolution is seen as the next step in the evolution of Next Generation Sequencing (NGS). These technologies have already produced unprecedented amounts of data at nucleotide-level resolution, and are transforming our ability to observe biological systems. NGS technology has had a particular impact in the study of transcriptomes through mRNA sequencing, or RNA-Seq. Offering a wide dynamic range and truly global view, this NGS application is quickly supplanting existing approaches for monitoring complex transcriptomes where both transcript lengths and concentrations are highly heterogeneous. The multi-faceted nature of RNA-Seq has enabled in-depth analysis of transcript abundance [9, 74, 75], alternative splicing [76-79], novel transcript detection [80], biomarker discovery [81-83], pathogen detection and characterization [84-86], and gene fusion discovery [70, 87, 88].

The first wave of 'next generation' sequencing platforms such as those from Applied Biosystems, Illumina, Ion Torrent, and Roche/454, utilize PCR based amplification steps in sample preparation and sequencing and are thus categorized as amplification based sequencing (AS) methods. A second set of platforms, described as 'single molecule sequencing' (SMS) [89] by Helicos and Pacific Biosciences, eliminate the amplification steps involved in the sample preparation and sequencing process and thus profess to provide a more accurate view of the transcriptome.

AS techniques typically involve two amplification steps; the first amplification occurs during the creation of the double-stranded cDNA library from the fragmented mRNA. The cDNAs are ligated to a pair of adapter molecules, and PCR amplified. A second amplification step is carried out with the adapter-ligated single cDNA strands hybridized to primers bound to a glass or silicon substrate to produce local clusters of identical molecules using isothermal amplification or emulsion PCR. Taken together, these two steps have the potential to selectively introduce over-represented segments and genes into AS data. It has been observed that this bias exists [90-93], however its effect on transcript coverage and quantification has not been thoroughly explored in complex samples with transcripts at variable concentration. The Helicos SMS protocol involves creation of single-stranded cDNA templates directly from mRNA and hybridization of these poly-adenylated templates to complementary oligomers bound to a glass slide for sequencing (**Figure S 3**).

Results

Assessment of SMS RNA-Seq through transcript profiling

To systematically assess the differences between the two sequencing technologies, we analyzed RNA-Seq results from amplification-based sequencing (AS) and single-molecule sequencing (SMS) across a set of twelve cancer cell lines and tissue samples. In particular, our approach attempted to discover recurrent biases that may be introduced by the amplification steps implicit in AS. Our initial dataset used to evaluate quantification performance is comprised of samples from the prostate cancer cell lines DU145, RWPE, VCaP, and LnCaP, and one prostate cancer tumor tissue with a matched adjacent normal sample. Out of our set, three samples each of VCaP and LnCaP were structured as a time course study with 0h, 24h, and 48h time points.

In our analysis of the two technologies, we chose to use the preferred alignment tool for each technology in a “best vs. best” approach. AS reads were aligned with the Bowtie aligner [94] while SMS reads were aligned with IndexDP [95] (**Figure S 4**). Reads aligning to known biological contaminants such as mitochondrial DNA, ribosomal RNA, and technology-specific contaminants such as adapter sequences and long oligomers, were filtered out of the data set prior to analysis.

To assess the variation between SMS and AS technologies, we adopted a simple read counting procedure similar to other RNA-Seq quantification methodologies [9, 74]. Reads from single lanes of AS and SMS technologies run in parallel, were aligned to 56,722 University of California Santa Cruz (UCSC) transcripts (version hg18). We then enumerated reads per-transcript and normalized based on the number of high quality, non-contaminant reads per sample to obtain values in reads per million (RPM). To avoid uncertainty associated with multi-mappings to gene isoforms, only single-best mapping methods were used to quantify the genes for comparison. Single best mappings were derived from AS reads by setting Bowtie to report only the single highest quality alignment per read. Single best alignments were derived from SMS reads by accepting alignments with the highest quality scores. Values from all gene transcript isoforms, as defined by UCSC, were summed to yield values in terms of alignments per million reads for each of the 29,416 genes. Coverage values in reads per kilobase per million (RPKM) were computed by summing RPKM values of the isoforms of each gene. Through a head to head comparison between AS and SMS reads of identical samples run in parallel on the two platforms, we observed a systematic over-representation of high expressing transcripts in AS as compared to SMS. This bias resulted in reduced coverage of mid- and lower-level expression genes leading to overall lower transcript detection sensitivity in AS. Reprocessing a subset of AS samples using IndexDP and repeating the analysis ruled out technical differences in read assignment as the cause of this representation bias. As the sequencing technologies and chemistries continue to advance, we expect AS platforms will overcome the limitation of low expressed transcript detection by enhanced throughput.

Global properties of AS and SMS results

Transcriptome sequencing was carried out in parallel on AS and SMS platforms for 12 samples including 10 prostate cancer cell lines and 2 prostate cancer tissues. Overall, we generated 2.8 to 19.7 million raw AS and SMS reads in each of the 12 samples. Approximately 30-60% of these reads passed initial filtering steps and aligned to our transcriptome reference. SMS reads were produced in two separate machine runs while AS reads were produced across 6 independent machine runs. This procedure resulted in 2.1 – 15 million and 2.8 – 8 million reads for SMS and AS, respectively, which aligned to our transcriptome reference. In 10 out of the 12 samples used in the evaluation, SMS produced more alignable reads in absolute terms, with a median of 1.39x across all 12 samples. SMS results contained more reads aligning to known contaminants, ranging from 12% to 51% of total reads, with a median of 22%. The fraction of reads aligning to contaminants in AS ranged from 2.6% to 14% with a median of 4.2%. SMS read length was variable and a filtering step restricted usable reads to a length range between 24bp and 57bp in the first run, and 25bp and 64bp in our second run, yielding a read count-weighted mean length of approximately 33bp in each of the twelve samples (**Table S1**). A median of 97% of all SMS reads had lengths between 25bp and 47bp across all 12 samples (**Figure S 5**). AS reads were generated at a minimum length of 36bp in each sample, although the first and last several bases were ignored to produce high quality reads at least 34bp in length. All AS reads were considered to have a maximum of 36bp length. Reproducibility between technical replicates of the DU145 cell line was high for both AS and SMS methods, with a Pearson correlation of *r*=.98 for both technologies (**Figure S 6**). Reads from both AS and SMS were also aligned allowing for 25 maximum mappings to assess the distribution between uniquely- and multiply- mapped reads at the gene level, although only single-best mappings were used for quantification and comparison purposes. Both technologies achieved very similar unique mapping rates of 72% and 75% in AS and SMS, respectively. From this raw aligned data, we examined the relative distribution of reads across genes observed in our samples by comparing their normalized read counts. As expected, we observed broad agreement in terms of gene expression values between the technologies (**Figure S 7**). However, we observed a recurrent pattern of over-representation of high-abundance transcripts by the AS methodology as compared to SMS.

Coverage bias in amplification-based sequencing

Comparison of transcriptome reads of the same samples quantified in parallel from AS and SMS platforms reveals a distinct bias in AS results towards a slight overrepresentation of highly expressed genes as compared to SMS, as shown in **Figure 3A**. This difference was qualitatively assessed by dividing the genes into quartiles of equal number, ordered by observed values in AS, with the first quartile representing the highest expressing genes, the second quartile representing mid-level expression genes, and the third and fourth quartile defining the genes with the lowest levels of transcripts (**Figure 3B**). Highly expressed transcripts tended to have more read coverage in AS, whereas SMS tended to cover the lower expressed transcripts more effectively. This additional coverage of high-concentration transcripts consistently appeared to be at the expense of lower-expressed transcripts, which tended to be more thoroughly sequenced using SMS (**Table S 6**).

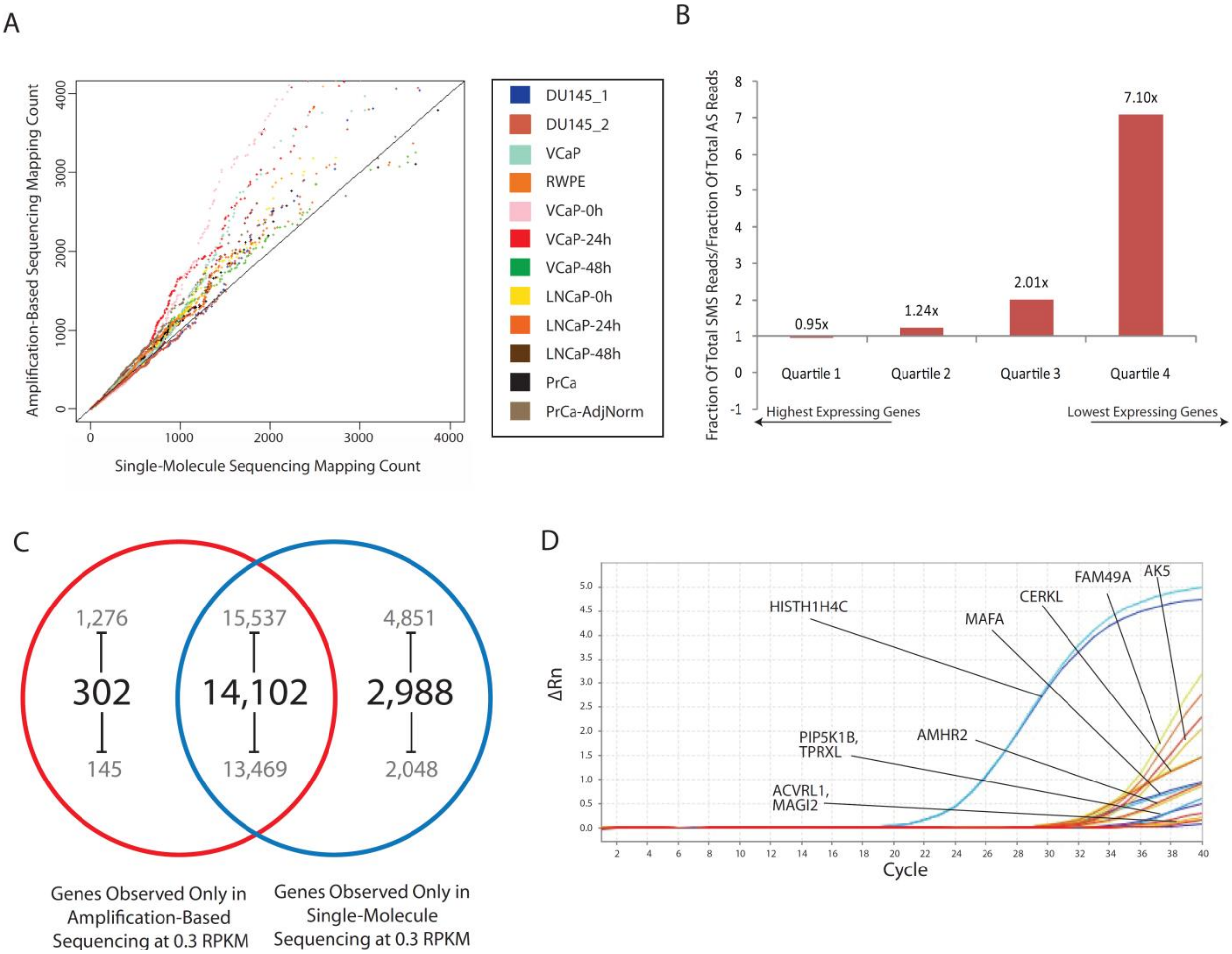


Figure 3: Observed bias in amplification-based sequencing. A. Single-best mapping method-based quantile-quantile plot demonstrates evidence of over-representation of highly expressed transcripts in amplification-based sequencing compared to single-molecule methods. **B.** Distribution of reads across genes by transcript concentration shows decreased SMS coverage of the most highly expressed genes, with those reads going to mid- and low-level expressors. **C.** Differences in the distribution of reads lead to increased sensitivity of low-expressing transcripts. **D.** Nine of the candidate genes seen above the 0.3 RPKM noise level demonstrated any amplification by RT-PCR, although only HIST1H4C showed high abundance.

In order to ensure that these biases were not the result of using a different aligner for each technology, AS reads were re-aligned using the IndexDP aligner used for SMS reads for a subset of the samples, composed of the VCaP-24h, VCaP-48h, LnCaP-24h, LnCaP-48h, and DU145_1 samples (**Figure S 8**). Very high correlation of gene-level values comparing Bowtie and IndexDP alignments for the set of AS reads ruled out differences between alignment tools as the source of the observed biases. For example, correlation of gene-level values in the LnCaP-24h sample was high between alignment methods at *r* = 0.97. Similarly high correlation levels above *r* = 0.95 were observed in the remaining samples. Similar patterns of high-expressor over-representation in AS were observed using IndexDP alignments of AS reads in place of standard alignments using Bowtie as shown in **Figure S 9**. With methodological differences essentially ruled out, we attempted to observe the effects of this high-concentration coverage bias by examining the detection of transcripts at low levels.

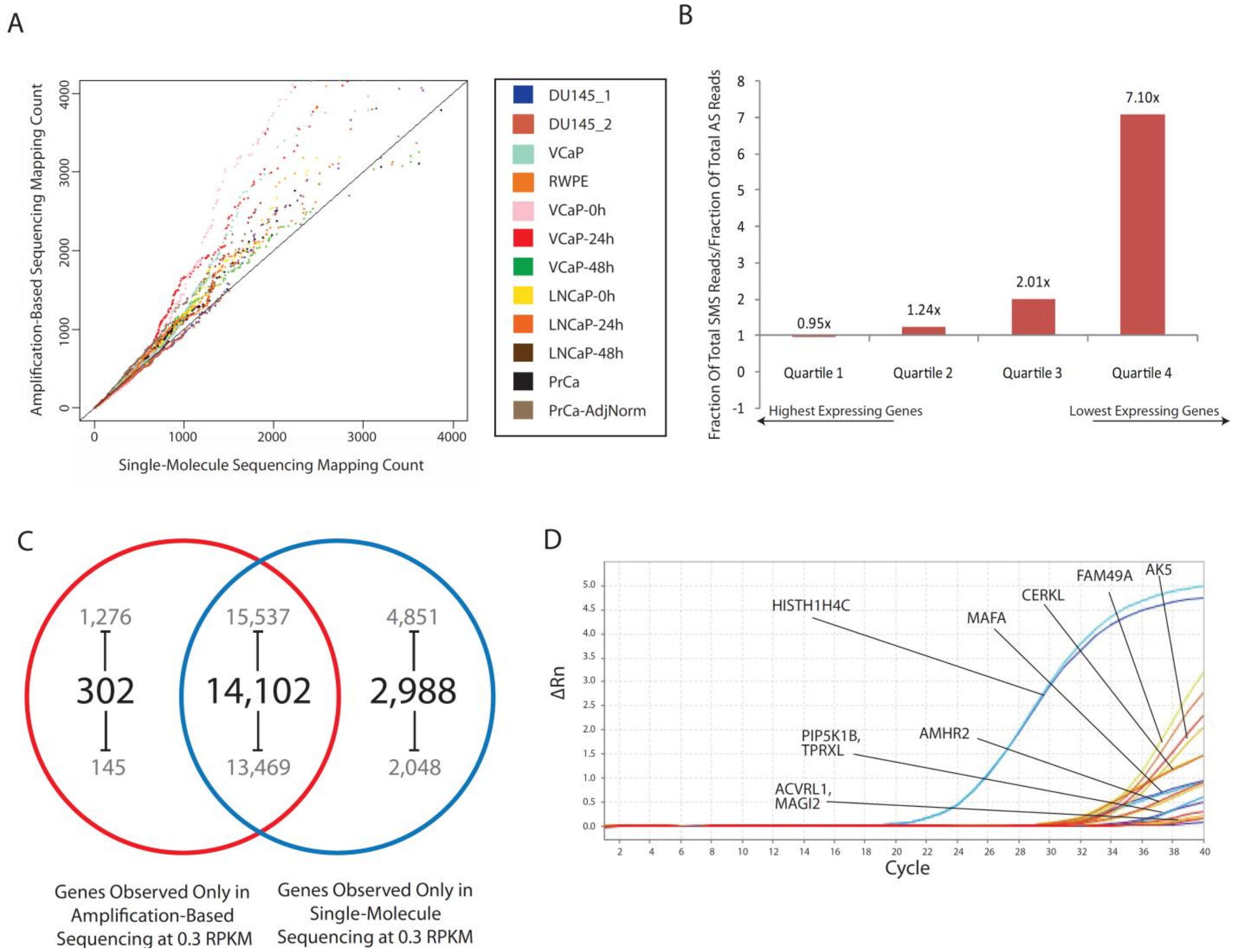
Increased SMS sensitivity results from high coverage of low-abundance transcripts

To evaluate the effects of increased coverage in mid- to low- level transcripts in SMS, we calculated the number of genes observed above a noise threshold in only one of the two technologies. Using the 0.3 RPKM noise level cutoff based on Ramskold, et al. [96], the number of genes detected in only a single technology varied between a high of 4,851 and a low of 2,048 and a high of 1,276 and a low of 145 in SMS and AS (**Figure 3C**), respectively, across the set of samples. A log-fold difference between the numbers of genes detected in only one of the SMS vs. AS technology was observed as we varied the cutoff value between 0.1 RPKM and 3.0 RPKM (**Figure S 10**) in 0.1 RPKM increments. These limits were chosen to examine the sensitivity of the two methods across a range of values starting from a near-zero noise level to an order of magnitude larger than previously reported. Stratification of the genes observed in a single technology into length classes of 0-300bp, 300-3000bp, and 3000+bp demonstrated that this was not due to differences in technology-specific sample preparation, as the AS protocol specifies a ~300bp size selection step that the SMS procedure does not require. This class shows relatively low representation across noise thresholds in both AS and SMS. We then took this evaluation one step further and examined the results from both SMS and AS techniques attempting to find genes detectable only in one technology.

Uniquely detected genes in SMS

In order to substantiate potential representation biases in the two platforms and the suggested additional sensitivity of SMS, we next queried for genes which were detected above a noise threshold by SMS, but were below that threshold in AS. We chose to analyze the DU145 sample as it was the most thoroughly sequenced sample with two replicates run using each technology. Using a 0.3 RPKM threshold, we chose to test the expression of 23 genes in our DU145 samples using RT-PCR, ten of which demonstrated detectable amplification.

Additionally, we sequenced the DU145 cell line much more thoroughly in order to ensure that our detections were not due to technical factors in a single machine run. As shown in **Figure S 11**, this set of genes had better sequencing coverage in SMS as compared to AS across the total 94,427,789 reads generated in our second set of runs. This list was generated by examining the distribution of reads and coverage maps of the top 50 genes whose RPKM coverage showed the largest difference between AS and SMS techniques and had official HUGO names [97]. Candidates were chosen for the presence of long (>36bp) mapping reads and well-distributed read alignments across the length of the transcripts. Of the validated genes detected only by SMS, only *HISTH1H4C* was found to be present in the DU145 sample with high confidence, as shown in **Figure** 3**D**. Nine other candidate genes *AK5*, *ACVRL1*, *AMHR2*, *CERKL*, *MAFA*, *MAGI2*, *PIP5K1B*, *FAM49A*, and *TPRXL* showed weak amplification. In this set of genes, amplification



was only seen beyond cycle 30 making it difficult to confirm their presence. This weak amplification makes it difficult to determine if their detection in SMS is due only to increased sensitivity, or are an artifact of ambiguous mapping. We next sought to examine the over-represented genes that may contribute to the reduction of sensitivity using amplification-based sequencing techniques.

Consistent over-representation of high-expression genes in amplification-based sequencing

Overall, 393 genes were found to be consistently within the set of the top 500 over-represented genes according to normalized read mapping count in at least 40% of our samples (**Table S 5**). Of these 393 genes, ten genes were found to be over-represented by normalized read mapping count across all 12 of the samples considered in the study. The coverage maps of *RPLP0* and *RPL31*, over-represented in all 12 samples, and *SPINT2*, over-represented in 11 samples, demonstrate this coverage bias in these three high expressing transcripts **(**

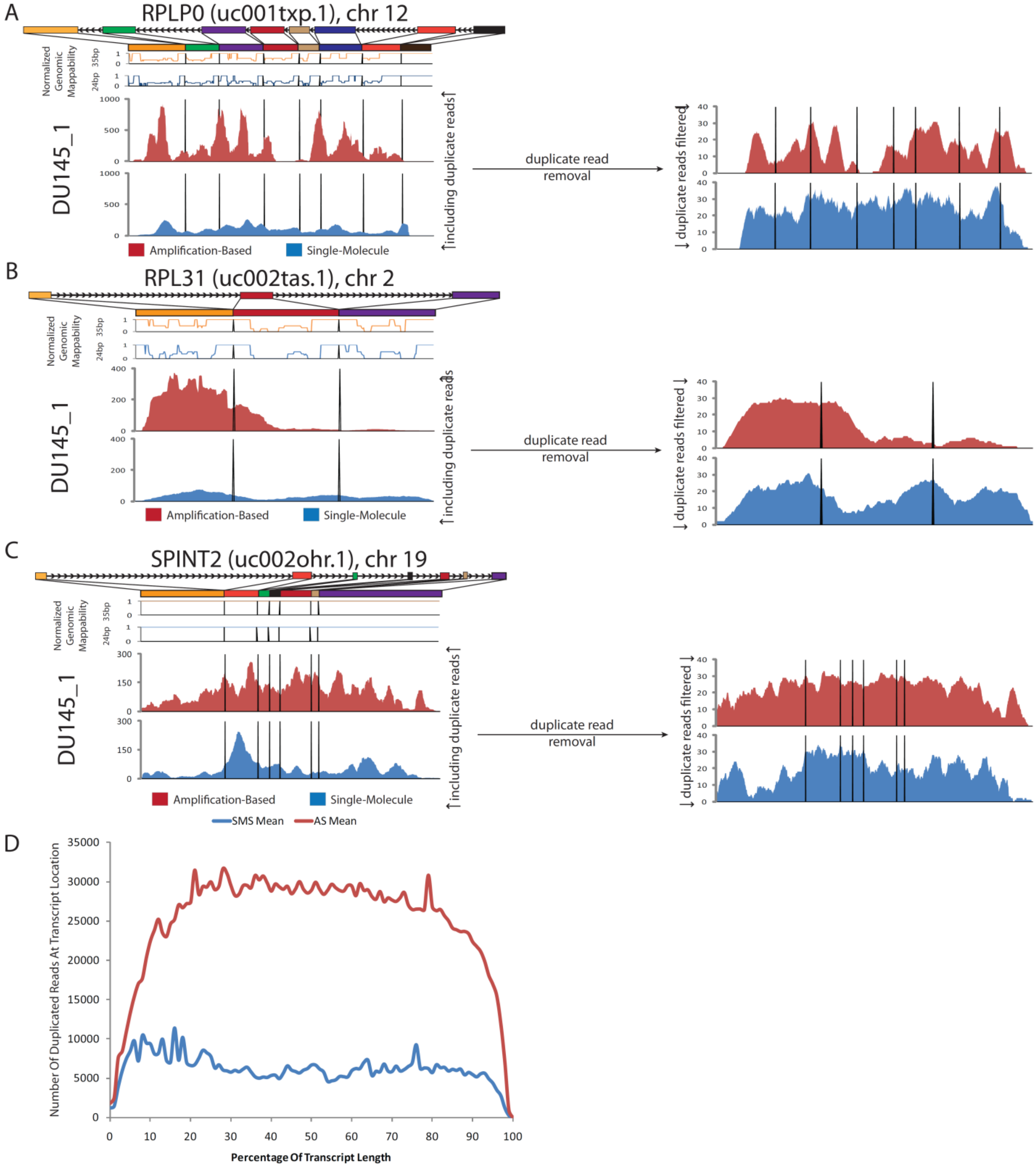


Figure 4**A-C)**. We then examined the composition and distribution of reads in some of these highly over-represented transcripts.

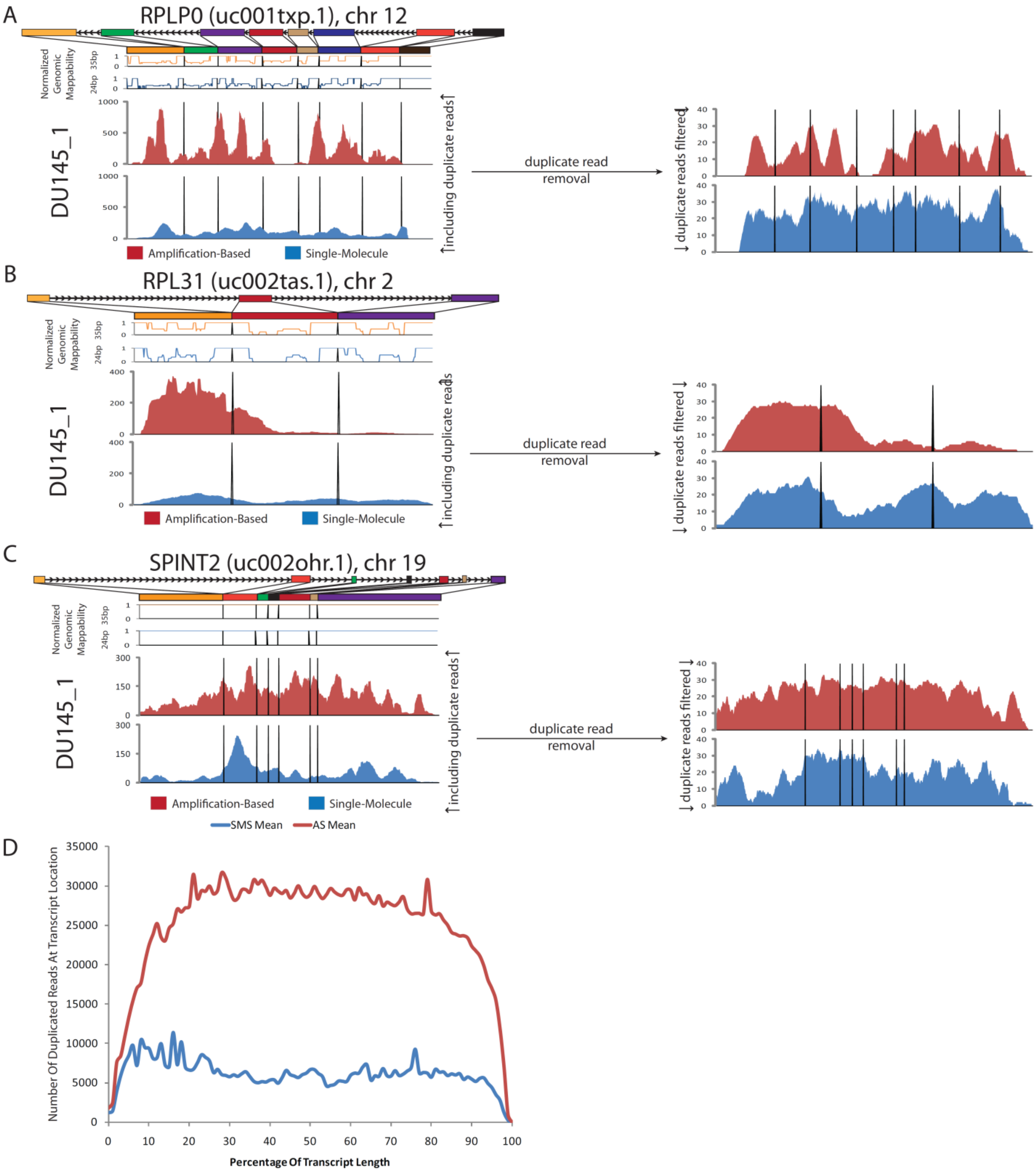


Figure 4: High-concentration transcript bias leads to differences in gene coverage in amplification-based sequencing. Coverage maps from amplification-based and single molecule sequencing demonstrate significantly greater coverage of **A.** RPLP0, **B.** RPL31, and **C.** SPINT2. Removal of reads with the same start positions, strictly suppressing amplification of specific mRNA fragments, significantly reduces the "spikiness" seen in these cases. **D.** Duplicate reads, defined as reads in excess of one per start locus and read length, are relatively evenly distributed along the length of all observed transcripts across all samples in our evaluation set.

Impact of duplicated reads in amplification-based sequencing

The gene *RPLP0* had much greater total mapping coverage in AS across all twelve samples (**Figure S 12**). To aggressively mitigate the effect of amplification in the coverage of this gene, duplicate reads were removed (allowing only 1 read per unique start location) for both technologies as done in previous studies [92, 93]. This resulted in suppression of many of the observed peaks in AS. In contrast, SMS coverage of the gene appeared to be relatively consistent across the length of the *RPLP0* transcript before and after this procedure. This substantial difference in behavior between pre- and post- duplicate read removal for AS in comparison to SMS suggests that amplification is a significant contributory factor in the observed bias. Similar behavior is observed in the *RPL31* and *SPINT2* genes as well.

We considered both alignment locus and read length in our definition of read duplication, allowing one read at each locus with a unique read length. Looking across the transcriptome using this definition of read duplication, we observed a roughly normal distribution along the length of all transcripts captured. A 3-fold difference in the median number of duplicate reads between AS and SMS across all transcripts observed in all samples was maintained across the majority of the transcript length (

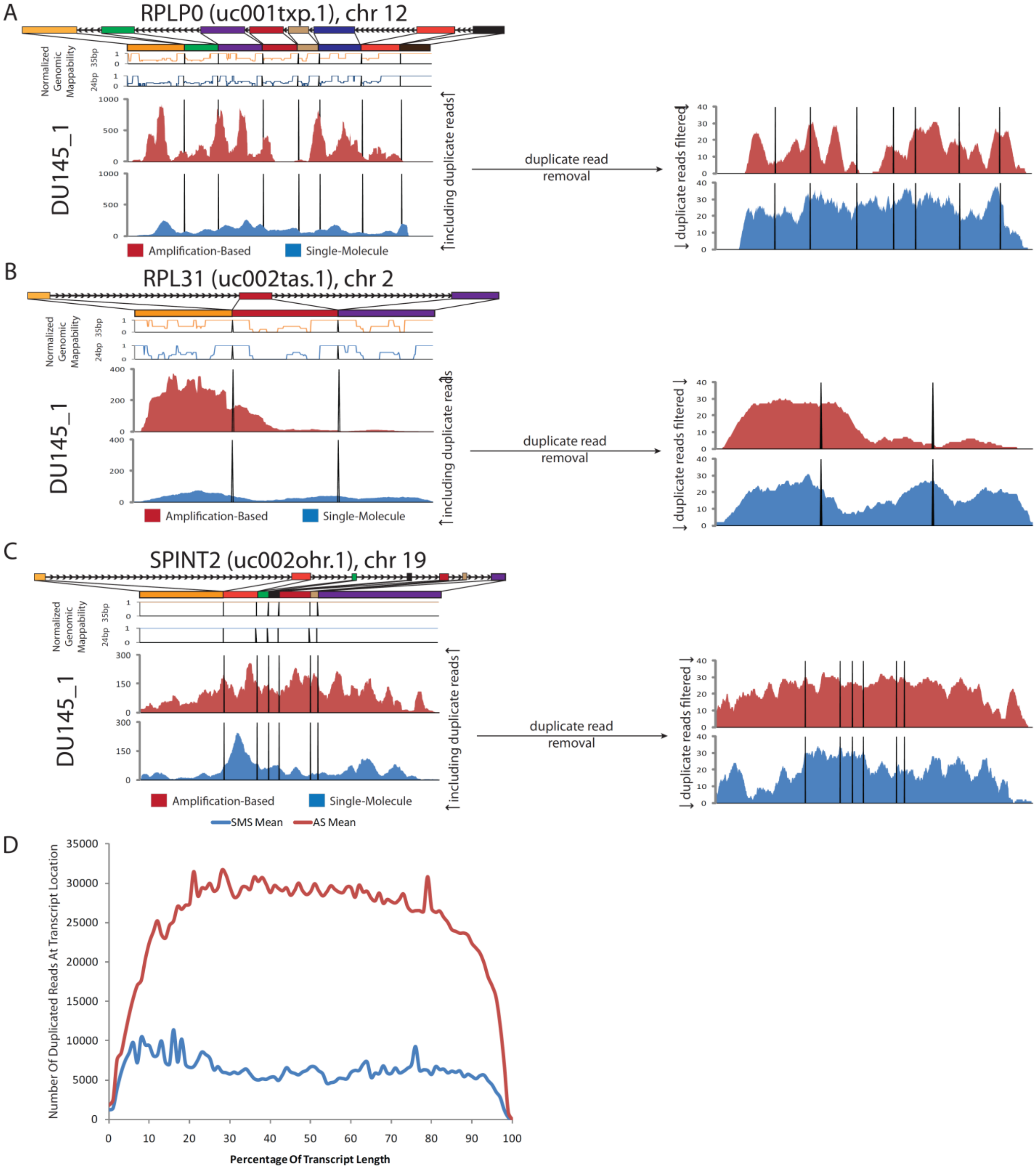


Figure 4**D**). This pattern of read duplication is similar to that observed in the literature between standard amplification-dependent and amplification-free sequencing methodologies [98]. Removal of duplicate reads, allowing only one read per locus, yielded inconsistent results across the sample set (**Figure S 13**). In some cases, the procedure reduced the over-representation in the highest expressing genes, however the bias appeared to remain in other samples. The procedure also drastically reduced the number of total usable reads in each sample by a median of 47% across the 12 sample dataset (**Figure S 14**). While this naïve methodology of duplicate read removal had some positive effect in reducing the discrepancies between AS and SMS in terms of transcript quantification, the drastic effects it has on the number of usable reads in AS suggests a different approach may be desirable. With this understanding of the impact of duplicated reads, we analyzed the set of recurrently over-represented genes to see if they sequenced biologically interesting categories of genes.

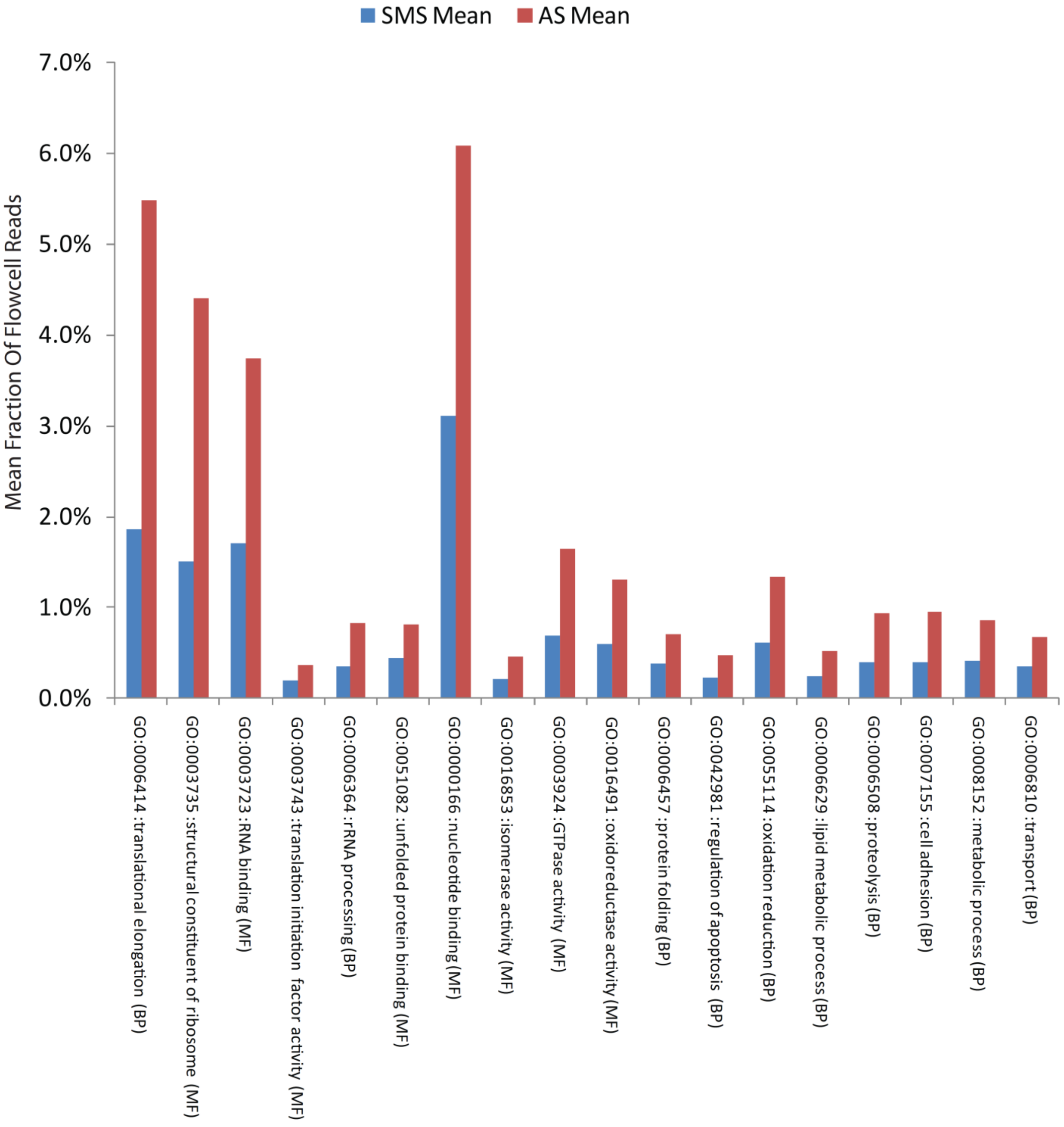


Figure 5: Global representation of Gene Ontology classes in Amplification-based sequencing. GO analysis of the 392 most over-represented genes found using our recurrence analysis in the Molecular Function (MF) and Biological Process (BP) subtrees demonstrates that translational processes and components of the ribosome are over-represented across samples in amplification-based sequencing.

Gene Ontology analysis of the set of 393 recurrently over-expressed genes

Across the samples, genes associated with the cell's replicative machinery comprised the largest portion of over-represented transcripts by total normalized number of mapping reads in most samples. Gene Ontology analysis of the set of 393 consistently over-represented genes shows that they are components of the cell's translational machinery (**Figure 5**), a class generally found at high levels in all twelve samples used in this evaluation. This again suggests that the amplification procedure implicit in AS library preparation exaggerates a particular bias towards these already-abundant transcripts. The total number of reads falling into each of the classes observed to be over-represented in AS was a mean of 2.23x higher as compared to SMS, although genes overlap between the classes. With less of a focus on high-concentration translational machinery and housekeeping genes, we then attempted to apply SMS in finding gene fusions in the transcriptome.

Re-discovery of known gene fusions using single-molecule sequencing

We evaluated the applicability of single read SMS in gene fusion discovery by attempting to re-discover known gene fusions in the VCaP cell line, known to harbor *TMPRSS2-ERG*, in a *de novo* process. As shown in **Figure S 15**, we first aligned all possible reads against the transcriptome and genome using IndexDP. The non-mapping reads, which harbor chimeras, were subsequently aligned against the transcriptome returning those reads that had a partial alignment of at least 18 nucleotides. The portion of the read that fails to align is defined as the overhang. All reads having the same partial alignments, suggesting a common breakpoint, were clustered. All clusters were then compared to determine if the overhang from one breakpoint region had similarity to the overhang of an independent breakpoint thereby reconstructing the fusion junction. Lastly, all remaining non-mapping reads were aligned against the novel fusion junctions.

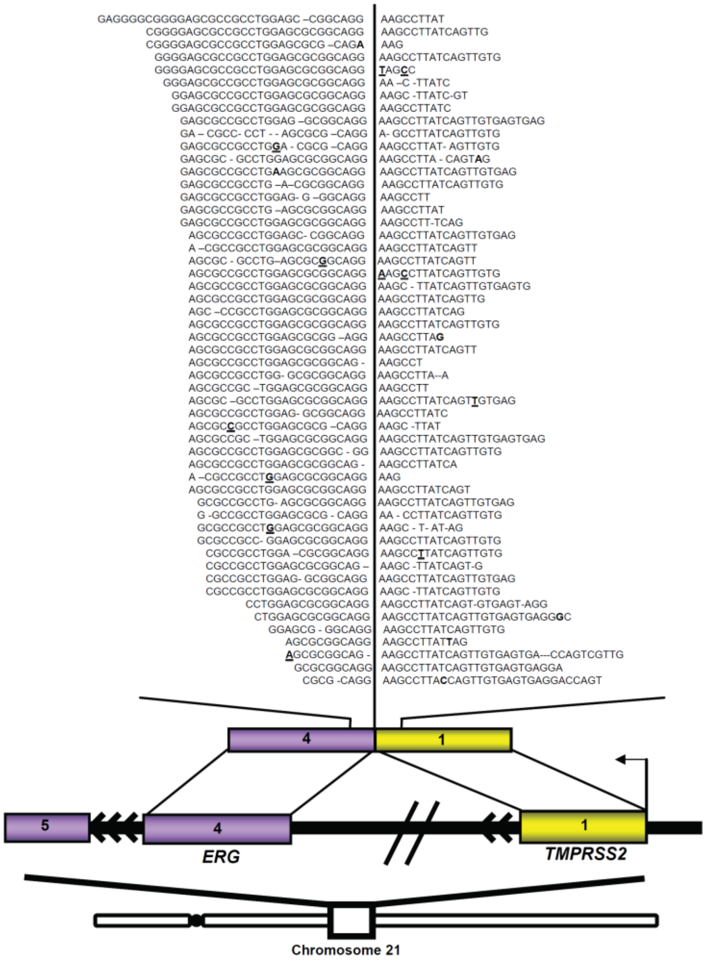

GAGGGGCGGGGAGCGCCGCCTGGAGC –CGGCAGG AAGCCTTAT
CGGGGAGCGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTTG
CGGGGAGCGCCGCCTGGAGCGCG –CAGA AAG
GGGGAGCGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTTGTG
GGGGAGCGCCGCCTGGAGCGCGGCAGG TAGCC
GGGAGCGCCGCCTGGAGCGCGGCAGG AA –C -TTATC
GGAGCGCCGCCTGGAGCGCGGCAGG AAGC -TTATC-GT
GGAGCGCCGCCTGGAGCGCGGCAGG AAGCCTTATC
GAGCGCCGCCTGGAG –GCGGCAGG AAGCCTTATCAGTTGTGAGTGAG
GA – CGCC- CCT - - AGCGCG –CAGG A- GCCTTATCAGTTGTG
GAGCGCCGCCTGGA - CGCG –CAGG AAGCCTTAT- AGTTGTG
GAGCGC - GCCTGGAGCGCGGCAGG AAGCCTTA - CAGTAG
GAGCGCCGCCTGAAGCGCGGCAGG AAGCCTTATCAGTTGTGAG
GAGCGCCGCCTG –A–CGCGGCAGG AAGCCTTATCAGTTGTG
GAGCGCCGCCTGGAG- G –GGCAGG AAGCCTT
GAGCGCCGCCTG –AGCGCGGCAGG AAGCCTTAT
GAGCGCCGCCTGGAGCGCGGCAGG AAGCCTT-TCAG
AGCGCCGCCTGGAGC- CGGCAGG AAGCCTTATCAGTTGTGAG
A –CGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTT
AGCGC - GCCTG–AGCGCGGCAGG AAGCCTTATCAGTT
AGCGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTTGTG
AGCGCCGCCTGGAGCGCGGCAGG AAGC - TTATCAGTTGTGAGTG
AGCGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTTG
AGC –CCGCCTGGAGCGCGGCAGG AAGCCTTATCAG
AGCGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTTGTG
AGCGCCGCCTGGAGCGCGG –AGG AAGCCTTAG
AGCGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTT
AGCGCCGCCTGGAGCGCGGCAG - AAGCCT
AGCGCCGCCTGG- GCGCGGCAGG AAGCCTTA--A
AGCGCCGC –TGGAGCGCGGCAGG AAGCCTT
AGCGC –GCCTGGAGCGCGGCAGG AAGCCTTATCAGTTGTGAG
AGCGCCGCCTGGAG- GCGGCAGG AAGCCTTATC
AGCGCCGCCTGGAGCGCG –CAGG AAGC -TTAT
AGCGCCGC –TGGAGCGCGGCAGG AAGCCTTATCAGTTGTGAGTGAG
AGCGCCGCCTGGAGCGCGGC - GG AAGCCTTATCAGTTGTG
AGCGCCGCCTGGAGCGCGGCAG - AAGCCTTATCA
A –CGCCGCCTGGAGCGCGGCAGG AAG
AGCGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGT
GCGCCGCCTG- AGCGCGGCAGG AAGCCTTATCAGTTGTGAG
G -GCCGCCTGGAGCGCG - CAGG AA - CCTTATCAGTTGTG
GCGCCGCCTGGAGCGCGGCAGG AAGC - T- AT-AG
GCGCCGCC- GGAGCGCGGCAGG AAGCCTTATCAGTTGTG
CGCCGCCTGGA –CGCGGCAGG AAGCCTTATCAGTTGTG
CGCCGCCTGGAGCGCGGCAG – AAGC -TTATCAGT-G
CGCCGCCTGGAG- GCGGCAGG AAGCCTTATCAGTTGTGAG
CGCCGCCTGGAGCGCGGCAGG AAGC -TTATCAGTTGTG
CCTGGAGCGCGGCAGG AAGCCTTATCAGT-GTGAGT-AGG
CTGGAGCGCGGCAGG AAGCCTTATCAGTTGTGAGTGAGGGC
GGAGCG - GGCAGG AAGCCTTATCAGTTGTG
AGCGCGGCAGG AAGCCTTATTAG
AGCGCGGCAG - AAGCCTTATCAGTTGTGAGTGA---CCAGTCGTTG
GCGCGGCAGG AAGCCTTATCAGTTGTGAGTGAGGA
CGCG -CAGG AAGCCTTACCAGTTGTGAGTGAGGACCAGT
4
1
5
4
ERG
1
TMPRSS2
Chromosome 21

Figure 6: Single molecule sequencing “re-discovers” known gene fusions. Schematic of the intra-chromosomal rearrangement on chromosome 21 fusing TMPRSS2 (yellow) to ERG (purple).

For this purpose, a sample of the VCaP cell line was sequenced more extensively in 2 channels, generating 31,198,128 reads aligned to the transcriptome or genome. The VCaP sample was prepared with one channel each with and without fragmentation. The benchmark fusion between prostate-specific gene *TMPRSS2* and ETS oncogenic family member, *ERG* [99], was found to be covered by 53 reads from generating 65 million reads in the VCaP cell line (**Figure 6**).

Discussion

This is the first study assessing the performance of RNA-Seq using single-molecule sequencing in comparison to existing amplification-based techniques. While the characteristics of the SMS reads will vary depending on platform, we expect that the distribution of reads across varying transcript concentrations to remain relatively consistent. The SMS technique was able to generate more usable reads in ten of the twelve samples considered in the RNA-Seq quantification and coverage evaluation, producing a mean 78% more reads in these 10 samples. More importantly, these reads tended to be less concentrated at the very highest abundance transcripts as shown in **Figure 3B**, where fraction of total reads mapping to the highest abundance transcripts in SMS are 4% below that of AS. Because the AS technique amasses a large fraction of reads sequencing high- abundance transcripts, detection of lower abundance genes are reduced. The large differences between the highest and second-highest quartile of expressed transcripts suggests that this effect is non-linear as transcript abundance increases in the sample. The wide range of transcript expression in biological samples makes this skewed read distribution of coverage an important factor when profiling mRNAs at the nucleotide level, departing from models that may assume a linear correlation between transcript abundance and sequencing coverage.

The number of duplicated reads observed in the samples across all transcripts was, not surprisingly, 3-fold higher in AS compared to SMS. The removal of duplicate reads is a well-defined procedure in experiments involving DNA sequencing but is less clear-cut when sequencing the transcriptome where varying transcript concentrations naturally lead to reads of identical mRNA segments. This caveat is due to highly expressed transcripts contributing false positive duplicate reads due to random sampling of read start locations along the transcript, where high coverage naturally leads to repeated sequencing of identical segments. However, highly expressed transcripts in SMS would likely generate a large number of these aberrant false positives as well. As a result, this source of false positive duplicated reads is unlikely to be the major factor behind the large observed differences in the number of duplicates between AS and SMS. The removal of duplicated reads by filtering out all reads in excess of a single read for a single locus appears to be an incomplete solution that introduces several confounding factors when using single reads. First, the process of removing duplicates is inconsistent, affecting the biased representation of reads in only a subset of the 12 samples in the dataset. Second, the duplicate removal process also reduced the usable sequence yield from each experimental run by nearly half, although this is an overestimation due to the naïve nature of the method. Finally, these duplicate removal methods impose a peak coverage limit for each transcript that is equivalent to the read length. The naïve process we applied for the elimination of duplicates is most certainly over-aggressive and the use of paired-end reads may be more effective, due to the production of additional mapping and sequence information that improves the process of duplicate identification and removal. However, the differences that result from the characteristics of these two methodologies can lead to disparities in the sequence coverage of genes along the spectrum of expression.

Small differences in the distribution of reads at the highest quartile of expressed genes have a large effect on the coverage of the remaining expressed genes. For example, the lowest quartile of all genes seen in both technologies in the VCaP-24h sample composes 0.4% of the sum total of normalized reads seen in the highest expressed quartile by AS. A 1% reduction in the number of reads used to sequence the highest expressing genes in the forth quartile can be used to triple the coverage of the lowest expressing genes when reads are applied within the set. The result of shifting the read distribution to lower expressing genes is seen between the VCaP-0h and VCaP AS samples. Both samples yielded a relatively similar number of reads, with 3,636,454 and 3,352,960 reads in VCaP-0h and VCaP, respectively. However, the VCaP-0h sample has more than twice the fraction of the total reads falling into the lowest 2 quartiles with 2.2% and 0.9%, in the respective VCaP-0h and VCaP samples. It comes as no surprise that in the VCaP-0h sample, we are able to observe 16,813 genes above the 0.3 RPKM noise threshold whereas in VCaP, we only observe 13,866 genes above this threshold. Similarly, the reduced high-abundance coverage bias across variable concentrations allows the SMS approach 2- to 6-fold more coverage in the lower half of all expressed genes. The variable read length of the SMS reads contributes to quantification noise, compared to AS, due to the number of short reads which map ambiguously. These mis-mappings may contribute to the larger number of genes observed at the very lowest expression levels. Examination of the reads mapping to genes only found in SMS shows the presence of more than 30% of long SMS reads (>36bp in length) in a median of 17% of the genes (approximating the read length distribution across all samples), leaving a 1.7-fold advantage in favor of SMS sensitivity if genes detected with only short 24- to 35-mer reads are all considered detections due to noise. While a significant proportion of this noise is directly attributable to ambiguities in accurately mapping short reads, the presence of long (>36bp) aligned reads is not a guarantee of transcript presence. In a large number of the cases where detected genes have long reads aligned to them, false positives were attributable to these long reads mapping to repetitive elements or low complexity regions within the transcripts.

Our PCR validation results suggest that using amplification to confirm transcripts exclusively detected by single-molecule sequencing (and missed by AS sequencing) is not ideal, since any sequence that is difficult to amplify will be hard to detect using AS RNA-Seq and hard to validate using an amplification-based system. Therefore, we cannot verify such transcripts unless an amplification-free technology is employed. Sample preparation differences may also contribute to differential representation of transcripts in the sequencing libraries, as AS involves a size selection step that SMS does not. In addition, the two protocols use differing fragmentation procedures which may affect the prevalence of detectable transcript fragments. This is one significant factor that may contribute to the detection of some genes above the noise threshold exclusively by AS. There may be other reasons for differences in the relative representation of transcripts in each technology. Some transcripts may be under-represented because they are hard to capture using SMS. Conversely, the amplification procedure may alter the apparent transcript abundance as some sequences may amplify highly leading to over-representation in AS, which may increase their candidate transcript counts above the noise threshold. For some candidates seen in only one technology, increasing sequencing depth may be the most straightforward solution to the lack of resolution for low abundance transcripts. Some candidates may require modification of the library preparation protocol to ensure sufficient library complexity to capture these low-abundance transcripts. For example, the use of a normalized AS RNA-Seq library preparation protocol or the introduction of a greater amount of input RNA may increase the complexity of the library, possibly enabling higher sensitivity as a result. However, the paucity of published data addressing these topics at this time precludes a thorough examination of potential solutions.

However, while SMS confers the advantages of higher sensitivity and abrogation of issues stemming from read duplication, the technology has a number of confounding characteristics. First, SMS produces reads that are, on average, shorter than their AS counterparts, magnifying the issue of accurately mapping reads to their correct positions. While the inclusion of long 64bp reads confers an advantage, these are the minority of all reads produced. Approximately 60% of all SMS reads were 36bp or smaller across all samples. Second, the SMS methodology used in this evaluation produces reads that include randomly introduced gaps due to the incorporation of "dark bases" which do not produce photo-detectable fluorescence. This characteristic requires the use of alignment algorithms that allow for the inclusion of insertions and deletions relative to the reference, and may complicate the detection of structural variation. We also observed a higher proportion of contaminant-alignable reads in SMS compared to AS, although it is unclear whether this is a product of either the sample preparation procedure or a characteristic of the sequencing process.

Altogether, these differences suggest that SMS has advantages in quantitative expression profiling and nucleotide-level assessment such as polymorphism detection in mid- to low- abundance transcripts although the lowest levels of detection are subject to noise due to mapping. However, the log-fold advantage SMS holds may be overcome as rapid advances in sequencing technology result in the production of increasing numbers of usable reads.

Methods

Preparation and sequencing of samples

Sequencing libraries for the RNA-Seq evaluation set were prepared from a DU145 cell line (ATCC; HTB-81), an RWPE cell line (ATCC; CRL-11609), an androgen-induced VCaP cell line time course at 0h, 24h, 48h, an identical time course in the LnCaP (ATCC; CRL-1740) cell line, and a tissue sample from a prostate tumor paired with an adjacent normal sample. Sample preparation of the entire 12-sample set included the RNA fragmentation step to ensure consistency. Two replicates of a normal untreated VCaP cell line were run for gene fusion discovery evaluation, one each of fragmented and un-fragmented RNA. The fragmented sample was included in the 12-sample evaluation set. The VCaP cell line was derived from a vertebral metastasis from a patient with hormone-refractory metastatic prostate cancer, and was provided by Ken Pienta (University of Michigan, Ann Arbor, MI). LNCaP or VCaP [100] cells were starved in phenol red free media supplemented with charcoal-dextran filtered FBS and 5% penicillin/streptomycin for 48 h before the addition of 1 nM synthetic androgen (R1881) as indicated. RNA was then isolated using the miRNeasy kit (Qiagen) according to the manufacturer’s instructions. Prostate tumor tissue was obtained from the University of Michigan tissue core. Identical samples were submitted for SMS and AS sequencing in all cases with the exception of the VCaP and LnCaP time course samples. The DU145, VCaP, RWPE, as well as the VCaP and LNCaP AS-sequenced time course samples were treated with DNAse. The VCaP and LNCaP time course samples submitted for SMS, as well as the PrCa and PrCa-Adjacent normal samples, were not treated with DNAse during sample preparation. Poly-A containing mRNA for these samples was isolated by two rounds of binding to Sera-Mag Magnetic Oligo(dT) beads, wash and elution in 10mM Tris buffer pH 7.5, according to manufacturer's instructions (Thermo Scientific, Indianapolis). The purified mRNA was immediately processed for library preparation. The VCaP and LNCaP time course AS sample mRNA was selected with oligodT linked beads according to manufacturer's instructions (Invitrogen).

Amplification-based sequencing was done in paired-end mode run to a minimum of 36bp per read and trimmed to a minimum of 34bp to remove low quality bases. For amplification-based sequencing, messenger RNA (2 µg) was fragmented at 85° C for 5 min in a fragmentation buffer (Ambion) and converted to single stranded cDNA using SuperScript II reverse transcriptase (Invitrogen), followed by second-strand cDNA synthesis using *Escherichia coli* DNA polymerase I (Invitrogen). The double stranded cDNA was further processed by Illumina mRNA sequencing Prep kit. Briefly, double-stranded cDNA was end repaired by using T4 DNA polymerase and T4 polynucleotide kinase, monoadenylated using an exo minus Klenow DNA polymerase I (3’to 5’ exonucleotide activity), and ligated with adaptor oligo mix (Illumina) using T4 DNA ligase. The adaptor-ligated cDNA library was then fractioned on a 3% agarose gel, and fragments corresponding to 280-320 bp were excised, purified, and PCR amplified (15 cycles) by Phusion polymerase (NEB). The PCR product was again size selected on a 3% agarose gel by cutting out the fragments in the 300 bp range. The library was then purified with the Qiaquick Minelute PCR Purification Kit (Qiagen) and quantified with the Agilent DNA 1000 kit on the Agilent 2100 Bioanalyzer following the manufacturer's instructions. Library (5-8 pM) was used to prepare flowcells for analysis on the Illumina Genome Analyzer II.

Single-molecule sequencing was done on a Helicos HeliScope in single-read mode, resulting in useful reads ranging between 24bp and 61bp for the first set and 25bp and 64bp in length in the second set. polyA+ RNA was purified on an RNeasy MinElute column (Qiagen). Then 100ng of RNA (on average, between 86ng – 130ng) was heat fragmented by incubation at 95C for 10 minutes or left un-fragmented. First strand cDNA was then made using the SuperScript III reagent kit (Invitrogen, Carlsbad CA) as follows: 500ng random hexamers, 2ul of 10mM dNTP, and DEPC water were added to the RNA up to a volume of 25ul. The mixture was then incubated at 65C for 5 min and placed directly on ice for 2 minutes. Next, 5ul 10X buffer, 5ul 0.1M DTT, and 10ul 25mM MgCl were added to each sample, and the, now 45ul, sample was incubated at 15C for 30 minutes. After this incubation time 2.5ul of RNaseOut (100U), and 2.5ul of SuperScript III (500U) were added to each sample and the samples were incubated at 42C for 30 minutes, 55C for 50 minutes, and 85C for 5 minutes. After the reverse transcription reaction, 1ul RNase H and 1ul of RNase I were added to each sample, followed by a 30 minute incubation at 37C.

Samples were twice purified on DyeEx columns (Qiagen). cDNA samples were then Poly-A tailed using the Helicos DGE assay reagent kit (Helicos, Cambridge MA), and the terminal transferase kit (NEB, Ipswich MA) as follows: 5ul Helicos Tailing control Oligonucleotide A was added to 20ul of each cDNA and the volume was adjusted to 35.5ul with water. This mixture was then denatured for 5 minutes at 95C and placed directly on ice for 2 minutes. Then, 5ul 2.5mM CoCl, 5ul 10X terminal transferase buffer, 2ul Helicos polyA tailing dATP, and 1.2ul terminal transferase (24U) were added to each samples, followed by incubation at 42C for 1hour, and then 70C for 10 minutes. After the tailing reaction the samples were 3' blocked as follows: samples were denatured for 5 minutes at 95C and placed directly on ice for 2 minutes, 300 pmoles biotin-dideoxy ATP (Perkin Elmer, Waltham MA) and 1.2ul terminal transferase

(24U) were then added, followed by 1 hour incubation at 37C, and a final 10 minute heat inactivation step at 70C. 3' biotinylation of samples was used to assess sample molarity to inform HeliScope sample-loading for the sequencing reaction (according to manufacturer's instructions).

Alignment of reads

The first read of AS read pairs was used in this study to compare to the single reads derived from SMS. SMS reads were aligned with the IndexDP aligner, while amplification-based sequencing reads were aligned with both the Bowtie and IndexDP aligners as shown in **Figure S 4**. IndexDP alignments were filtered by NScore, defined as (5*#_match-4*#_error)/read_length) with a minimum of score 4, reporting at most 25 alignments per read. Reads between 24bp and 57bp and 25bp and 64bp in length were used for sets 1 and 2, respectively. Bowtie was set to report alignments with at most two mismatches within a 32-base seed region, reporting at most 25 multiple alignments per read. The first base of all AS reads was trimmed to maximize quality. Single-best quality alignments were derived using Bowtie by setting the –best and –k 1 parameters to report only the single highest quality alignment per read. Reads were aligned to the set of UCSC transcripts defined in hg18, downloaded from the UCSC Genome Browser at http://genome.ucsc.edu. Known contaminants were also included in the set of references. Bowtie alignments included references for mitochondrial DNA, adapter sequence, and ribosomal RNA. IndexDP alignments included references for poly-A, poly-T, poly-C, and poly-G oligomers. Re-alignment of AS reads using IndexDP was done using the same parameters as SMS reads, using the full length of the read. Reads from the PrCa sample were trimmed to 50bp from 75bp to meet technical limitations of the alignment program. Sequence reads from this study have been deposited into the NCBI Short Read Archive with accession number SRA028835.1.

Duplicate read removal

Duplicate reads were removed from the data by analyzing the alignments to each UCSC transcript in the transcriptome reference. One read was allowed to align at each start locus (with and without consideration of read length). Reads with alignments to locations along the reference transcript in excess to those were marked as duplicates and removed from the data set.

Relative quantification of genes and coverage calculation

Reads aligning to each UCSC transcript were counted at transcript level resolution and then summarized at the gene level using transcript to gene symbol mappings from the kgXref table downloaded from the UCSC Genome Browser at http://genome.ucsc.edu. Reads aligning to the known contaminant references were marked and not considered in the analysis. Genes were quantified using only the single-best mapping methodology. Single-best mappings were derived from IndexDP alignments by choosing alignments with the highest NScore, or an alignment randomly picked from the set of highest scores when multiple alignments are present with the same NScore value. Gene-level RPM values were derived by summing the number of aligned reads from each gene’s constituent transcript isoforms and dividing by the total number of usable reads. Read sums were calculated using R Statistical Environment [101]. RPKM values were computed for each observed UCSC transcript and summed for all isoforms of a gene to derive a gene-level RPKM expression value. Coverage levels were calculated by summing the read lengths of all reads aligning to all isoforms of each gene and dividing by the mean isoform length.

Detection of genes observed in a single technology

We derived a list of genes observed in only SMS or AS for the DU145 samples in this study by comparing the mean gene-level RPKM expression values of each pair of samples run on AS and SMS. A list of candidates was nominated by then sorting the list of genes with expression values above the noise threshold in SMS and below the threshold in AS by the observed differences. These genes were evaluated for mis-mappings by examining secondary and alternate alignments of the reads aligning to each candidate as shown in **Figure S 16**. The list was filtered to remove genes detected only by short reads and the top 50 remaining genes manually evaluated to have well-defined HUGO names, diffuse read distribution along the transcript length, and the presence of long (>36bp) reads in both SMS technical replicates.

Validation of Detected Single-Technology Transcripts by PCR

RNA was extracted from the cells using Qiazol based on Qiagen's miRNeasy Minikit following the manufacturer's instructions (Qiagen). 1 µg of total RNA was reverse transcribed into cDNA using SuperScript III (Invitrogen) in the presence of oligo dT and random primers. Quantitative PCR was carried out by Taqman assay method using gene specific primers and probes from the Universal Probe Library (UPL), Human (Roche) as the internal oligonucleotide, according to manufacturer's instructions. GAPDH was used as housekeeping control gene for UPL based Taqman assay (Roche), as per manufacturer's instructions.

All assays were performed in duplicate using the primer sequences in **Table S 8**.

Gene Ontology analysis of reads

Gene Ontology (GO) analysis of over-represented genes was done in order to assess the most highly represented GO classes and determine the relative abundance of reads attributable to each GO class. This analysis was done with GeneCoDis2 tool [102]. Single GO classes resulting from this process were evaluated for their representation in terms of fraction of total sequenced reads across the 12-sample set. Relative representation of reads attributable to each GO class was done by summing the number of single-best mapping alignments for each gene in each GO class as defined in the GO annotations for *Homo Sapiens*, downloaded from http://www.geneontology.org and dividing the total by the total number of reads in each sample.

Gene fusion discovery in single-molecule sequencing

The VCaP cell line was sequenced in two additional channels to evaluate the suitability of single molecule sequencing for the task of gene fusion detection. This was done by mining the reads in an effort to re-discover known gene fusions. All possible reads were first aligned against the transcriptome and genome using IndexDP. Non-mapping reads, which harbor chimeras, were subsequently aligned against the transcriptome returning those reads that had a partial alignment of at least 18 nucleotides. All reads having the same partial alignments, suggesting a common breakpoint, were clustered. All clusters were then compared to see determine if the overhang (portion of the read that fails to align) from one breakpoint region had similarity to the overhang of an independent breakpoint, thereby reconstructing the fusion junction. Finally, all remaining non-mapping reads were aligned against the novel fusion junctions. This de novo approach enabled the re-discovery of the TMPRSS2-ERG gene fusion across two channels of SMS reads.

Chapter 3: A Framework for integrating transcriptome and proteome data

Introduction

mRNA transcript levels are often used in research studies as a rapid and simple measure of biological activity in cells, a proxy for protein abundance and activity. However, this relationship is complex – complicated by numerous external factors such as RNA translation rates and decay, degradation through the microRNA pathway, non-coding RNAs, and numerous post-translational modifications of protein products [103, 104].

In the study described in the next two chapters, we use the VCaP and RWPE human prostate cell lines to study the transcript-protein relationship and extended the analysis to examine how that relationship is dysregulated in a cancer context. VCaP is derived from a vertebral metastatic lesion of a patient with castrate-resistant prostate cancer and serves as a model of prostate carcinoma; it expresses a large quantity of Prostate Specific Androgen (PSA) and Androgen Receptor (AR) and is known to be androgen-responsive [100]. RWPE serves as a model of normal prostate epithelium; it is derived from non-neoplastic prostatic epithelial cells, and is known to possess the characteristics of normal tissue [105].

In this chapter, we focus on addressing the challenges of quantification and integration of data from transcriptomic and proteomic experiments carried out using mRNA sequencing ("RNA-Seq") and tandem mass spectrometry ("MS/MS"), respectively, and describe a novel methodology using a common sequence reference database with which we quantify relative abundance of transcript and protein in the VCaP and RWPE cell lines and analyze their relationship. With this methodology, we demonstrate how differing short read alignment and spectral counting methods and filtering processes impact the measurement of the transcript-protein relationship.

Background

The correlation of protein and transcript levels is confounded by varying methodologies for identification, quantification, and data integration. The process of integrating this data effectively is itself a topic of study [42, 47, 48, 106]. Early work studying this relationship relied on gel electrophoresis or liquid chromatography coupled with mass spectrometry (LC-MS) and microarrays to quantify transcript and protein abundances. Due to technical constraints, these studies were limited by the dynamic range of assay methodology and a small sample set of assay genes. The small, pre-selected gene sample sets typically found in these studies resulted in highly variable correlation measurements.

Recent research examining the relationship between transcript and protein abundances has leveraged advances in next-generation sequencing to profile the transcriptome and higher throughput methods for proteome assessment to observe a more complete landscape of the cellular transcriptome and proteome [40, 107]. These studies have observed correlation between mRNA and protein abundance ranging from r = 0.3 to r = 0.6, examining 12,000-16,000 mRNAs and 7,000-9,000 proteins in each sample. Previous studies focusing on smaller subsets of genes and other methods have shown more varied correlation values [43, 44, 108, 109]. A study combining previously published isotope-labeled protein abundance values with separate RNA sequencing data in three cancer cell lines, A431, U251MG, and U2OS found transcript-protein relationships in nearly 5,500 genes correlated at levels from r = 0.55-0.61., G-protein coupled receptors demonstrated the most disagreement in a focused examination of U2OS, a characteristic which they attributed to their limited ability to assay the protein products of this class of genes due to detection limitations [46, 110]. A similar study focusing on deeply profiling both the transcriptome and proteome in the HeLa cell line assembled an integrated dataset of approximately 8,600 genes, from which a correlation of r = 0.6 between transcript and protein levels was observed [40]. From this data, the authors estimated that a complete proteome comprised 10-12,000 genes in total. Most recently, a study profiled the proteome of the NCI-60 set of cancer cell lines and compared these profiles to microarray-derived mRNA abundance levels finding similar correlation levels [111]. The inclusion of all 59 NCI60 cell lines resulted in the largest dataset, comprising 10,350 genes, with an observed global correlation of r = 0.76. However, the number of proteins profiled in individual cell lines was much lower, with only 6,003 protein products seen in at least 5 samples, and the correlation between transcript and protein was noted to be lower in cancers with higher cellular heterogenity.

Results and Discussion

RNA-seq and Proteomics Results

To profile the transcriptome and proteome on a genome-wide scale, we use next-generation mRNA sequencing (RNA-seq) and label-free tandem mass spectrometry (MS/MS). Our study is based on three replicate profiles of mRNA and protein of the VCaP and RWPE cell lines that were independently processed before integration in a database. To do an “apples-to-apples” comparison of transcript and protein abundances, we assembled a common reference database derived from RefSeq containing 34,728 transcripts and matching protein sequences. This database of transcripts and corresponding peptide sequences was used to align RNA-seq reads and quantify peptides and proteins from MS/MS data (**Figure 7**). All data was collapsed to gene level granularity using a single representative transcript and protein isoform for each gene. This representative isoform was chosen as the highest abundance isoform observed in the proteome data across both cell lines, with transcript abundance used to break ties. Transcript data was summarized as the sum of all isoform read counts as a Reads Per Kilobase Million (**RPKM**) measure. Protein spectral counts were normalized by the length of these representative isoforms to produce a Normalized Spectral Count (**NSpC**) value for each gene.

The RNA sequencing of three technical replicates of the VCaP and RWPE cell lines yielded a total of 15,998,482 and 14,887,668 reads for each cell line, respectively. MS/MS of three replicates each of VCaP and RWPE yielded a total of 557,642 and 606,145 peptide matching spectra, respectively. Our RNA-seq quantification had high correlation between technical replicates with $r \approx 0.9$ in each case.

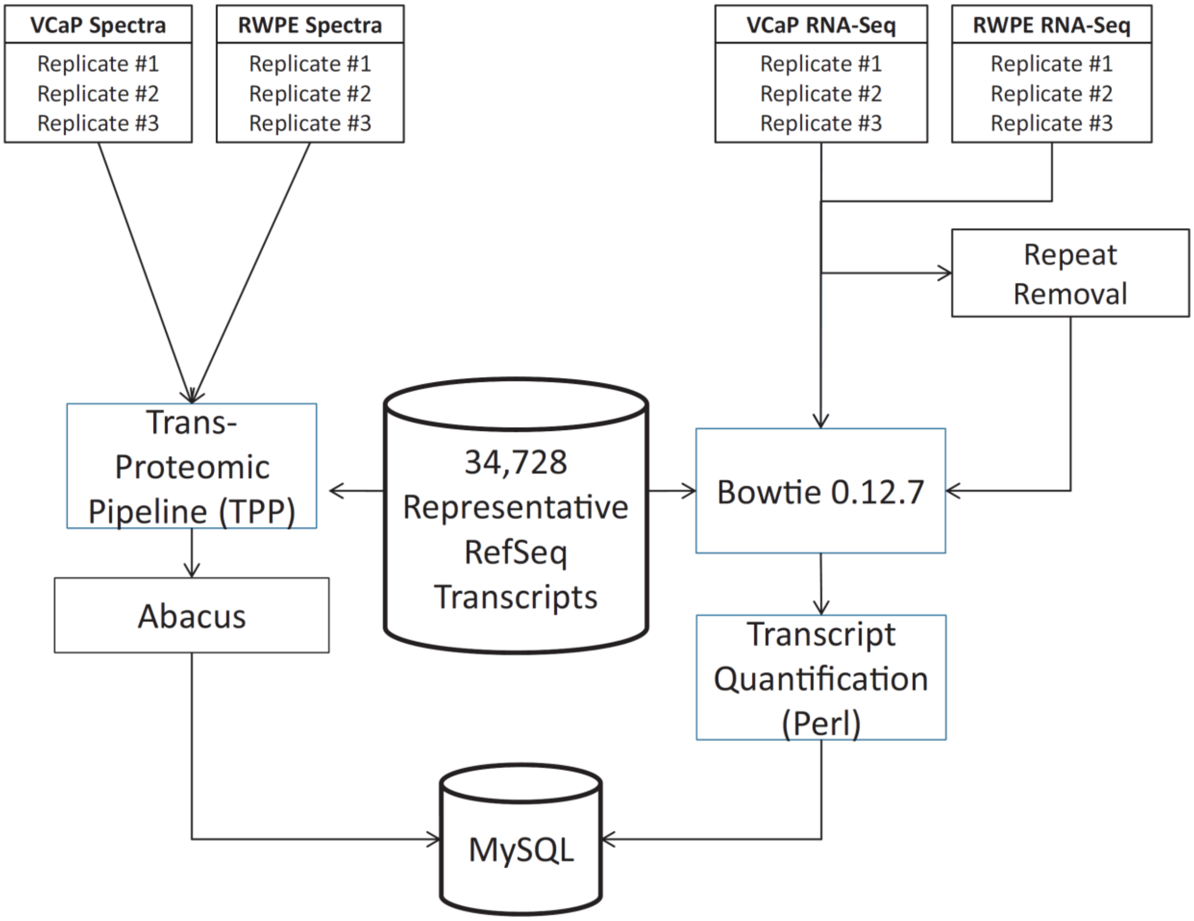


Figure 7: Data Processing and Integration Pipeline. Three replicates of each sample were generated using MS/MS and RNA-seq and quantified against a common reference library of mRNA and protein sequences. Tandem mass spectrometry data were processed with the TPP and post-processed with Abacus to yield spectral count data. RNA-seq reads were aligned to the common reference using Bowtie and post-processed with in-house Perl scripts to yield RPKM (reads per Kilobase million) quantification.

Similarly, pair-wise correlations of spectral count between replicates in our MS/MS data show correlations of *r* = 0.97 in both the VCaP and RWPE replicates (**Figure S 17**). Quantification of protein abundance was performed using the Trans-Proteomic Pipeline (TPP) [112] and Abacus [57]. Quantification of RNA-seq data was computed using in-house Perl scripts (**Figure 7**).

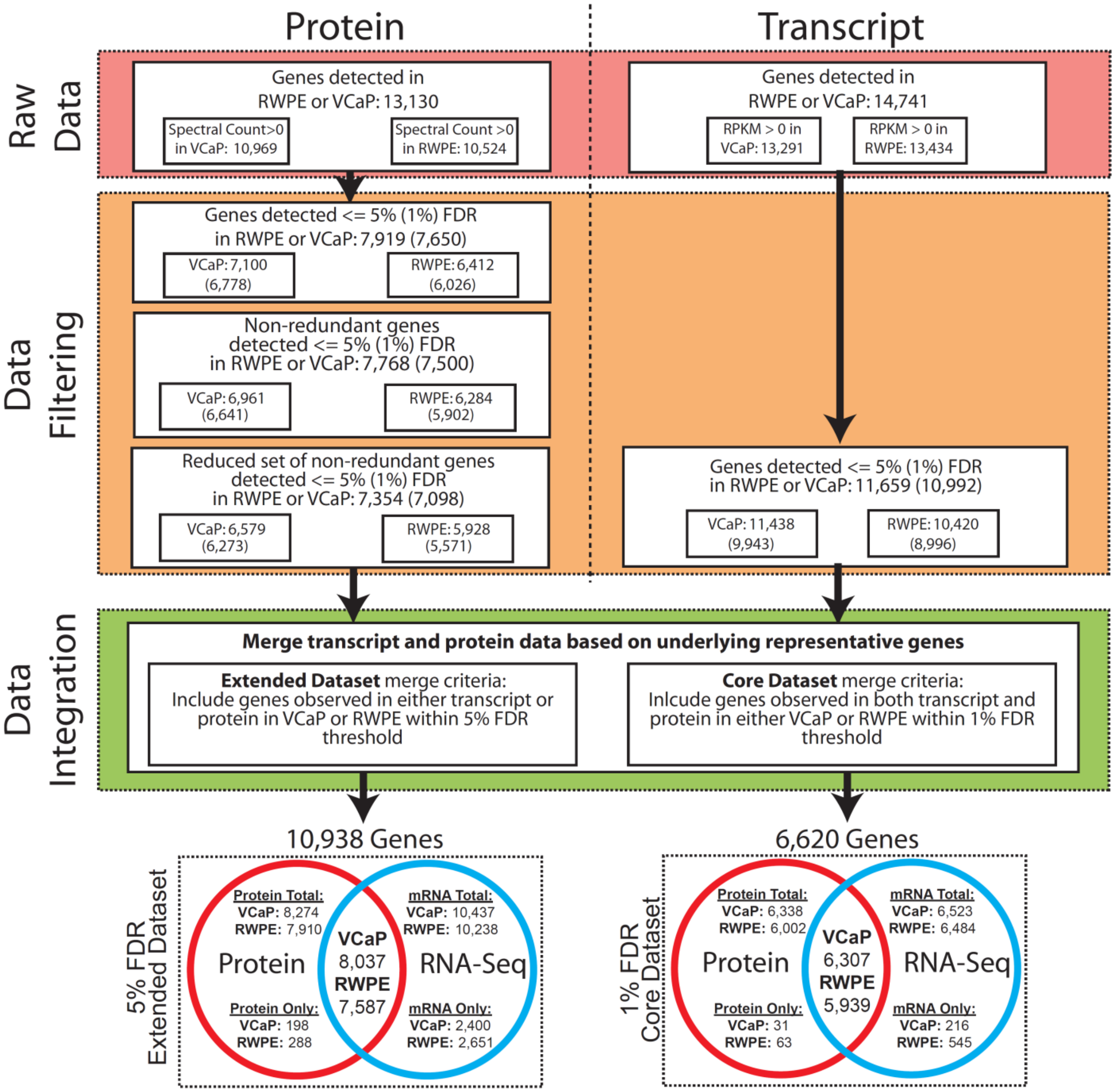
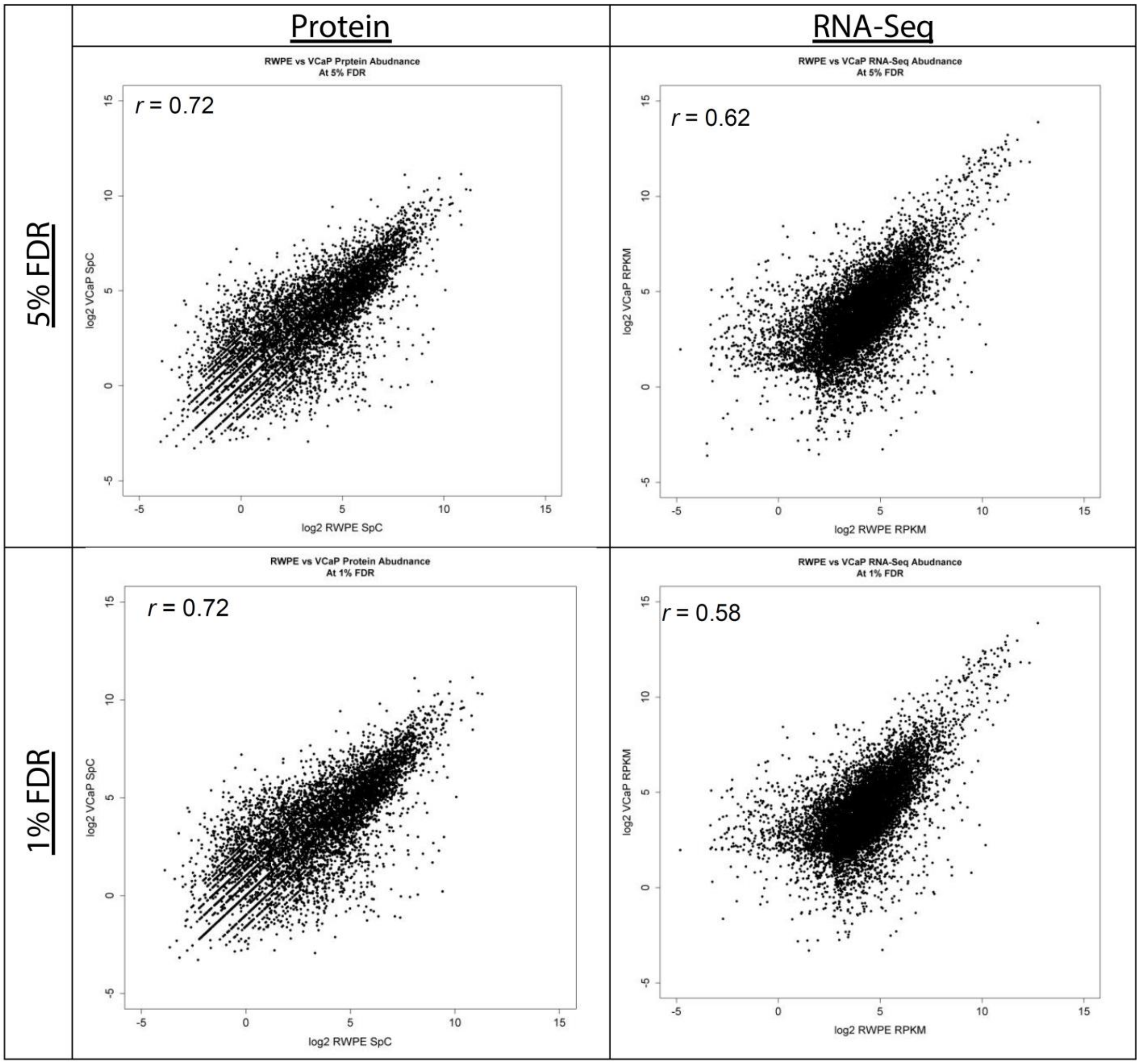

Protein
Transcript
Raw Data
Genes detected in RWPE or VCaP: 13,130
Spectral Count>0 in VCaP: 10,969
Spectral Count >0 in RWPE: 10,524
Genes detected in RWPE or VCaP: 14,741
RPKM > 0 in VCaP: 13,291
RPKM > 0 in RWPE: 13,434
Data Filtering
Genes detected <= 5% (1%) FDR in RWPE or VCaP: 7,919 (7,650)
VCaP: 7,100 (6,778)
RWPE: 6,412 (6,026)
Non-redundant genes detected <= 5% (1%) FDR in RWPE or VCaP: 7,768 (7,500)
VCaP: 6,961 (6,641)
RWPE: 6,284 (5,902)
Reduced set of non-redundant genes detected <= 5% (1%) FDR in RWPE or VCaP: 7,354 (7,098)
VCaP: 6,579 (6,273)
RWPE: 5,928 (5,571)
Genes detected <= 5% (1%) FDR in RWPE or VCaP: 11,659 (10,992)
VCaP: 11,438 (9,943)
RWPE: 10,420 (8,996)
Data Integration
Merge transcript and protein data based on underlying representative genes
Extended Dataset merge criteria: Include genes observed in either transcript or protein in VCaP or RWPE within 5% FDR threshold
Core Dataset merge criteria: Inlcude genes observed in both transcript and protein in either VCaP or RWPE within 1% FDR threshold
10,938 Genes
5% FDR Extended Dataset
Protein Total: VCaP: 8,274 RWPE: 7,910
mRNA Total: VCaP: 10,437 RWPE: 10,238
Protein
VCaP 8,037 RWPE 7,587
RNA-Seq
Protein Only: VCaP: 198 RWPE: 288
mRNA Only: VCaP: 2,400 RWPE: 2,651
6,620 Genes
1% FDR Core Dataset
Protein Total: VCaP: 6,338 RWPE: 6,002
mRNA Total: VCaP: 6,523 RWPE: 6,484
Protein
VCaP 6,307 RWPE 5,939
RNA-Seq
Protein Only: VCaP: 31 RWPE: 63
mRNA Only: VCaP: 216 RWPE: 545

Figure 8: The data filtering and integration statistics producing the core and extended datasets. Data is merged to the gene level before filtering by FDR and integrated using 1% and 5% FDR thresholds resulting in the core and extended datasets, respectively.

We used our False Discovery Rate (FDR) estimation procedure to threshold with which we filtered our data into two sets: a high-accuracy core dataset at 1% FDR to use for individual and set level analysis of genes, and a larger extended dataset at 5% FDR to use for correlation analysis (**Figure 8**). In our raw dataset, 13,130 and 14,741 genes were detected in either cell line at any level in the protein and transcript data, respectively. To achieve a 5% FDR in our extended dataset, we thresholded the minimum abundance to 1.8 and 3.6 RPKM for RNA-seq data and minimum peptide probabilities to 0.9475 and 0.958 for protein data in VCaP and RWPE, respectively. Due to the high correlation between RWPE and VCaP in terms of both mRNA (*r* = 0.79, Spearman) and protein (*r* = 0.70, Spearman) abundance and common tissue of origin (**Figure S 18**), we chose the extended dataset to include all genes which met the 5% FDR threshold in either cell line by either RNA-seq or MS/MS and was uniquely quantifiable by our proteomics approach. For our high-confidence 1% FDR core dataset, the abundance thresholds were set at 3.7 RPKM and 6.5 RPKM for RNA-seq data and minimum peptide probabilities of 0.9855 and 0.9885 for VCaP and RWPE, respectively (**Figure 9B and C**,**Figure S 19** and **Figure S**



20). To decrease noise in our fine-grained analysis, our core dataset also required detection of candidate genes at the 1% threshold in both the mRNA and protein data in either cell line. This process yielded a total of 10,938 unique genes in the 5% FDR extended dataset and 6,620 unique genes in our 1% FDR core dataset. Overall, the large majority of genes were filtered out by lack of protein data passing our filtering thresholds and detection criteria.

We used the extended dataset to examine the correlation between protein and transcript in both cell lines and observe how different alignment and counting methods for transcriptome and proteome data affect the relationship.

Calculating FDR in transcriptome and proteome data

To assess the false discovery rate (FDR) in our dataset, we followed a method similar to that of Ramskold, et al. [96] for the RNA-seq component of the study. Corresponding decoy intergenic sequences were sampled without replacement for each representative transcript in our database, for a total of 34,728 decoys. We aligned reads to the merged total set of these decoy and real mRNA transcripts. Abundance data was summarized at the gene level using the same transcript-gene mappings for both the real and decoy transcript set. FDR was calculated as the number of decoy genes detected divided by the number of non-decoy genes detected (**Figure S 19**). Across experiments in both cell lines, the decoy and real genes showed separated normal distributions, with decoys at a mean measured abundance of 0.46 RPKM and non-decoy genes at a mean abundance of 22.52 RPKM (**Figure 9A**). We did not find a bimodal distribution of transcriptome abundances as previously observed in other studies [113]. These studies have noted that a majority of the transcripts occupying the lower abundance peak are non-coding and small RNAs. Hence we do not observe this due to our inclusion of only protein-coding genes in our common reference database. Using this methodology, we went on to examine how technical and methodological factors could be optimized in order to achieve the most accurate correlation betweenrelative transcript and protein abundances.

We used the protein and peptide probability estimates provided by TPP and Abacus to control FDR in our protein data (**Figure S 20**). We used the peptide probability (independently set according to our FDR thresholds) in combination with the protein probability (held at 0.9) to filter out noise in our proteome datasets (**Figure 9B**). Additionally, we marked and removed all keratin genes in an effort to reduce the number of known common contaminants in our data.

Analyzing the impact of data processing methodology on correlation

We used our extended dataset to characterize our experimental output in an effort to avoid including experimental noise in our analysis. In general, our RNA-seq data were more sensitive to low abundance elements than our protein dataset (**Figure 9E**). In both cell lines, of the transcripts detected at 4-8 RPKM in our extended dataset, we detect approximately 60% of the protein products from their cognate transcript. This number rises to 90% of the protein products for transcripts with relatively high abundance of more than 16 RPKM. Although both RNA-seq and MS/MS methods produce similar data with a similar distribution, this observation is expected, as our ability to observe the transcriptome at depth surpasses our ability to observe its corresponding proteome. Some of the detection characteristics are explained by the differing dynamic range of the MS/MS and RNA-seq methods and their efficiency at assessing the protein and transcripts in our sample, in particular at low abundances.

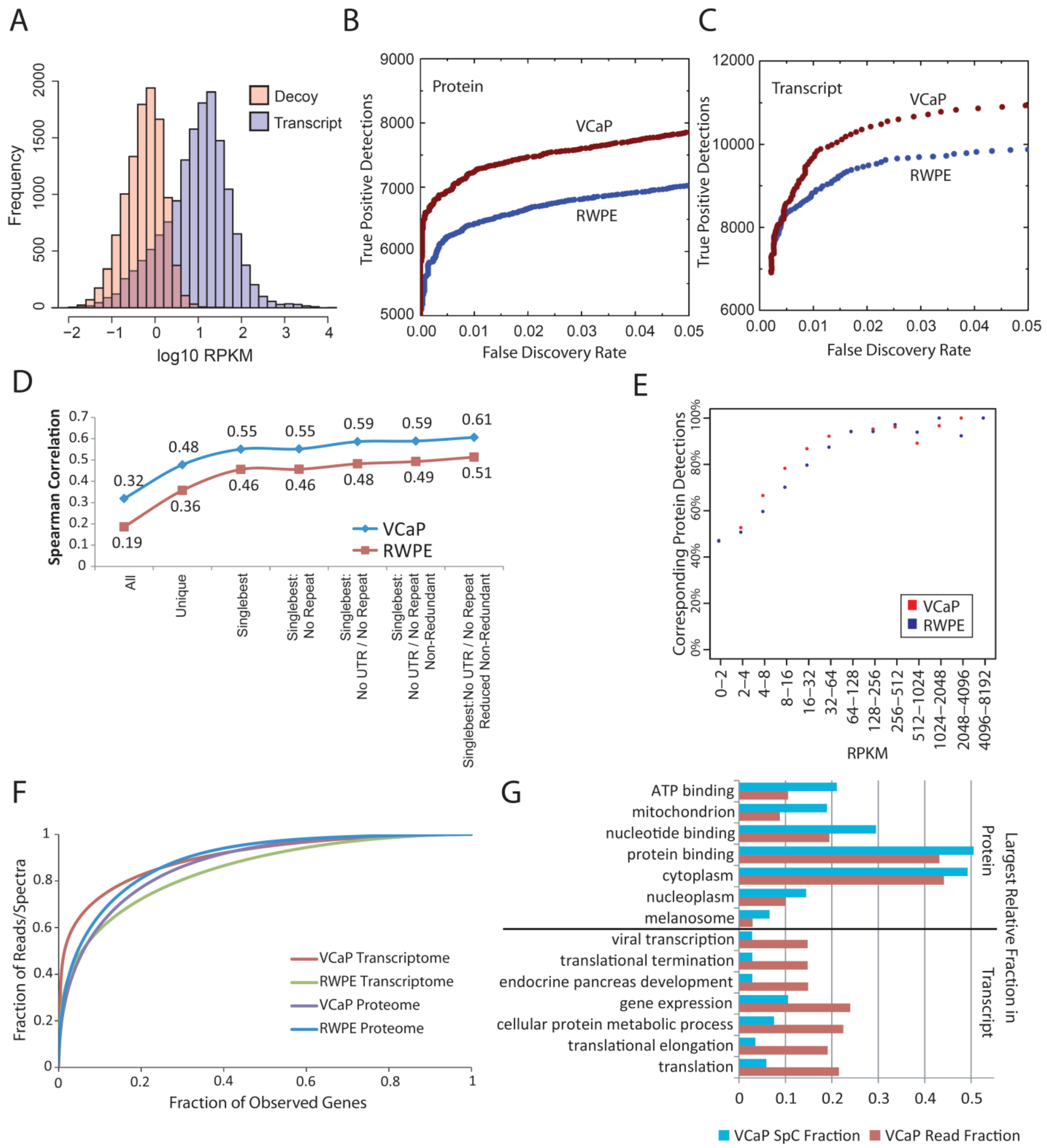


Figure 9: Analysis of transcript and protein datasets. A. The distribution of real and decoy gene values. The mean abundance of all decoy genes is 22.52 RPKM while decoys have a mean abundance of 0.46 RPKM. **B** and **C**. True positive detections across FDR values in protein and transcript data. **D.** Spearman correlation coefficient values for each alignment method and protein data reduction step in extended dataset. **E.** Protein detection at increasing transcript abundance levels. **F.** Distribution of RNA-seq reads and tandem MS spectra across all genes detected in the extended dataset for VCaP and RWPE. **G.** Most over-represented (as a fraction of all reads) Gene Ontology classes in transcript and protein data.

The impact of alignment and quantification methodology in RNA-seq

Called *multireads*, a fraction of all transcriptome reads map to multiple locations [9]. While most of these reads map to a small number of locations, a few have a muchgreater number of candidate mapping loci. We speculated that the presence of repetitive elements found in many transcripts may confound accurate quantification of transcript abundance from RNA-seq through the highly ambiguous alignment of this subset of reads. In an effort to reduce this effect, we removed reads coming from known repeat elements in the human genome. We removed all reads aligning to RepBase *H. Sapiens* and simple repeat elements and repeated the analysis for all genes. On average, this process removed a mean of 9.6% and 12.6% of all reads across replicates of VCaP and RWPE, respectively.

We also examined how our alignment and quantification methodology affected our correlation. Three transcriptome read alignment methods were evaluated to determine which best captured the relationship between transcript-protein with the hypothesis that concordance correlates with the performance of each methodology. Each of these alignment and counting methods resulted in a different number of reads assigned to each transcript **(Figure S 22).** The "unique" alignment policy is the most restrictive; where reads are required to map uniquely to a single position in the reference database. The "single best" alignment policy assigns reads to the best quality alignment among all found alignments as determined by the Bowtie aligner. The "all" alignment policy assigns reads to the best alignments up to 255 locations. In general, correlations between mRNA and protein were better in VCaP than in RWPE and are used for comparison. Correlations were computed from log2-tranformed values (**Figure 9D**). The Spearman correlation derived from the gene abundances determined using the unique method was the poorest at *r* = 0.32. While the "all" policy produced a reasonable correlation level of *r* = 0.48, the false discovery rate of the method was extremely high. The "single best" method we chose to use produced results yielding *r* = 0.55 and *r* = 0.46 for VCaP and RWPE, respectively. The high FDR of the "all" method led us to hypothesize that repetitive elements in the transcriptome confounded the abundance calculation by dispersing many non-unique reads onto many transcripts. Removal of these elements from the underlying transcriptome data yielded a negligible increase in correlation. Recent studies have suggested that UTR elements in the transcriptome also confound accurate quantification of mRNA transcript abundances [96]. With that consideration, we also evaluated the effect of removing UTR sequences and reads that align to these regions from our abundance calculations. Removal of UTR reads in addition to repeat elements led to a small increase in correlation in both VCaP and RWPE with increases to *r* = 0.59 and *r* = 0.48, respectively.

Processing of the RNA-seq data using the TopHat and Cufflinks suite of tools [55] yielded somewhat lower correlation, with Spearman *r* = 0.57 and *r* = 0.48 for VCaP and RWPE, respectively. Further analysis was carried out using our in-house tools as it allowed for more fine grained measurement and control over read mapping and counting.

Filtering out multiple assignment of spectra in proteomics data

Proteome data was filtered in an attempt to exclude quantification artifacts due to proteins that were indistinguishable based on the observed peptides, which leads to the assignment of spectra to both candidates. We used our process to eliminate the double counting of spectra inherent in this type of quantification. This was done by choosing the protein with the highest peptide probability between proteins which are otherwise indistinguishable. Ties were broken by choosing proteins with the highest mRNA abundance value. The same was done between indistinguishable isoforms of a given protein based on spectral count with ties broken by mRNA abundance. Removal of quantification artifacts between like proteins to create a non-redundant set increased the correlation to *r* = 0.59 and *r* = 0.49 in VCaP and RWPE, respectively. We then created a reduced non-redundant set through the removal of artifacts from isoform uncertainty, which further increased the Spearman correlation to *r* = 0.61 and *r* = 0.51 in VCaP and RWPE, respectively. This non-redundant set was used for further analysis.

Distribution of reads and spectra in the VCaP and RWPE datasets

In both the transcriptome and proteome, a relatively small number of genes encompass the large majority of transcripts and proteins in a cell. As a result, a majority of machine dynamic range is concentrated on this small number of genes. In both cases, this concentrated allocation of dynamic range also means that the majority of transcripts and proteins are poorly covered by reads and peptides (**Figure S 23, Figure S 24**). However, there were distinct differences between the two cell lines in our study in both proteome and transcriptome data. The distribution of reads in the VCaP and RWPE transcriptome data show marked differences (**Figure 9F**, **Figure S 25**); the top 30 highest abundance genes in VCaP comprised 25% of the total read density. To reach the same approximate total read density, the RWPE transcriptome data comprises the 75 highest abundance genes. This difference is attributable to the over-expression of a set of genes associated with the VCaP cell line's cancer origin. In contrast, the protein data showed a more similar distribution of abundance between the two cell lines. The difference is smaller in the protein data; in both VCaP and RWPE, the top 25% of peptide density is attributable to the top 71 and 102 genes, respectively. This is partially attributable to the use of dynamic exclusion in our protein data set, used to increase proteome coverage at the cost of reducing measurement accuracy at the highest end of the quantitative dynamic range. To examine the impact of the highest abundance genes in the context of dynamic range, we removed the top 100 most abundant genes and examined the distribution of remaining reads and spectra across the genes. This removal process had a much larger effect in the transcriptome data, bringing the distribution of reads over the dataset genes for VCaP closer to that observed in RWPE.

We then sought to examine how the relative makeup of the underlying proteome and transcriptome data may better explain these distributions. By ranking Gene Ontology classes in terms of relative read and spectral fraction without normalization, we observed a compositional divergence in the underlying data in our transcript and protein datasets (**Figure 9G, Table S 9**). The set of classes where transcript reads composed the largest relative proportion of the underlying data was dominated by translation associated classes such as translational elongation (GO:0006414) and gene expression (GO:0010467). A large number of reads in our transcriptome data were ribosomal in origin, explaining the heavy representation of these classes in the ranked list. To see if these ribosomal genes explained a large proportion of the observed differences, we removed them and reassessed the distribution (**Figure S 25**, **Table S 10**).

Our proteome data shows the use of a whole-cell lysate, with the largest fraction of genes annotated to the cytoplasm. With a mean of 1,594 observed genes, the topmost seven classes found in the high relative protein representation list tend to be significantly larger than those found on the high relative transcript enrichment list (which have a mean observed size of only 164 genes). This is consistent with the observation that proteins in general have longer half-lives in the cell, and are therefore more likely to be observed [60].

Conclusion

In this work, we describe a methodology for integrating transcriptome and proteome data in a manner that matches the reference transcriptome to the reference proteome, resolving a fundamental data mismatch issue that affects a number of previous studies to date. This is among the first studies to analyze the impacts of methodological differences in the quantification and filtering of transcriptome and proteome data. We demonstrate some of the sources of uncertainty that may degrade the fidelity of the observed transcript-protein relationship. Focusing on transcriptome data, we show that the treatment of ambiguously mapping multireads has significant effects on the derived transcript abundances, and downstream protein correlation. Looking at protein data, we show how filtering out artifacts stemming from the multiple assignment of spectra leads to a modest increase in the transcript-protein correlation.

This study is limited by a small number of samples, and the lack of biological replicates with which we can better define the inter-sample variances in transcript and protein. Future investigation into optimizing the integration of transcriptome and proteome data can leverage the increasing availability of publically accessible RNA-seq and tandem mass spectrometry data for large cohorts of samples. This increased sample heterogeneity will allow for better assessment of aberrations that arise from the highly variable transcriptomic and proteomic landscapes ultimately leading to more optimal methodologies.

Materials and Methods

Cell Lines

The benign immortalized prostate cell line RWPE was obtained from the American Type Culture Collection (ATCC). VCaP cell line was derived from a vertebral metastasis from a patient with hormone-refractory metastatic prostate cancer [114], and was provided by Ken Pienta (University of Michigan, Ann Arbor, MI).

Protein sample preparation

Collection of VCaP and RWPE whole cellular protein extract was done in RIPA complete buffer supplemented with HALT Protease and Phosphatase Inhibitor Cocktail (Peirce Biotechnology). Total protein extract was quantified with bicinchoninic acid. 50 mg aliquots of total cellullar proteins were first separated by 1D SDS-PAGE (4-12 % Bis-Tris Novex-Invitrogen, Carlsbad, CA). Forty equal sized gel bands were excised and subjected to in-gel digestion as previously described **[115]**. Extracted peptides were reconstituted with mobile phase A prior to on-line reverse phase nanoLC-MS/MS (LTQ-Velos with Proxeon nanoHPLC, ThermoFinnigan). Peptides were eluted on-line to the mass spectrometer with a reverse phase linear gradient from 97 % A (0.1 % Formic acid in water) to 45 % B (0.1 % formic acid in acetonitrille).   Peptides were detected and fragmented in the mass spectrometer in a data dependent manner sending the top 12 precursor ions, excluding singly charged ions, for collisional induced dissociation. Raw spectra files were converted into mzXML by an in-house version of ReAdW.

Parsing of transcript and protein sequence data

The Genbank formatted flat files for the Human transcripts and proteins of RefSeq release 47 were parsed into a MySQL relational database using in-house software. For this extraction, only entries that had both a transcript and a corresponding protein product were considered. The data extracted included paired transcript and protein identifiers along with the gene symbol of each pair. Sequence information for both transcripts and their protein products were also extracted.

Mass spectrometry and subsequent proteomic analysis

ThermoFisher RAW files for all replicates were converted to mzXML file using msconvert.exe from the Proteo-Wizard suite [116]. Protein searches were performed using X!Tandem with the K-score plugin [117, 118]. The data was searched against the proteins of Human RefSeq 47 along with common proteomics contaminant proteins. Reversed protein sequences were also included as decoy entries. The X!Tandem results were post-processed using PeptideProphet and ProteinProphet (version 4.4.1). [56, 119, 120].

Bioinformatics analysis of proteomics data

A summary ProteinProphet XML file was generated from all of the independent PeptideProphet results as described for Abacus [57]. All of the PeptideProphet and ProteinProphet XML files were subsequently parsed into a MySQL relational database using in-house software.

Abacus was used to obtain a gene-centric summary of the total spectral counts across all three replicates of each cell line. The Abacus results were then imported into the MySQL database. Parameters used for Abacus were: iniProbTH >= 0.5, minCombinedFilePw >= 0 and maxIniProb >= 0.5. Gene Symbol mappings for each protein were obtained from the RefSeq flat files described above.

Decoy protein matches were also imported into the database as "decoy-gene" entries. These entries were used to compute false discovery rates (FDR) of the gene-centric proteomics data.

Three probabilities were examined to determine which one provided the best discriminatory power between real genes and decoys. The FDR was computed using: bestMaxIniProb, bestMaxPw, and bestLocalPw. bestMaxIniProb is the maximum maxIniProb value observed among the all of the replicates. bestMaxPw is the maximum group probability observed for the gene from among all of the replicates. The bestLocalPw is the maximum protein probability observed for the gene from among each replicate. bestMaxIniProb was selected as the best discriminator at FDR cut offs of 0.05 and 0.01.

RNA-seq expression data

RNA-Seq library generation and sequencing

Messenger RNA (2 µg) was fragmented at 85°C for 5 min in a fragmentation buffer (Ambion) and converted to single stranded cDNA using SuperScript II reverse transcriptase (Invitrogen), followed by second-strand cDNA synthesis using ***Escherichia coli*** DNA polymerase I (Invitrogen). The double stranded cDNA was further processed by Illumina mRNA sequencing Prep kit. Briefly, double-stranded cDNA was end repaired by using T4 DNA polymerase and T4 polynucleotide kinase, monoadenylated using an exo minus Klenow DNA polymerase I (3'to 5' exonucleotide activity), and ligated with adaptor oligo mix (Illumina) using T4 DNA ligase. The adaptor-ligated cDNA library was then fractioned on a 3% agarose gel, and fragments corresponding to 280–320 bp were excised, purified, and PCR amplified (15 cycles) by Phusion polymerase (NEB). The PCR product was again size selected on a 3% agarose gel by cutting out the fragments in the 300 bp range. The library was then purified with the Qiaquick Minelute PCR Purification Kit (Qiagen) and quantified with the Agilent DNA 1000 kit on the Agilent 2100 Bioanalyzer following the manufacturer's instructions. The resulting library (5–8 pM) was used to prepare flowcells. Sequencing was done on an Illumina Genome Analyzer to produce single reads of 36 to 40bp.

Transcript quantification by RNA-Seq

We constructed a reference database composed of representative RNA sequences from RefSeq v47, matching decoy sequences, and known contaminants. Reads were aligned to this reference database using Bowtie version 0.12.5 using three alignment parameter sets, all allowing for two mismatches within a 32 base pair seed region. The "unique" alignment policy is the most restrictive; we require reads to map uniquely to a single position in the reference database using the arguments "--best -k 1 -m 1." The "single best" alignment policy assigns reads to the best quality alignment among all found alignments as determined by Bowtie using the arguments "--best -k 1." The "all" alignment policy uses the arguments "--best -k 255" to yield alignments to the best 255 locations.

Reads mapping to repeat regions were removed by alignment to the set of RepBase v16.05 Human and simple repeats with Bowtie 0.12.5 without allowing for alignment mismatches within a 32bp seed region. This process yielded a set of FASTQ files with reads stringently mapping to these known repeats removed.

Expression values in terms of Reads per Kilobase per Million reads (RPKM) were computed for each transcript including and excluding reads that mapped to 5' and 3' untranslated regions (UTRs) and adjusting for the presence and absence of these UTR regions in the total length of the transcript. Lengths for 5' and 3' UTRs were computed by counting UTR sequence lengths downloaded from the UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables) for each representative RefSeq transcript.

Reads were mapped to the hg19 genome build using TopHat 1.4.0 using an annotation file containing all refSeq transcripts. FPKM measures were generated using Cufflinks 1.4.0 with multi-read correction enabled, using the same annotation as supplied to TopHat, masking out all genes on chrM as well as all rRNA and tRNA genes in the genome. Cufflinks was run using the "-G" option to limit quantification to genes in the annotation file. FPKM values for genes with the same name in the genes.fpkm_tracking file were summed to yield a gene-level list of abundance values. These results were merged with proteomics data on the basis of gene name or RefSeq isoform id. Abundance data were thresholded to exclude genes with FPKM values less than 0.3.

Generation of Decoy Transcripts and Computation of False Discovery Rate

Using a method similar to that of Ramskold, et al [96], a False Discovery Rate (FDR) was computed by aligning reads to transcripts and decoy sequences of matched lengths and computing the difference between the number of transcripts and decoys seen at varying RPKM expression thresholds. Decoy sequences were derived from sub-sampling intergenic regions of hg19 outside of gene annotations from RefSeq, UCSC, and Ensembl and outside known sequencing gaps.

Comparison of RNA-seq data with spectral counts

Consolidation of ambiguous transcriptional evidence

All transcripts had five expression values calculated for them: raw, reads-per-million, RPKM, RPKM excluding repeats, and RPKM excluding reads mapping to UTRs and repeat regions. Expression data for multiple transcripts sharing a common gene symbol were collapsed into a single gene entry. Two different methods of collapsing the expression data were employed. For the 'all' data category, the maximum observed value (for each expression calculation) was selected from all transcripts of a gene. For the 'unique' or 'single best' data categories, the sum of the observed values for all shared transcripts was taken.

Assignment of representative transcript and protein identifiers

For each gene symbol with valid spectral count data, a representative protein identifier was selected. In cases where a gene symbol had multiple proteins associated with it, the protein with the largest number of unique spectral counts was selected. Ties were broken based upon the alphanumeric sorting of the remaining candidate protein identifiers and selecting the first one. The representative transcript for a gene symbol was taken to be the parent transcript of each representative corresponding protein.

Selection of candidate genes common to transcript and proteomics data sets

Two data sets were derived from unfiltered data at FDR levels of 1% and 5% using different metrics for quantification accuracy. The 5% FDR dataset was derived by filtering for genes with a bestLocalPw >= 0.9 and a bestMaxIniProb >= 0.9475 and 0.958 for protein data, or a minimum RNA-seq abundance of 1.8 and 3.6 RPKM, in VCaP and RWPE, respectively. Candidates in this data set had to match one of these 5% thresholds in either the protein or mRNA data to be included in the dataset. The 1% FDR dataset was derived by requiring the gene to meet the 1% FDR criteria in both our RNA-seq and protein data in either cell line. A simple filter for keratins (a common artifact in tandem MS experiments) was applied by marking and removing genes with names matching “KRT” followed by a number.

Correlation of transcript expression with spectral count data

The final data sets used for analysis were derived from the repeat-removed sequence files, aligned using the single-best policy, and excluding UTR reads from quantification. For all candidate genes, the spectral counts were averaged together for each cell line using the mean.

The averaged spectral counts for each gene were then converted to NSpC values (normalized to the length of the representative protein identifier) using R [121]. Correlations were computed between log2-transformed RPKM and NSpC values, excluding values that were incomplete (where log transformation of either protein or transcript values resulted in an NA value)

Chapter 4: The transcript-protein relationship in human prostate cancer

Introduction

Prostate cancer is the most common cancer afflicting men, with a 1 in 6 lifetime risk of the disease in the United States [122]. Its prevalence has made prostate cancer a subject of extensive molecular profling at the genome, transcriptome, and proteome levels. However, few studies have investigated the transcript-protein relationship in prostate cancer. Previous research in various human cancers using lower-throughput methods have specifically noted discordance in relative mRNA and protein abundance [41], including numerous dysregulated pathways in prostate cancer with mRNA-protein correlation at varying levels up to $r = 0.68$ [123]. These pathways include functionally important molecular networks and pathways such as nF-kB, which mediates immune response, apoptosis, and inflammation and insulin signaling. More focused studies on specific genes have noted discordance in transcript-protein relationships in a number of cancers; endometrial carcinoma where urokinase and tissue plasminogen activators were noted to diverge [124], acute myeloid leukemia where the transcript and protein expression of the breast cancer resistance protein *ABCG2* was observed to be uncorrelated [125], and colorectal cancers where the transcription factor AP-2 (*TFAP2A*) was observed at moderate abundance at the transcript level while showing no protein detection [126]. Altogether, these studies suggest that dysregulation of the transcript-protein relationship may be a marker for the establishment and/or progression of cancer.

The majority of previous genome-scale studies examined the relationship between transcript and protein within the context of single cell lines. In general, there has been a paucity of research leveraging ultra-high throughput technologies to assess the relationship between mRNA and protein abundance in the context of human cancers.

Here, we use the VCaP and RWPE prostate cell lines as models for cancer and normal prostate epithelium. Using the abundance values derived from the method described in **Chapter 3**, we classify genes into functionally discrete categories based on their relative transcript-protein relationships within each of these cell lines and examined the impact of protein and transcript half-life on this relationship. We then compared the relative transcript-protein relationships across our two cell lines. Through this process, we identified genes where this relationship becomes dysregulated in our cancer model using novel discordance and concordance index values. We coupled the results of this analysis with the human protein interaction network and demonstrate how several biological processes closely interlinked with the Akt signaling pathway show transcript-protein relationship dysregulation in our cancer model. Additionally, we demonstrate how the integrative analysis of both transcriptome and proteome leads to insights about the variance in the proteome and transcriptome, and how these changes lead to discordance in the transcript-protein relationship.

Results and discussion

Transcript and Protein abundance in each of the cell line models

To determine some of the biological factors that affect the transcript-protein relationship, we chose to separate the genes into approximate subsets based on their relative transcript and protein abundance. In order to capture genes that are otherwise ignored due to measurements of zero, we added a small adjustment factor of 0.2 to the RPKM and NSpC values before log2 transformation in this analysis. We divided the genes in VCaP and RWPE into four broad subsets based on the relationship we observed between protein and mRNA abundance (**Figure 10A, Figure S 26, Table S 13**). This was done by choosing genes 1.5 standard deviations away from the best fit line (ignoring points that have values less than 0.3 RPKM or NSpC) between mRNA and protein, and sub-selecting sets of genes that also showed a log2 normalized spectral count or RPKM values less than 0.3 RPKM or NSpC. These thresholds ensured that we selected for two conceptual classes of genes: those that had significant differences in transcript and protein abundance and those which were only detected by a single method. Genes in each of these broad subsets were analyzed with DAVID to examine their functional composition with GO.

The first two of these four subsets consisted of genes with higher protein or mRNA abundance with both detected. The latter two subsets yielded genes with either high protein or mRNA abundance but little or no observed abundance of the corresponding transcript or protein. While many of the genes segregating into each of these broad sets are driven by functional biology, a subset are mis-categorized due to technical factors; in particular our limited ability to capture, detect, and quantify some genes. For example, genes with relatively high protein but very low mRNA abundance tended to be contaminants such as keratin typical of MS experiments of this kind, usually introduced during the sample handling process common to MS experiments [127]. This is reflected by the presence of classes such as keratinocyte differentiation in this class, even despite our effort to filter out the effect of these contaminants from our dataset (see **Methods**). An additional example is the well-known bias against membrane proteins [128, 129] in MS/MS experiments due to their low solubility resulting in their under-representation in the data. Biological factors that underlie some of these observed differences include different rates of transcript and protein turnover, which affects our ability to measure these genes.

Genes with high mRNA abundances and little or no observed proteins tended to fall into GO classes such as regulation of transcription, composed of genes with low relative abundance or short half-lives, such as transcription factors. This class is larger than the corresponding class of genes with high protein but almost no detectable transcript, likely due to our ability to probe the transcriptome more deeply than the corresponding proteome with our data.

Genes with higher levels of mRNA than protein encompassed some of the same transcription-associated classes, although this group also composes a large abundance of ribosomal genes commonly found in mRNA as well as transporters. This category shows some of the classes not elucidated by previous studies using array-based techniques for profiling the transcriptome, due to the limited dynamic range of those methods. Genes with high protein but more modest amounts of mRNA were largely contained genes associated with the cytoskeleton and microtubules – this is expected as these proteins tend to be highly stable. This association of metabolic and structural component class genes with higher relative protein to transcript levels (compared to the association of regulatory genes to the converse group), along with the observation that these proteins are more long-lived than their associated transcripts, is consistent with the concept that this core functionality of the cell is less subject to variation than genes with regulatory function [110, 130].

Transcript-protein relationships within biological classes

The correlation between transcript and protein abundance in cells is affected by many intermediary factors involving transcript and protein structure [131], translational delay [132], stability, and degradation. Correlation was calculated using the 5% FDR extended dataset. From the baseline Spearman correlation of r = 0.61 and r = 0.51 in the VCaP and RWPE cell lines, respectively, we attempted to find biological classes which exhibit relatively high and low correlation between protein and mRNA abundance.

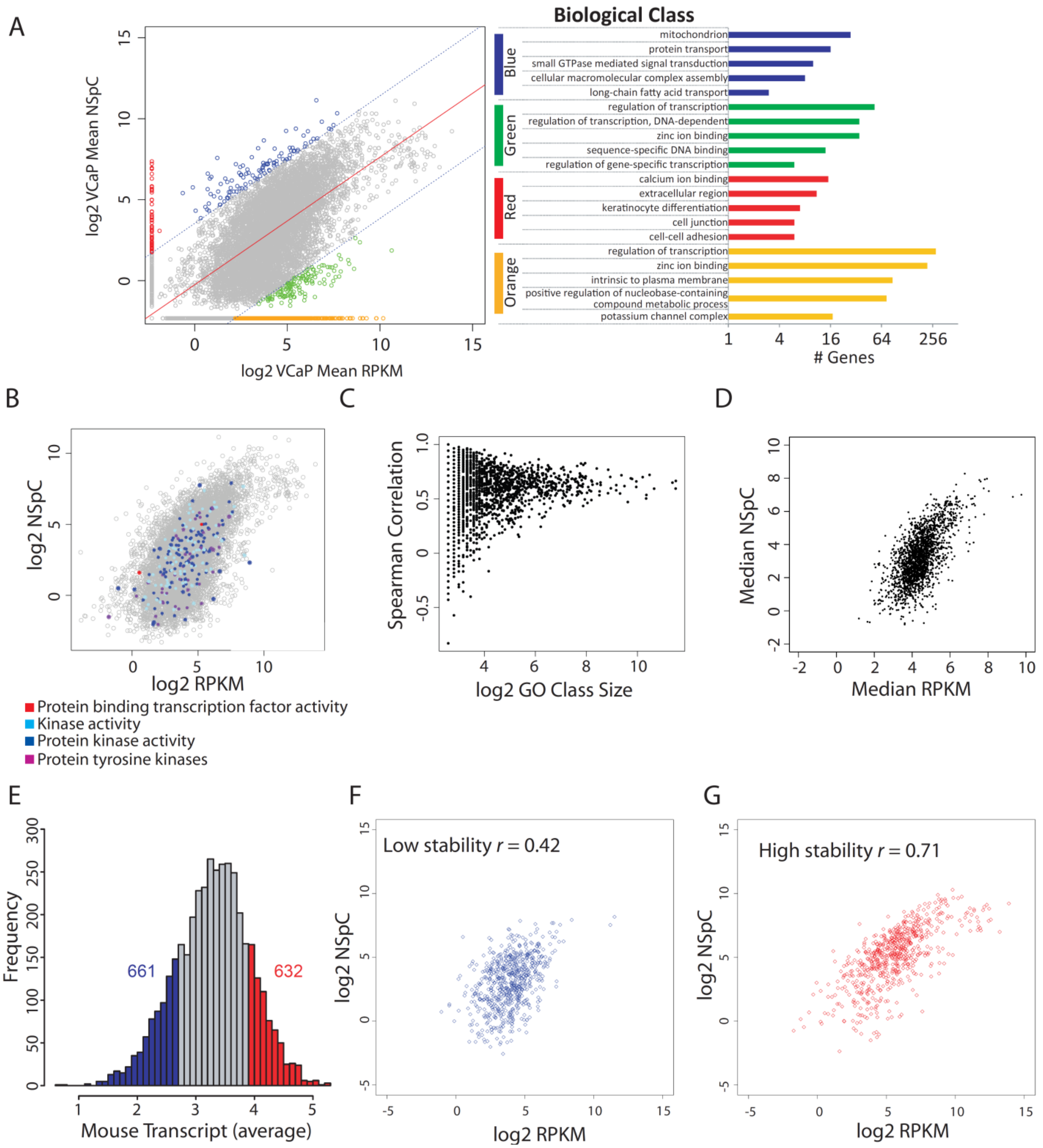


Figure 10: Analysis of the transcript-protein relationship in VCaP and RWPE. A. Division of genes by relative protein-transcript relationship with zero values. **B.** Plot of cancer-related GO class genes of interest. **C.** Relationship between GO class size and transcript-protein Spearman correlation coefficient. **D.** Relationship between transcript and protein abundance per Gene Ontology class. **E.** Distribution of transcript stability used to segment extended dataset into high and low stability sets. **F and G.** Correlation of transcript and protein levels for low and high stability genes in VCaP.

We first examined the distribution of genes in several cancer-related classes of interest - in particular kinases (and subclasses thereof) which often act as drivers in cancer (**Figure 10B, Table S 11, Table S 12**). As mediators of the cell cycle, kinases are frequently altered in cancers and can drive oncogenic processes. As a result, they are the focus of many targeted cancer therapies [133]. We examined the correlation and distribution of this class of genes (defined as genes mapped to GO Class GO:0016301 “kinase activity”), and obtained a correlation of *r* = 0.61 and *r* = 0.44 with observations of 87 and 88 genes (out of 581 total annotated to the GO class) in VCaP and RWPE, respectively. The difference in correlation between the cell lines is likely attributable to the greater mean abundance of protein products in VCaP, measured at 2.89 NSpC compared to 1.78 NSpC in RWPE, leading to better quantification accuracy.

We also examined the class of transcription factors that affect cell signaling and proliferation. Relatively few genes observed in our extended dataset are annotated in GO as transcription factors, with a total of only five genes annotated to protein binding transcription factor activity (GO:0000988). This is likely due to the relatively low abundance of both transcript and protein of many of the genes in this class, resulting in their exclusion from our datasets. This small number of observations therefore led to the class being excluded from our GO analysis. The genes in this class that we observed; *PITX1*, *HMGA2*, *HEY1*,*SMAD4*,and *LHX2*, were expressed relatively higher in our RWPE cells compared to our VCaP cancer cells, with a mean of 27.14 and 12.82 RPKM, respectively. At first glance, we might expect that the increased transcriptional activity in cancer cells would imply increased abundance of transcription factors. However, these results are consistent with the observations in a number of published studies; the gene *PITX1* [134] was noted to be lower in prostate cancer cells compared to normal, *HEY1* is excluded from the nucleus in prostate cancer tissues [135], and *SMAD4* acts as a barrier to the growth and progression of prostate cancers [136]. Since the mechanism of action of these genes implies that they act as transcriptional repressors, it is not surprising that their levels are down-regulated in cancers.

We then conducted a more unbiased analysis of the dataset using all Gene Ontology classes with 10 genes or greater observed in our dataset and calculated the Spearman correlation, the associated p-values, and the mean abundance of transcript and protein for the genes in each class. The correlation within these classes scales with class size approaching the mean dataset correlation coefficient of *r* = 0.61 (**Figure 10C**) in VCaP as class sizes become large. Much of the variation in the data is seen in GO classes containing 16 or fewer observed genes. Previous studies have noted that the proteome exhibits a larger number of significantly differential genes in cancer than the transcriptome [137]. To examine whether the large differences in correlation in small GO classes was driven by larger proportional membership of significantly differential genes, we compared the mean protein abundance and the protein-transcript correlation p-value. The large variation in transcript-protein correlation within each GO class appeared to be an effect of sample size, as the Pearson correlation between the two factors in each class was *r* = -0.03 in VCaP and *r* = -0.05 in RWPE. A similar observation was made that higher transcript-protein correlations are seen in gene subsets with higher abundance [138]. Pearson correlation between the mean protein and transcript abundance in each GO class and Spearman correlation value yielded *r* = 0.12 and *r* = 0.04 in VCaP and *r* = 0.07 and *r* = 0.07 in RWPE, respectively.

To study the genes in our dataset on the basis of biological function and localization, we analyzed the values by the median abundance of protein and transcript within individual Gene Ontology categories (**Figure 10D**). The most obvious outliers are members of ribosomal small subunit biogenesis and cytosolic small ribosomal subunit classes, with a very high relative transcript/protein ratio that reflect the large abundance of ribosomal gene mRNAs in our transcript data.

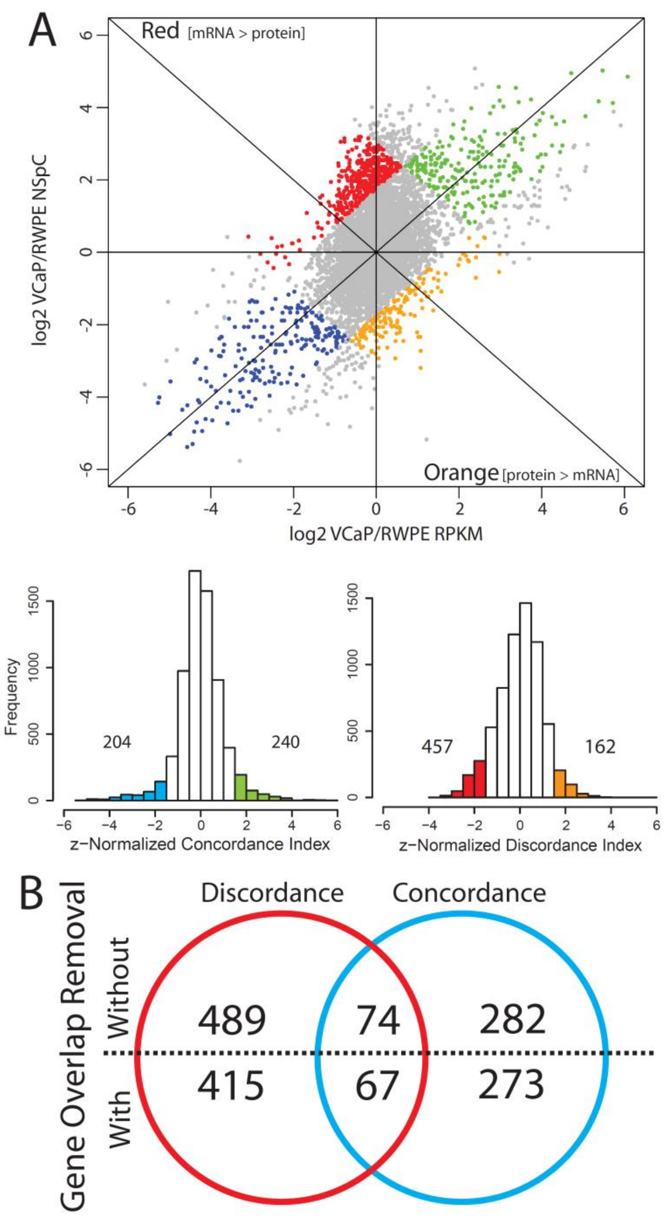
The impact of stability in the transcript-protein abundance relationship

Protein and transcript stability have been noted in the literature to have a significant impact on the relationship between transcript and protein abundance levels, and the effect is clearly visible in our dataset (**Figure 10E**). Using transcript and protein stability data from Schwanhäusser B, et al. [60] derived from NIH 3T3 mouse fibroblast cells, we assigned transcript and protein stability to the genes in our dataset through orthology. We separated the genes in our dataset into high and low transcript and protein stability groups by selecting one standard deviation tails of the z-normalized stability distribution in each of the two cell lines. The differences in transcript-protein correlation between the high and low stability groups on the basis of transcript stability are the most marked. In VCaP, the correlation for the low and high stability transcripts is *r* = 0.404 and *r* = 0.71, respectively (**Figure 10F-H**). The difference is similar in RWPE where the correlation is *r* = 0.288 and *r* = 0.543 for the low and high stability transcripts, respectively. When comparing low and high stability groups on the basis of protein stability in the two cell lines, the difference is smaller with *r* = 0.419 and *r* = 0.572 for RWPE and *r* = 0.441 and *r* = 0.799 for VCaP for low and high stability proteins, respectively (**Figure S 27 and Figure S 28)**.

Comparison of VCaP and RWPE cell lines

To examine aberrations in the protein-mRNA abundance relationship specific to cancer, we compared the relative transcript - protein ratios between the VCaP and RWPE cell lines. We applied our core dataset to ensure accurate gene level quantification.

For functional analysis, we selected the most concordant and discordant genes. In the large majority of cases, the relationship between transcript and protein abundance between the two cell lines is unchanged. For this purpose, the data were analyzed along two axes to measure the genes with the most concordant and most discordant transcript-protein relationships (**Figure 11A**). The genes with the highest transcript-protein concordance were found by using an index value derived from adding the normalized RPKM transcript fold change value to the normalized protein fold change abundance value between VCaP and RWPE (as described in **Methods**). The most discordant genes were found by the derivation of a similar index value of the normalized protein abundance subtracted from the normalized RPKM transcript value. The index values where both z-transformed and p-values were computed using these scaled distributions of concordance and discordance for use with LRPath for Gene Ontology and pathway analysis.



protein > transcript

Class/Pathway	Type	# Genes	OddsRatio	P-Value
cytoskeletal part	GO CC	169 (385)	0.161	2.67E-16
cell cycle phase	GO BP	125 (272)	0.152	1.8E-13
regulation of signal transduction	GO BP	141 (282)	0.156	1.98E-13
focal adhesion	GO CC	22 (51)	0.135	0.000189
regulation of cell cycle	GO BP	90 (220)	0.283	2.06E-05
macromolecule catabolic process	GO BP	132 (342)	0.435	0.000968
hydrolase activity, acting on ester bonds	GO MF	115 (308)	0.500	0.009015
chromosome	GO CC	104 (300)	0.487	0.007775
kinase activity	GO CC	136 (314)	0.270	1.52E-07
regulation of kinase activity	GO BP	59 (149)	0.342	0.003006
thiolester hydrolase activity	GO MF	20 (50)	0.117	4.39E-05
Apoptosis	KEGG	24 (42)	0.068	7.91E-06
cell cycle checkpoint	GO BP	25 (65)	0.185	0.000674
Pathways in cancer	KEGG	58 (128)	0.147	1.69E-06
regulation of binding	GO BP	36 (97)	0.267	0.002206
regulation of ARF protein signal transduction	GO BP	13 (21)	0.053	3.58E-05
Fructose and mannose metabolism	KEGG	7 (27)	0.092	0.001768
I-kappaB kinase/NF-kappaB cascade	GO BP	36 (76)	0.219	0.001286
positive regulation of response to stimulus	GO BP	28 (71)	0.194	0.000627
small GTPase mediated signal transduction	GO BP	85 (194)	0.321	0.000346
regulation of localization	GO BP	102 (239)	0.428	0.004232
regulation of cellular component organization	GO BP	111 (283)	0.433	0.002289
Cell cycle	KEGG	26 (76)	0.245	0.007466
meiotic cell cycle	GO BP	17 (35)	0.173	0.00809
polyol metabolic process	GO BP	17 (22)	0.055	3.45E-05
ubiquitin ligase complex	GO CC	27 (81)	0.196	0.000319
PTEN dependent cell cycle arrest and apoptosis	Biocarta	8 (15)	0.051	0.00146
guanyl-nucleotide exchange factor activity	GO MF	32 (52)	0.092	1.68E-06

transcript > protein

Class/Pathway	Type	# Genes	OddsRatio	P-Value
ectoderm development	GO BP	34 (70)	4.316	0.009157
transmembrane transporter activity	GO MF	114 (227)	9.868	3.01E-13
endoplasmic reticulum part	GO CC	210 (353)	8.829	6.27E-17
cellular lipid metabolic process	GO BP	123 (266)	3.385	6E-05
lipid catabolic process	GO BP	38 (78)	5.706	0.000806
cellular hormone metabolic process	GO BP	16 (22)	14.388	0.001213
membrane fraction	GO CC	150 (317)	2.879	0.000146
positive regulation of epithelial cell proliferation	GO BP	6 (11)	18.909	0.006493
regulation of blood pressure	GO BP	11 (20)	22.177	9.7E-05
antigen processing and presentation	GO BP	12 (19)	28.204	1.82E-05
cofactor metabolic process	GO BP	69 (142)	3.437	0.002372
nucleoplasm part	GO CC	186 (375)	1.992	0.007628
hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	GO MF	22 (51)	5.280	0.009673
DNA-directed RNA polymerase II, holoenzyme	GO CC	41 (65)	5.294	0.003141
glycoprotein metabolic process	GO BP	48 (86)	9.585	1.61E-06
Golgi membrane	GO CC	108 (227)	3.412	0.000145
cellular ketone metabolic process	GO BP	150 (325)	2.546	0.000786

Figure 11: Detecting Dysregulation of transcript-protein relationships in prostate cancer. A. z-transformed fold changes of transcript and protein observed between VCaP and RWPE and major Gene Ontology class clusters in red and orange classes, corresponding to enrichment in protein and transcript abundance, respectively. Representative GO classes for each annotation cluster were chosen by examining overlap between DAVID clusters and LRPath results with FDR ≤ 0.05, in order of observed genes **B.** Number of GO classes resulting from LRPath analysis before and after removal of genes overlapping between concordance and discordance classes.

Ontology and pathway analysis with concordance and discordance indices

LRPath was used due to its ability to evaluate enrichment of classes and pathways in aggregate without requiring the use of cutoff values [139]. This is particularly important in our analysis of genes nominated by our index values as significance cutoff thresholds are not well defined. Although it relies on the use of cutoff values, we also applied DAVID [140, 141] to our dataset to leverage it's clustering of resultant classes. For consistency with our previous analysis looking within each cell line, we selected genes using a 1.5 standard deviation cutoff from perfect concordance and discordance.

Analysis using our discordance and concordance indices in LRPath produced 727 and 619 GO classes and 114 and 56 Biocarta and KEGG pathways with a p-value ≤ 0.05 in VCaP and RWPE, respectively (**Table S 14, Table S 15**). The selection of identical classes in both of these classes led us to evaluate the effects of this gene overlap between high correlation and high anti-correlation due to our selection process. Genes typically considered significant (p <= 0.05) by the discordance index had their p-values adjusted to non-significant values when they were also in the significant tail distributions of the genes in the concordance index. The dataset with these modified p-values was re-analyzed with LRPath. Major themes from our primary analysis remained statistically significant, suggesting that the effect of genes in overlapping regions is relatively small. The majority of classes we observed to be significant in our initial analysis demonstrated increased statistical significance. With this subtraction of overlapping genes, the discordance and concordance indices produced a respective 619 and 599 GO classes as well as 114 and 56 KEGG and Biocarta pathways with a p-value ≤ 0.05 . While the reduction of overlapping classes between the highly concordant and discordant class sets was relatively modest (**Figure 11B**), we chose to use this dataset for further analysis in an effort to minimize noise data in our analysis.

DAVID clustering of the same GO classes and pathways from Biocarta and KEGG revealed that many of our resultant classes fall in a small number of biological themes. We used our DAVID results to guide our manual classification of discordance-derived LRPath results into broad biological categories. In the case of where protein levels are overabundant, we see broad classes of protein metabolism and modification, cell cycle and structure, ion binding, and GTPase regulation. In the converse case, we see the broad clusters including classes associated with the cell membrane, mitochondria and energy metabolism, phosphorylation, and lipid metabolism.

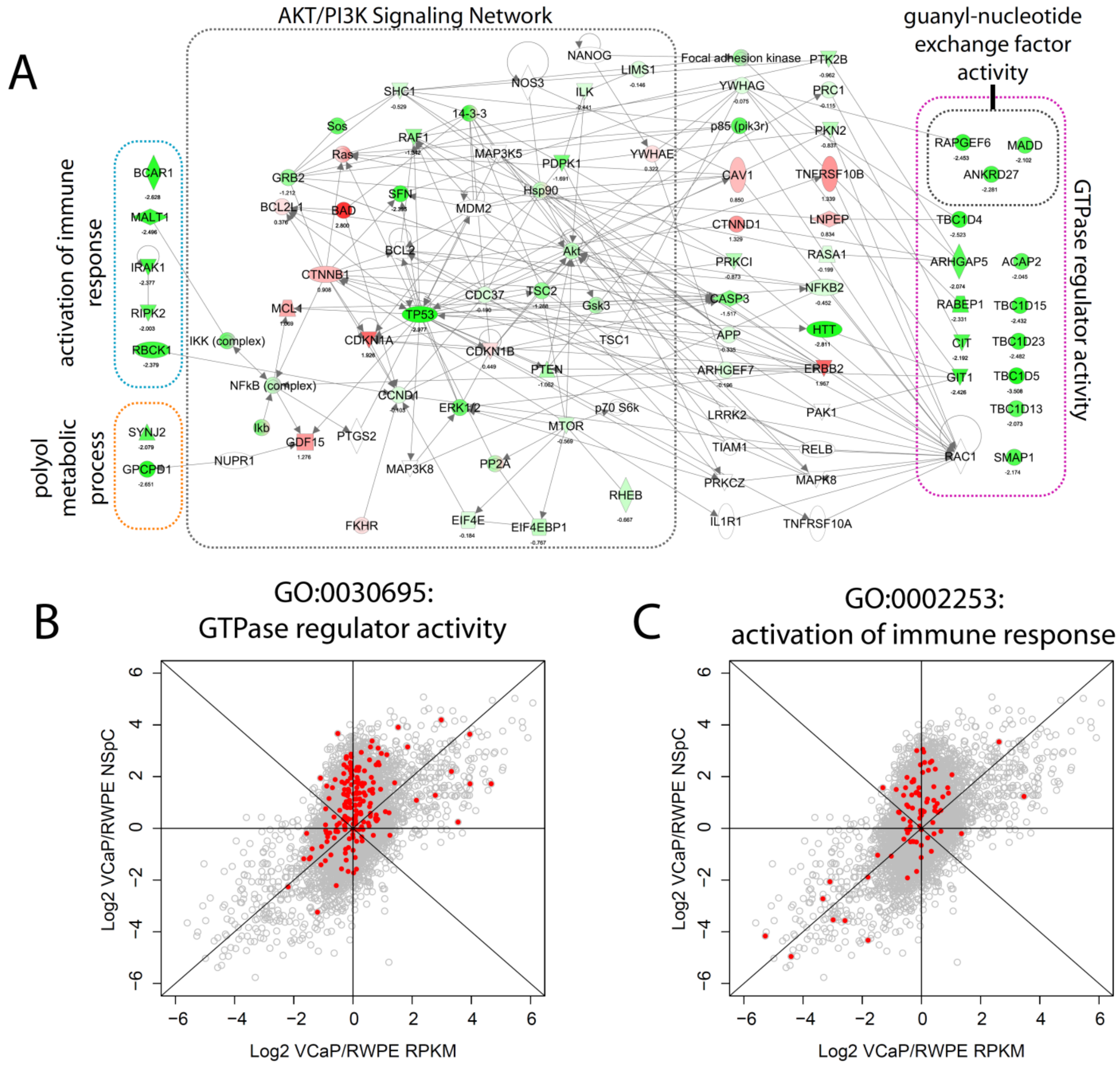


Figure 12: The dysregulated networks surrounding Akt. A. Network closely linked to Akt formed by genes constituting the GTPase regulator activity, polyol metabolic process, guanyl-nucleotide exchange factor activity, and regulation of immune response classes. B and C. Distribution of genes in the GTPase regulator activity and activation of immune response Gene Ontology classes highlighted in red in the context of all dataset genes.

The sets of discordant genes were of particular interest as they suggest a number of biological processes are dysregulated on a post-transcriptional level. Correlation analysis of fold changes by Gene Ontology class in this set of genes finds a number of processes in the two cases where the transcript-protein relationship is significantly dysregulated in VCaP compared to RWPE. Similar to our correlation analysis of transcript and protein abundance within each cell line, we observed a correlation between GO class size and Spearman correlation coefficient with higher variance coming from smaller classes. As expected, the set of correlation values has a roughly normal distribution centered on the correlation coefficient of *r* = 0.42 for the broader dataset. This pattern holds when we separate the resultant classes by GO tree as well. Because of the correlation between GO class size and the resultant observed correlation, we thresholded the classes to exclude those with less than 10 observed genes. The most concordant and discordant GO classes by tree can often be explained by the differing stabilities of the protein and mRNA, such as the classes involving genes associated with cell spindle and microtubules which have very long half-lives. In these two examples, structural role of the protein yields higher stability leading to higher abundance than the comparatively shorter-lived complementary mRNA.

Using this correlation data, we examined GO classes with below-mean correlation coefficients between cell lines and ranked them by the difference in correlation between VCaP and RWPE. This nominated candidate classes that showed large differences in correlation where the transcript-protein relationship was drastically altered. This set included a number of voltage-gated ion channel classes and the smoothened signaling pathway. The presence of the ion channel classes reflects the activity of voltage-gated potassium and sodium channels observed to play a role in the growth and metastasis of prostate cancer cells [142-144]. The genes in these classes go from highly correlated in RWPE, with Spearman correlation coefficients in the *r* = 0.8 range, to essentially uncorrelated (Spearman correlation falling to the r = 0.1 to 0 range) in VCaP. Similarly, correlation of the genes in the smoothened signaling pathway goes from *r* = 0.77 in RWPE to *r* = -0.03 in VCaP. The Smoothened (*SMO*) gene itself is known to act as an oncogene, and this pathway is known to stimulate hedgehog signaling, which is noted to be activated in advanced and metastatic prostate cancer [145, 146] and is associated with aggressiveness [147].

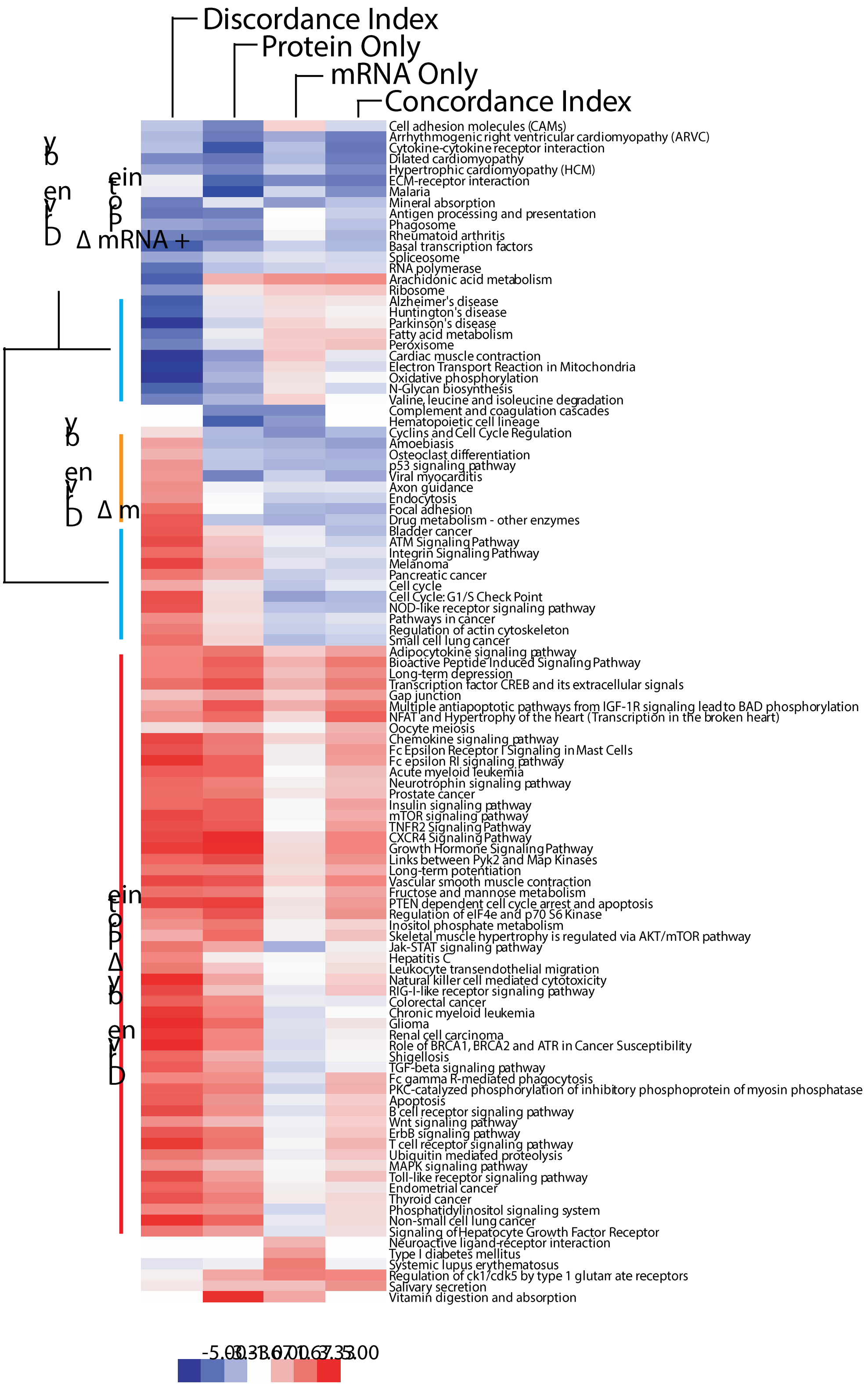


Figure 13: Insights from joint transcriptome-proteome analysis. Heatmap of log-odds ratios for KEGG and Biocarta pathways with p-values ≤ 0.05 in each of four categories – results derived from protein only, transcript only, discordance index, and concordance index data.

Dysregulation of the transcript-protein relationship surrounding Akt

A common thread we observe in our results is the association of a number of deranged pathways and processes upstream and adjacent to the PI3K/*AKT* signaling pathway, which plays a role in several biological processes central to the development of cancer including apoptosis, differentiation, and cellular metabolism. These adjoining and connected pathways and biological classes include members of the GTPase regulation and signaling classes, metabolism of the polyol substrate central to Akt signaling, immune response, and the set of guanyl nucleotide exchange factors. The genes that underlie these associated pathways and processes form a tightly interconnected network with the Akt pathway (**Figure 12A**).

In many cases, the activation of the Akt pathway is owed to a number of factors, such as the deletion or inactivation of the PTEN gene, which acts as a phosphatase on the PIP3 substrate, or the activation of upstream kinases. While PTEN is not deleted in either VCaP or RWPE, we observe that a number of genes in the PI3K/Akt pathway, such as PIK3R1, a regulatory subunit of PI3K, and SOS1, an inositol phosphatase, show significant mRNA-protein discordance with observed within the top 4% most discordant genes in the core dataset (**Figure S 29**).

The transcript-protein relationship for the genes in these four pathways and classes are not all uniformly dysregulated – in many cases, only a small subset of genes in the set show large changes in the relationship (**Figure 12B,C**). Because only genes meeting the p ≤ 0.05 significance threshold from LRPath analysis are included in the network, not all interactions are included. GTPase regulator activity, which includes 59 genes, although only 9 of these genes have discordance-based $p \leq 0.05$. This class includes SMAP1, which has been implicated in oncogenesis, and is thought to act as a tumor suppressor in intestinal cells [148]. The Polyol metabolic process class is intrinsically associated with genes that mediate the processing of PIP2 to PIP3 substrate of the Akt pathway, a substrate that is central to Akt signaling. In this class is an example of the direct interactions to the Akt signaling pathway, specifically the activation of IKK by MALT1, which contributes to T-cell activation[149]. The immune response class relates Akt to the effect its activation has on immune resistance – providing tumors with immune resistance and apoptotic escape [150]. Upstream of Akt, the set of guanyl nucleotide exchange factors are known to activate Ras though the removal of GDP [151]. This process then leads to Akt induction [152]. Not all of these networks dignal directly into the PI3K/Akt pathway; much of the connectivity of the GTPase regulator activity class is though intermediates, including notable genes such as ERBB2, PRC1, and YWHAG which are often aberrant in regulation or structure in cancers. The class of guanyl nucleotide exchange factors is a subset of genes with GTPase regulator activity, and is similarly attached to the network.

Analysis of joint analysis results

Our pathway and Gene Ontology analysis nominated a number of biological pathways which showed significant numbers of member genes with dysregulated transcript-protein relationships in our prostate cancer model that were brought to the forefront by joint analysis. To examine if this joint analysis provided insights that neither transcript- or protein-based analysis could alone, we compared our results against these isolated analyses.

We took the mRNA and protein data individually and derived the set of significant GO classes to observe the significant classes from the viewpoint of mRNA and protein in isolation. This process yielded 684 and 814 total classes with p-values ≤ 0.05 in mRNA and protein, respectively. We then compared these classes to the ones derived using our discordance and concordance indices. The joint analysis class sizes are comparable to those in the isolated analysis examples, but they are compositionally quite different. The joint analyses nominated 265 and 96 categories that were exclusive to the discordance and concordance methods alone.

In a more focused examination of these results, we examined the level of association between the biological processes we identified and the methodology we used looking at only the highest significance ($p \leq 0.01$) KEGG and Biocarta pathways (**Figure 13**). Using the odds ratio as a metric for evaluating the association strength of these pathways with the variance in their underlying genes, we clustered these classes in an attempt to dissect the drivers of discordance. In the context of our discordance index, a negative log-odds indicates an association whereby the protein abundance is greater than the corresponding transcript in VCaP relative to RWPE. A positive log-odds value indicates the contrary; that the transcript abundance is enriched in RWPE compared to VCaP without a corresponding increase in protein abundance. In a majority of cases, the discordance between protein and mRNA abundance is driven by differences at the protein level with little change at the transcript level. A smaller, corresponding set of pathways is nominated chiefly by changes in the transcriptome. These observations agree with previous observations that the proteome is more dynamic than the transcriptome [110]. More interestingly, there is a set of pathways whose association by the discordance index is driven by more subtle changes in both their transcript and protein abundances moving in opposite directions - we observe this because of the directionality of the LRPath enrichment test.

This set of pathways is split into two broad categories – pathways where transcript levels are decreased in opposite of increasing protein abundance, and the converse where transcript levels are increased while protein abundances are decreased. In the first case where transcripts are observed to be higher in abundance while there is a decrease in protein abundance, is composed of ten pathways. These include pathways associated with neurodegenerative diseases and cellular metabolism. While the three neurodegenerative diseases included in this category might first appear to be noise, likely nominated together since they contain many of the same genes, recent observations linking an inverse relationship between the incidence of Alzheimer's disease and cancer [153] as well as the application of cancer drugs to treat Alzheimer's disease [154] suggests that they may share some common molecular dysfunction. The second instance is composed of eleven pathways, broadly covering cell cycle-related classes and pathways involved in specific cancers. The observation of cell cycle and cancer specific classes is expected, as the evasion of apoptosis and insensitivity to anti-growth signals is a major hallmark of cancers [61] with many cancers sharing similar molecular dysregulation. While the pathways in both of these classes are nominated as interesting candidates in both analyses, the use of joint analysis brought them to the forefront of our analysis.

Conclusion

In summary, we show that the transcript-protein relationship is affected by a number of biological factors. Our classification of genes by their relative transcript and protein abundance in VCaP and RWPE demonstrates that the relationship is subject to the stability and resulting half-life of the constituent transcript and protein, with a large number of short-lived transcription factors picked up exclusively in transcriptome data. The effect of transcript and protein stability is reinforced by our integration of transcript and protein stability data from the literature and the observation that genes with above-average protein or transcript stability have higher abundance correlation between transcript and protein levels. Other studies have suggested that sequence features also contribute quite significantly to the transcript-protein relationship, and this is a clear path for further examination.

We then examine the transcript-protein relationship to find both individual genes and biological classes of genes where this relationship is dysregulated in a cancer context by comparing VCaP to RWPE. To achieve this goal, we derive novel discordance and concordance index values for all candidate genes in our dataset. Focusing on the candidates nominated through GO class and pathway enrichment analysis based on our discordance index, we find that several biological pathways surrounding the PI3K/Akt signaling pathway exhibit significant discordance. Coupled with evidence in the literature that modifying the stability of genes serving a regulatory role, these results may suggest an alternate pathway for the induction of functional networks conferring growth, survival, and immune and apoptotic escape in cancer. Furthermore, analysis of the pathways uniquely nominated through our joint analysis, in particular cases where transcript and protein levels move in opposite directions, elucidated possible mechanisms that may underlie the inverse relationship observed between cancer and neurodegerative disease.

Our study is limited by the small number of samples involved and the use of relatively immature RNA-seq technology. The inclusion of additional biological replicates and additional cancer types may lead to broader insights about the nature of the dysregulation of the transcript-protein relationship in cancer and its larger implications in the establishment and progression of the disease. The application of even higher throughput transcriptome sequencing techniques (optimally in concert with higher mass accuracy MS/MS profiling) will help improve the accuracy of transcript-level measurement, and increased transcript coverage will provide nucleotide level information that allows for the attribution of dysregulation to known sequence level aberrations such as disruptive single-nucleotide polymorphisms.

Methods

Derivation of index values

We computed several index values to quantify the transcript-protein abundance relationship in our two cell lines. Using the transcript and protein abundances, we computed the ratios of transcript and protein between the cell lines. We added a value of 0.2 to the values in our fold change calculations in order in order to avoid division by zero values. This value was chosen because it is below the threshold for a single spectra detected in many of the three replicates for each cell line as well as falling below the abundance cutoffs for transcriptome data, and should not significantly alter any results.

$$\boldsymbol{logratio_{transcript} = log2\left(\frac{RPKM_{VCaP} + 0.2}{RPKM_{RWPE} + 0.2}\right)} \qquad \textbf{Equation 1}$$

$$\boldsymbol{logratio_{protein} = log2\left(\frac{SpC_{VCaP} + 0.2}{SpC_{RWPE} + 0.2}\right)} \qquad \textbf{Equation 2}$$

These values were then z-transformed using the *scale* function in R to derive ***z_logratio****transcript* and ***z_logratio****protein*. Values for the concordance and discordance index for each gene were computed from these z-transformed transcript and protein log ratios

$$\boldsymbol{idx_{concordance} = z\_logratio_{transcript} + z\_logratio_{protein}} \qquad \textbf{Equation 3}$$

$$\boldsymbol{idx_{discordance} = z\_logratio_{transcript} - z\_logratio_{protein}} \qquad \textbf{Equation 4}$$

The index values were then z-normalized.

p-values were derived from these two log-ratio index values, based on a fit of the normal distribution. A correction was applied in an attempt to remove noise from genes that overlapped between highly concordance and high discordant genes. We used the commonly significant $p \leq 0.05$ level as a basis cutoff value where these genes would have their p-values adjusted to non-significant values.

Gene Ontology analysis of correlation data

Gene Ontology analysis was carried out using DAVID (http://david.abcc.ncifcrf.gov/) [140] and LRPath (http://lrpath.ncibi.org) [139] as well as the Gene Ontology (GO.db) and KEGG (KEGG.db) Bioconductor packages in R[155]. The entirety of the core and extended datasets were used as backgrounds in DAVID analysis when analyzing genes derived from those respective datasets. DAVID analysis of broad categories in each cell line was done using GO FAT terms. Analysis of genes comparing the cell lines in DAVID was done using all GO terms as well as KEGG and Biocarta pathway entries. LRPath was used to analyze candidate genes in the comparison of VCaP and RWPE and between protein and mRNA levels in each cell line. Gene identifiers were converted from RefSeq to Entrez IDs using mappings in Bioconductor for submission to LRPath. LRPath parameters were left at their default values.

Chapter 5: Conclusion

In this work, I have compared the strengths and weaknesses of emerging single-molecule NGS technology in contrast to an established method employing amplification in the context of a cancer gene expression study. From this, we note several broad conclusions; single-molecule methods appear to better sample the low-abundance genes in the transcriptome, and the experimental results may better represent the underlying distribution of abundances in the transcriptome. However, these advantages are quickly degraded by the rapid increase in sequencing capacity from competitive amplification based methods. Additionally, the Helicos methodology used remained at an average read length and yield disadvantage. It is not clear whether other single molecule methods may improve on these disadvantages. While we were able to clearly distinguish the TMPRSS2-ERG fusion prevalent in prostate cancers in the VCaP cell line, the longer read lengths of other methods are likely more advantageous in comparison.

The broader impact of this research in gene expression estimation from mRNA sequencing is most apparent in the contributions of gene expression data to other studies, as noted in the introduction. While many methodologies exist for the derivation of expression levels from RNA-seq data today [9, 156-160], this early methodology for RNA-seq data provided expression estimation before the majority of other methods had been made public.

The constantly evolving nature of massively parallel sequencing, and continuing development of single molecule methods, makes the insights extensible to future sequencing methods. While the SMS method applied in our study had read length and read volume disadvantages, future methods may sidestep these issues while retaining a sampling methodology free of amplification. Although the specifics of future developments in sequencing remain to be tested in depth, we can infer that future single-molecule based methods may exhibit more even coverage of transcripts being sequenced in addition to less concentrated sequencing of the very highest abundance transcripts.

Employing the knowledge and techniques developed for transcriptome profiling, we then aimed to integrate parallel transcriptome and proteome data. Specifically, to construct a standardized framework and methodology for the integration of knowledge from these two scales of biology, ensuring like comparisons between transcript and protein levels, with measurable parameters. This work addresses the issue of highly varying methods for integrating transcriptome and proteome data, a significant source of ambiguity in many previous integrative studies of these scales of biology.

The framework we developed utilizes a novel common reference of corresponding transcript and protein sequences such that enables the direct comparison of the transcript and protein abundances across thousands of genes while reducing the noise from isoform uncertainty. The development of a decoy transcript sequence based method for estimating false discovery rate in RNA-seq data as part of this effort enables direct management of noise levels from transcriptome data. By setting forth this standardized pipeline and methodology, we hope to increase comparability of these integrative experiments across multiple studies by reducing the cumulative effect of methodological variation.

We apply this framework to characterize the transcript-protein relationship in the VCaP and RWPE human prostate cell lines often used in cancer research. This research demonstrated the significant impact of transcript and protein stability on the transcript-protein relationship, and showed how this relationship is dysregulated in a number of functionally significant biological networks in our VCaP cancer model compared to our RWPE model of normal prostate epithelium. Several of these networks closely interact with the PI3K/Akt signaling pathway commonly seen to be deranged in cancers, where it is known to confer survival and growth advantage. Coupled with emerging knowledge that stabilization of transcripts or proteins are a pathway by which cancer cells sidestep regulatory mechanisms, we suggest that dysregulation of the transcript-protein relationship constitutes a possible mechanism by which cancers attain some of the hallmarks of cancer enabled by factors not related to genome stability.

While this work is a contribution to our understanding of the mechanisms that govern the transcriptome-proteome relationship and an examination of how well we interrogate the transcriptome, it is only a small step in building a comprehensive understanding of the complex interactions underlying the changes we see in biology and disease, and many challenges remain to be solved.

Assessment of the proteome still remains a challenge. While MS/MS technologies have made great strides in sensitivity, dynamic range, and capacity, they are still insufficient for the characterization of the proteome with its tremendous number of post-translational modifications. For example, our study (and other studies utilizing label-free MS/MS) do not measure phosphorylated versions of the proteins under study, despite phosporylation being a crucial component of molecular activation in cancers.

The ultimate goal of much of integrative bioinformatics is the derivation of methods and knowledge that can be translated to patients to optimize and improve the treatment of disease. This is already being explored in the realm of personalized medicine. While patient care has always been to a great extent personalized to individuals, we are now beginning to leverage the tremendous knowledge brought on by high throughput technologies. Next generation sequencing is seeing particular use in the cancer field, with an increasing focus on guiding patient therapy using insights from sequencing [73]. This is an result of the increasing application of multiple molecular "omic" technologies in cancer, referring to the genomic, transcriptomic, proteomic, and other methods of molecular characterization. The previous product of this trend was the development of prospective genomic signatures for cancer prognosis [161-164] seen in the past decade, a number of which have been commercialized and applied to patient care [165, 166].

The massive throughput and plummeting costs of these new technologies has also sparked increased efforts to quantify individuals' genetic backgrounds to profile disease risk. The most exhaustive of these is the iPOP, or Personal Omics Profiling effort, which fuses data from the transcriptome, genome, proteome, and metabolome [167]. Using data sampled over the course of 14 months from a single individual, the authors developed a disease risk profile based on observed variants in the patient's genome which ultimately revealed a high susceptibility to Type 2 Diabetes. This is only one example of how data from the multiple scales of biology can be fused to affect health outcomes positively, with the promise of improving the quality and lowering the cost of health care.

As genomic and proteomic profiling increase in capacity and become cheaper and more reliable, such exhaustive profiling of individuals will become commonplace. The proliferation of these molecular profiling efforts underscores the need for the development of methods for integrating diverse data such as those from the genome, transcriptome, and proteome. With advances in both technology and methods, we can begin to fully leverage the power of this multi-scale, multi-“omics” data revolution.

APPENDICES

APPENDIX A: Chapter 1 Supplementary Methods, Figures, and Tables

Methods

Data Extraction

The disease and biological process associated subnetworks are built from two fundamental components. First, a protein interaction network is used to define the relationships and interactions between the proteins considered in the study. The second is a database of genes relating them to diseases and biological processes.

Protein interaction data was retrieved from the Michigan Molecular Interaction Index (MiMi) [168], which integrates interaction and annotation data from BIND , the Gene Ontology, HPRD, DIP, the BioGRID, IntAct, InterPro, IPI, the Max-Delbrueck Center for Molecular Medicine protein interaction database, Pfam, ProtoNet, SwissProt, and RefSeq. This process yielded 12,318 unique protein-protein interactions involving 6199 unique Entrez Gene identifiers. Gene-disease relationships were derived from two sources; the Online Mendelian Inheritance in Man (OMIM) [169] and the PhenoGO database [170]. Gene-Disease associations in PhenoGO not using Entrez Gene identifiers were translated using mappings from HUGO [171]. Diseases in these two resources were defined in terms of coded Medical Subject Heading (MeSH) [172] and Unified Medical Language System (UMLS) [173] identifiers. The unfiltered, translated data set resulted in 3469 Entrez identifiers associated to 2325 phenotype codes. OMIM mappings found in the mim2gene file supplied by NCBI already employ Entrez Gene identifiers and no translation was necessary for the OMIM data. Entries in the OMIM database were filtered to include only gene-disease references, resulting in 1846 distinct Entrez indentified genes annotated to OMIM-defined diseases. 708 of the identifiers found in the OMIM mappings are also present in the MiMi interaction data set. Gene Ontology [174] data and biological annotation was extracted from BioMart [175] using data from Ensembl version 47 built from the NCBI36 release of the human genome. MeSH and UMLS term descriptors were retrieved directly from the NLM.

Data was extracted from MiMi using SQL queries for human-specific interactions from the National Center for Integrative Biomedical Informatics SQL server using SQL Server Management Studio Express.

Subnetwork Generation

The generated results were split into three distinct classes. A "background" set was generated from *a priori* knowledge from the Gene Ontology, consisting of the subnetworks formed by the classes represented in the "Biological Process" and "Molecular Function" trees of the Gene Ontology. This process resulted in the generation of 6,606 GO-associated subnetworks. A "single gene disease" (**SGD**) subnetwork set was generated from the contents of OMIM, producing 2,079 subnetworks. A "complex disease" (**CD**) set was built from the PhenoGO annotations, composed of 2,317 subnetworks in total.

We separate the OMIM and PhenoGO sets for two reasons. The primary factor for the separation is the drastically different underlying focus of both of these resources, although they do share some commonly annotated diseases. PhenoGO contains data describing both single gene and multi-gene complex disease, whereas OMIM is primary focused on single gene diseases. The secondary factor is curation; the OMIM data is manually curated while PhenoGO is a computationally derived data source.

Derivation of the subnetworks was done using the Boost Library version 1.43.1 (http://www.boost.org/) and version .9 of the Boost Graph Library bindings to Python (http://osl.iu.edu/~dgregor/bgl-python/) using ActiveState ActivePython version 2.4.3 (http://www.activestate.com/).

Subnetworks that resulted in errors in the software were removed from the set, as the memory requirements for processing a number of large, dense networks was beyond the memory capacity of our workstation.

Data Characterization and Filtering

Resulting subnetworks in each of the three data sets was topologically characterized using a set of Perl scripts employing the Boost Graph Library interface. Subnetworks are topologically characterized based on node count, clustering coefficient, observed edge fraction, average degree, maximum degree, radius, diameter, cyclicity, and biconnectivity. Biological characteristics noted for each subgraph include mean gene start location, mean gene end location, mean length, strand, mean PFAM domain annotation count, mean ProSite annotation count, mean number of signal domains, mean number of transmembrane domains, and mean G-C content fraction. The networks are filtered for size, imposing a minimum of three nodes found in the interaction network. 79 and 278 subnetworks passed this filter from the SGD and CD sets, respectively. 2590 of the subnetworks generated from the Gene Ontology passed this filter. This final filtered set was used to train and test the classifier.

Because the data in the PhenoGO resource spans drugs, cell types, and other biological contexts not directly associated with disease, the subnetworks formed by this resource were filtered using the UMLS metathesaurus. Therefore, only genes associated with MeSH and UMLS terms are used to create the subnetworks. To restrict the set, a list of UMLS and MeSH codes was derived using a Perl script containing a total of unique terms. Of the 423,550 terms in the UMLS and MeSH that met these rules, the UMLS composed 419,087 terms and MeSH composed 5,563 terms. This process of restricting the set yielded a dramatic reduction in the number of subnetworks in the disease set.

The data from the biological and topological characterization for each of the classes was then filtered for size using a perl script, constraining the set to networks of size between 3 and 9999 nodes. 79 and 278 subnetworks passed this filter from the OMIM and PhenoGO sets, respectively. 2590 of the subnetworks generated from the Gene Ontology passed this filter.

Parameterization/Characterization of Subnetworks

To characterize subnetworks structurally, we chose a number of well-defined metrics to measure their size, density, and connectivity. Subnetworks are characterized based on node count, clustering coefficient, average degree, maximum degree, radius, diameter, cyclicity, and biconnectivity. Cyclicity and biconnectivity are handled as Boolean variables with values of either 1 (True) or 0 (false). To account for the biological characteristics of the constituent genes of these subnetworks, we use biological characteristics for the constituent genes extracted from BioMart. These factors accounted for positional and orientation effects, biological role of the protein product, and physical stability. Factors include mean gene start location, mean gene end location, mean length, strand, mean PFAM domain annotation count, mean ProSite annotation count, mean number of signal domains, mean number of transmembrane domains, and mean G-C content fraction.

Parameterization of subnetworks was done using a series of Perl scripts using the Perl-Graph library version .84 (http://search.cpan.org/dist/Graph/) as well as the Boost Graph Library Bindings for Perl version 1.4 (http://search.cpan.org/~dburdick/Boost-Graph-1.4/). These libraries were used to determine the topological characteristics of each of the subnetworks. Factors include the average degree, maximum degree, node count, radius, and diameter for each subnetwork. Each subnetwork was also tested for cyclicity and biconnectivity.

During the parameterization process, a number of entries were removed from the set as the subnetworks they formed were not computable within the memory limits of our workstation. These classes are GO:0007218 : “neuropeptide signaling pathway”, GO:0045893: “positive regulation of transcription, DNA-dependent”, and GO:0006937: “regulation of muscle contraction”.

Machine Learning and Classification

The Waikato Environment for Knowledge Analysis (Weka), version 3.4.12 [176] was used to train and test a random forest classifier with a stratified 10-fold cross validation methodology using the built-in weka.classifiers.trees.RandomForest component. In this case, the cross-validation approach was chosen due to the relative paucity of data from the disease subsets. Each random forest was composed of 100 trees, each taking into account four random parameters from the data. In all, a total of nine classifications were done in an attempt to discretize the three sets of subnetworks using varying parameter sets and amalgamations of the two disease sets. Because the Weka random forest classifier did not provide variable importance measures, the analysis was repeated using the randomForest package in R 2.7.1, which provided nearly identical results. Principal components analysis of the data was done using PAST [177].

The parameterized data was split into 3 sets for the biological and topological groups. The first set composed of all three data sources comprising three distinct classes. The second set assigned “normal” and “disease” flags to the subnetworks derived from the Gene Ontology, and OMIM and PhenoGO, respectively. The third subset was composed of only disease subnetworks derived from OMIM while maintaining the GO background set.

The first classification was done on a set combining all SGD and CD subnetworks into a single larger disease class in comparison to the GO-derived background set. The second classification used only the SGD subset of the data in comparison to the GO data. The third classification used each subset of data in its own discrete class. These subsets were further separated into three groups depending on the underlying parameters available to the classifier. These groups used parameters exclusively from the topological and biological parameter sets, as well as the combined parameterization.

It can be seen that overall the biological characteristics prove more informative than the topological ones and achieve a lower misclassification error rate, ranging between 2.89 and 3.70%. On the other hand, for the topological characteristics the misclassification error rate was around 10% for the three class problem. However, when the CD class was excluded, the topological characteristics matched the performance of the biological ones. Further, an inspection of **Sup. Tables 2e and 2f** suggests that the presence of the SGD class is the source of the significantly higher misclassification error rate with respect to the topological features. In most cases, the presence of the large number of representative GO subnetworks leads to a high classification accuracy. However, it is useful to examine the true positive (**TP**) rate of classification between the combined “disease” set, a combination of the SGD and CD sets, and the GO background. In the combined parameterization and biological parameter only cases, the TP rate of this combined set is relatively good, at 61% and 72%, respectively. Examination of the TP rates for classifying into the three distinct classes revels that the subnetworks in the SGD set appear to be poorly distinguishable from the background GO set. However, the CD set appears to have predictive power setting it apart from the GO background. This similarity between the GO and SGD sets likely leads to the poor classification accuracy seen between the two sets as reflected in the poor TP values for the SGD set in **Sup. Tables 2e, 2f, 2h, and 2i**.

Feature Analysis

A factor analysis was done using the RandomForest package in R 2.7.1 in each of the biological parameter only, topological parameter only, and combined parameter groups to determine the relative influence of each of the parameters in determining class membership in each of the classification sets. The random forest was set to use 4 variables per tree and 100 total trees for the classification task.

Figures and Tables

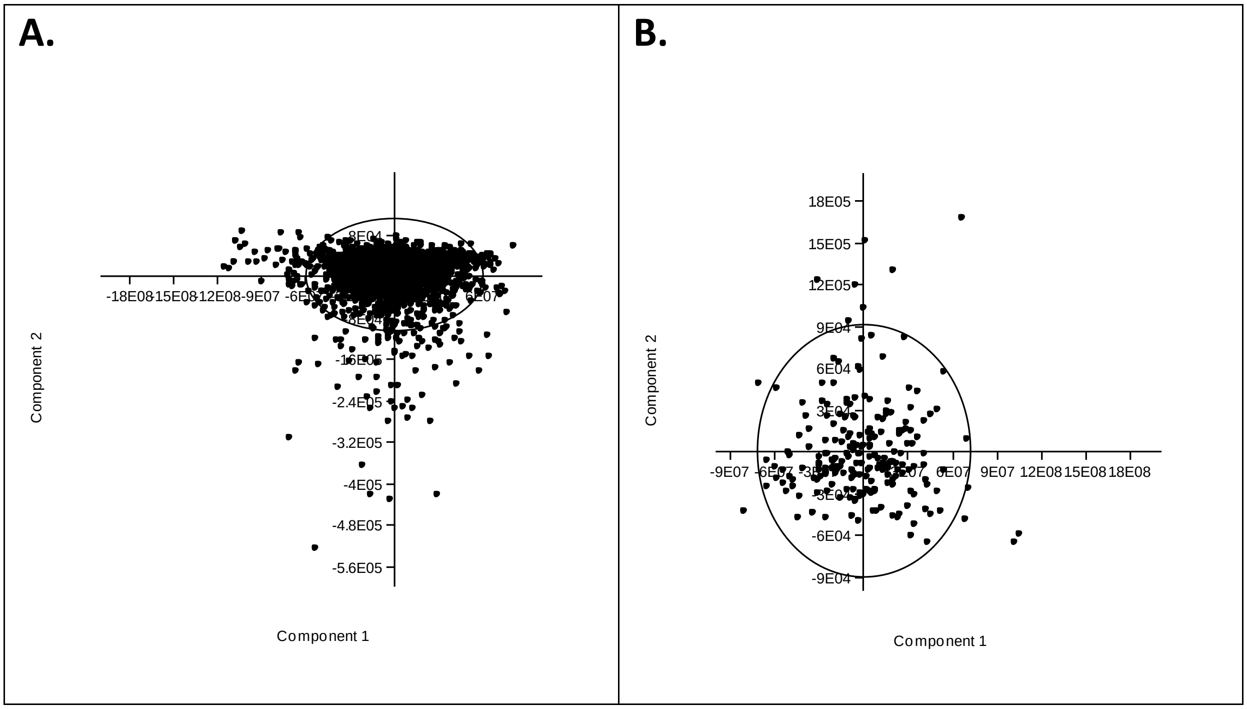


Figure S 1: Principal Components Analysis demonstrates the poor separability of the data. A principal components analysis of the combined sets using all the parameters, suggests that the difference between disease-related subnetworks and the GO baseline subnetworks are subtle and not easily derived. When the PCA is done over just the CD and SGD sets, we see a similar pattern where there is no clear separation. However the non-continuous nature of the features may be a confounding factor when applying the PCA approach. With that in mind, a simple k-means clustering approach was taken where k = 3 to represent the three source types. **A.** Principal components analysis of all sets using all parameters. 95% of data points fall within the ellipse. **B.** Principal components analysis of SGD and CD sets using all parameters. 95% of data points fall within the ellipse.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 3 -S 10
Relation:     combined_data
Instances:    2944
Attributes:   20
              average gene start
              average gene end
              average length
              average gene strand
              average pfam count
              average prosite count
              average # of singnal domains
              average # transmembrane domains
              average GC content
              observed edges/total possible edges
              average node degree
              max node degree
              radius
              diameter
              node count
              cyclicity
              biconnectivity
              clustering coefficent
Ignored:
            source
            phenotype code
Test mode:    Classes to clusters evaluation on training data
=== Model and evaluation on training set ===

kMeans
======
Number of iterations: 6
Within cluster sum of squared errors: 1660.859140812153

Cluster centroids:
```

Variables	Cluster 0		Cluster 1		Cluster 2	
Variable	**Mean/Mode**	**Std Devs**	**Mean/Mode**	**Std Devs**	**Mean/Mode**	**Std_Devs**
average gene start	**70562607**	21895436	**72069986**	8353007	**71199760**	12743762
average gene end	**70623972**	21898365	**72141696**	8355198	**71264921**	12743801
average length	**61364.07**	39100.1	**71710.6**	34510.64	**65160.52**	32673.26
average gene strand	**0.2259**	0.4311	**0.0898**	0.1649	**0.1195**	0.2538
average pfam count	**26.3999**	48.5588	**26.908**	16.715	**25.0051**	20.9655
average prosite count	**26.3999**	48.5588	**26.908**	16.715	**25.0051**	20.9655
average # of singnal domains	**0.1312**	0.1977	**0.156**	0.1162	**0.1235**	0.1314
average # transmembrane domains	**0.1335**	0.2008	**0.1715**	0.1121	**0.1415**	0.1412
average GC content	**43.1182**	3.0456	**41.7223**	1.2326	**42.2927**	2.0194

observed edges/total possible edges	**0.318**	0.1051	**0.0559**	0.0477	**0.1336**	0.0608
average node degree	**2.173**	0.6153	**4.417**	1.2629	**3.3774**	0.9891
max node degree	**4.2338**	1.9778	**60.4362**	70.3251	**12.8042**	13.2817
radius	**2**	N/A	**4**	N/A	**3**	N/A
diameter	**3.199**	0.8274	**7.0021**	1.1488	**5.0434**	0.724
node count	**5.6166**	3.0982	**151.7489**	185.4197	**26.2334**	29.3995
cyclicity	**0.7564**	0.4294	**0.9936**	0.0797	**0.9711**	0.1677
biconnectivity	**0.0237**	0.152	**0.0064**	0.0797	**0.0222**	0.1473
clustering coefficent	**0.0207**	0.0386	**0.0204**	0.0391	**0.0202**	0.0387
Clustered Instances	**1437 (49%)**		**470 (16%)**		**1037 (35%)**	

Class attribute: source

		Assigned to Cluster		
		Cluster 0 <-- GO	**Cluster 1 <-- OMIM**	**Cluster 2 <-- PhenoGO**
Source	**SGD/OMIM**	59	4	16
	GO	1220	435	932
	CD/PhenoGO	158	21	89

Incorrectly clustered instances : 1631.0 55.4008 %

Table S 1: Complete results of unsupervised k-means clustering of the data.

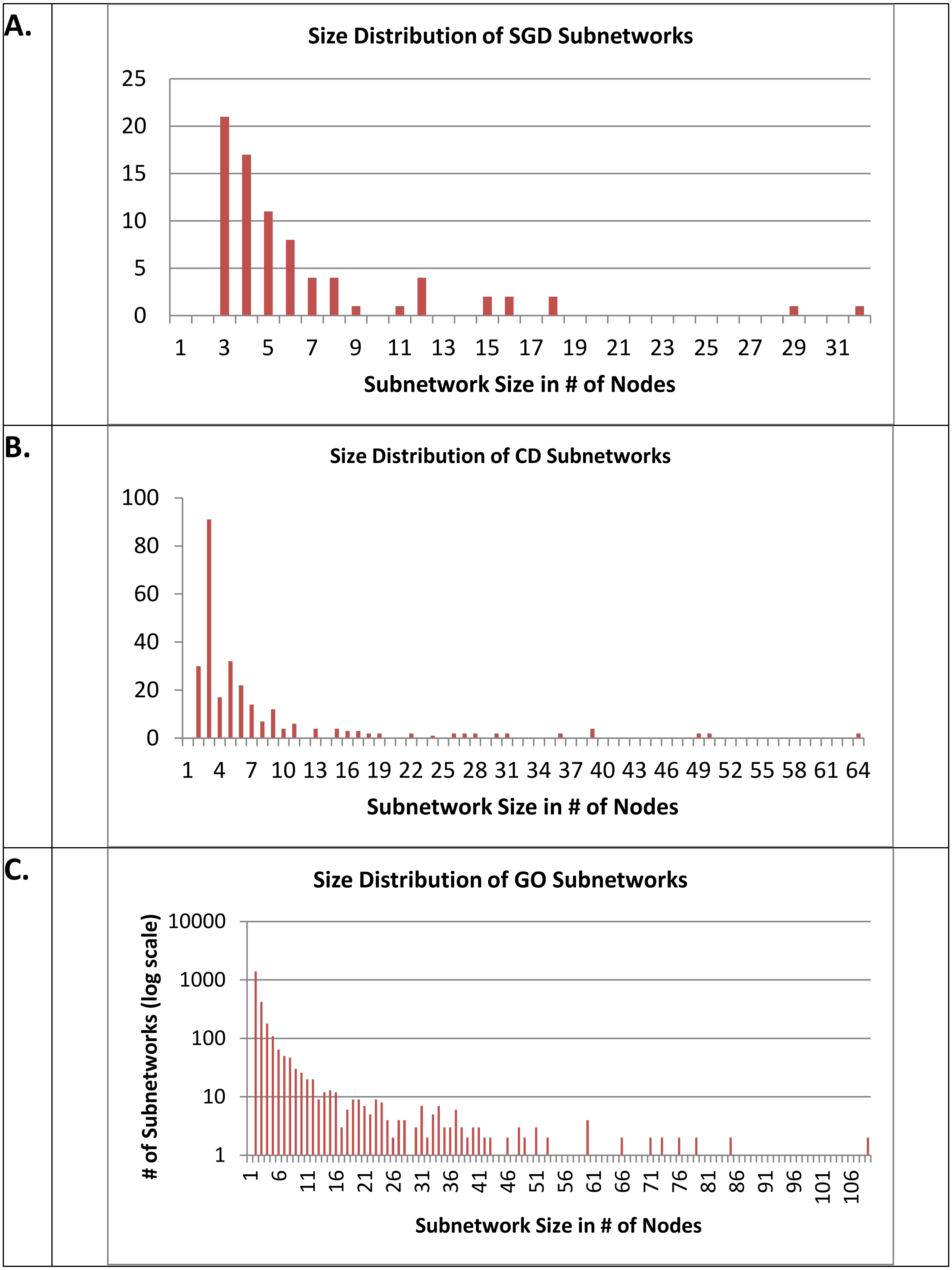


Figure S 2: A. Size Distribution of SGD Subnetworks **B**. Size Distribution of CD Subnetworks **C.** Size Distribution of GO Subnetworks

Table S 2: Classification results from each of nine classification attempts using complete GO set

Biological Parameters Only

A. Biological parameters only: dataset split into “disease” and “normal” classes

Out of bag error: 0.0309

Correctly Classified Instances		2836		96.2988 %	
Incorrectly Classified Instances		109		3.7012 %	
Kappa statistic		0.8064			
Mean absolute error		0.1287			
Root mean squared error		0.216			
Relative absolute error		60.339 %			
Root relative squared error		66.1667 %			
Total Number of Instances		2945			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.995	0.272	0.964	0.995	0.979	GO
0.728	0.005	0.956	0.728	0.827	Disease

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
2576	12	**a = GO/Normal**
97	260	**b = Disease**

B. Biological parameters only: dataset split into CD, SGD, and GO classes

Out of bag error: 0.0309

Correctly Classified Instances		2832		96.163 %	
Incorrectly Classified Instances		113		3.837 %	
Kappa statistic		0.8008			
Mean absolute error		0.0893			
Root mean squared error		0.1801			
Relative absolute error		61.2569 %			
Root relative squared error		66.7931 %			
Total Number of Instances		2945			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.165	0.003	0.565	0.165	0.255	SGD
0.867	0.001	0.992	0.867	0.925	CD
0.996	0.283	0.962	0.996	0.979	GO

Confusion Matrix:

Classified as:			
a	**b**	**c**	**Actual assignment**
2578	1	9	**a = GO**
36	241	1	**b = CD**
65	1	13	**c = SGD**

C. Biological parameters only: SGD and GO classes

Out of bag error: 0.0274

Correctly Classified Instances		2590		97.1129 %	
Incorrectly Classified Instances		77		2.8871 %	
Kappa statistic		0.1974			
Mean absolute error		0.0527			
Root mean squared error		0.1661			
Relative absolute error		91.1176 %			
Root relative squared error		97.9961 %			
Total Number of Instances		2667			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.127	0.003	0.556	0.127	0.206	SGD
0.997	0.873	0.974	0.997	0.985	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
2580	8	**a = GO**
69	10	**b = SGD**

Topological Parameters Only

D. Topological Parameters Only: dataset split into “disease” and “normal” classes

Out of bag error: 0.0853

Correctly Classified Instances	2675	90.8628 %
Incorrectly Classified Instances	269	9.1372 %
Kappa statistic	0.4646	
Mean absolute error	0.1475	
Root mean squared error	0.2732	
Relative absolute error	69.1481 %	
Root relative squared error	83.7012 %	
Total Number of Instances	2944	

TP Rate	FP Rate	Precision	Recall	f-Measure	class
0.392	0.02	0.729	0.392	0.51	Disease
0.98	0.608	0.921	0.98	0.95	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
2535	52	**a = GO/Normal**
217	140	**b = Disease**

E. Topological Parameters Only: dataset split into CD, SGD, and GO classes

Out of bag error: 0.0832

Correctly Classified Instances	2688	91.30%
Incorrectly Classified Instances	256	8.70%
Kappa statistic	0.4863	
Mean absolute error	0.1016	
Root mean squared error	0.2241	
Relative absolute error	69.7015 %	
Root relative squared error	83.1102 %	
Total Number of Instances	2944	

TP Rate	FP Rate	Precision	Recall	f-Measure	class
0.038	0.004	0.214	0.038	0.065	SGD
0.493	0.011	0.83	0.493	0.619	CD
0.985	0.608	0.922	0.985	0.952	GO

Confusion Matrix:

Classified as:			
a	**b**	**c**	**Actual assignment**
2548	28	11	**a = GO**
141	137	0	**b = CD**
76	0	3	**c = SGD**

F. Topological Parameters Only: SGD and GO classes

Out of bag error: 0.0315

Correctly Classified Instances	2581	96.81%
Incorrectly Classified Instances	85	3.19%
Kappa statistic	0.0586	
Mean absolute error	0.0543	
Root mean squared error	0.1716	
Relative absolute error	93.8315 %	
Root relative squared error	101.201 %	
Total Number of Instances	2666	

TP Rate	FP Rate	Precision	Recall	f-Measure	class
0.038	0.003	0.25	0.038	0.066	SGD
0.997	0.962	0.971	0.997	0.984	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
2578	9	**a = GO**
76	3	**b = SGD**

Combined Parameterization

G. All parameters: dataset split into “disease” and “normal” classes

Out of bag error: 0.0452

Correctly Classified Instances		2791		94.803 %	
Incorrectly Classified Instances		153		5.197 %	
Kappa statistic		0.7128			
Mean absolute error		0.1269			
Root mean squared error		0.2191			
Relative absolute error		59.5021 %			
Root relative squared error		67.1287 %			
Total Number of Instances		2944			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.611	0.005	0.94	0.611	0.74	Disease
0.995	0.389	0.949	0.995	0.971	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
218	139	**a = Disease**
14	2573	**b = GO/Normal**

H. All parameters: dataset split into CD, SGD, and GO classes

Out of bag error: 0.0438

Correctly Classified Instances		2795		94.9389 %	
Incorrectly Classified Instances		149		5.0611 %	
Kappa statistic		0.7225			
Mean absolute error		0.0886			
Root mean squared error		0.1815			
Relative absolute error		60.7398 %			
Root relative squared error		67.2984 %			
Total Number of Instances		2944			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.101	0.003	0.5	0.101	0.168	SGD
0.997	0.387	0.949	0.997	0.972	GO
0.752	0.001	0.986	0.752	0.853	CD

Confusion Matrix:

Classified as:			
a	**b**	**c**	**Actual assignment**
8	70	1	**a = SGD**
7	2578	2	**b = GO**
1	68	209	**c = CD**

I. All parameters: SGD and GO classes

Out of bag error: 0.0281

Correctly Classified Instances		2591		97.1868 %	
Incorrectly Classified Instances		75		2.8132 %	
Kappa statistic		0.2332			
Mean absolute error		0.0498			
Root mean squared error		0.1594			
Relative absolute error		86.0831 %			
Root relative squared error		93.9883 %			
Total Number of Instances		2666			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.152	0.003	0.6	0.152	0.242	SGD
0.997	0.848	0.975	0.997	0.986	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
12	67	**a = SGD**
8	2579	**b = GO**

J. All parameters: SGD and CD classes

```
=== Run information ===
Scheme:      weka.classifiers.trees.RandomForest -I 100 -K 4 -S 1
Relation:    OMIM-PhenoGO-weka.filters.unsupervised.attribute.Remove-R2
Instances:   357
Attributes:  19
             source
             average gene start
             average gene end
             average length
             average gene strand
             average pfam count
             average prosite count
             average # of signal domains
             average # transmembrane domains
             average GC content
             observed edges/total possible edges
             average node degree
             max node degree
             radius
             diameter
             node count
             cyclicity
             biconnectivity
             clustering coefficient
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===
Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.1232
```

Correctly Classified Instances	315	88.2353 %
Incorrectly Classified Instances	42	11.7647 %
Kappa statistic	0.5965	
Mean absolute error	0.1785	
Root mean squared error	0.2972	
Relative absolute error	51.6603 %	
Root relative squared error	71.5991 %	
Total Number of Instances	357	

TP Rate	FP Rate	Precision	Recall	f-Measure	class
0.519	0.014	0.911	0.519	0.661	SGD
0.986	0.481	0.878	0.986	0.929	CD

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
274	4	**a = CD**
38	41	**b = SGD**

Table S 3: Ranked Features By Parameter Type. A. Biological Parameters Only **B.** Topological Parameters Only **C.** Combined Parameterization

A.

```
                                       GO  SGD/OMIM CD/PhenoGO MeanDecreaseAccuracy MeanDecreaseGini
averageGeneStart                0.2783482 1.0280783 0.9059960            0.2757494         84.56684
averageGeneEnd                  0.2768157 0.9394527 0.8925733            0.2747467         82.32455
averageLength                   0.2644807 1.2301754 0.9510359            0.2876197         89.97404
averageGeneStrand               0.1758904 0.1357294 0.9539724            0.2776031         63.51283
averagePfamCount                0.2730130 0.5254745 0.8856997            0.2717815         68.71366
averagePrositeCount             0.2732054 0.7780531 0.8667791            0.2729219         71.44485
averageSingnalDomainCount       0.2126032 1.1321489 0.9215645            0.2744301         46.04487
averageTransmembraneDomainsCount 0.2369126 0.7511460 0.9107138            0.2746473         41.26618
averageGCContent                0.2527932 1.1863229 0.9633071            0.2872784         90.52120
```

B.

```
                              GO   SGD/OMIM  CD/PhenoGO MeanDecreaseAccuracy MeanDecreaseGini
observedEdgeFraction 0.23001163  0.5940764 0.90312482           0.24675347        93.847995
averageNodeDegree    0.18907358 -0.1896722 0.92494854           0.25118579        73.325193
maxNodeDegree        0.23248537 -0.0195584 0.75146118           0.23964507        45.595834
radius               0.14363009  0.3341730 0.73797260           0.17620126        10.558500
diameter             0.16504637  0.3258433 0.89950612           0.21990106        24.283709
nodeCount            0.24716779  0.1174077 0.62814917           0.24756213        47.349672
cyclicity            0.07668406  0.1599157 0.05666838           0.08233893         2.229017
biconnectivity       0.05281318  0.2182699 0.47637630           0.10961336         3.538654
clusteringCoefficent 0.28966769  0.9925351 0.96101890           0.28810431        97.553541
```

C.

```
                                         GO  SGD/OMIM CD/PhenoGO MeanDecreaseAccuracy MeanDecreaseGini
averageGeneStart                 0.25577147 0.6187922 0.8782965            0.2631096        58.025555
averageGeneEnd                   0.24189366 0.8649050 0.8823725            0.2517155        54.866536
averageLength                    0.21860181 1.0476172 0.9157395            0.2702029        53.928221
averageGeneStrand                0.21222727 0.4779712 0.8899448            0.2613027        37.971447
averagePfamCount                 0.24589871 0.7138733 0.8139401            0.2557329        51.837923
averagePrositeCount              0.24653767 0.8026352 0.8288924            0.2553449        51.873560
averageSingnalDomainCount        0.17608440 0.8725259 0.8494207            0.2504462        28.695867
averageTransmembraneDomainsCount 0.17643006 0.8016404 0.8587388            0.2398903        25.758543
averageGCContent                 0.20630777 1.0249891 0.9042456            0.2621500        57.568889
observedEdgeFraction             0.22721854 0.9567640 0.8553682            0.2424491        39.992423
averageNodeDegree                0.24357245 0.6044357 0.8350696            0.2586311        33.451690
maxNodeDegree                    0.23311884 0.5687222 0.7704013            0.2418791        23.089282
radius                           0.19372018 0.5507024 0.5725879            0.1942285         9.303571
diameter                         0.22263432 0.7683270 0.7232573            0.2295851        16.473967
nodeCount                        0.23954530 0.7925986 0.8041791            0.2430081        25.501844
cyclicity                        0.11050759 0.2201386 0.5157050            0.1559355         3.013125
biconnectivity                   0.07642597 0.1890160 0.2993280            0.1074229         1.420956
clusteringCoefficent             0.26042896 1.4008805 0.8991804            0.2705517        61.914586
```

APPENDIX B: Chapter 2 Supplementary Figures and Tables

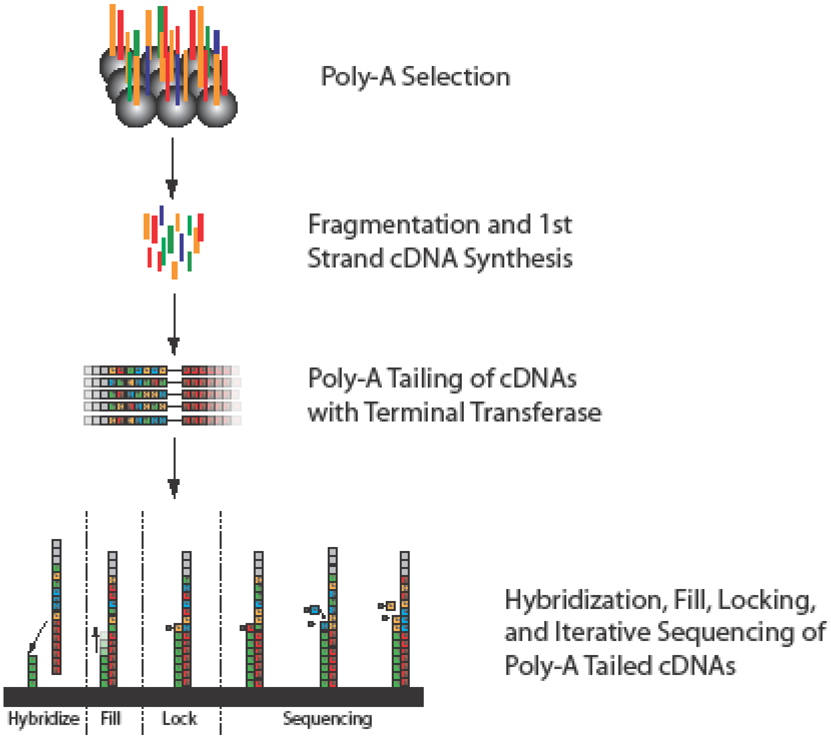
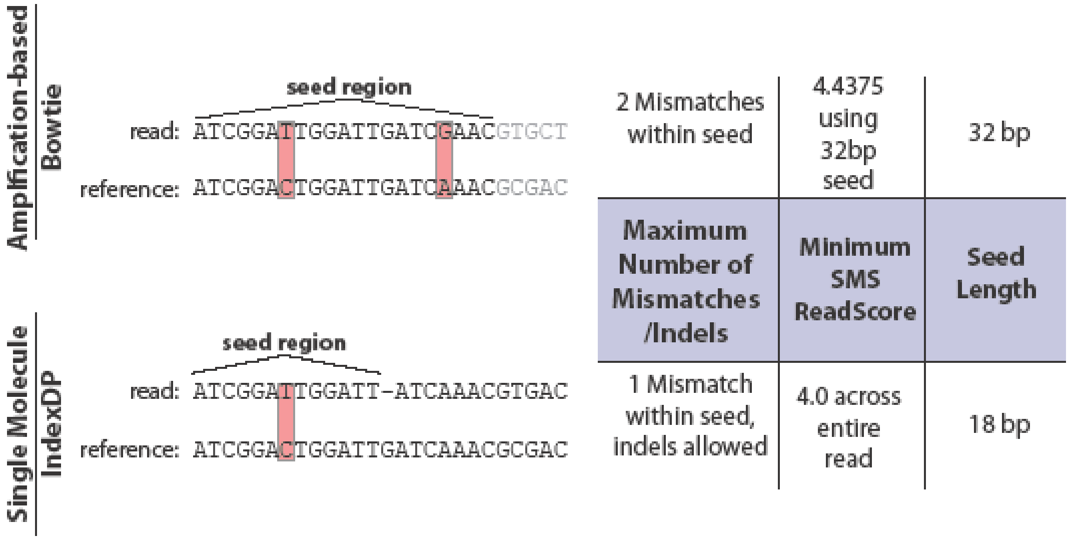

Poly-A Selection
Fragmentation and 1st Strand cDNA Synthesis
Poly-A Tailing of cDNAs with Terminal Transferase
Hybridization, Fill, Locking, and Iterative Sequencing of Poly-A Tailed cDNAs
Hybridize
Fill
Lock
Sequencing

Figure S 3: Single-molecule mRNA-sequencing. mRNAs are purified using poly-A selection and then fragmented. 1st-strand cDNA is synthesized from the fragmented mRNA, and then poly-A tailed using terminal transferase. Polyadenylated cDNA fragments are hybridized to poly-T oligomers bound to a glass substrate, excess A bases are "filled ," and then "locked" with an A, C, or G base attached to a virtual terminator. The sequencing process then occurs with repeated cycles of virtual terminator cleavage, bases addition, and image readout.


Amplification-based
Bowtie
seed region
read: ATCGGATTGGATTGATCGAACGTGCT
reference: ATCGGACTGGATTGATCAAACGCGAC
Single Molecule
IndexDP
seed region
read: ATCGGATTGGATT-ATCAAACGTGAC
reference: ATCGGACTGGATTGATCAAACGCGAC

Maximum Number of Mismatches /Indels	Minimum SMS ReadScore	Seed Length
2 Mismatches within seed	4.4375 using 32bp seed	32 bp
1 Mismatch within seed, indels allowed	4.0 across entire read	18 bp

Figure S 4: Read alignment with Bowtie and IndexDP. Bowtie was used for amplification-based sequencing read alignment and IndexDP for single molecule read alignment. While different in their parameters, the effective alignments and specificity between the aligners are similar, although Bowtie has a slightly higher cutoff

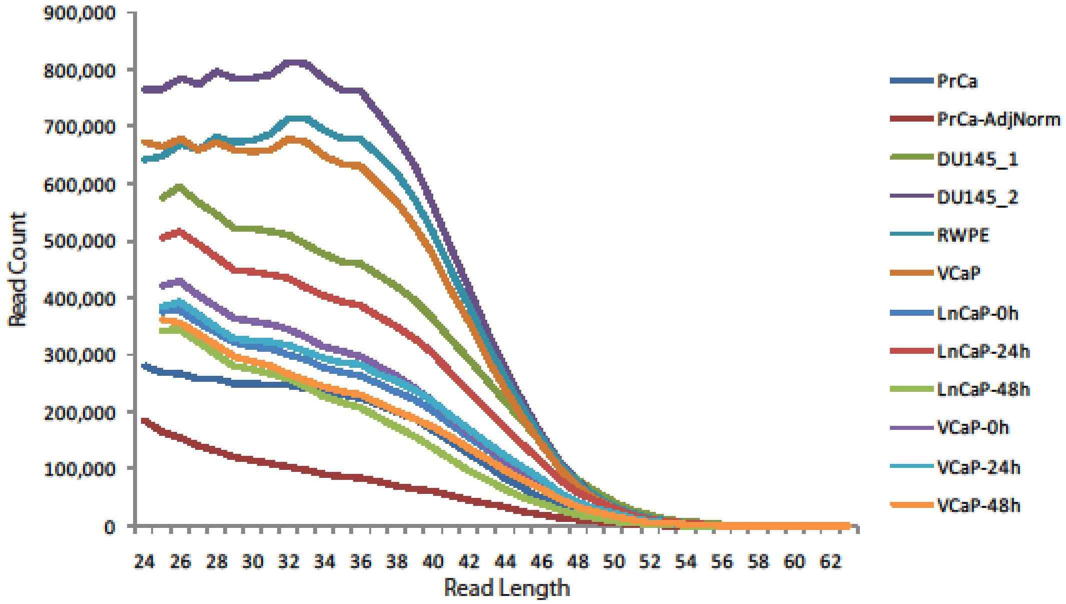


Figure S 5: Length distribution of aligned SMS reads. Aligned SMS read lengths varied between 24bp to 57bp in our first set of samples and 25bp to 63bp in our second set. The majority of reads are between 25bp and 45bp in length.

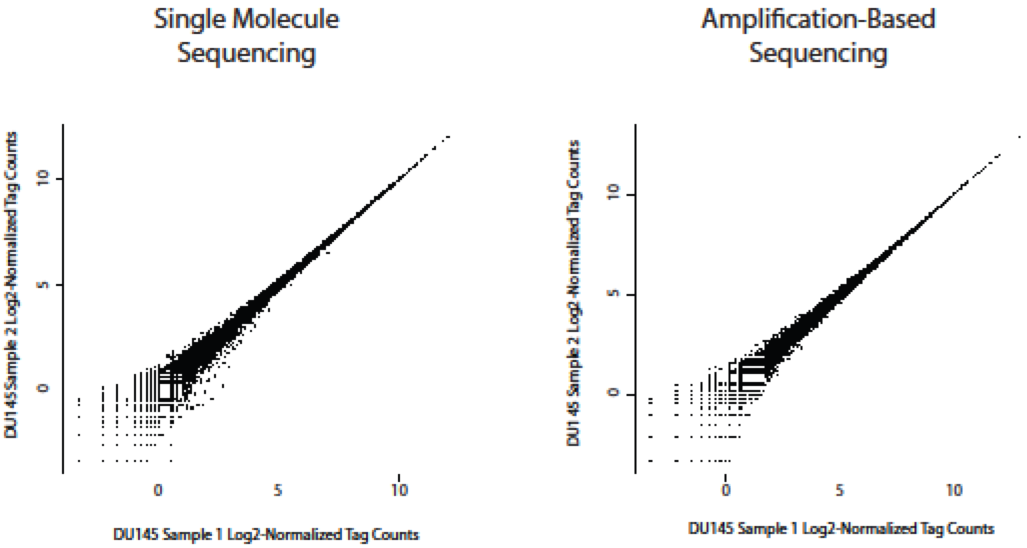


Figure S 6: Sample Profiling Reproducibility in SMS and AS. Bowtie was used for amplification-based sequencing read alignment and IndexDP for single molecule read alignment. Pearson correlation for log2-transformed, normalized tag counts is r=0.98 for both SMS and AS.

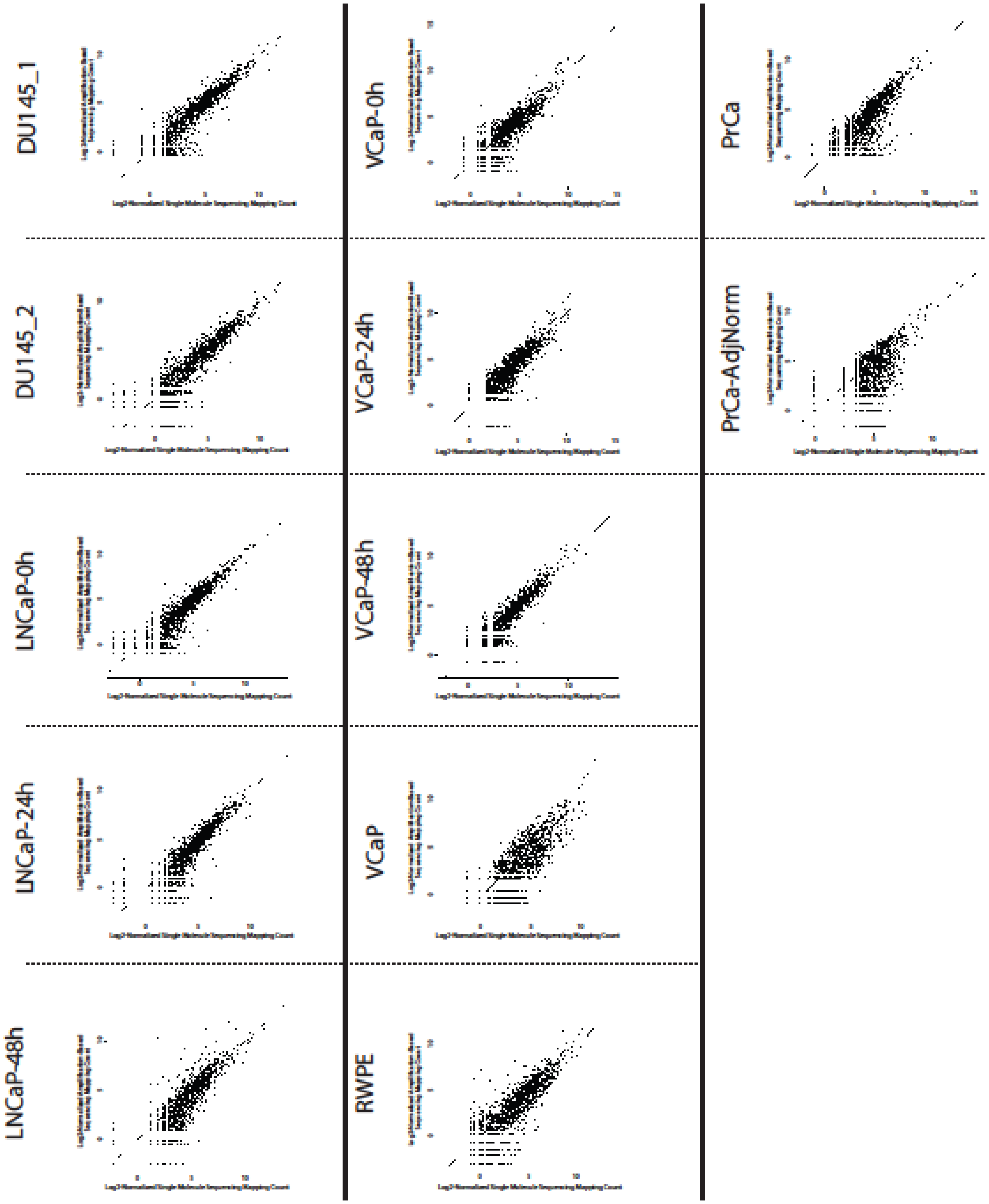


Figure S 7: Log2 correlation between amplification-based and single-molecule sequencing. Log2 correlation between single-molecule and amplification-based RNA-Seq single-best read mappings in these samples show that in broad terms the two sequencing methods yield similar results, suggesting the observed bias is not due to sample differences.

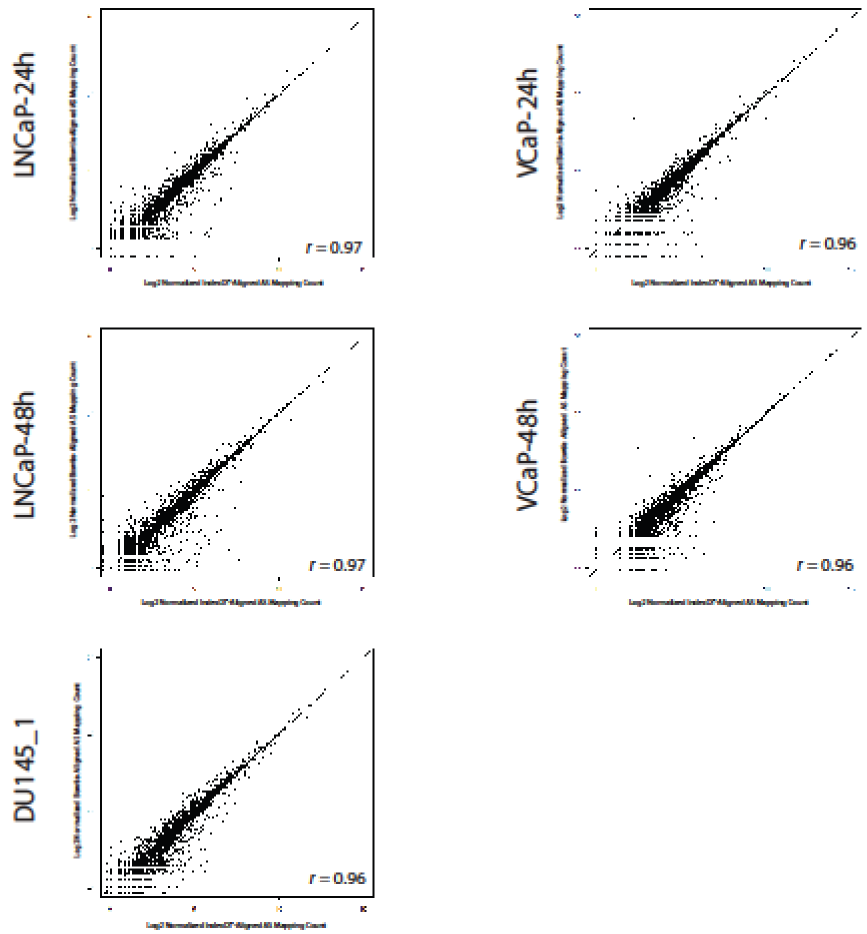


Figure S 8: Correlation between IndexDP and Bowtie alignment of amplification-based sequencing reads. The correlation between Bowtie and IndexDP within the subset of samples was relatively high, with Pearson correlation values above r=0.95 in all samples.

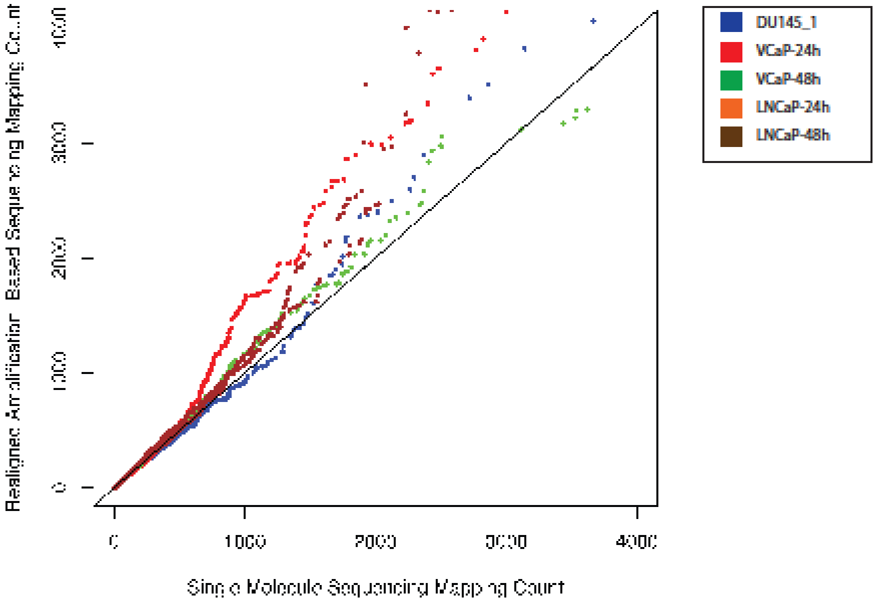


Figure S 9: IndexDP realignment of amplification-based sequencing reads. Alignment of amplification-based sequencing reads using the IndexDP alignment tool used to align single-molecule reads shows persistence of the observed bias in amplification-based technology. This provides evidence that the alignment method is not responsible for this bias towrds high-concentration transcripts.

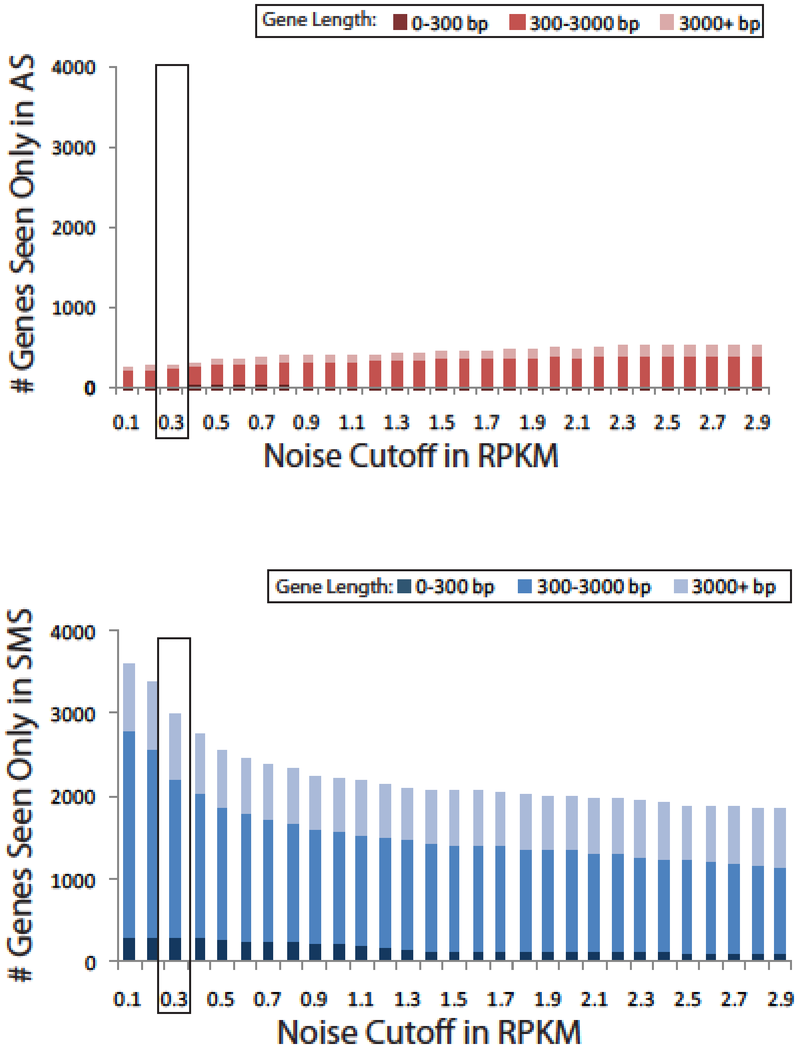


Figure S 10: Unique gene detection in AS and SMS across threshold values, by transcript length. The pattern of increased sensitivity in SMS is uniform as the baseline noise level is varied from 0.1 to 3.0 RPKM. Low representation by short transcripts show that this effect is not due to the lack of a size-selection step in SMS.

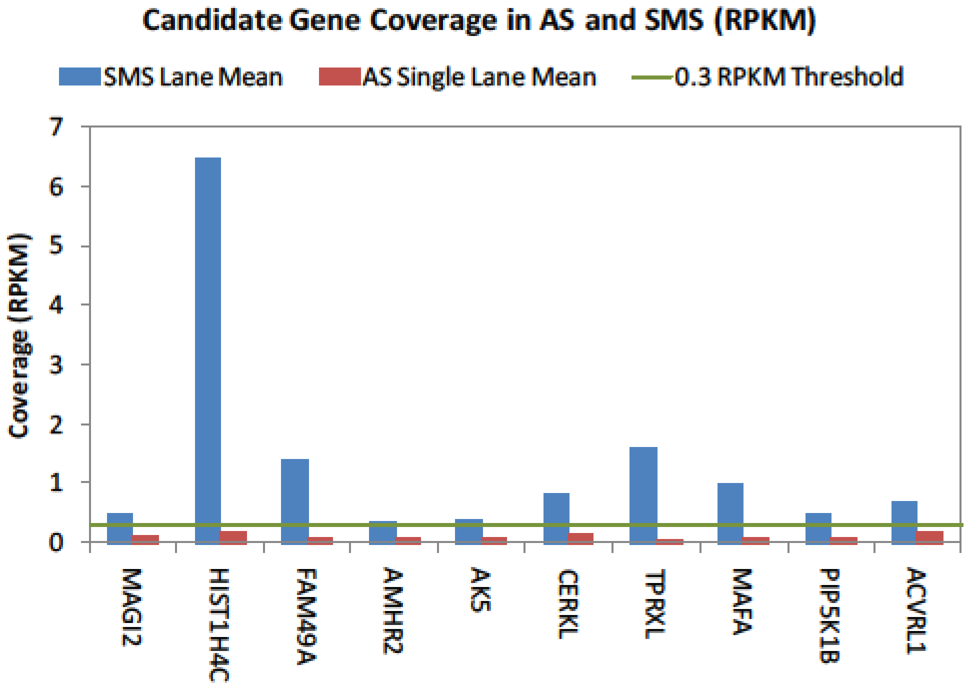


Figure S 11: Expression values of validation candidate genes showing amplification. Out of the set of genes chosen for RT-PCR validation for their detection over the 0.3 RPKM noise threshold by only SMS, diffuse read alignment pattern, and the presence of long reads aligned to their transcripts, these ten genes showed detectable amplification.

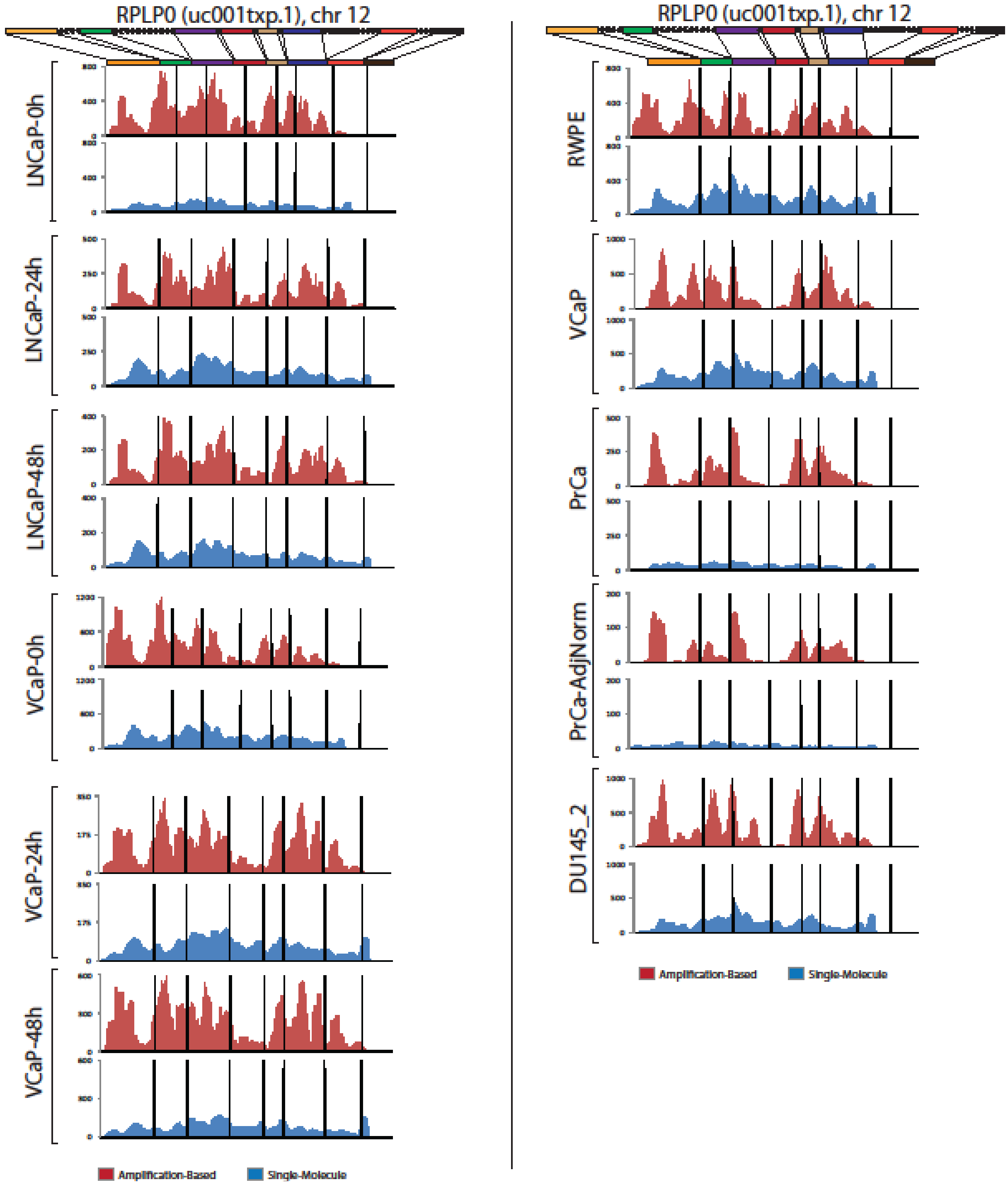


Figure S 12: RPLP0 coverage in other samples. Coverage plots of the over-represented gene RPLP0 in the LNCaP-24h, LNCaP-48h, VCaP-24h, VCaP-48h, and PrCa-Met samples show that this gene is often more highly sequenced using the amplification-based method.

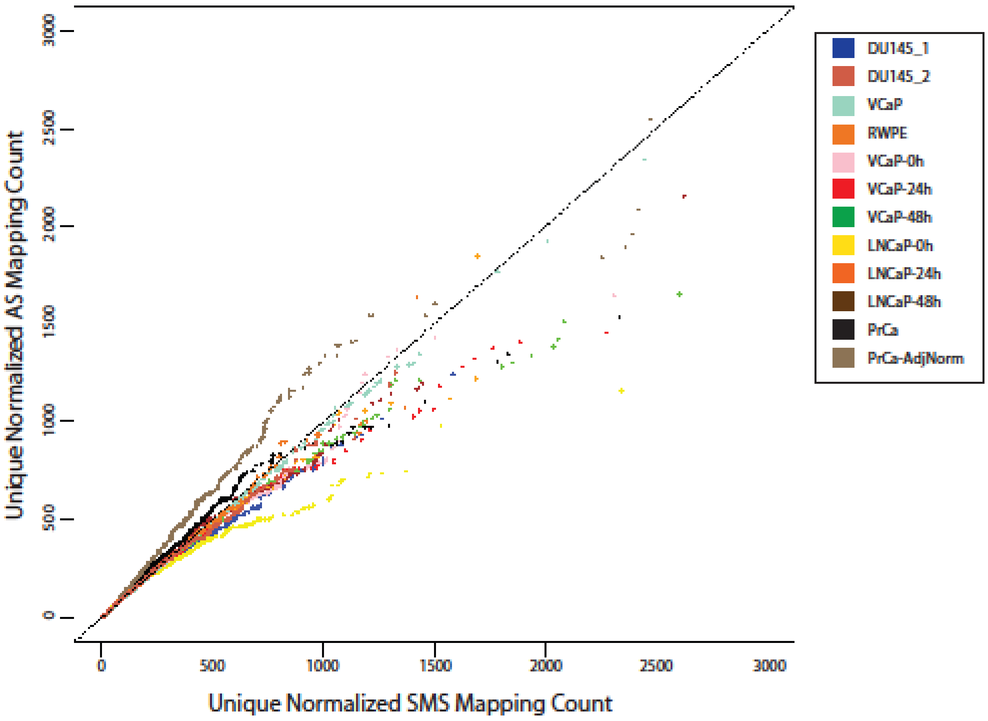
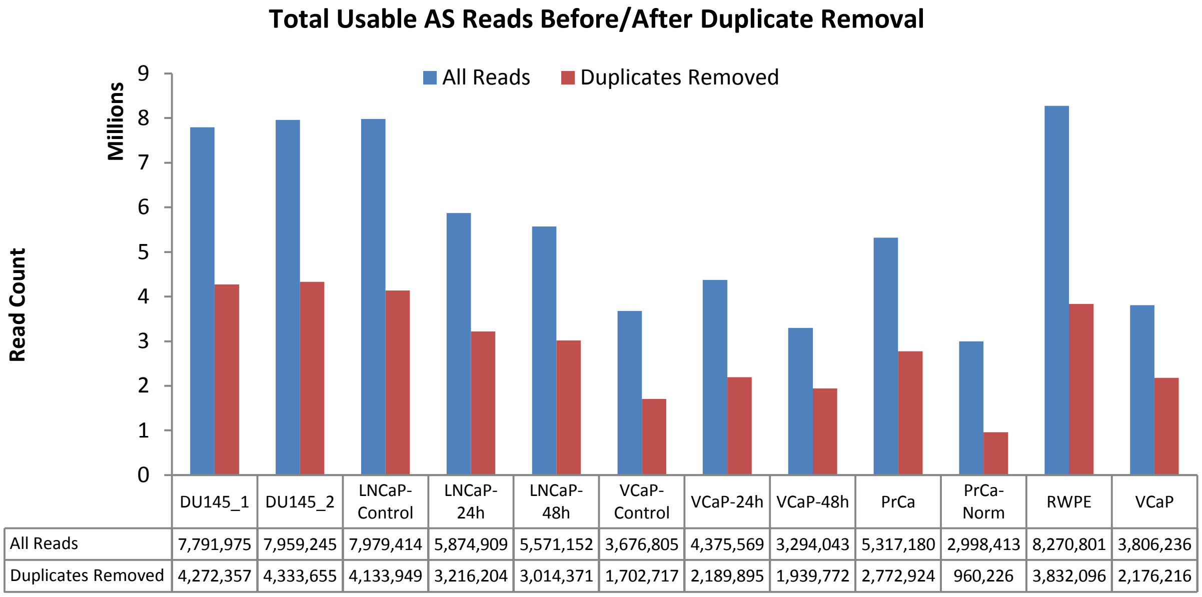


Figure S 13: Quantile-quantile plot of AS and SMS reads with duplicates removed. Reads in excess of a single read per aligned locus were removed from both AS and SMS data sets. The result of this procedure was inconsistent across the data set; some samples saw reduced representation of high expressing genes while the high-concentration bias remained in others

A.



	DU145_1	DU145_2	LNCaP-Control	LNCaP-24h	LNCaP-48h	VCaP-Control	VCaP-24h	VCaP-48h	PrCa	PrCa-Norm	RWPE	VCaP
All Reads	7,791,975	7,959,245	7,979,414	5,874,909	5,571,152	3,676,805	4,375,569	3,294,043	5,317,180	2,998,413	8,270,801	3,806,236
Duplicates Removed	4,272,357	4,333,655	4,133,949	3,216,204	3,014,371	1,702,717	2,189,895	1,939,772	2,772,924	960,226	3,832,096	2,176,216

B.

Coverage	DU145_1	DU145_2	LNCaP-0h	LNCaP-24h	LNCaP-48h	VCaP-0h	VCaP-24h	VCaP-48h	PrCa	PrCa-AdjNorm	RWPE	VCaP
(0-1]	3,659	3,621	3,659	3,987	4,019	5,953	5,082	5,192	4,588	6,089	3,738	5,045
(1-10]	3,806	3,826	3,802	3,206	3,120	2,066	2,418	2,235	2,986	1,199	3,681	2,137
(10-100]	370	377	363	286	255	117	172	132	187	63	309	143

Figure S 14: Effect of duplicate removal in AS. Reads in excess of a single read per aligned locus were removed from both AS and SMS data sets, resulting in (A) a median 47% drop in the number of usable reads across the 12 samples in the evaluation set and (B) the loss of dynamic range for genes in with high coverage levels.

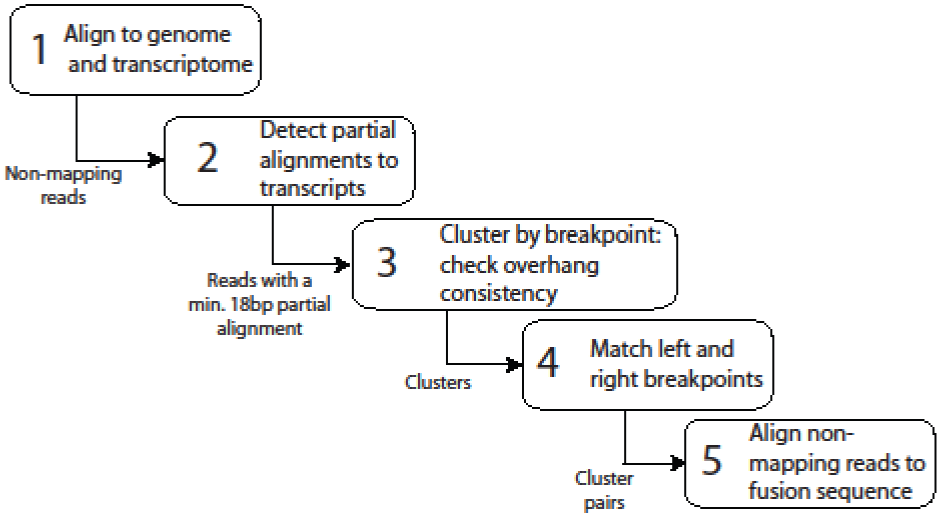

1 Align to genome and transcriptome
Non-mapping reads
2 Detect partial alignments to transcripts
Reads with a min. 18bp partial alignment
3 Cluster by breakpoint: check overhang consistency
Clusters
4 Match left and right breakpoints
Cluster pairs
5 Align non-mapping reads to fusion sequence

Figure S 15: Gene Fusion Discovery Using SMS Reads. All possible reads were aligned against the transcriptome and genome using IndexDP. The set of non-mapping reads (some of which harbor chimeras) were subsequently aligned against the transcriptome, returning reads that had a partial alignment of at least 18 nucleotides. All reads having the same partial alignments, suggesting a common breakpoint, were clustered. All clusters were then compared to determine if the non-aligning "overhang" portion of the read from one breakpoint region had similarity to the overhang of an independent breakpoint, thereby reconstructing the fusion junction. Finally, all remaining non-mapping reads were aligned against the candidate novel fusion junctions.

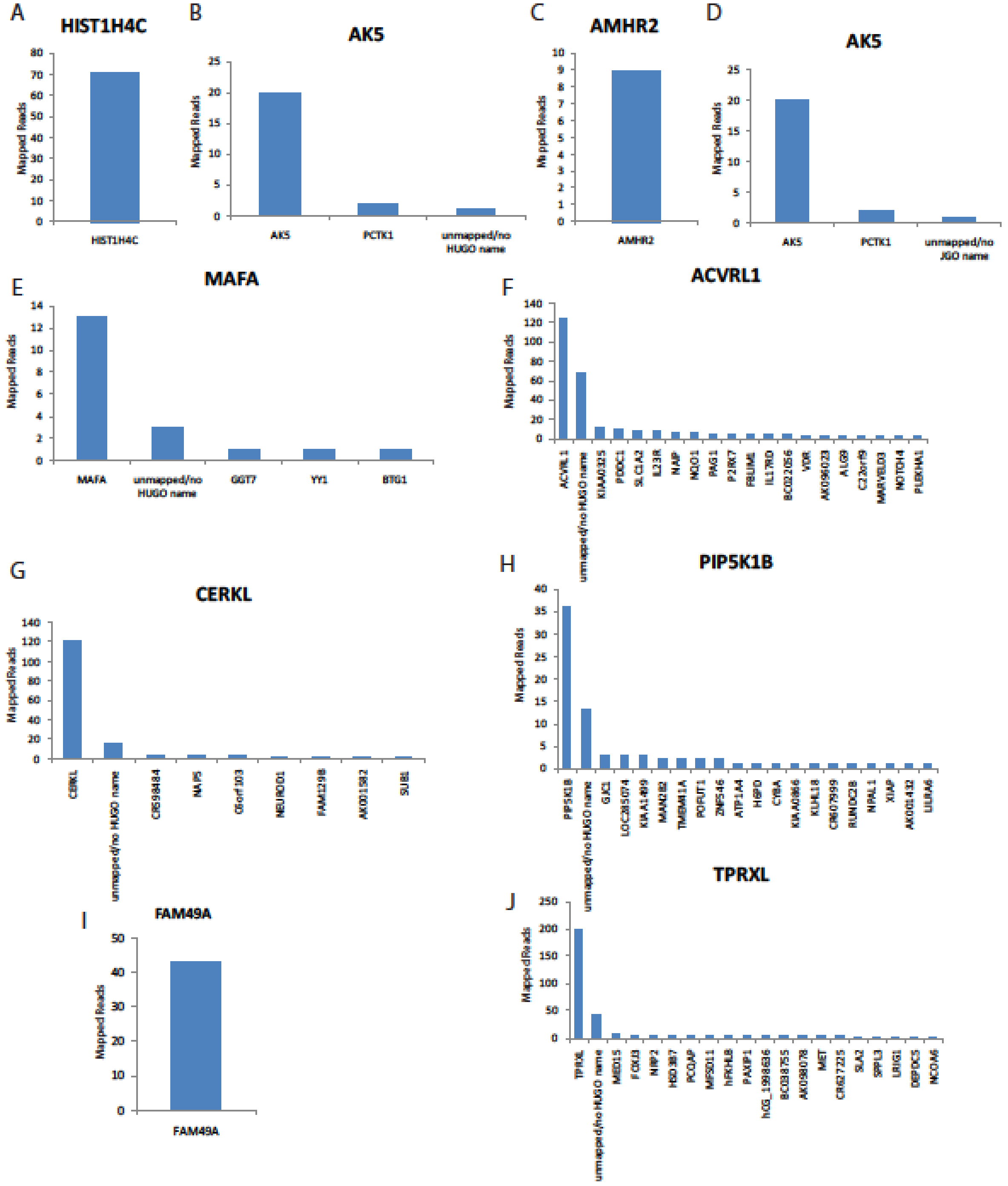


Figure S 16: Alternate mappings for genes detected by SMS only in DU145. We analyzed alternate mappings for the reads attributable to each of the nine genes we observed to be detectable only by SMS in DU145 using reads from both replicates. In all nine cases, reads mapped most strongly to the genes of interest, suggesting that the detection of these genes is not an artifact of mis-mapping. The top 20 alternate mappings, ordered by mapping read count, are shown in the graph.

A.	**Amplification-Based Sequencing**											
Sample Name	**DU145_1**	**DU145_2**	**LnCaP-0h**	**LNCaP-24h**	**LNCaP-48h**	**VCaP-0h**	**VCaP-24h**	**VCaP-48h**	**PrCa**	**PrCa-AdjNorm**	**RWPE**	**VCaP**
Total Reads	8,042,864	8,218,352	8,230,250	6,020,775	5,662,644	3,769,409	4,528,584	3,361,076	8,085,146	3,340,699	8,935,465	3,466,908
Non-Contaminant Reads:	7,709,472	7,875,168	7,820,929	5,790,207	5,513,448	3,636,454	4,297,981	3,272,200	7,748,986	2,871,802	8,038,501	3,352,960
Contaminant Reads:	333,392	343,184	409,321	230,568	149,196	132,955	230,603	88,876	336,160	468,897	896,964	113,948
Unique Mapping Reads:	5,818,322	5,942,613	5,530,586	4,126,612	3,959,711	2,281,328	2,931,544	2,325,189	5,849,652	2,303,763	5,590,474	2,477,323
Percent Unique	75.5%	75.5%	70.7%	71.3%	71.8%	62.7%	68.2%	71.1%	75.5%	80.2%	69.5%	73.9%
Multi-Mapping Reads:	1,891,150	1,932,555	2,290,343	1,663,595	1,553,737	1,355,126	1,366,437	947,011	1,899,334	568,039	2,448,027	875,637
Percent Multimapping	24.5%	24.5%	29.3%	28.7%	28.2%	37.3%	31.8%	28.9%	24.5%	19.8%	30.5%	26.1%
Total Mappings:	50,647,727	51,736,429	58,619,091	41,011,335	36,987,487	33,344,839	34,593,079	23,253,763	50,903,051	21,075,169	69,586,355	21,769,265
Max Read Mength:	36	36	36	36	36	36	36	36	36	36	36	36
Mean Read Length:	36	36	36	36	36	36	36	36	36	36	36	36
Min Read Length:	36	36	36	36	36	36	36	36	36	36	36	36

B.	**Single-Molecule Sequencing**											
Sample Name	**DU145_1**	**DU145_2**	**LnCaP-0h**	**LNCaP-24h**	**LNCaP-48h**	**VCaP-0h**	**VCaP-24h**	**VCaP-48h**	**PrCa**	**PrCa-AdjNorm**	**RWPE**	**VCaP**
Total Reads	12,605,568	19,741,065	7,257,338	9,917,739	5,700,598	7,399,104	6,959,550	5,760,821	9,630,377	2,848,185	18,628,241	16,254,121
Non-Contaminant Reads:	9,665,231	15,012,289	5,633,863	8,120,878	4,489,176	6,266,115	5,957,786	5,067,698	4,709,102	2,130,950	13,294,348	12,713,722
Contaminant Reads:	2,940,337	4,728,776	1,623,475	1,796,861	1,211,422	1,132,989	1,001,764	693,123	4,921,275	717,235	5,333,893	3,540,399
Unique Mapping Reads:	7,543,462	11,719,852	4,263,248	6,232,105	3,377,187	4,663,004	4,564,718	3,862,102	3,085,428	1,542,533	10,214,631	9,737,305
Percent Unique	78.0%	78.1%	75.7%	76.7%	75.2%	74.4%	76.6%	76.2%	65.5%	72.4%	76.8%	76.6%
Multi-Mapping Reads:	2,121,769	3,292,437	1,370,615	1,888,773	1,111,989	1,603,111	1,393,068	1,205,596	1,623,674	588,417	3,079,717	2,976,417
Percent Multimapping	22.0%	21.9%	24.3%	23.3%	24.8%	25.6%	23.4%	23.8%	34.5%	27.6%	23.2%	23.4%
Total Mappings:	35,091,411	54,881,579	20,016,203	26,912,318	15,869,700	20,525,342	18,151,466	15,215,471	32,088,861	7,702,501	51,058,659	44,504,047
Max Read Mength:	63	57	63	63	63	63	63	63	57	57	57	57
Mean Read Length:	34.23	33.5	33.7	33.98	32.83	33.54	33.89	33.55	33.18	32.13	33.73	33.5
Min Read Length:	25	24	25	25	25	25	25	25	24	24	24	24

Table S 4: Sample statistics in **A.** amplification-based and **B.** single-molecule sequencing technologies

Gene	# Samples Overrepresented	DU145_1	DU145_2	VCaP_0h	VCaP_24h	VCaP_48h	LnCaP_0h	LnCaP_24h	LnCaP_48h	PrCa	PrCa-AdjNorm	RWPE	VCaP
RPS18	12	·	·	·	·	·	·	·	·	·	·	·	·
GNAS	12	·	·	·	·	·	·	·	·	·	·	·	·
RPLP0	12	·	·	·	·	·	·	·	·	·	·	·	·
RPL8	12	·	·	·	·	·	·	·	·	·	·	·	·
RPL31	12	·	·	·	·	·	·	·	·	·	·	·	·
RPS8	12	·	·	·	·	·	·	·	·	·	·	·	·
CYC1	12	·	·	·	·	·	·	·	·	·	·	·	·
GNB2	12	·	·	·	·	·	·	·	·	·	·	·	·
rpl10a	12	·	·	·	·	·	·	·	·	·	·	·	·
OK/SW-cl.12	12	·	·	·	·	·	·	·	·	·	·	·	·
RPS14	12	·	·	·	·	·	·	·	·	·	·	·	·
SLC25A3	11	·	·	·	·	·	·	·	·	·	·	·	
UBB	11	·	·	·	·	·	·	·	·	·	·	·	
RPS5	11	·	·	·	·	·	·	·	·	·	·	·	
RPL7A	11	·	·	·	·	·	·	·	·	·	·	·	
EIF1	11	·	·	·	·	·	·	·	·	·	·	·	
UQCRC1	11	·	·	·	·	·	·	·	·	·	·	·	
PFKL	11	·	·	·	·	·	·	·	·	·	·	·	
RPS10	11	·	·	·	·	·	·	·	·	·	·	·	
RPL18A	11	·	·	·	·	·	·	·	·	·	·	·	
RPL3	11	·	·	·	·	·	·	·	·	·	·	·	
ENO1	11	·	·	·	·	·	·	·	·	·	·	·	
RPL37A	11	·	·	·	·	·	·	·	·	·	·	·	
EIF3I	11	·	·	·	·	·	·	·	·	·	·	·	
RPL29	11	·	·	·	·	·	·	·	·	·	·	·	
EEF2	11	·	·	·	·	·	·	·	·	·	·	·	
PSAP	11	·	·	·	·	·	·	·	·	·	·	·	
PPP2R1A	11	·	·	·	·	·	·	·	·	·	·	·	
GNB2L1	11	·	·	·	·	·	·	·	·	·	·	·	
RPL13A	11	·	·	·	·	·	·	·	·	·	·	·	
SPINT2	11	·	·	·	·	·	·	·	·	·	·	·	
SHISA5	11	·	·	·	·	·	·	·	·	·	·	·	
RPS11	10	·	·	·	·	·	·	·	·	·	·		
BTF3	10	·	·	·	·	·	·	·	·	·	·		
GPX4	10	·	·	·	·	·	·	·	·	·	·		
ATP5A1	10	·	·	·	·	·	·	·	·	·	·		
KRT18	10	·	·	·	·	·	·	·	·	·	·		
KIAA0088	10	·	·	·	·	·	·	·	·	·	·		
ACTG1	10	·	·	·	·	·	·	·	·	·	·		
RPS9	10	·	·	·	·	·	·	·	·	·	·		
RPL10A	10	·	·	·	·	·	·	·	·	·	·		
RPL12	10	·	·	·	·	·	·	·	·	·	·		
ATP5B	10	·	·	·	·	·	·	·	·	·	·		
RPS3	10	·	·	·	·	·	·	·	·	·	·		
RPSA	10	·	·	·	·	·	·	·	·	·	·		
MTCH1	10	·	·	·	·	·	·	·	·	·	·		
PYCR1	10	·	·	·	·	·	·	·	·	·	·		
LRP10	10	·	·	·	·	·	·	·	·	·	·		
ALDOA	10	·	·	·	·	·	·	·	·	·	·		
EEF1A1	10	·	·	·	·	·	·	·	·	·	·		
NME2	10	·	·	·	·	·	·	·	·	·	·		
RPS20	10	·	·	·	·	·	·	·	·	·	·		
C19orf48	10	·	·	·	·	·	·	·	·	·	·		
RPL18	10	·	·	·	·	·	·	·	·	·	·		
RPL13	10	·	·	·	·	·	·	·	·	·	·		
P4HB	10	·	·	·	·	·	·	·	·	·	·		
OAZ1	10	·	·	·	·	·	·	·	·	·	·		
EEF1G	10	·	·	·	·	·	·	·	·	·	·		
UBC	10	·	·	·	·	·	·	·	·	·	·		

Table S 5: Recurrently over-represented genes in amplification-based sequencing in ten or more samples. Of the 393 genes are recurrently within the top 500 over-represented genes by total read count in five (40%) or more samples, these 59 are seen most often, occurring in at least 10 samples.

A.

	Single-Molecule			
	Q1	**Q2**	**Q3**	**Q4**
DU145_1	678,716.20	82,650.17	15,312.21	1,756.09
DU145_2	682,755.30	80,340.18	13,794.63	1,455.81
VCaP-0h	679,511.10	83,560.23	20,066.82	2,400.37
VCaP-24h	681,607.20	77,752.04	18,903.67	2,624.63
VCaP-48h	688,925.40	75,791.22	18,493.60	2,544.35
LnCaP-0h	658,971.00	97,176.66	21,592.82	2,512.66
LnCaP-24h	678,955.60	86,868.69	17,595.02	2,044.48
LnCaP-48h	653,867.40	100,158.20	24,825.49	3,424.91
PrCa	610,787.00	107,986.40	33,056.41	4,683.91
PrCa-AdjNorm	581,561.70	121,731.20	41,077.92	7,530.44
RWPE	688,043.60	83,353.84	16,642.86	1,740.82
VCaP	674,367.00	83,563.10	19,041.08	2,192.36

B.

	Amplification-Based			
	Q1	**Q2**	**Q3**	**Q4**
DU145_1	673,250.40	86,579.46	11,608.48	299.67
DU145_2	676,079.70	84,926.77	10,734.94	233.56
VCaP-0h	724,435.80	67,473.25	15,836.85	1,688.69
VCaP-24h	731,422.10	63,380.33	10,563.88	519.48
VCaP-48h	713,318.90	68,976.03	11,508.96	602.30
LnCaP-0h	683,466.00	83,738.35	11,073.12	440.38
LnCaP-24h	703,435.10	70,455.56	8,175.45	192.68
LnCaP-48h	705,382.50	70,566.91	7,889.93	170.70
PrCa	610,253.20	70,574.82	9,521.59	293.01
PrCa-AdjNorm	592,366.00	64,313.02	12,772.42	278.15
RWPE	680,671.10	71,476.63	10,868.36	245.32
VCaP	711,504.06	52,559.02	6,843.76	0.00

Table S 6: Sum of normalized expression values per quartile by sample in AS and SMS. We observe that the number of reads aligning to transcripts seen in the third and fourth quartiles is consistently greater in **A.** SMS than **B.** AS across the sample set.

A.

	Single-Molecule Sequencing											
Coverage	**DU145_1**	**DU145_2**	**LnCaP-0h**	**LNCaP-24h**	**LNCaP-48h**	**VCaP-0h**	**VCaP-24h**	**VCaP-48h**	**PrCa**	**PrCa-AdjNorm**	**RWPE**	**VCaP**
(0 -1]	3,875	3,420	4,459	4,109	5,008	4,691	4,920	5,169	4,971	6,841	3,432	3,513
(1 - 10]	3,871	3,917	3,638	3,765	3,270	3,579	3,321	2,937	3,514	1,745	4,314	4,316
(10 -100]	922	1,410	432	752	331	401	423	365	289	110	1,054	992
(100 - 1000]	66	105	35	59	27	49	31	31	23	3	80	83
(1000 - 10000]	0	1	0	0	0	1	0	0	0	0	4	2

B.

	Amplification-Based Sequencing											
Coverage	**DU145_1**	**DU145_2**	**LnCaP-0h**	**LNCaP-24h**	**LNCaP-48h**	**VCaP-0h**	**VCaP-24h**	**VCaP-48h**	**PrCa**	**PrCa-AdjNorm**	**RWPE**	**VCaP**
(0 -1]	3,299	3,261	3,421	3,767	3,804	5,656	4,776	4,964	4,300	5,033	3,433	4,879
(1 - 10]	3,721	3,731	3,537	3,069	2,982	2,213	2,508	2,288	3,033	2,077	3,525	2,290
(10 -100]	752	769	790	598	561	208	332	272	395	221	695	282
(100 - 1000]	63	63	73	45	47	56	54	35	33	19	72	40
(1000 - 10000]	0	0	3	0	0	3	2	0	0	1	3	0

Table S 7: Gene-level read coverage of observed transcripts. A. and **B.** illustrate the number of genes with coverage values at various depths in single molecule and amplification-based sequencing, respectively.

Sequence	UPL Probe #
agacccccaccatcccta	71
cgcatcatctgagctaggc	
cgcaatgtgctggtcaag	76
gttgccgatgtccaggtaat	
tgttattgatggatttccaagaga	61
ccaaatcgggggtacagatt	
ctgattatgaagatcagggtgatg	55
tctcaaatcttccatgaaacctc	
ccctacatcccatccacct	30
ggtctgcatcccaacagtct	
ctatgggccttggcagtg	55
gagctccctcagcccatc	
acccattccggattatgga	62
ttgcttgcacagacctttga	
ttttcatgggtggcctct	71
tgccaatgatgttactcagacc	
gcatgcaaacgttagaacca	22
ggctacttcgctagcagatcc	
gaagttgcatcagaggtccat	69
aaacaattacatgttactttggaatca	
ctgcaagacatccaagatcg	57
aacctgagggcatttagcag	
ctgcaagacatccaagatcg	68
aacctgagggcatttagcag	
aagaggtggcaacaacctaca	17
gatgcaataattgtctttagtgtcct	
aagaggtggcaacaacctaca	75
gatgcaataattgtctttagtgtcct	
ttctacaagcgcagcaagg	58
cagggtccagtaattgccttt	
cgtaaggtgctccgggata	37
gagccaaacggcgaatag	
ctatacggagcacgccaag	76
cctgacgttttagggcatatactac	
tactggccttggctgtgc	71
cacagggttttcaccaacct	
agcgagaagtgccaactcc	39
ttgtacaggtcccgctcttt	
gcagaagatggaccagcaat	88
tgtgctttccccattgattt	
gggacaggtcccagaatatg	70
gcctacttccggcagacc	
tcctagctgaatgctataacctctg	15
ggcatccttcagggtcttc	
gtcattgaaaatccccagtacttt	9
aattattatcaggcggtcttgg	
gcttctgtgcttgacgtctattt	53
ggataattctggtgcggaga	
agcagccttgatgaagaagc	38
gaagaagatgaaattgtggttgc	
tgggctcaaacaatccttct	13
atcctgggtcctgctctgta	
tgggctcaaacaatccttct	16
atcctgggtcctgctctgta	

Table S 8: Primers used for validating transcripts seen only by SMS. All experiments were performed in duplicate using two primer pairs per candidate gene when possible.

APPENDIX C: Chapter 3 and 4 Supplementary Figures and Tables

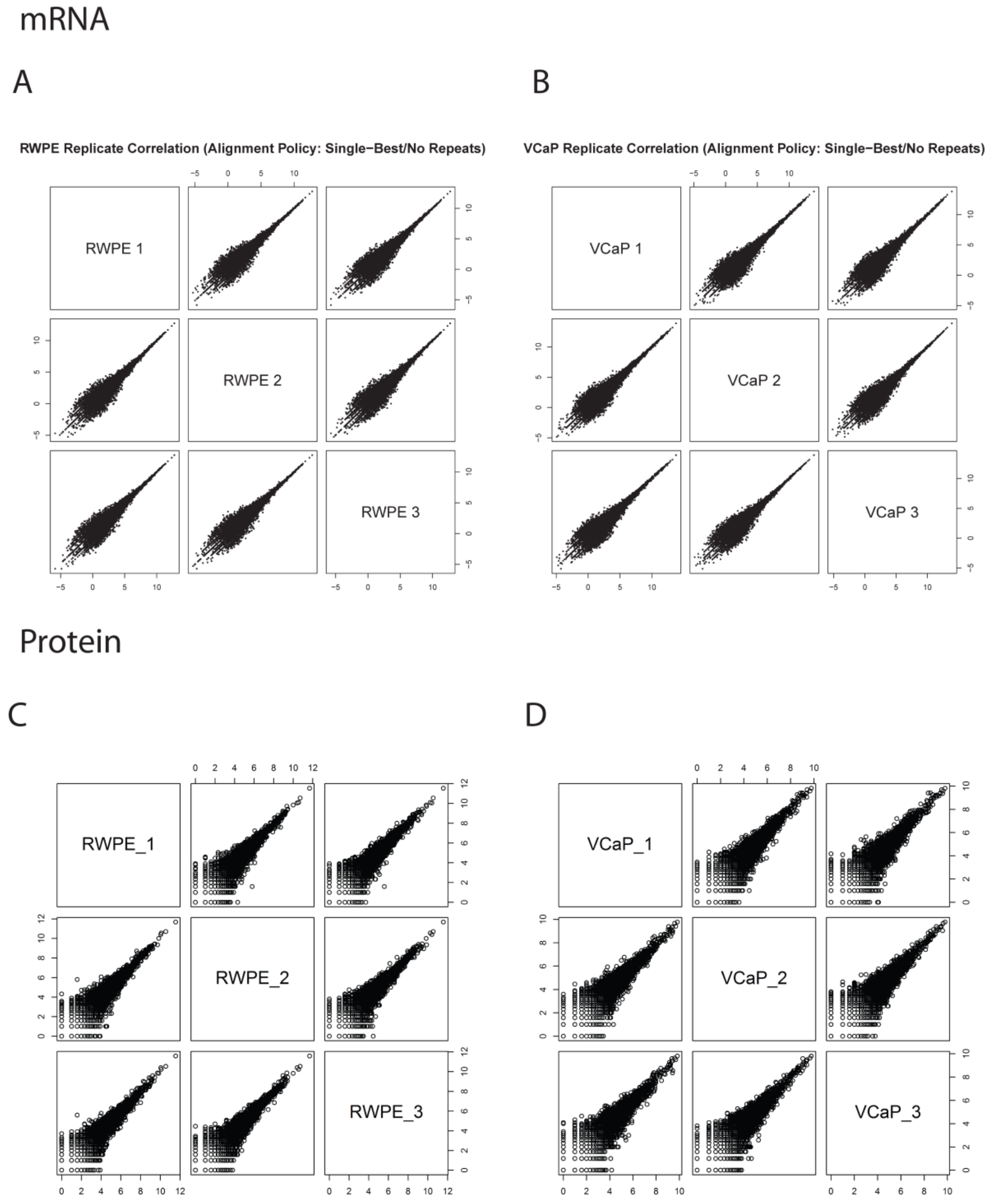


Figure S 17: Reproducibility between replicates. A. RWPE RNA-seq, B. VCaP RNA-seq, C. RWPE tandem MS, and D. VCaP tandem MS. Data are derived from the extended dataset.

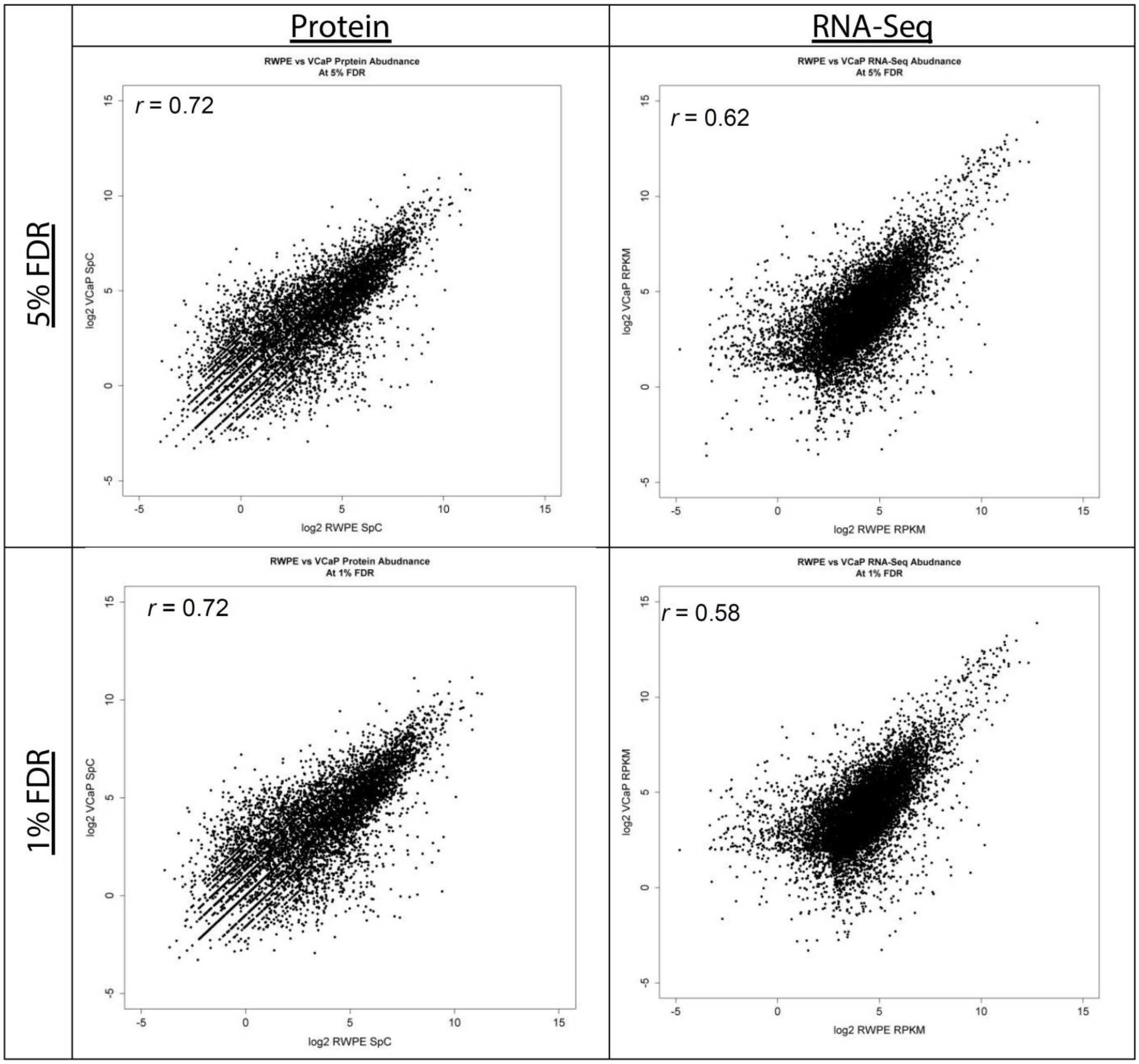


Figure S 18: Correlation between VCaP and RWPE by RNA-seq and tandem MS. Within both the transcriptome and proteome data, both cell lines showed relatively high similarity in abundance profile. This is expected, owing to their common prostate tissue origin.

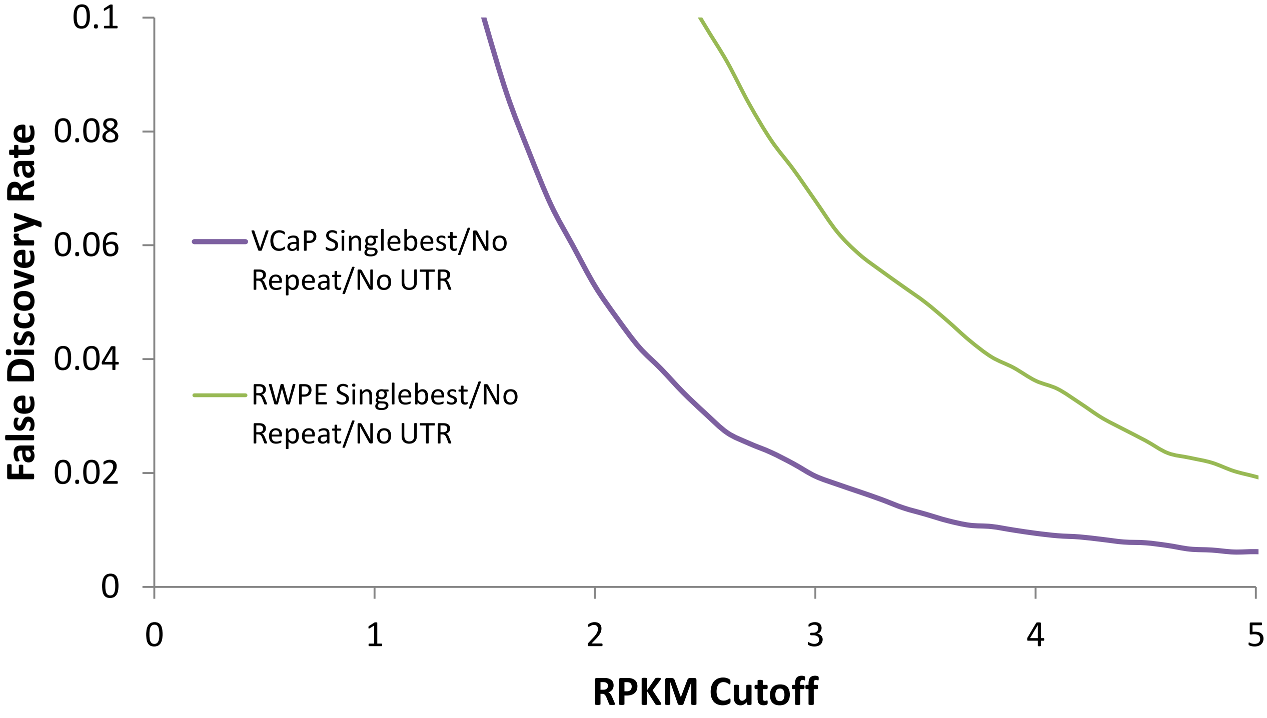
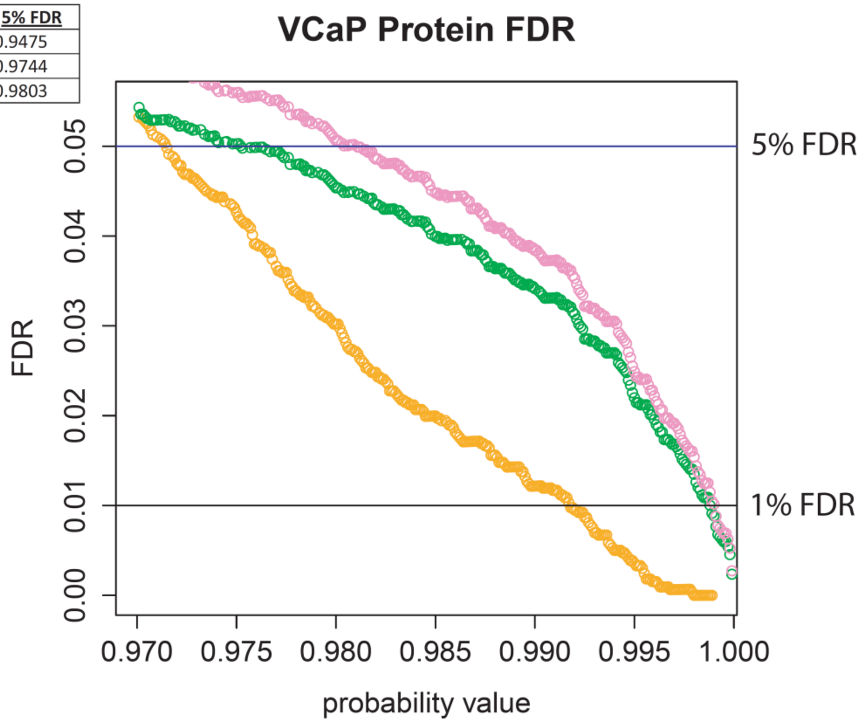


Figure S 19: False Discovery Rate estimation in RNA-seq data. We measured the FDR at increasing RPKM cutoffs to determine the 1% and 5% FDR levels in each cell line in our RNA-seq data

	VCaP	1 % FDR	5% FDR
🟨	Peptide Probability	0.9855	0.9475
🟩	Protein Group Probability	0.9988	0.9744
🟪	Protein Probability	0.9989	0.9803



	RWPE	1 % FDR	5% FDR
🟨	Peptide Probability	0.9885	0.9580
🟩	Protein Group Probability	0.9994	0.9771
🟪	Protein Probability	0.9995	0.9809

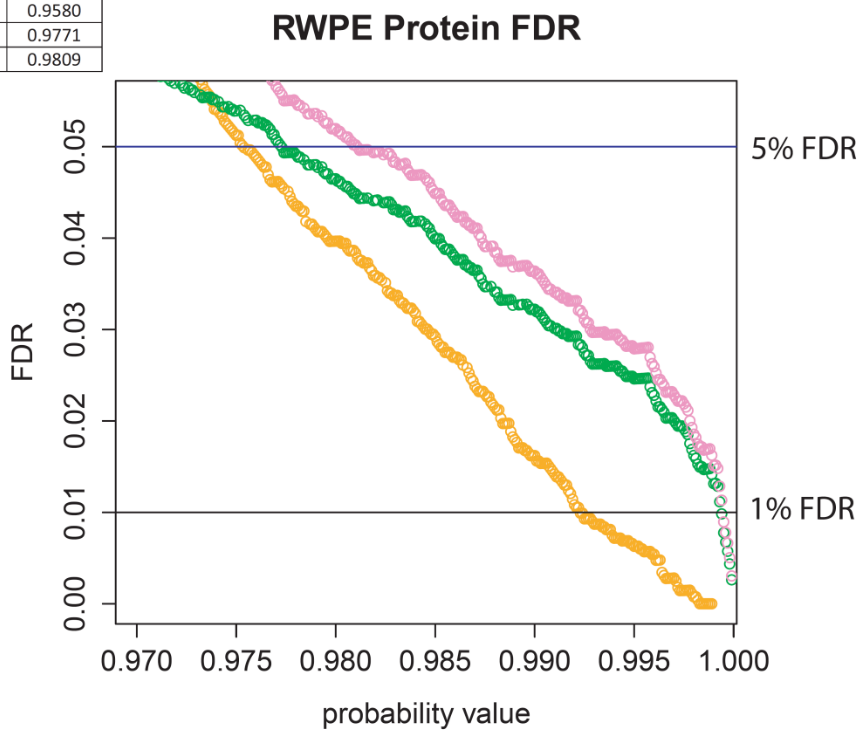
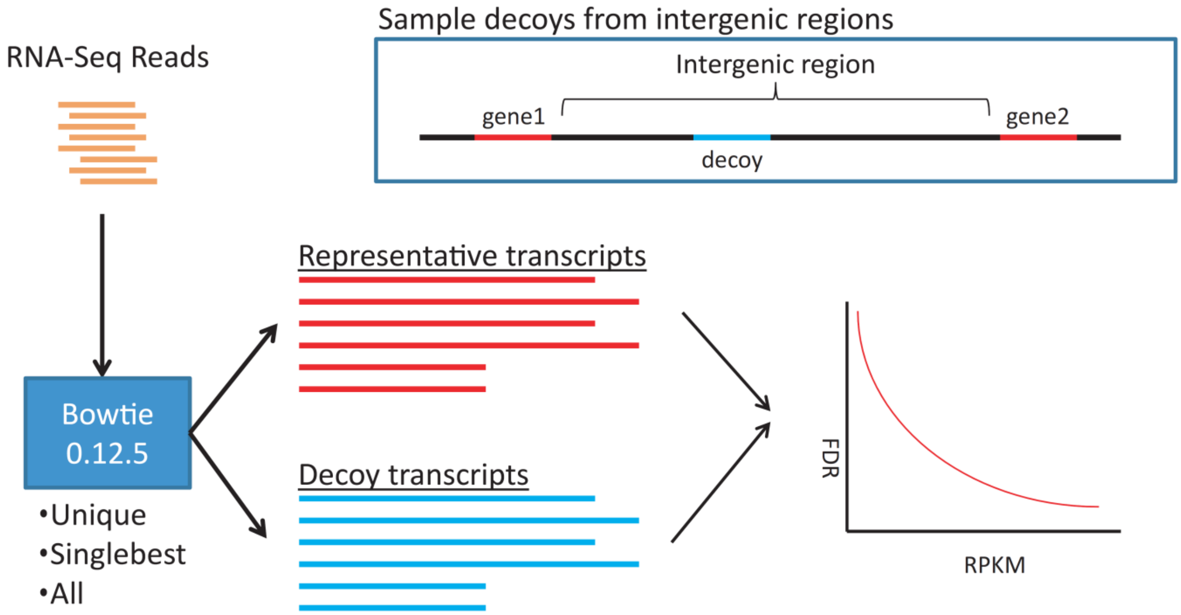
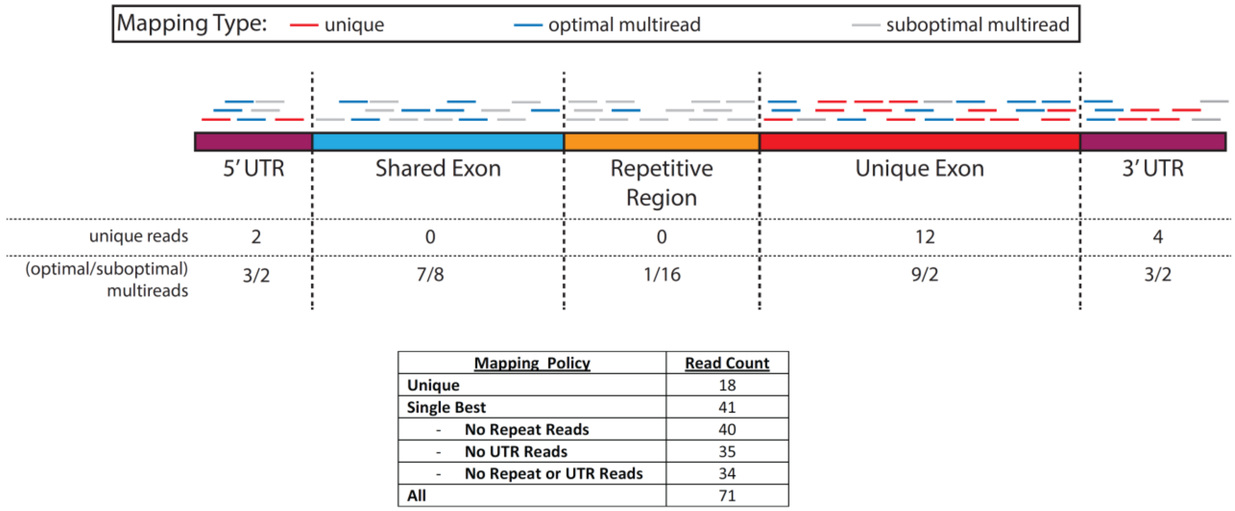


Figure S 20: False Discovery Rate estimation in protein data. We used the output of TPP and Abacus to determine appropriate parameter values for controlling FDR in our protein data. Three parameters were considered to control FDR at 1% and 5% FDR levels; peptide probability, protein group probability, and protein probability.



$$FDR(RPKM_{threshold}) = \frac{\#(decoy \geq RPKM_{threshold})}{\#(transcripts \geq RPKM_{threshold}) - \#(decoy \geq RPKM_{threshold})}$$

Figure S 21: RNA-seq False Discovery Rate estimation methodology. We used a methodology similar to that of Ramskold, et al. to estimate FDR in our RNA-seq data. Corresponding decoy sequences were sampled without replacement from the intergenic regions in hg19 for each representative transcript in our database, for a total of 34,728 decoys. These decoy sequences were of equal length as the real transcripts. We aligned reads to the merged total set of these decoy and real mRNA transcripts. Abundance data was summarized at the gene level using the same transcript-gene mappings for both the real and decoy transcript set. FDR was calculated as the number of decoy genes detected divided by the number of non-decoy genes detected at each threshold value.



Mapping Policy	Read Count
Unique	18
Single Best	41
- **No Repeat Reads**	40
- **No UTR Reads**	35
- **No Repeat or UTR Reads**	34
All	71

Figure S 22: Comparison of mapping methodologies. The number of reads assigned to this hypothetical gene, reflecting its abundance, is highly variable based on both the mapping and counting parameters can constraints.

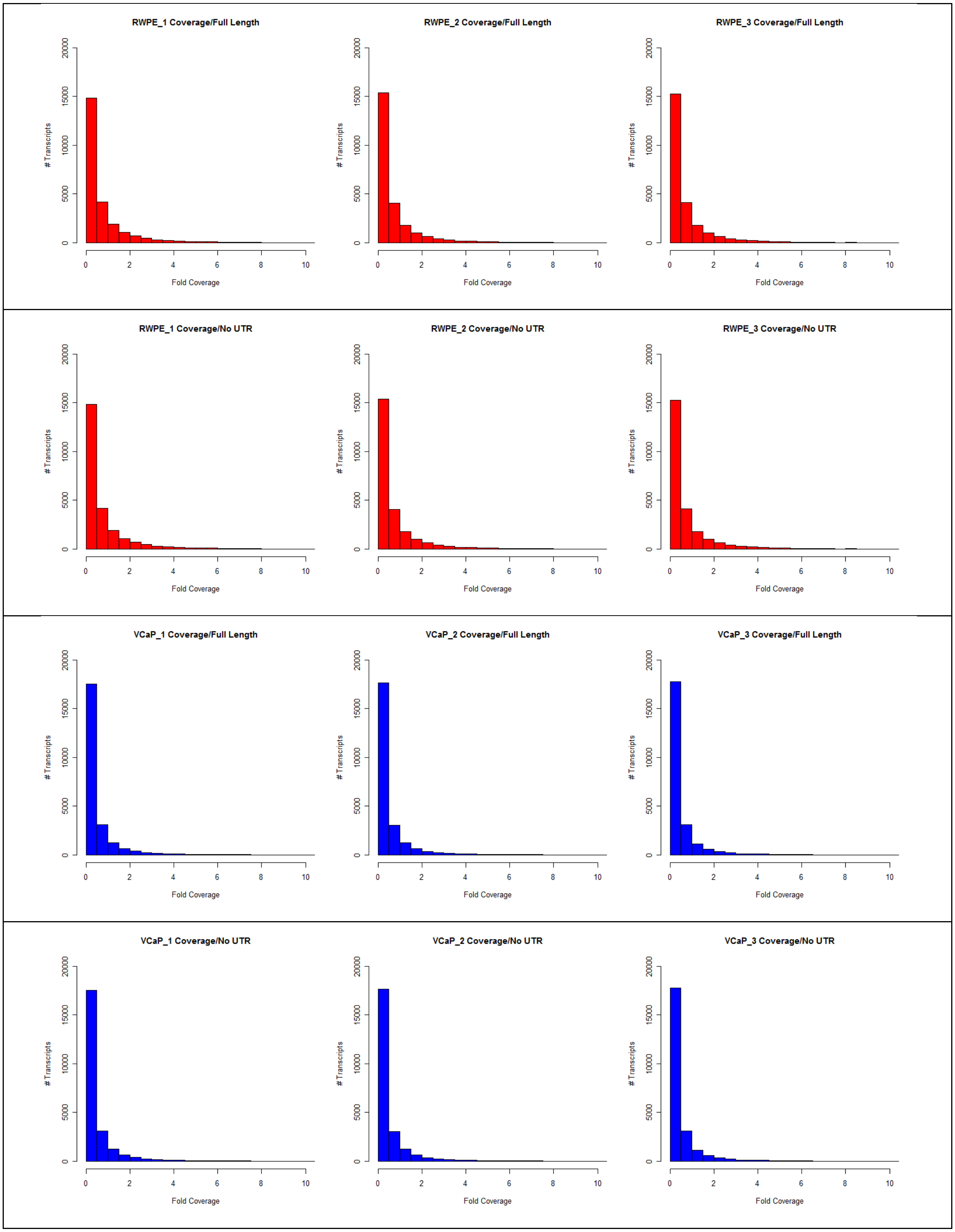


Figure S 23: Coverage of the transcriptome by observed RNA-seq reads.

Histogram reports the number of proteins having a certain percent of their sequence covered by peptides. In this figure, all peptides from both cell lines are considered. Only peptides with a probability >= 0.5 were used.

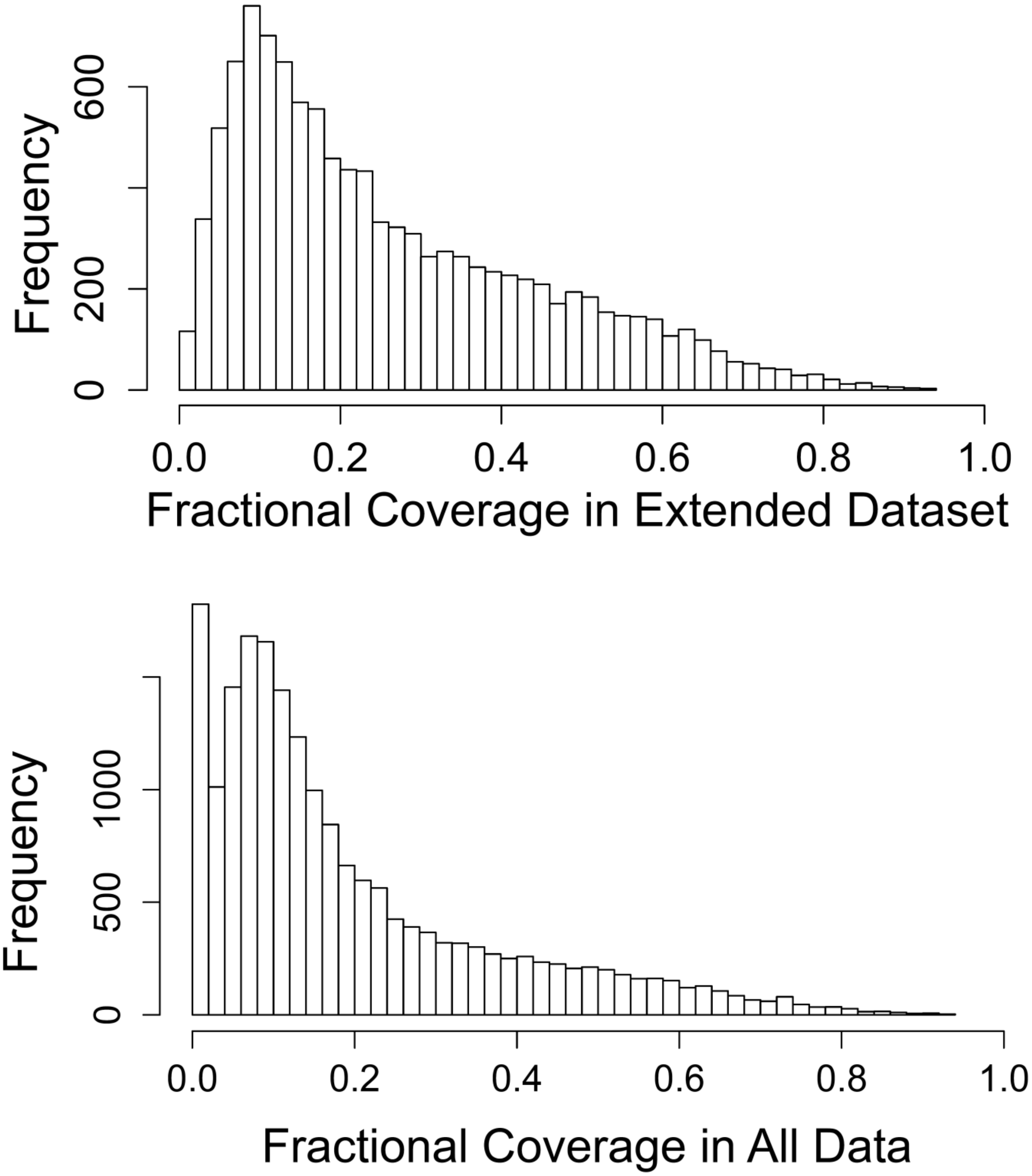


Figure S 24: Coverage of the proteome by observed peptides.

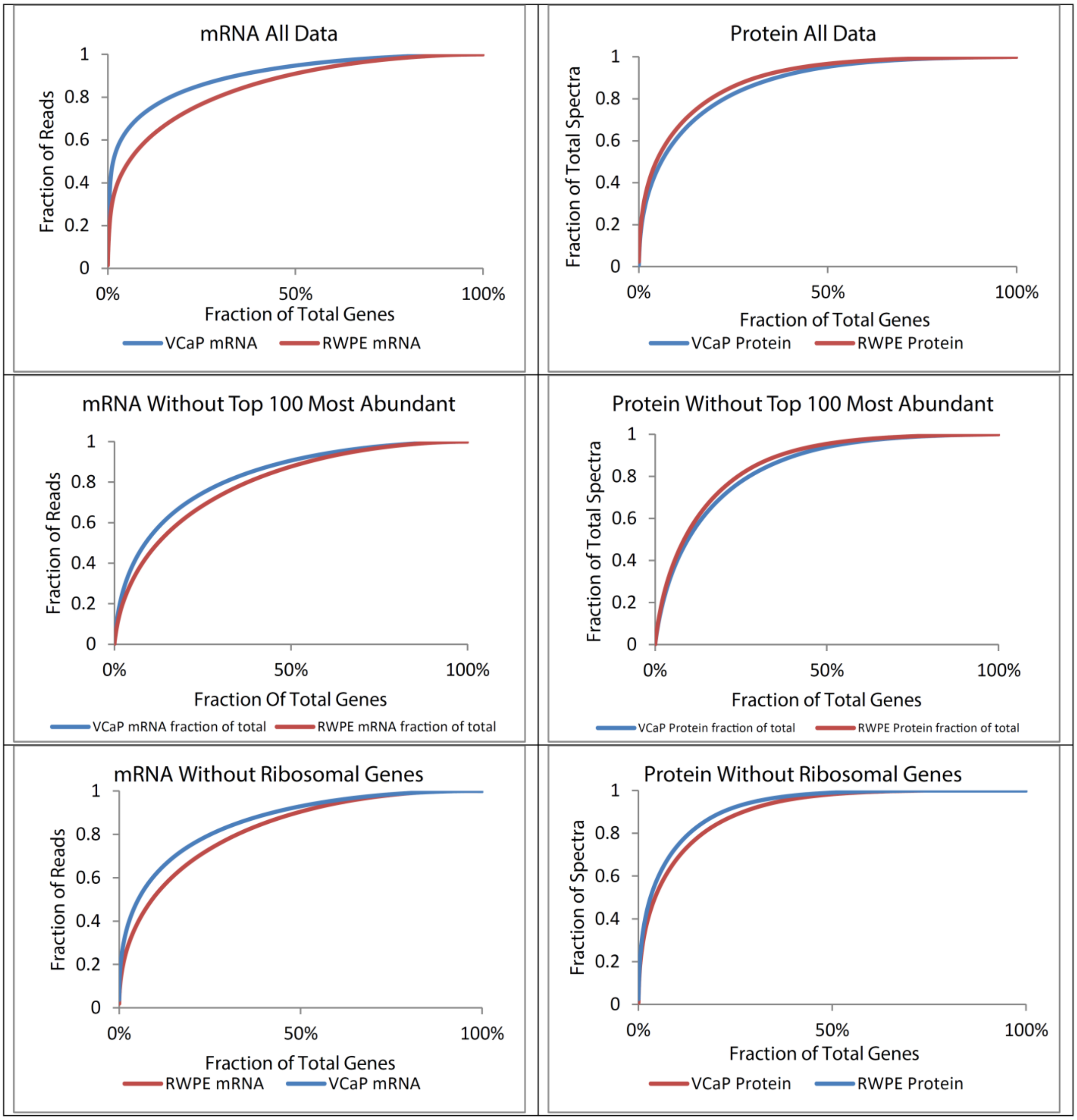


Figure S 25: Distribution of reads and spectra among observed genes in VCaP and RWPE. A and B. Distribution of reads and spectra, respectively, across extended dataset. **C and D**. Distribution of reads and spectra, respectively, across extended dataset after removal of top 100 most abundantly observed genes. **E and F.** Distribution of reads and spectra, respectively, across extended dataset after removal of ribosome-associated genes.

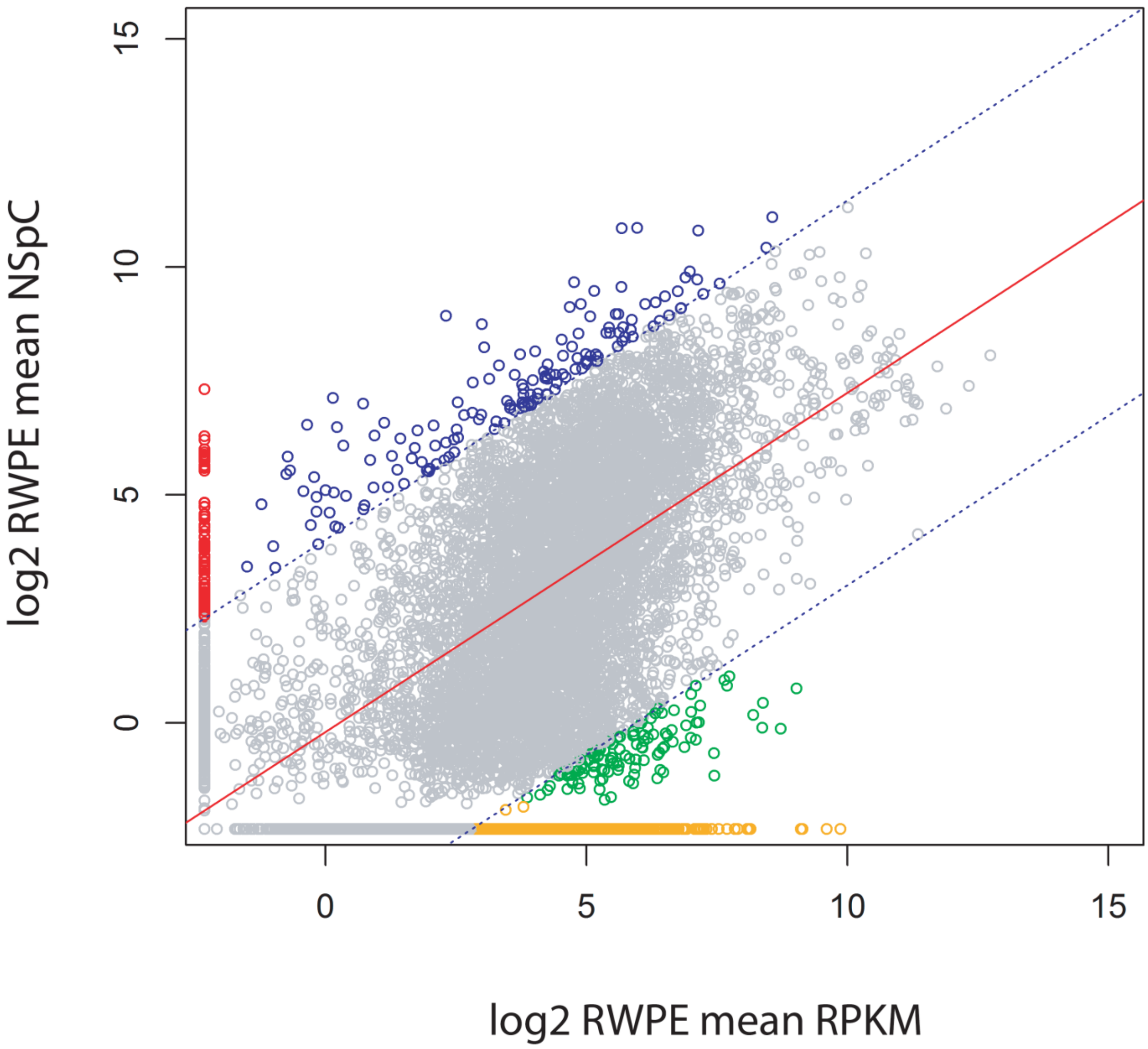


Figure S 26: Segregation of RWPE genes into broad categories by transcript-protein relationship

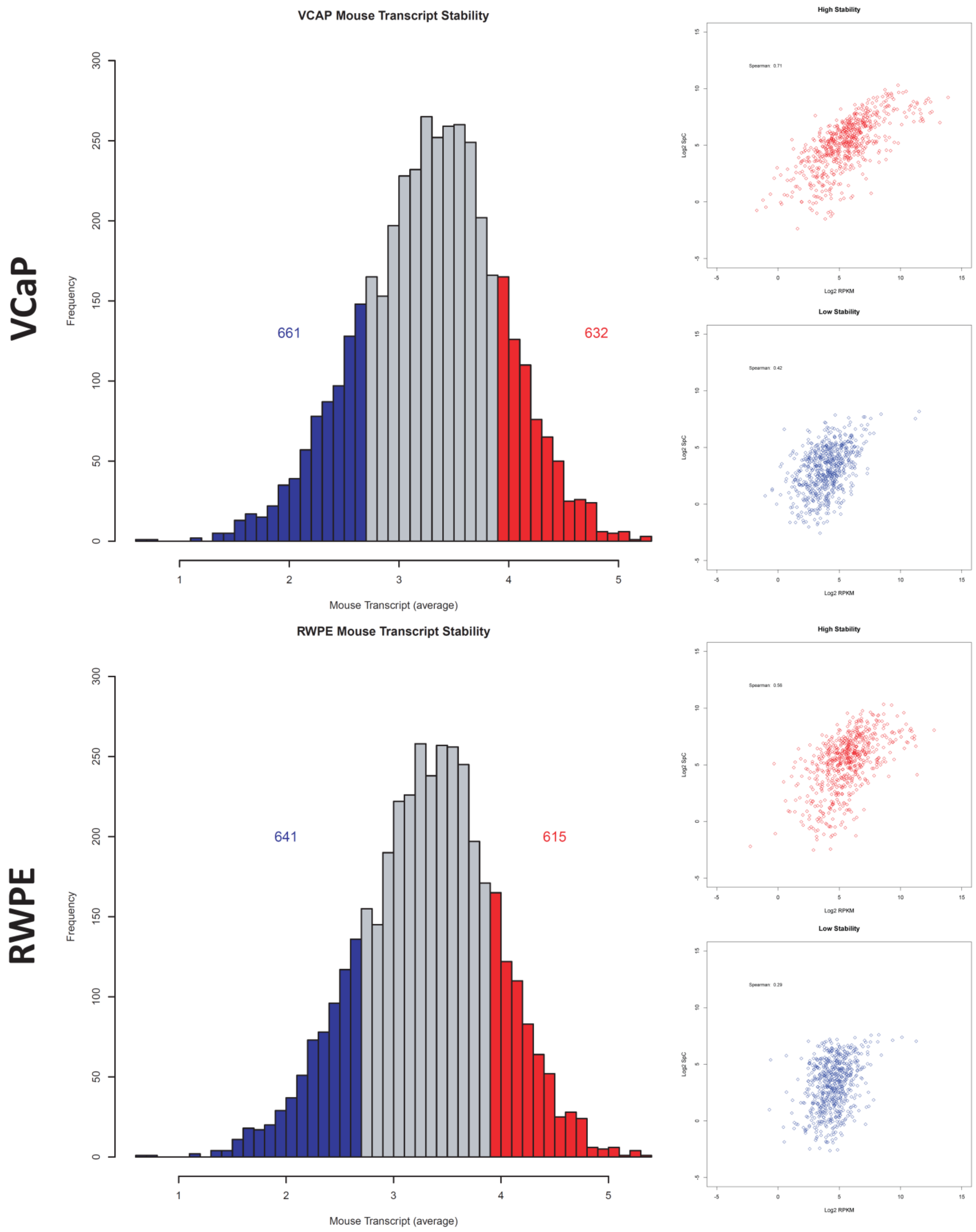


Figure S 27: Correlation of protein and transcript in high and low stability groups chosen using transcript half-life

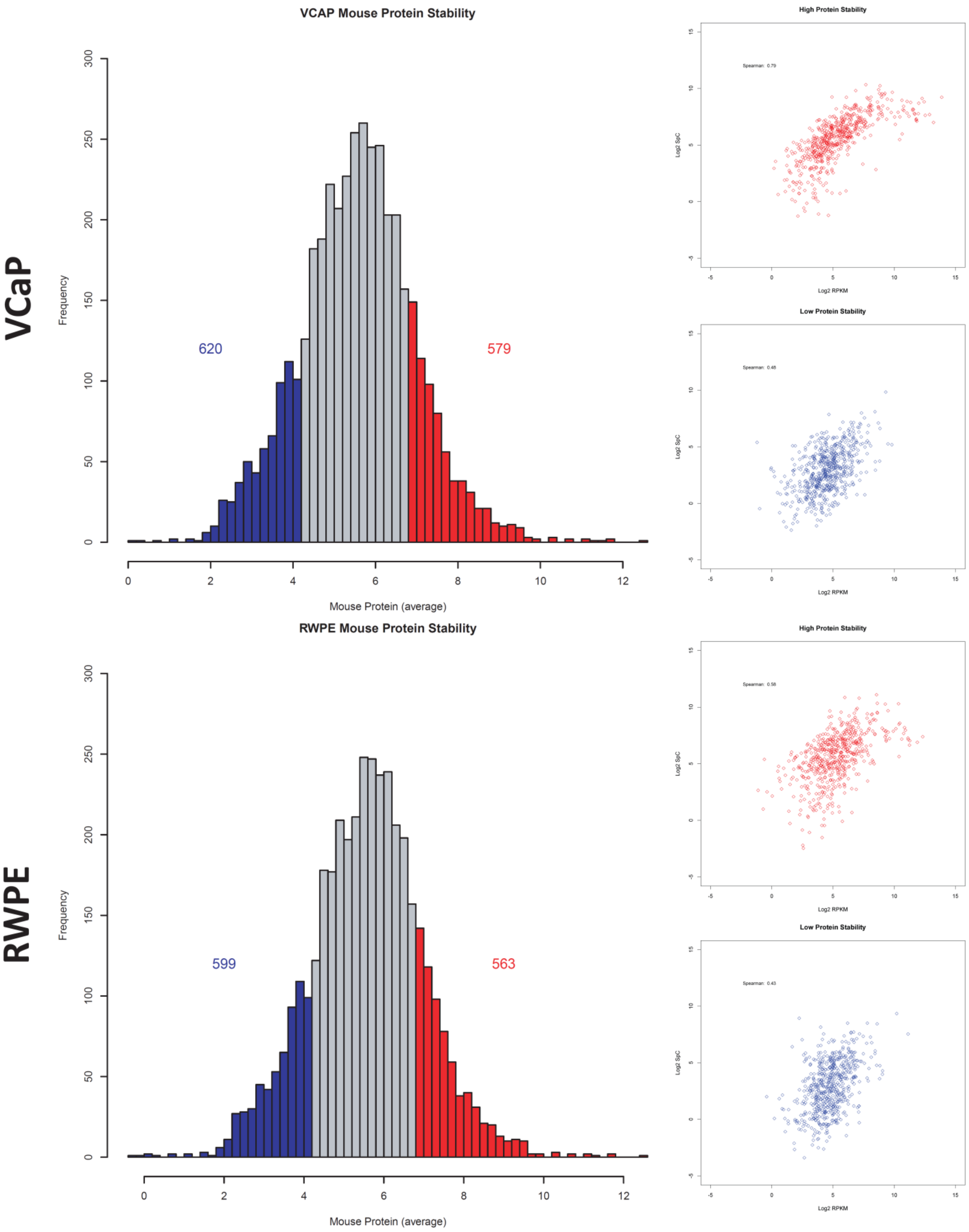
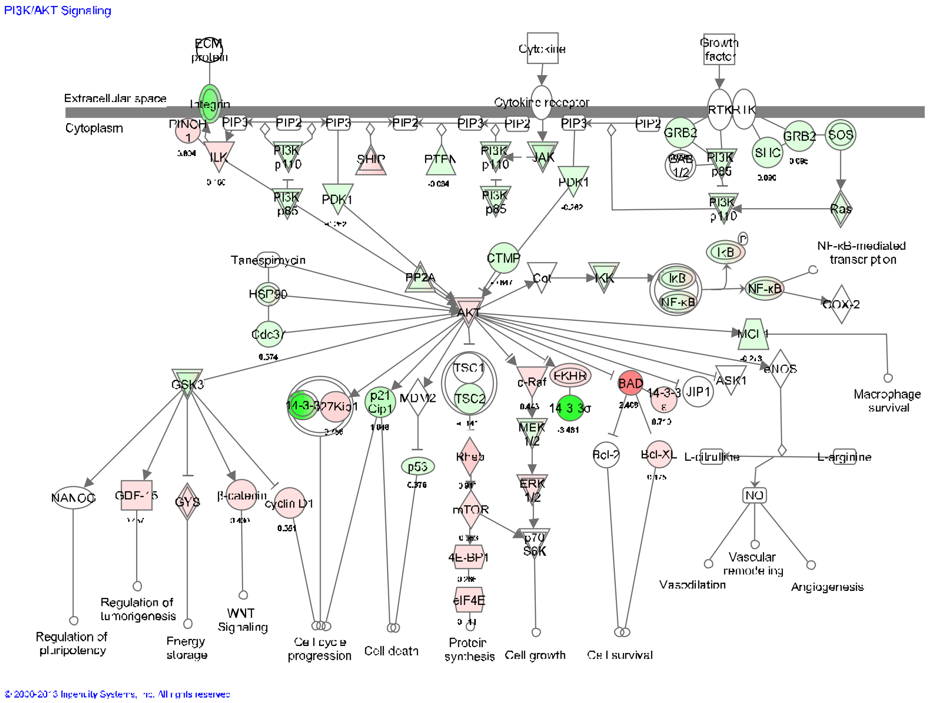
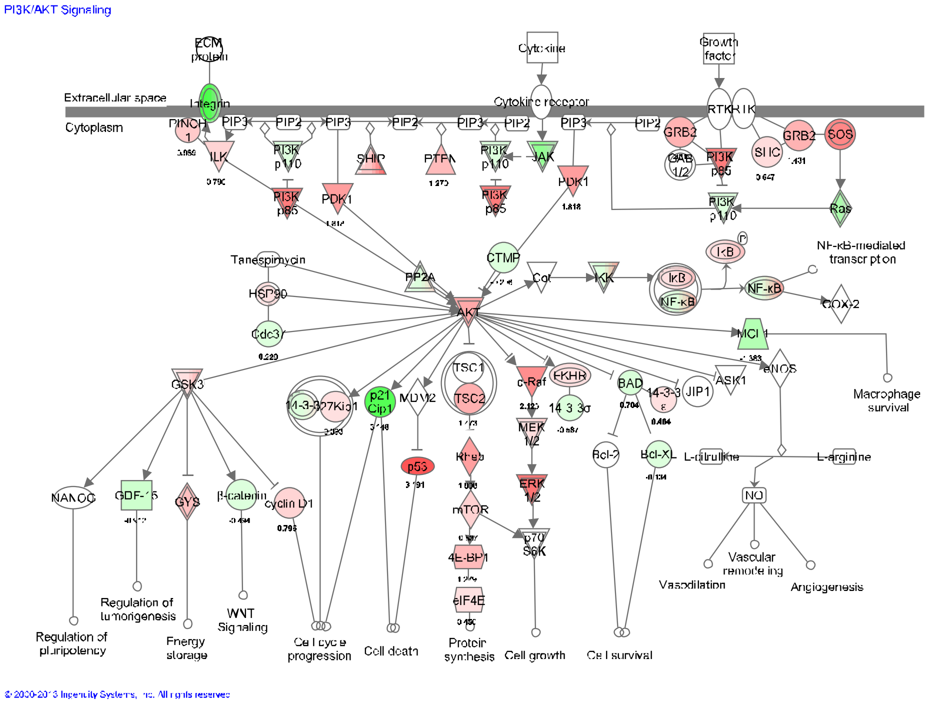


Figure S 28: Correlation of protein and transcript in high and low stability groups chosen using protein half-life

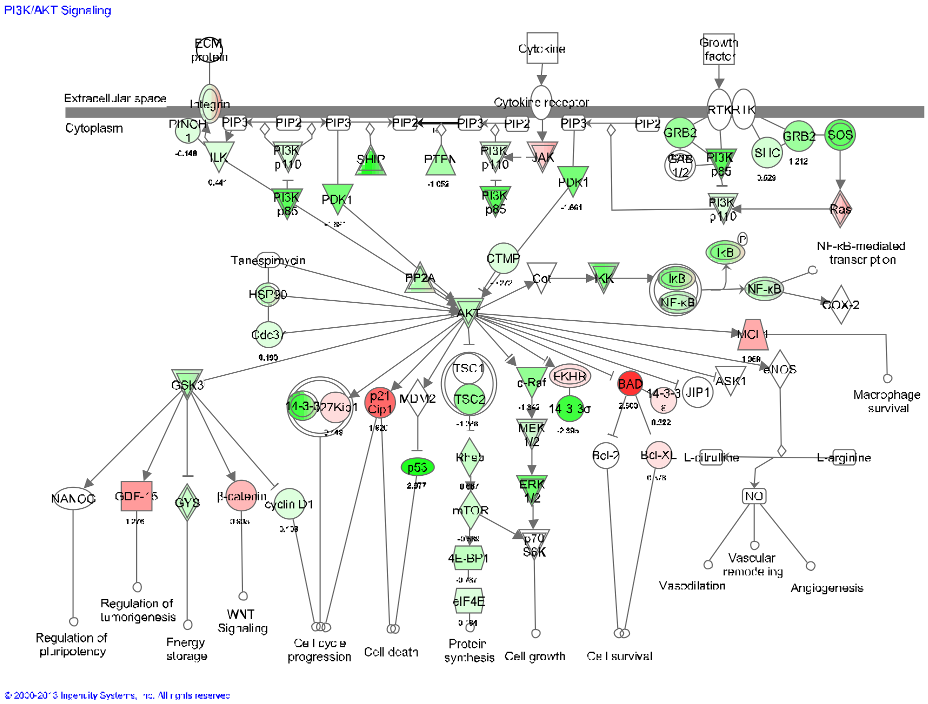
A. PI3K/Akt Pathway colored by transcript fold change



B. PI3K/Akt Pathway colored by protein fold change



C. PI3K/Akt Pathway colored by discordance index



D. PI3K/Akt Pathway colored by concordance index

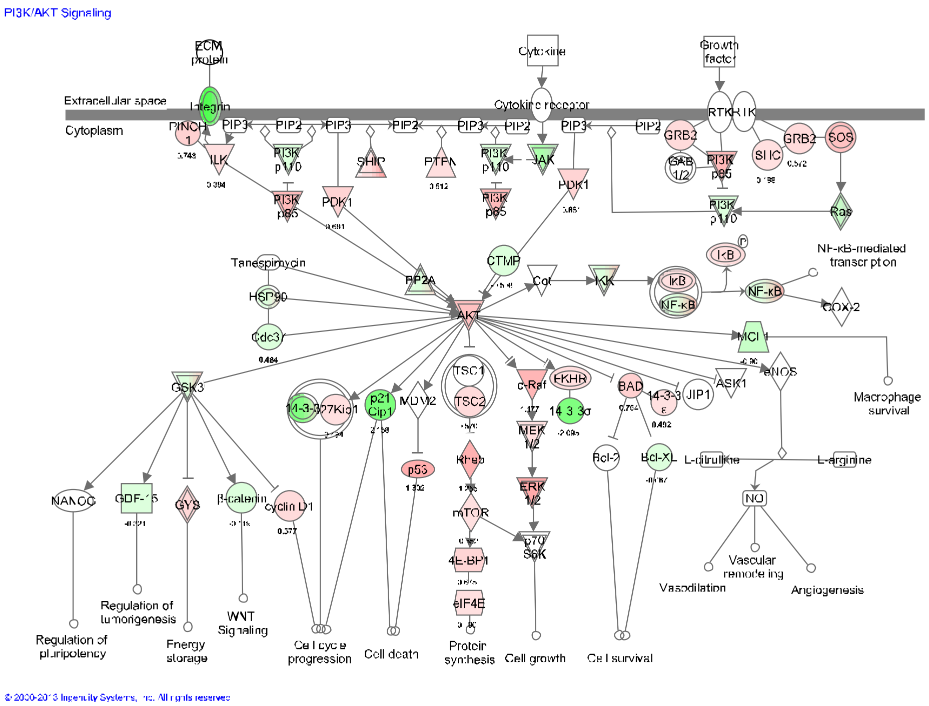


Figure S 29: PI3K/Akt Signaling Pathway colored by VCaP/RWPE transcript fold change, protein fold change, discordance index, and concrodance index data.

go_id	Observed genes	Total # Genes in Class	vcap_readfraction	vcap_rpkmfraction	rwpe_readfraction	rwpe_rpkmfraction	vcap_spcfraction	vcap_pnorm_frac	rwpe_spcfrac	rwpe_pnorm_frac	Term	go_tree
GO:0005739	1109	4669	0.087974	0.095673	0.07436	0.092678	0.189368	0.239181	0.220921	0.283798	mitochondrion	CC
GO:0000166	1371	7290	0.194515	0.126364	0.190091	0.148335	0.295258	0.226834	0.287902	0.230131	nucleotide binding	MF
GO:0016020	2213	12613	0.153828	0.12507	0.143564	0.1261	0.176585	0.191634	0.207733	0.238873	membrane	CC
GO:0005524	1037	5332	0.105549	0.062988	0.120108	0.083077	0.210816	0.128989	0.199519	0.127243	ATP binding	MF
GO:0005515	3215	17433	0.43172	0.38374	0.453007	0.400261	0.505604	0.44195	0.530704	0.452411	protein binding	MF
GO:0005743	251	980	0.026316	0.036553	0.016199	0.026615	0.060533	0.08263	0.086684	0.119398	mitochondrial inner membrane	CC
GO:0005634	3676	19673	0.391271	0.342794	0.437047	0.391448	0.423686	0.383929	0.442599	0.379974	nucleus	CC
GO:0006810	361	2081	0.033341	0.030051	0.036993	0.034407	0.061443	0.06892	0.065442	0.074691	transport	BP
GO:0005654	744	3572	0.099669	0.09049	0.109085	0.106917	0.144546	0.128082	0.165531	0.140452	nucleoplasm	CC
GO:0005783	702	3602	0.067749	0.049968	0.06281	0.051301	0.090773	0.08754	0.106616	0.111975	endoplasmic reticulum	CC
GO:0005737	3603	19916	0.441512	0.423029	0.451041	0.438006	0.492567	0.460282	0.409522	0.388043	cytoplasm	CC
GO:0005886	1462	12903	0.138727	0.097976	0.175349	0.125197	0.166759	0.135185	0.195319	0.158953	plasma membrane	CC
GO:0015031	340	1395	0.027568	0.019526	0.031146	0.02519	0.054205	0.056471	0.047257	0.047685	protein transport	BP
GO:0005759	175	707	0.015747	0.014125	0.011658	0.013385	0.048689	0.049274	0.063661	0.065725	mitochondrial matrix	CC
GO:0008380	241	928	0.040935	0.029537	0.045753	0.041021	0.066799	0.063403	0.088998	0.082265	RNA splicing	BP
GO:0007264	214	1077	0.016956	0.011659	0.016244	0.012738	0.025088	0.04462	0.017212	0.033747	small GTPase mediated signal transduction	BP
GO:0000398	157	540	0.032946	0.024053	0.03559	0.032637	0.05655	0.055835	0.07599	0.07214	nuclear mRNA splicing, via spliceosome	BP
GO:0016787	690	3423	0.047288	0.030796	0.052173	0.040299	0.079969	0.061638	0.078641	0.058643	hydrolase activity	MF
GO:0005625	242	1442	0.033228	0.025399	0.032718	0.030957	0.058825	0.056057	0.042506	0.046765	soluble fraction	CC
GO:0007596	250	1917	0.029639	0.021224	0.047444	0.035119	0.047667	0.051767	0.055077	0.053919	blood coagulation	BP
GO:0042470	77	349	0.029361	0.021334	0.02447	0.021969	0.065558	0.050236	0.065648	0.059203	melanosome	CC
GO:0006457	144	531	0.033489	0.028962	0.024752	0.025576	0.061372	0.057592	0.057928	0.053851	protein folding	BP
GO:0005794	648	3435	0.061752	0.039819	0.061565	0.04393	0.078185	0.067553	0.050552	0.057415	Golgi apparatus	CC
GO:0006915	438	2403	0.042704	0.032036	0.058945	0.047421	0.061623	0.059639	0.080756	0.064667	apoptosis	BP
GO:0016491	299	1501	0.02977	0.030428	0.02361	0.025715	0.05671	0.057085	0.046014	0.051143	oxidoreductase activity	MF
GO:0006184	102	550	0.010289	0.00823	0.010677	0.010086	0.020955	0.034412	0.017564	0.029229	GTP catabolic process	BP
GO:0051082	98	344	0.024217	0.020695	0.01991	0.020996	0.052397	0.046814	0.051097	0.045701	unfolded protein binding	MF
GO:0005525	231	1254	0.060679	0.039976	0.038858	0.031914	0.052462	0.063898	0.039386	0.050232	GTP binding	MF

GO:0005975	200	1020	0.024052	0.021633	0.017535	0.016672	0.050279	0.043944	0.046034	0.035814	carbohydrate metabolic process	BP
GO:0005789	444	2314	0.041613	0.029469	0.039817	0.031186	0.056872	0.051719	0.071712	0.07196	endoplasmic reticulum membrane	CC
GO:0016021	1943	13912	0.12101	0.087015	0.131349	0.102517	0.113958	0.107536	0.157464	0.162976	integral to membrane	CC
GO:0016887	98	452	0.012205	0.009056	0.017111	0.012567	0.040621	0.028672	0.040845	0.031311	ATPase activity	MF
GO:0055085	356	2525	0.026425	0.018163	0.035169	0.026055	0.044846	0.037765	0.073076	0.064187	transmembrane transport	BP
GO:0030168	118	1016	0.018543	0.014129	0.025386	0.0188	0.03097	0.033428	0.026838	0.021531	platelet activation	BP
GO:0006200	57	250	0.010276	0.008265	0.01488	0.011231	0.037708	0.027206	0.037888	0.029462	ATP catabolic process	BP
GO:0071013	76	253	0.020642	0.013563	0.020474	0.01616	0.035179	0.032437	0.049203	0.044187	catalytic step 2 spliceosome	CC
GO:0005488	444	2312	0.036697	0.025843	0.043748	0.033315	0.063443	0.044471	0.050021	0.040373	binding	MF
GO:0016192	140	650	0.015684	0.011262	0.012281	0.010297	0.029401	0.02962	0.020429	0.021862	vesicle-mediated transport	BP
GO:0007411	187	1316	0.026703	0.017955	0.033987	0.02185	0.047809	0.036	0.03761	0.026808	axon guidance	BP
GO:0006006	67	398	0.014176	0.0149	0.00765	0.008658	0.032216	0.032131	0.018375	0.019913	glucose metabolic process	BP
GO:0016740	443	2029	0.035523	0.028693	0.036035	0.031104	0.053771	0.045869	0.044094	0.044486	transferase activity	MF
GO:0003779	198	1222	0.02335	0.014653	0.041177	0.026135	0.051253	0.031213	0.052974	0.018345	actin binding	MF
GO:0022904	78	282	0.010255	0.020979	0.005435	0.015013	0.020438	0.0374	0.023662	0.042602	respiratory electron transport chain	BP
GO:0005856	606	3308	0.058544	0.049213	0.068653	0.059094	0.075452	0.065535	0.058636	0.057366	cytoskeleton	CC
GO:0019904	115	612	0.016441	0.010636	0.018995	0.014849	0.024966	0.026482	0.025474	0.025061	protein domain specific binding	MF
GO:0003924	141	752	0.051805	0.033419	0.031734	0.026263	0.037428	0.048955	0.029556	0.040638	GTPase activity	MF
GO:0006096	29	178	0.012121	0.013327	0.005838	0.007203	0.028142	0.028531	0.01505	0.015198	glycolysis	BP
GO:0007165	628	4542	0.055654	0.035406	0.060299	0.047832	0.057791	0.050545	0.052423	0.057563	signal transduction	BP
GO:0030971	11	114	0.008649	0.013232	0.00639	0.007568	0.002349	0.002947	0.002149	0.002198	receptor tyrosine kinase binding	MF
GO:0005102	104	961	0.019482	0.025224	0.011788	0.015031	0.013495	0.014749	0.012046	0.013139	receptor binding	MF
GO:0030335	52	409	0.012279	0.014638	0.008741	0.00927	0.004928	0.003818	0.004085	0.003286	positive regulation of cell migration	BP
GO:0043204	17	147	0.007913	0.013147	0.003291	0.006161	0.001894	0.00223	0.001287	0.001591	perikaryon	CC
GO:0032436	22	111	0.010396	0.014827	0.004345	0.0073	0.004213	0.003417	0.002667	0.002262	positive regulation of proteasomal ubiquitin-dependent protein catabolic process	BP
GO:0043547	37	254	0.010056	0.013974	0.005959	0.007288	0.003666	0.002546	0.003122	0.002163	positive regulation of	BP

											GTPase activity	
GO:0008284	169	1279	0.024954	0.027856	0.022127	0.020966	0.013436	0.016408	0.013145	0.014832	positive regulation of cell proliferation	BP
GO:0042593	31	234	0.008001	0.013105	0.004395	0.008346	0.001536	0.001405	0.000556	0.000758	glucose homeostasis	BP
GO:0042169	20	156	0.009194	0.013515	0.00502	0.00719	0.001926	0.001807	0.001364	0.001288	SH2 domain binding	MF
GO:0030425	90	621	0.014277	0.016652	0.009139	0.009908	0.006296	0.004904	0.004103	0.00385	dendrite	CC
GO:0030178	15	91	0.008028	0.012949	0.003788	0.006333	0.001007	0.001104	0.000798	0.000806	negative regulation of Wnt receptor signaling pathway	BP
GO:0048511	10	50	0.007609	0.01282	0.002664	0.005756	0.000733	0.000941	0.000422	0.000589	rhythmic process	BP
GO:0051726	41	220	0.010682	0.014616	0.007175	0.008948	0.004053	0.002682	0.003899	0.002658	regulation of cell cycle	BP
GO:0017148	17	103	0.012083	0.017088	0.004837	0.007757	0.004038	0.005022	0.00348	0.004328	negative regulation of translation	BP
GO:0040008	49	250	0.012202	0.016526	0.006266	0.009479	0.004183	0.004333	0.003603	0.004451	regulation of growth	BP
GO:0001934	42	278	0.010414	0.014067	0.006491	0.007706	0.001911	0.001579	0.001418	0.001399	positive regulation of protein phosphorylation	BP
GO:0008270	1143	7162	0.08386	0.056167	0.097387	0.071519	0.063901	0.03945	0.052274	0.036237	zinc ion binding	MF
GO:0043065	111	646	0.023501	0.031212	0.017939	0.022638	0.017138	0.014452	0.023832	0.019535	positive regulation of apoptosis	BP
GO:0015934	10	38	0.01072	0.020035	0.003991	0.009229	0.001625	0.003015	0.0019	0.003331	large ribosomal subunit	CC
GO:0003674	386	2105	0.032772	0.041671	0.029448	0.037419	0.024899	0.024256	0.01994	0.021453	molecular_function	MF
GO:0003729	41	211	0.018466	0.03554	0.010563	0.020871	0.011923	0.015733	0.012499	0.0158	mRNA binding	MF
GO:0006413	40	130	0.02094	0.034275	0.015848	0.022013	0.01182	0.014019	0.008007	0.009614	translational initiation	BP
GO:0042273	10	27	0.00958	0.025192	0.008277	0.029077	0.002437	0.004166	0.002072	0.003698	ribosomal large subunit biogenesis	BP
GO:0042254	28	85	0.012562	0.027733	0.006095	0.015637	0.003273	0.004885	0.003691	0.005367	ribosome biogenesis	BP
GO:0003746	17	79	0.045285	0.031812	0.019935	0.017412	0.009052	0.008462	0.00547	0.005422	translation elongation factor activity	MF
GO:0007275	443	3388	0.041516	0.038392	0.045535	0.037053	0.022771	0.01477	0.023387	0.014736	multicellular organismal development	BP
GO:0042274	11	34	0.01533	0.044131	0.00912	0.034037	0.003152	0.007661	0.002373	0.00585	ribosomal small subunit biogenesis	BP
GO:0019843	21	58	0.019626	0.05317	0.007958	0.028844	0.003456	0.007446	0.003089	0.007005	rRNA binding	MF
GO:0005730	429	1900	0.075173	0.114545	0.077831	0.109151	0.06936	0.06276	0.087932	0.077141	nucleolus	CC
GO:0006364	84	236	0.029006	0.073513	0.023985	0.070573	0.014488	0.019575	0.017418	0.021468	rRNA processing	BP
GO:0015935	16	62	0.028931	0.075786	0.013817	0.049943	0.005064	0.011855	0.003696	0.008872	small ribosomal subunit	CC

GO:0030529	95	416	0.083171	0.118695	0.050943	0.080217	0.053682	0.051568	0.062281	0.061747	ribonucleoprotein complex	CC
GO:0022625	34	109	0.056064	0.125895	0.02712	0.086786	0.010801	0.025034	0.007495	0.01728	cytosolic large ribosomal subunit	CC
GO:0003723	504	2292	0.15269	0.240052	0.12157	0.191674	0.119267	0.127315	0.135635	0.138667	RNA binding	MF
GO:0022627	32	100	0.061937	0.173238	0.026142	0.097605	0.013409	0.031887	0.008975	0.022414	cytosolic small ribosomal subunit	CC
GO:0005840	143	469	0.10628	0.26614	0.047984	0.156328	0.032801	0.071278	0.027883	0.060417	ribosome	CC
GO:0005829	1634	8692	0.369316	0.532487	0.280537	0.403723	0.328059	0.332868	0.253661	0.257336	cytosol	CC
GO:0005622	1180	6697	0.21673	0.365107	0.157222	0.266877	0.11459	0.14503	0.092671	0.113584	intracellular	CC
GO:0016032	268	1111	0.16874	0.38648	0.085608	0.241387	0.068255	0.111621	0.064276	0.091472	viral reproduction	BP
GO:0003735	144	479	0.148243	0.363859	0.067451	0.223718	0.034281	0.082215	0.028148	0.067506	structural constituent of ribosome	MF
GO:0016070	238	871	0.177833	0.390017	0.092348	0.245842	0.062642	0.107733	0.049697	0.08822	RNA metabolic process	BP
GO:0016071	203	746	0.174226	0.386783	0.08873	0.242156	0.058775	0.100212	0.045669	0.080137	mRNA metabolic process	BP
GO:0010467	361	1446	0.239667	0.432129	0.13933	0.28462	0.105771	0.139605	0.109123	0.129832	gene expression	BP
GO:0019058	83	284	0.152063	0.37148	0.06692	0.221226	0.033965	0.074026	0.023085	0.051966	viral infectious cycle	BP
GO:0006415	77	273	0.147869	0.366752	0.064603	0.218601	0.028355	0.068282	0.019052	0.047151	translational termination	BP
GO:0019083	74	259	0.14759	0.366594	0.06421	0.218302	0.028125	0.068055	0.018949	0.047024	viral transcription	BP
GO:0031018	92	374	0.148621	0.367162	0.065323	0.219685	0.02855	0.068491	0.020632	0.049269	endocrine pancreas development	BP
GO:0044267	240	929	0.224395	0.423099	0.111249	0.262106	0.075603	0.111339	0.065452	0.089585	cellular protein metabolic process	BP
GO:0006412	220	775	0.215326	0.418099	0.10781	0.260924	0.058624	0.10269	0.044418	0.081367	translation	BP
GO:0006414	85	309	0.190688	0.39615	0.083582	0.235132	0.035123	0.074307	0.023796	0.051742	translational elongation	BP

Table S 9: Sum of reads and spectra for top and bottom 50 in RWPE and VCaP by Gene Ontology class, ordered by VCaP relative enrichment

go_id	obs_genes	totalgenes	vcap_readfrac	vcap_rpkmfrac	rwpe_readfrac	rwpe_rpkmfrac	vcap_spcfrac	vcap_pnorm_frac	rwpe_spcfrac	rwpe_pnorm_frac	Term	go_tree
GO:0005739	1032	4669	0.09926	0.142869	0.076408	0.111656	0.188687	0.243711	0.216344	0.280198	mitochondrion	CC
GO:0000166	1356	7290	0.224941	0.188664	0.200768	0.180257	0.305077	0.246353	0.295978	0.246213	nucleotide binding	MF
GO:0005524	1027	5332	0.123656	0.099365	0.128403	0.106872	0.218386	0.140819	0.205702	0.137004	ATP binding	MF
GO:0005743	250	980	0.03097	0.058022	0.017356	0.034335	0.062727	0.090161	0.089242	0.12816	mitochondrial inner membrane	CC
GO:0005737	3565	19916	0.445929	0.438844	0.454505	0.45111	0.496185	0.470572	0.412106	0.39499	cytoplasm	CC

GO:0005759	175	707	0.018556	0.022439	0.012511	0.017287	0.050565	0.053971	0.065701	0.07087	mitochondrial matrix	CC
GO:0015031	340	1395	0.032485	0.031017	0.033424	0.032532	0.056293	0.061854	0.048772	0.051418	protein transport	BP
GO:0007264	214	1077	0.019981	0.01852	0.017432	0.016451	0.026054	0.048872	0.017763	0.036389	small GTPase mediated signal transduction	BP
GO:0006810	360	2081	0.039265	0.047674	0.039643	0.044258	0.063702	0.075229	0.067415	0.080216	transport	BP
GO:0005515	3178	17433	0.448299	0.433091	0.462981	0.43902	0.513955	0.459592	0.539346	0.468704	protein binding	MF
GO:0005654	735	3572	0.106254	0.109415	0.112952	0.124094	0.147627	0.134567	0.169323	0.147688	nucleoplasm	CC
GO:0006184	102	550	0.012124	0.013074	0.011458	0.013026	0.021762	0.037691	0.018127	0.031517	GTP catabolic process	BP
GO:0007596	250	1917	0.034926	0.033716	0.050915	0.045355	0.049503	0.056701	0.056842	0.05814	blood coagulation	BP
GO:0000398	157	540	0.038823	0.038209	0.038194	0.04215	0.058729	0.061156	0.078425	0.077788	nuclear mRNA splicing, via spliceosome	BP
GO:0008380	241	928	0.048237	0.046921	0.0491	0.052977	0.069372	0.069446	0.091851	0.088705	RNA splicing	BP
GO:0005829	1557	8692	0.262183	0.26837	0.232019	0.240008	0.311367	0.290212	0.242185	0.226858	cytosol	CC
GO:0005625	240	1442	0.038842	0.039951	0.034975	0.039852	0.061067	0.06138	0.043863	0.050421	soluble fraction	CC
GO:0003723	452	2292	0.08184	0.076481	0.089095	0.090659	0.106192	0.097684	0.126881	0.118968	RNA binding	MF
GO:0042470	77	349	0.034599	0.03389	0.02626	0.028373	0.068084	0.055024	0.067752	0.063837	melanosome	CC
GO:0016787	688	3423	0.055667	0.048812	0.055952	0.051953	0.082878	0.067271	0.080744	0.062608	hydrolase activity	MF
GO:0005783	701	3602	0.077613	0.076899	0.066882	0.065562	0.093984	0.095303	0.109859	0.120361	endoplasmic reticulum	CC
GO:0051082	98	344	0.028536	0.032875	0.021367	0.027116	0.054415	0.051275	0.052735	0.049279	unfolded protein binding	MF
GO:0006457	144	531	0.039462	0.046008	0.026563	0.033031	0.063736	0.063081	0.059785	0.058066	protein folding	BP
GO:0016887	98	452	0.014382	0.014386	0.018363	0.01623	0.042186	0.031404	0.042155	0.033762	ATPase activity	MF
GO:0006200	57	250	0.012109	0.01313	0.015968	0.014505	0.039161	0.029799	0.039102	0.031769	ATP catabolic process	BP
GO:0016192	140	650	0.018482	0.01789	0.013179	0.013299	0.030534	0.032442	0.021084	0.023574	vesicle-mediated transport	BP
GO:0016491	299	1501	0.03508	0.048337	0.025337	0.03321	0.058895	0.062526	0.047489	0.055147	oxidoreductase activity	MF
GO:0030168	118	1016	0.021851	0.022445	0.027243	0.02428	0.032163	0.036614	0.027698	0.023216	platelet activation	BP
GO:0071013	76	253	0.024324	0.021545	0.021972	0.020871	0.036535	0.035529	0.05078	0.047646	catalytic step 2 spliceosome	CC
GO:0005975	200	1020	0.028343	0.034365	0.018818	0.021531	0.052216	0.048132	0.04751	0.038618	carbohydrate metabolic process	BP
GO:0042645	34	112	0.006444	0.006803	0.00621	0.007291	0.022262	0.02055	0.036773	0.038771	mitochondrial nucleoid	CC
GO:0000278	272	1081	0.028805	0.029037	0.031756	0.033226	0.051933	0.041821	0.038406	0.03328	mitotic cell cycle	BP
GO:0055085	356	2525	0.031139	0.028854	0.037742	0.033649	0.046573	0.041364	0.075418	0.069212	transmembrane transport	BP
GO:0002576	45	357	0.011946	0.014557	0.016129	0.015933	0.021802	0.02656	0.015792	0.013425	platelet degranulation	BP

GO:0005741	84	462	0.008351	0.008846	0.008403	0.008253	0.016479	0.020752	0.02238	0.0317	mitochondrial outer membrane	CC
GO:0006915	435	2403	0.049705	0.049205	0.062406	0.058532	0.062395	0.060944	0.082184	0.06648	apoptosis	BP
GO:0019904	113	612	0.019319	0.016825	0.020337	0.019094	0.025661	0.028521	0.02573	0.026039	protein domain specific binding	MF
GO:0006006	67	398	0.016705	0.023669	0.00821	0.011182	0.033457	0.035193	0.018964	0.021472	glucose metabolic process	BP
GO:0016032	195	1111	0.026019	0.036531	0.023095	0.030517	0.041766	0.048009	0.046819	0.048067	viral reproduction	BP
GO:0016020	2211	12613	0.181184	0.198637	0.153947	0.162782	0.183354	0.209874	0.214391	0.257573	membrane	CC
GO:0005634	3648	19673	0.409774	0.388218	0.448025	0.431147	0.42985	0.399201	0.449588	0.394075	nucleus	CC
GO:0007411	182	1316	0.031131	0.028281	0.03622	0.02803	0.049295	0.039214	0.038685	0.028824	axon guidance	BP
GO:0003779	198	1222	0.027515	0.023277	0.04419	0.033753	0.053228	0.034188	0.054672	0.019781	actin binding	MF
GO:0005681	70	268	0.012993	0.013471	0.011549	0.013466	0.022802	0.024298	0.029582	0.031289	spliceosomal complex	CC
GO:0003697	51	172	0.00989	0.01	0.011558	0.014558	0.021256	0.020783	0.025645	0.023001	single-stranded DNA binding	MF
GO:0005794	648	3435	0.072767	0.063254	0.066069	0.056734	0.081197	0.073992	0.052172	0.06191	Golgi apparatus	CC
GO:0005694	184	874	0.021503	0.017247	0.023532	0.018393	0.03011	0.027744	0.027563	0.023922	chromosome	CC
GO:0005198	103	749	0.016371	0.012043	0.024662	0.0176	0.038161	0.022215	0.043212	0.031903	structural molecule activity	MF
GO:0006397	160	770	0.023358	0.020199	0.025587	0.02499	0.034313	0.030336	0.045759	0.040588	mRNA processing	BP
GO:0006096	29	178	0.014283	0.021171	0.006265	0.009302	0.029227	0.03125	0.015532	0.016388	glycolysis	BP
GO:0048471	277	1631	0.053092	0.065101	0.04363	0.048169	0.055118	0.054442	0.049994	0.047978	perinuclear region of cytoplasm	CC
GO:0009615	84	497	0.018879	0.032182	0.013248	0.024227	0.014377	0.021446	0.013418	0.018223	response to virus	BP
GO:0006874	22	218	0.008425	0.015768	0.005154	0.011775	0.003491	0.00339	0.003735	0.004583	cellular calcium ion homeostasis	BP
GO:0030496	59	305	0.017821	0.027685	0.012322	0.016528	0.016963	0.014623	0.01982	0.017406	midbody	CC
GO:0008134	178	1054	0.027869	0.028639	0.030417	0.031061	0.026542	0.015191	0.030668	0.013764	transcription factor binding	MF
GO:0043565	248	2052	0.023547	0.020015	0.021701	0.020183	0.00921	0.006563	0.009088	0.007237	sequence-specific DNA binding	MF
GO:0003674	375	2105	0.032016	0.035818	0.027151	0.027919	0.02398	0.021681	0.01864	0.018491	molecular_function	MF
GO:0042995	98	638	0.018776	0.030108	0.013554	0.016287	0.010364	0.015822	0.004928	0.006366	cell projection	CC
GO:0006351	338	2278	0.038969	0.036239	0.038198	0.038142	0.022869	0.0219	0.017758	0.016409	transcription, DNA-dependent	BP
GO:0030308	67	445	0.014944	0.026705	0.008565	0.013658	0.010609	0.011599	0.008288	0.010009	negative regulation of cell growth	BP
GO:0043065	110	646	0.022072	0.031978	0.016832	0.020268	0.01747	0.015256	0.024327	0.020565	positive regulation of apoptosis	BP
GO:0003700	489	3558	0.047693	0.039547	0.050307	0.047403	0.024664	0.022753	0.020303	0.018451	sequence-specific DNA	MF

											binding transcription factor activity	
GO:0019903	33	216	0.01557	0.025841	0.00907	0.012494	0.013619	0.008153	0.014578	0.00833	protein phosphatase binding	MF
GO:0030971	11	114	0.010192	0.02102	0.006858	0.009774	0.002439	0.003228	0.002218	0.00237	receptor tyrosine kinase binding	MF
GO:0005080	26	170	0.013196	0.023987	0.007296	0.012311	0.006216	0.006098	0.004532	0.005658	protein kinase C binding	MF
GO:0043025	106	723	0.019945	0.030821	0.015667	0.015833	0.01164	0.012804	0.007669	0.006857	neuronal cell body	CC
GO:0032880	17	114	0.010869	0.021926	0.004917	0.009165	0.002589	0.003901	0.001982	0.002928	regulation of protein localization	BP
GO:0043204	17	147	0.009325	0.020885	0.003532	0.007956	0.001967	0.002443	0.001328	0.001716	perikaryon	CC
GO:0030335	52	409	0.01447	0.023254	0.00938	0.011973	0.005118	0.004182	0.004216	0.003544	positive regulation of cell migration	BP
GO:0048511	10	50	0.008967	0.020365	0.002859	0.007434	0.000761	0.001031	0.000436	0.000635	rhythmic process	BP
GO:0030178	15	91	0.00946	0.020571	0.004065	0.00818	0.001046	0.001209	0.000823	0.000869	negative regulation of Wnt receptor signaling pathway	BP
GO:0043547	37	254	0.01185	0.022199	0.006395	0.009413	0.003807	0.002789	0.003222	0.002332	positive regulation of GTPase activity	BP
GO:0042169	20	156	0.010834	0.021469	0.005387	0.009285	0.002	0.001979	0.001408	0.001388	SH2 domain binding	MF
GO:0032436	22	111	0.012251	0.023554	0.004663	0.009427	0.004375	0.003743	0.002752	0.002439	positive regulation of proteasomal ubiquitin-dependent protein catabolic process	BP
GO:0051726	41	220	0.012588	0.023218	0.0077	0.011557	0.004209	0.002938	0.004024	0.002866	regulation of cell cycle	BP
GO:0005622	1074	6697	0.113946	0.107319	0.109151	0.099066	0.0908	0.086897	0.073798	0.067046	intracellular	CC
GO:0016021	1943	13912	0.142596	0.138229	0.140959	0.132399	0.118348	0.117785	0.162512	0.175735	integral to membrane	CC
GO:0001934	42	278	0.012272	0.022345	0.006966	0.009953	0.001985	0.001729	0.001464	0.001509	positive regulation of protein phosphorylation	BP
GO:0030425	90	621	0.016824	0.026453	0.009808	0.012796	0.006538	0.005371	0.004234	0.004152	dendrite	CC
GO:0046872	1775	10580	0.140818	0.112463	0.147036	0.126006	0.111399	0.091297	0.100019	0.084938	metal ion binding	MF
GO:0040008	49	250	0.014378	0.026252	0.006724	0.012242	0.004344	0.004746	0.003719	0.0048	regulation of growth	BP
GO:0017148	17	103	0.014239	0.027146	0.005191	0.010018	0.004194	0.0055	0.003591	0.004666	negative regulation of translation	BP
GO:0005102	104	961	0.022957	0.040069	0.01265	0.019413	0.014015	0.016154	0.012432	0.014168	receptor binding	MF
GO:0006355	931	5999	0.09235	0.077149	0.10105	0.090573	0.060527	0.051787	0.062389	0.052149	regulation of transcription, DNA-dependent	BP
GO:0008270	1140	7162	0.096292	0.070918	0.103422	0.083522	0.066123	0.041933	0.053771	0.038123	zinc ion binding	MF
GO:0010467	288	1446	0.109598	0.109048	0.080747	0.086352	0.080727	0.07866	0.093105	0.08943	gene expression	BP
GO:0007275	442	3388	0.045729	0.052317	0.048106	0.045414	0.023155	0.01536	0.023711	0.015141	multicellular organismal	BP

											development	
GO:0003746	17	79	0.053363	0.050535	0.021394	0.022488	0.009401	0.009269	0.005645	0.005846	translation elongation factor activity	MF
GO:0006414	12	309	0.051882	0.051894	0.02092	0.02244	0.007357	0.007139	0.005042	0.005226	translational elongation	BP
GO:0044267	167	929	0.091601	0.094704	0.050612	0.057277	0.049396	0.0477	0.048033	0.046032	cellular protein metabolic process	BP
GO:0006412	80	775	0.076759	0.077957	0.043805	0.04771	0.025011	0.021246	0.017179	0.014978	translation	BP

Table S 10: Sum of reads and spectra for top and bottom 50 in RWPE and VCaP by Gene Ontology class with ribosomal genes removed, ordered by VCaP relative enrichment

go_id	num_genes_obs	num_genes_class	pearson_cor	pearson_p	spearman_cor	spearman_p	prot_mean	tx_mean	prot_median	tx_median	cor_pearson_bh	cor_spearman_bh	Term	go_tree
GO:0000155	6	56	-0.14516	0.783795	0.028571	1	1.706637	4.125767	2.85886	4.322668	0.810463	1	two-component sensor activity	MF
GO:0005123	7	68	-0.13188	0.778058	0	1	1.885487	3.284669	2.069755	2.950255	0.805816	1	death receptor binding	MF
GO:0006264	6	34	0.274419	0.598705	-0.02857	1	2.968612	3.337416	2.914576	3.463296	0.640528	1	mitochondrial DNA replication	BP
GO:0006809	6	54	0.533425	0.275753	-0.02857	1	4.49339	3.884449	5.127782	4.151512	0.321354	1	nitric oxide biosynthetic process	BP
GO:0006978	6	105	0.077446	0.884063	-0.02857	1	1.58638	4.46804	0.96182	4.910636	0.897912	1	DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	BP
GO:0007096	7	23	0.126086	0.787647	0	1	2.919312	4.08101	2.94712	4.026579	0.814014	1	regulation of exit from mitosis	BP
GO:0030217	7	83	-0.10794	0.817813	0	1	1.06325	3.954415	1.741783	3.563928	0.840723	1	T cell differentiation	BP
GO:0032402	6	32	-0.35827	0.48559	0.028571	1	1.943608	2.525552	1.977075	2.876346	0.531068	1	melanosome transport	BP
GO:0043240	6	33	0.776588	0.069294	0.028571	1	0.312925	2.929093	0.033833	2.449059	0.099714	1	Fanconi anaemia nuclear complex	CC
GO:0051974	6	55	0.035615	0.9466	0.028571	1	2.91232	3.448586	2.944651	3.483562	0.954255	1	negative regulation of telomerase activity	BP

GO:0030897	12	40	0.383211	0.218846	0.006993	0.99123	2.950357	3.619346	2.904523	3.716765	0.262104	0.996353	HOPS complex	CC
GO:0045022	9	34	-0.15323	0.693881	-0.01667	0.98157	2.466359	3.382254	2.745233	3.604485	0.728725	0.987153	early endosome to late endosome transport	BP
GO:0009953	8	106	0.080765	0.849222	-0.02381	0.976786	-0.23555	1.660737	-0.66064	1.640695	0.868421	0.98285	dorsal/ventral pattern formation	BP
GO:0022857	11	127	0.058339	0.864713	-0.01818	0.967576	7.787493	7.93148	7.96804	7.943038	0.881019	0.974087	transmembrane transporter activity	MF
GO:0042776	11	56	0.058339	0.864713	-0.01818	0.967576	7.787493	7.93148	7.96804	7.943038	0.881019	0.974087	mitochondrial ATP synthesis coupled proton transport	BP
GO:0000780	7	52	-0.03206	0.945605	-0.03571	0.963492	1.087055	3.712705	1.251539	3.794842	0.954255	0.97098	condensed nuclear chromosome, centromeric region	CC
GO:0014047	7	112	-0.03184	0.945977	-0.03571	0.963492	3.744536	4.540525	4.603214	5.134869	0.954255	0.97098	glutamate secretion	BP
GO:0034361	7	57	0.348357	0.443831	0.035714	0.963492	3.101188	3.933392	2.979942	3.871388	0.490484	0.97098	very-low-density lipoprotein particle	CC
GO:0001947	12	112	0.112892	0.726846	0.020979	0.956169	0.781019	3.707599	0.205568	3.844372	0.757618	0.965101	heart looping	BP
GO:0021510	9	48	-0.01271	0.97411	-0.03333	0.948391	2.1024	4.196527	1.798366	4.22595	0.977507	0.957747	spinal cord development	BP
GO:0005665	10	30	-0.19658	0.586209	0.030303	0.94571	4.88888	6.077029	5.301099	5.665269	0.628196	0.955812	DNA-directed RNA polymerase II, core complex	CC
GO:0007076	11	56	0.744044	0.008648	0.027273	0.945984	3.336622	4.020329	3.72547	4.004515	0.017781	0.955812	mitotic chromosome condensation	BP
GO:0007224	10	98	-0.01976	0.956792	-0.0303	0.94571	0.423537	2.937403	0.099753	2.827569	0.962732	0.955812	smoothened signaling pathway	BP
GO:0001578	8	66	0.199837	0.635162	0.047619	0.934871	2.471787	3.858481	2.194879	3.995346	0.674339	0.946058	microtubule bundle formation	BP
GO:0005885	6	24	0.060362	0.909568	0.085714	0.919444	5.862062	6.215573	5.845399	6.01438	0.922372	0.930932	Arp2/3 protein complex	CC
GO:0006402	6	42	0.154606	0.769939	-0.08571	0.919444	3.600288	3.537144	3.18834	3.402525	0.798258	0.930932	mRNA catabolic process	BP
GO:0008430	6	46	0.331697	0.520702	0.085714	0.919444	5.001446	5.287518	5.188867	5.312317	0.567057	0.930932	selenium binding	MF
GO:0019370	6	82	0.178169	0.735575	0.085714	0.919444	4.720371	5.335455	4.712433	5.367478	0.765076	0.930932	leukotriene biosynthetic process	BP
GO:0019509	6	14	0.329718	0.523346	0.085714	0.919444	4.536792	4.931061	5.370887	4.85833	0.569618	0.930932	L-methionine salvage from methylthioadenosine	BP
GO:0032012	6	51	-0.12423	0.814612	-0.08571	0.919444	2.946582	4.808346	4.049818	4.762404	0.838318	0.930932	regulation of ARF protein signal transduction	BP

GO:0042719	6	15	0.214519	0.683157	0.085714	0.919444	6.258884	5.022933	6.200121	4.800543	0.718193	0.930932	mitochondrial intermembrane space protein transporter complex	CC
GO:0050873	6	63	0.254166	0.626961	0.085714	0.919444	3.766627	4.328653	3.900917	3.734092	0.667089	0.930932	brown fat cell differentiation	BP
GO:0051775	6	33	0.260688	0.617826	-0.08571	0.919444	3.57828	3.666702	2.98081	4.502902	0.658089	0.930932	response to redox state	BP
GO:0005249	13	204	-0.0197	0.949074	0.038462	0.906202	0.897892	3.154025	0.655172	3.0916	0.955954	0.921992	voltage-gated potassium channel activity	MF
GO:0005852	14	36	0.000727	0.998031	0.037363	0.903515	5.654854	7.128447	5.656926	7.336547	0.998031	0.921992	eukaryotic translation initiation factor 3 complex	CC
GO:0017016	7	70	0.216712	0.640676	0.071429	0.906349	3.34035	4.908538	2.823527	5.431044	0.679081	0.921992	Ras GTPase binding	MF
GO:0042994	7	28	0.086827	0.853152	-0.07143	0.906349	1.593033	4.282385	0.807175	4.045346	0.871982	0.921992	cytoplasmic sequestering of transcription factor	BP
GO:0046902	7	42	-0.02229	0.962174	0.071429	0.906349	4.105835	5.641096	4.102837	5.547346	0.967147	0.921992	regulation of mitochondrial membrane permeability	BP
GO:0060170	7	51	-0.37459	0.407732	0.071429	0.906349	0.520899	2.96779	-0.10098	3.868971	0.453943	0.921992	cilium membrane	CC
GO:0070776	7	20	0.62817	0.130868	0.071429	0.906349	1.316469	4.098003	0.953359	3.789622	0.171095	0.921992	MOZ/MORF histone acetyltransferase complex	CC
GO:0006595	10	41	0.056151	0.877556	0.054545	0.891639	4.255821	5.781861	5.265638	6.148583	0.892235	0.910361	polyamine metabolic process	BP
GO:0005838	8	25	-0.04318	0.919138	-0.07143	0.881994	5.325698	5.717427	5.488394	5.702749	0.930621	0.900987	proteasome regulatory particle	CC
GO:0030216	14	143	0.053804	0.855052	-0.05055	0.867596	4.11148	3.986381	4.967306	4.158697	0.873349	0.886745	keratinocyte differentiation	BP
GO:0004540	10	49	0.0664	0.855388	0.066667	0.864754	3.044213	3.402404	3.177262	3.300741	0.873349	0.884304	ribonuclease activity	MF
GO:0035035	10	65	0.056663	0.876447	0.066667	0.864754	1.114761	4.349615	1.192009	4.341432	0.891574	0.884304	histone acetyltransferase binding	MF
GO:0005762	12	53	-0.10149	0.753637	-0.06294	0.851682	5.99713	5.341713	5.797267	5.602616	0.782608	0.871853	mitochondrial large ribosomal subunit	CC
GO:0000060	8	48	0.295673	0.477076	0.095238	0.840129	3.499068	5.574924	3.960934	5.353057	0.522493	0.86048	protein import into nucleus, translocation	BP
GO:0004180	8	64	0.172182	0.683483	0.095238	0.840129	4.566439	4.856274	4.754083	4.633416	0.718193	0.86048	carboxypeptidase activity	MF
GO:0004708	7	56	-0.16974	0.71598	-0.10714	0.839683	2.976688	4.748909	4.189957	4.67014	0.748141	0.86048	MAP kinase kinase activity	MF
GO:0004812	8	18	0.045462	0.914876	-0.09524	0.840129	3.988127	4.759914	4.225371	4.81864	0.926788	0.86048	aminoacyl-tRNA ligase activity	MF

GO:0007613	7	125	0.208374	0.653882	0.107143	0.839683	1.379606	3.500382	1.184834	3.592548	0.690446	0.86048	memory	BP
GO:0007616	8	56	-0.08513	0.841159	-0.09524	0.840129	0.666376	3.591943	0.709482	3.530432	0.861081	0.86048	long-term memory	BP
GO:0016574	7	16	-0.04364	0.925981	-0.10714	0.839683	4.264455	5.277682	4.573467	5.030299	0.937062	0.86048	histone ubiquitination	BP
GO:0034045	7	40	-0.17632	0.705304	-0.10714	0.839683	0.840781	3.935966	1.327348	3.780694	0.739524	0.86048	pre-autophagosomal structure membrane	CC
GO:0051298	7	24	0.403053	0.369952	0.107143	0.839683	0.509864	2.244409	0.101317	2.542946	0.418104	0.86048	centrosome duplication	BP
GO:0070652	7	25	-0.07939	0.865641	0.107143	0.839683	1.726521	2.996759	1.876996	3.306382	0.881042	0.86048	HAUS complex	CC
GO:0005763	18	51	0.004591	0.985576	0.05676	0.824197	5.749246	5.460994	5.526888	5.648383	0.987607	0.84863	mitochondrial small ribosomal subunit	CC
GO:0016758	10	76	0.219678	0.541981	0.090909	0.811417	1.127205	3.601567	0.914634	3.885428	0.586619	0.835914	transferase activity, transferring hexosyl groups	MF
GO:0031080	10	32	0.344109	0.33024	0.090909	0.811417	4.472784	4.096785	4.450308	4.133036	0.377168	0.835914	Nup107-160 complex	CC
GO:0000127	6	18	-0.14164	0.788964	0.142857	0.802778	3.802158	5.203798	4.033205	5.121666	0.814077	0.827891	transcription factor TFIIIC complex	CC
GO:0005869	6	14	0.146653	0.781598	0.142857	0.802778	3.93185	5.354414	3.658466	5.329118	0.808622	0.827891	dynactin complex	CC
GO:0006970	6	60	0.146807	0.781372	0.142857	0.802778	3.959046	4.787542	4.124674	4.797692	0.808622	0.827891	response to osmotic stress	BP
GO:0008045	6	55	0.525475	0.284335	0.142857	0.802778	1.054204	3.8056	0.963417	3.690749	0.330761	0.827891	motor axon guidance	BP
GO:0009967	6	40	0.150969	0.775266	0.142857	0.802778	1.979504	3.022653	2.059615	3.072051	0.803353	0.827891	positive regulation of signal transduction	BP
GO:0016597	6	69	0.758267	0.08059	0.142857	0.802778	5.681765	5.033042	5.265131	4.799985	0.113175	0.827891	amino acid binding	MF
GO:0032040	6	18	0.643882	0.167649	-0.14286	0.802778	5.125832	4.970863	4.949758	4.303314	0.209292	0.827891	small-subunit processome	CC
GO:0033197	6	24	0.036412	0.945407	-0.14286	0.802778	3.727987	5.07015	4.210584	4.861769	0.954255	0.827891	response to vitamin E	BP
GO:0035329	6	40	0.192038	0.715484	0.142857	0.802778	1.335753	3.886391	0.906091	3.938212	0.748141	0.827891	hippo signaling cascade	BP
GO:0042791	6	18	-0.14164	0.788964	0.142857	0.802778	3.802158	5.203798	4.033205	5.121666	0.814077	0.827891	5S class rRNA transcription from RNA polymerase III type 1 promoter	BP
GO:0042797	6	18	-0.14164	0.788964	0.142857	0.802778	3.802158	5.203798	4.033205	5.121666	0.814077	0.827891	tRNA transcription from RNA polymerase III	BP

													promoter	
GO:0045821	6	46	0.109416	0.836531	0.142857	0.802778	1.622886	3.225132	0.953598	3.465567	0.857246	0.827891	positive regulation of glycolysis	BP
GO:0050872	6	32	0.305533	0.555962	0.142857	0.802778	3.470047	4.050053	4.327572	3.433872	0.600414	0.827891	white fat cell differentiation	BP
GO:0051787	6	17	0.704671	0.11795	0.142857	0.802778	3.480283	4.52336	3.558056	3.518262	0.155957	0.827891	misfolded protein binding	MF
GO:0072321	6	23	0.311223	0.548238	0.142857	0.802778	5.890992	5.273052	5.519437	5.168467	0.59306	0.827891	chaperone-mediated protein transport	BP
GO:0004860	8	83	-0.37973	0.353493	-0.11905	0.793006	2.451951	4.469697	1.94218	4.225808	0.401837	0.82437	protein kinase inhibitor activity	MF
GO:0015450	8	32	-0.01361	0.974492	0.119048	0.793006	4.864972	5.559191	4.84356	5.573749	0.977507	0.82437	P-P-bond-hydrolysis-driven protein transmembrane transporter activity	MF
GO:0005385	7	63	-0.26329	0.568351	-0.14286	0.78254	1.906459	3.302833	1.979942	3.121176	0.61108	0.81436	zinc ion transmembrane transporter activity	MF
GO:0045039	7	17	0.291589	0.525756	0.142857	0.78254	5.904888	4.894105	5.6878	4.427176	0.571922	0.81436	protein import into mitochondrial inner membrane	BP
GO:0048010	7	143	0.224831	0.627888	0.142857	0.78254	2.036957	4.292898	2.629081	4.548084	0.66771	0.81436	vascular endothelial growth factor receptor signaling pathway	BP
GO:0009303	9	48	-0.06875	0.860493	0.116667	0.775628	3.037248	4.014229	3.778786	4.467472	0.8781	0.808466	rRNA transcription	BP
GO:0009434	9	84	-0.21274	0.582624	-0.11667	0.775628	1.939946	2.586505	1.744197	2.760648	0.625388	0.808466	microtubule-based flagellum	CC
GO:0080008	9	49	0.279484	0.466411	0.116667	0.775628	1.868996	4.498818	1.576045	4.807762	0.512814	0.808466	CUL4 RING ubiquitin ligase complex	CC
GO:0005753	12	72	0.170314	0.596662	0.097902	0.766288	7.632982	7.860837	7.926093	7.830071	0.638694	0.800017	mitochondrial proton-transporting ATP synthase complex	CC
GO:0008076	13	303	-0.02124	0.945086	0.093407	0.764582	1.0571	3.513766	0.986053	3.273362	0.954255	0.798663	voltage-gated potassium channel complex	CC
GO:0007093	11	57	0.042351	0.901603	0.118182	0.734252	2.617307	3.788773	2.445362	3.9771	0.914772	0.767394	mitotic cell cycle checkpoint	BP
GO:0001078	6	47	0.107315	0.839645	-0.2	0.713889	1.42974	1.48496	1.239537	1.787778	0.859984	0.746513	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription	MF

GO:0005786	6	24	0.379296	0.45834	0.2	0.713889	6.498161	6.241037	6.452568	6.019001	0.505083	0.746513	signal recognition particle, endoplasmic reticulum targeting	CC
GO:0006699	7	80	0.222599	0.631397	0.178571	0.713095	4.865813	4.340004	4.890856	4.431447	0.671075	0.746513	bile acid biosynthetic process	BP
GO:0016601	6	43	-0.0092	0.986205	0.2	0.713889	3.823909	5.341234	2.615023	5.908155	0.987728	0.746513	Rac protein signal transduction	BP
GO:0016705	6	125	-0.02907	0.956404	0.2	0.713889	2.737547	4.24624	3.438236	4.468448	0.962732	0.746513	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	MF
GO:0030132	6	34	-0.18242	0.72941	-0.2	0.713889	5.888782	5.723061	5.45786	6.291793	0.759477	0.746513	clathrin coat of coated pit	CC
GO:0030658	6	67	0.293838	0.571927	0.2	0.713889	3.852178	3.319619	3.693018	3.616373	0.614585	0.746513	transport vesicle membrane	CC
GO:0032886	7	36	0.668333	0.100772	0.178571	0.713095	2.572378	3.988237	3.836501	4.165429	0.137064	0.746513	regulation of microtubule-based process	BP
GO:0033189	6	44	-0.30745	0.553355	-0.2	0.713889	1.901814	2.491207	1.735519	2.491306	0.597931	0.746513	response to vitamin A	BP
GO:0042220	6	66	-0.30999	0.549914	-0.2	0.713889	2.830842	3.892832	2.878781	3.800731	0.594543	0.746513	response to cocaine	BP
GO:0045749	6	54	0.574724	0.232832	0.2	0.713889	2.353655	3.885501	2.84048	4.155438	0.277317	0.746513	negative regulation of S phase of mitotic cell cycle	BP
GO:0045861	7	56	0.068475	0.884025	-0.17857	0.713095	2.594123	3.461883	4.22095	3.205879	0.897912	0.746513	negative regulation of proteolysis	BP
GO:0046856	7	28	0.060771	0.897023	-0.17857	0.713095	2.309937	3.756066	2.372793	4.228631	0.9106	0.746513	phosphatidylinositol dephosphorylation	BP
GO:0046965	6	68	0.136091	0.797124	0.2	0.713889	2.123397	4.430161	1.397027	4.530316	0.822061	0.746513	retinoid X receptor binding	MF
GO:0048844	6	104	0.410202	0.419208	0.2	0.713889	1.891024	4.184359	1.873121	4.363168	0.465122	0.746513	artery morphogenesis	BP
GO:0070628	7	39	0.84539	0.01658	0.178571	0.713095	3.722316	4.27441	4.7994	4.802679	0.030309	0.746513	proteasome binding	MF
GO:0004857	10	116	0.138793	0.702171	0.139394	0.707204	3.694002	4.469526	4.465802	3.632776	0.737034	0.745939	enzyme inhibitor activity	MF
GO:0005086	8	55	-0.01779	0.966656	0.166667	0.703323	2.834204	4.452866	3.694324	4.648291	0.97115	0.742249	ARF guanyl-nucleotide exchange factor activity	MF
GO:0010039	8	61	0.24615	0.556773	0.166667	0.703323	2.685636	4.22205	2.714877	4.076957	0.600956	0.742249	response to iron ion	BP
GO:0033205	8	24	0.301171	0.468522	0.166667	0.70332	2.466489	4.175228	2.313894	4.512244	0.514554	0.742	cell cycle cytokinesis	BP

						3						249		
GO:0035176	8	88	-0.10291	0.808409	-0.16667	0.703323	0.730298	2.763651	0.380696	2.011826	0.832815	0.742249	social behavior	BP
GO:0048589	11	86	0.129165	0.705057	0.136364	0.693511	2.094279	4.349534	2.037414	4.482276	0.739524	0.733485	developmental growth	BP
GO:0005798	9	40	0.177061	0.648585	0.166667	0.677745	4.270202	4.745904	4.11933	4.555844	0.686716	0.7172	Golgi-associated vesicle	CC
GO:0010388	9	32	-0.19562	0.613976	-0.16667	0.677745	5.212592	5.454356	5.152003	5.185979	0.654347	0.7172	cullin deneddylation	BP
GO:0043015	9	44	0.267954	0.485743	0.166667	0.677745	0.765909	2.905093	0.397828	3.16564	0.531068	0.7172	gamma-tubulin binding	MF
GO:0071479	13	82	-0.06873	0.823468	0.131868	0.669269	3.754099	4.544498	4.154754	4.667042	0.845642	0.709389	cellular response to ionizing radiation	BP
GO:0006342	8	45	-0.09394	0.824903	0.190476	0.664583	1.68228	3.320104	1.711693	3.321277	0.846668	0.704806	chromatin silencing	BP
GO:0034968	8	40	-0.09873	0.816088	0.190476	0.664583	1.758078	5.185802	2.559593	5.425431	0.839393	0.704806	histone lysine methylation	BP
GO:0000028	6	18	0.86197	0.027264	0.257143	0.658333	7.336446	10.69022	7.633733	11.44902	0.045911	0.702311	ribosomal small subunit assembly	BP
GO:0002089	7	42	0.619982	0.137499	0.214286	0.661508	1.564496	2.79103	0.873027	3.265312	0.177817	0.702311	lens morphogenesis in camera-type eye	BP
GO:0004497	7	176	0.324888	0.477094	0.214286	0.661508	3.120909	4.353972	3.347399	4.186589	0.522493	0.702311	monooxygenase activity	MF
GO:0004691	6	57	0.56001	0.247798	0.257143	0.658333	3.366666	4.509872	3.361766	4.371856	0.292278	0.702311	cAMP-dependent protein kinase activity	MF
GO:0005655	7	27	0.023671	0.959826	0.214286	0.661508	3.927042	3.736024	3.354022	3.411782	0.965285	0.702311	nucleolar ribonuclease P complex	CC
GO:0005675	10	30	0.279228	0.434625	0.163636	0.656721	2.172685	3.92879	2.099752	3.981999	0.481131	0.702311	holo TFIIH complex	CC
GO:0006002	6	25	0.337619	0.512814	0.257143	0.658333	6.450405	5.89908	6.466776	6.070004	0.559093	0.702311	fructose 6-phosphate metabolic process	BP
GO:0006024	6	34	0.561501	0.246264	0.257143	0.658333	2.90425	4.301344	2.332168	4.247647	0.290822	0.702311	glycosaminoglycan biosynthetic process	BP
GO:0006613	7	26	0.171048	0.713848	-0.21429	0.661508	5.610698	5.650862	5.775539	5.71471	0.746872	0.702311	cotranslational protein targeting to membrane	BP
GO:0006614	6	26	0.371912	0.467853	0.257143	0.658333	6.229818	6.227187	6.452568	5.977454	0.51411	0.702311	SRP-dependent cotranslational protein targeting to membrane	BP
GO:0006777	6	30	0.418393	0.409031	0.257143	0.658333	3.517156	3.667566	3.661834	3.39619	0.454608	0.702311	Mo-molybdopterin cofactor biosynthetic process	BP

GO:0007628	7	80	-0.54028	0.210579	-0.21429	0.661508	3.071974	3.646511	3.541618	3.650579	0.253451	0.702311	adult walking behavior	BP
GO:0008053	7	37	0.327517	0.473325	0.214286	0.661508	4.502082	4.585344	4.371054	4.256274	0.519243	0.702311	mitochondrial fusion	BP
GO:0008121	6	18	-0.03542	0.946896	-0.25714	0.658333	7.287192	8.099733	7.22068	8.042117	0.954255	0.702311	ubiquinol-cytochrome-c reductase activity	MF
GO:0008198	7	63	-0.21721	0.639893	-0.21429	0.661508	3.244306	3.849059	2.836501	4.195741	0.678621	0.702311	ferrous iron binding	MF
GO:0008312	6	24	0.405308	0.425329	0.257143	0.658333	6.376076	6.212075	6.452568	5.932117	0.471645	0.702311	7S RNA binding	MF
GO:0008378	7	73	0.380211	0.400158	0.214286	0.661508	1.148832	3.113289	1.50226	3.333375	0.447303	0.702311	galactosyltransferase activity	MF
GO:0008535	7	28	-0.09189	0.844654	-0.21429	0.661508	2.624227	3.580697	2.586406	3.741468	0.864204	0.702311	respiratory chain complex IV assembly	BP
GO:0016303	6	24	-0.02094	0.968588	0.257143	0.658333	1.165125	2.800004	0.830173	2.610031	0.972588	0.702311	1-phosphatidylinositol-3-kinase activity	MF
GO:0030127	7	38	0.209832	0.651568	0.214286	0.661508	4.498504	4.467927	5.09278	4.947864	0.688604	0.702311	COPII vesicle coat	CC
GO:0033327	7	74	0.239356	0.605195	0.214286	0.661508	4.698243	5.465353	3.816037	5.664749	0.646406	0.702311	Leydig cell differentiation	BP
GO:0045494	6	107	-0.37973	0.45778	-0.25714	0.658333	1.330218	3.134156	1.267936	3.306436	0.504752	0.702311	photoreceptor cell maintenance	BP
GO:0046934	6	30	0.246694	0.637466	0.257143	0.658333	0.138123	3.440317	-0.11229	3.855756	0.676417	0.702311	phosphatidylinositol-4,5-bisphosphate 3-kinase activity	MF
GO:0048193	6	20	0.325252	0.529327	0.257143	0.658333	3.883264	4.755331	3.755475	4.632732	0.57452	0.702311	Golgi vesicle transport	BP
GO:0048469	6	124	0.284398	0.584904	0.257143	0.658333	2.831301	4.241518	2.286359	3.198199	0.627143	0.702311	cell maturation	BP
GO:0048813	7	90	0.444746	0.317376	0.214286	0.661508	2.479469	4.166878	2.249371	4.403025	0.364746	0.702311	dendrite morphogenesis	BP
GO:0048839	6	81	0.26652	0.609685	0.257143	0.658333	2.449564	3.753587	2.459437	3.848262	0.650131	0.702311	inner ear development	BP
GO:0070330	6	78	0.059672	0.910598	0.257143	0.658333	1.419744	3.911336	1.422053	3.624144	0.922935	0.702311	aromatase activity	MF
GO:0071577	6	49	-0.18976	0.718783	0.257143	0.658333	1.452174	4.650851	0.223212	4.561511	0.75042	0.702311	zinc ion transmembrane transport	BP
GO:0008138	15	116	0.253647	0.361668	0.132143	0.638933	2.08685	3.869725	1.556393	3.840429	0.409694	0.689254	protein tyrosine/serine/threonine phosphatase activity	MF
GO:0005774	8	33	0.343713	0.404496	0.214286	0.61909	2.44747	4.7893	2.630444	4.506983	0.450866	0.668	vacuolar membrane	CC

						7						226		
GO:0006939	8	59	0.424742	0.294207	0.214286	0.619097	1.502986	3.867751	1.7486	3.566193	0.340818	0.668226	smooth muscle contraction	BP
GO:0000080	14	93	0.183091	0.530971	0.147415	0.615026	2.646912	4.406066	2.435286	3.871792	0.575662	0.664569	G1 phase of mitotic cell cycle	BP
GO:0005978	9	55	0.205222	0.596322	0.2	0.613404	4.007587	3.366678	4.440263	3.550875	0.638682	0.663185	glycogen biosynthetic process	BP
GO:0046326	9	108	0.124223	0.750167	0.2	0.613404	2.486307	4.07297	2.629081	4.018281	0.779421	0.663185	positive regulation of glucose import	BP
GO:0090263	12	174	0.198823	0.535598	0.167832	0.603728	1.870801	4.63801	1.375893	4.575197	0.580032	0.65345	positive regulation of canonical Wnt receptor signaling pathway	BP

Table S 11: Top 150 Highest Transcript-protein correlation by Gene Otology class in VCaP

go_id	num_genes_obs	num_genes_class	pearson_cor	pearson_p	spearman_cor	spearman_p	prot_mean	tx_mean	prot_median	tx_median	cor_pearson_bh	cor_spearman_bh	Term	go_tree
GO:0000083	11	79	0.578206	0.062422	-0.00909	0.98919	0.943491	4.405098	0.690824	3.859411	0.106015	1	regulation of transcription involved in G1/S phase of mitotic cell cycle	BP
GO:0000780	6	52	0.069057	0.89658	0.028571	1	1.092294	3.490959	1.534296	2.941705	0.908358	1	condensed nuclear chromosome, centromeric region	CC
GO:0001932	6	90	-0.09384	0.859648	0.028571	1	4.081225	5.319534	4.233211	4.877704	0.878866	1	regulation of protein phosphorylation	BP
GO:0005484	16	63	0.04252	0.875754	-0.00294	0.991368	4.082291	4.527798	4.394332	4.368086	0.892412	1	SNAP receptor activity	MF
GO:0005885	6	24	-0.30081	0.562395	-0.02857	1	5.977803	6.504942	6.106591	6.536278	0.612719	1	Arp2/3 protein complex	CC
GO:0006402	6	42	-0.07763	0.883792	0.028571	1	3.857025	5.245389	4.625575	4.830376	0.899137	1	mRNA catabolic process	BP
GO:0006744	8	22	0.073752	0.862215	0	1	4.552738	3.912085	4.497401	3.833662	0.88101	1	ubiquinone	BP

													biosynthetic process	
GO:0006836	6	123	0.310487	0.549236	-0.02857	1	1.159222	4.647837	0.969639	5.019907	0.601954	1	neurotransmitter transport	BP
GO:0008378	9	73	0.3119	0.413885	0	1	1.505785	3.813266	0.780909	3.818202	0.479033	1	galactosyltransferase activity	MF
GO:0015813	7	39	0.06358	0.892282	0	1	5.923861	4.097933	6.052496	4.273186	0.904829	1	L-glutamate transport	BP
GO:0017015	6	72	0.045808	0.931336	-0.02857	1	0.847658	4.771944	0.46429	4.715752	0.938846	1	regulation of transforming growth factor beta receptor signaling pathway	BP
GO:0033205	8	24	-0.14781	0.726859	0	1	0.992269	3.63886	0.776114	3.478839	0.763936	1	cell cycle cytokinesis	BP
GO:0043097	6	34	0.22016	0.675096	-0.02857	1	2.959259	4.683752	2.985357	3.950088	0.71717	1	pyrimidine nucleoside salvage	BP
GO:0043406	7	161	0.153797	0.741984	0	1	3.516117	5.343588	4.165809	5.619048	0.776784	1	positive regulation of MAP kinase activity	BP
GO:0045600	9	56	0.038286	0.922099	0	1	1.015886	4.173472	1.058894	4.467799	0.931539	1	positive regulation of fat cell differentiation	BP
GO:0048193	6	20	0.007624	0.988565	-0.02857	1	3.94513	5.331832	4.714244	5.026644	0.989621	1	Golgi vesicle transport	BP
GO:0048255	7	80	0.133501	0.775376	0	1	5.958022	6.59122	6.614494	6.359974	0.806786	1	mRNA stabilization	BP
GO:0048706	8	70	0.083914	0.843399	0	1	0.691883	3.490251	-0.00106	3.168178	0.866333	1	embryonic skeletal system development	BP
GO:0048839	6	81	0.074912	0.887842	0.028571	1	1.630889	4.80235	1.160322	4.55511	0.902005	1	inner ear development	BP
GO:0051457	9	42	0.130211	0.738455	0	1	2.470677	4.417102	2.272297	4.370446	0.773521	1	maintenance of protein location in nucleus	BP
GO:0071479	12	82	0.260734	0.413066	0	1	3.169443	4.710873	3.157797	4.255418	0.478381	1	cellular response to ionizing radiation	BP
GO:0090200	7	72	0.077148	0.869419	0	1	4.369794	3.527011	3.888969	3.58635	0.887403	1	positive regulation of release of cytochrome c from mitochondria	BP
GO:0001578	9	66	-0.08575	0.826376	0.016667	0.98157	0.210464	3.905076	-0.08095	3.943825	0.851349	0.99322	microtubule bundle	BP

												3	formation	
GO:0001707	9	87	0.228075	0.555042	-0.01667	0.98157	1.193942	3.977355	0.910502	3.942787	0.606471	0.993223	mesoderm formation	BP
GO:0032784	9	36	0.134074	0.730926	0.016667	0.98157	3.326442	4.824548	3.344051	4.212112	0.76762	0.993223	regulation of transcription elongation, DNA-dependent	BP
GO:0045860	10	108	0.196375	0.586616	-0.01818	0.972841	1.774997	4.939309	1.263712	4.85641	0.634794	0.985988	positive regulation of protein kinase activity	BP
GO:0048589	11	86	0.025755	0.940084	-0.01818	0.967576	0.51194	3.952207	-0.02006	4.019218	0.947156	0.981182	developmental growth	BP
GO:0006071	7	53	-0.17168	0.712829	-0.03571	0.963492	2.22236	2.754802	2.439611	2.751458	0.750874	0.977569	glycerol metabolic process	BP
GO:0006613	7	26	0.138917	0.766436	0.035714	0.963492	5.735365	6.214166	5.831593	6.338304	0.798471	0.977569	cotranslational protein targeting to membrane	BP
GO:0016574	7	16	-0.19176	0.680403	0.035714	0.963492	3.677796	5.129518	3.988504	5.250958	0.721582	0.977569	histone ubiquitination	BP
GO:0019773	7	37	0.435758	0.328421	-0.03571	0.963492	7.000844	6.04337	7.028403	5.956266	0.395244	0.977569	proteasome core complex, alpha-subunit complex	CC
GO:0070403	7	46	0.087757	0.851593	-0.03571	0.963492	3.847249	3.860134	3.910502	3.657623	0.872534	0.977569	NAD+ binding	MF
GO:0003950	12	78	0.129216	0.688982	0.020979	0.956169	2.218197	3.820041	1.739887	3.959964	0.729442	0.972771	NAD+ ADP-ribosyltransferase activity	MF
GO:0031572	12	103	0.189909	0.554402	0.020979	0.956169	1.978818	3.735724	2.090401	3.877485	0.606125	0.972771	G2/M transition DNA damage checkpoint	BP
GO:0051865	12	129	0.042062	0.896731	0.020979	0.956169	2.116269	4.038644	1.880118	4.409363	0.908358	0.972771	protein autoubiquitination	BP
GO:0006695	24	125	0.148698	0.488026	0.013043	0.952979	4.701021	5.095515	5.17854	5.425903	0.547933	0.971106	cholesterol biosynthetic process	BP
GO:0043473	9	94	0.292413	0.44513	0.033333	0.948391	0.359067	2.735137	-0.76001	2.880714	0.508112	0.966956	pigmentation	BP
GO:0045022	9	34	0.065087	0.867873	-0.03333	0.948391	0.648975	3.778828	-0.06212	4.268493	0.886308	0.966956	early endosome to late endosome transport	BP
GO:0002053	8	153	0.361239	0.379294	0.047619	0.934871	1.299036	5.49012	0.182053	5.098177	0.445321	0.95421	positive regulation	BP

													of mesenchymal cell proliferation	
GO:0004812	8	18	0.371342	0.365094	0.047619	0.934871	4.479887	5.742679	5.565166	5.409695	0.431034	0.95421	aminoacyl-tRNA ligase activity	MF
GO:0006829	8	59	0.282455	0.497891	0.047619	0.934871	2.552103	3.831671	2.91061	4.125763	0.556344	0.95421	zinc ion transport	BP
GO:0007062	8	44	0.016412	0.969233	0.047619	0.934871	3.445465	5.338565	5.112441	5.491767	0.972344	0.95421	sister chromatid cohesion	BP
GO:0007093	8	57	0.110806	0.793933	0.047619	0.934871	2.567786	4.205014	2.867301	4.076333	0.822899	0.95421	mitotic cell cycle checkpoint	BP
GO:0015450	8	32	0.271531	0.51535	-0.04762	0.934871	6.144671	5.874437	6.010794	6.22926	0.571088	0.95421	P-P-bond-hydrolysis-driven protein transmembrane transporter activity	MF
GO:0021510	8	48	0.245037	0.558616	-0.04762	0.934871	3.034725	5.017556	3.432397	5.04351	0.60931	0.95421	spinal cord development	BP
GO:0030374	30	137	0.163573	0.387756	0.01624	0.93261	2.738105	5.082293	3.07926	5.130238	0.452986	0.95421	ligand-dependent nuclear receptor transcription coactivator activity	MF
GO:0051898	8	64	0.219962	0.600681	-0.04762	0.934871	2.080266	5.208602	1.796763	5.176546	0.648777	0.95421	negative regulation of protein kinase B signaling cascade	BP
GO:0005742	6	22	0.116444	0.826123	-0.08571	0.919444	6.772297	6.046374	7.165902	5.926692	0.851349	0.943084	mitochondrial outer membrane translocase complex	CC
GO:0007043	6	44	-0.10802	0.838605	0.085714	0.919444	2.772256	3.389679	2.965327	3.858168	0.862053	0.943084	cell-cell junction assembly	BP
GO:0014065	6	52	0.325617	0.528836	0.085714	0.919444	1.174255	4.615857	1.638061	4.913686	0.582932	0.943084	phosphatidylinositol 3-kinase cascade	BP
GO:0017091	6	42	-0.2186	0.677319	-0.08571	0.919444	4.094554	5.236286	5.517274	4.980946	0.718717	0.943084	AU-rich element binding	MF
GO:0030325	6	120	0.028232	0.957663	-0.08571	0.919444	3.387157	4.542376	3.683162	4.250216	0.962798	0.943084	adrenal gland development	BP
GO:0030330	6	30	0.072879	0.890875	0.085714	0.919444	2.735388	4.593422	2.814179	4.405801	0.903891	0.943084	DNA damage response, signal transduction by p53 class mediator	BP
GO:0031016	6	106	0.711467	0.112866	-0.08571	0.919444	1.243299	4.21182	-0.06199	3.996898	0.167424	0.943084	pancreas development	BP

GO:0032040	6	18	0.633561	0.176814	0.085714	0.919444	5.518951	5.730261	5.576771	4.830694	0.236438	0.943084	small-subunit processome	CC
GO:0043195	6	138	0.428252	0.396892	0.085714	0.919444	1.186655	3.506194	0.670776	4.022345	0.462507	0.943084	terminal button	CC
GO:0045070	6	36	-0.12585	0.812226	0.085714	0.919444	3.864616	4.619701	4.435874	5.360881	0.838614	0.943084	positive regulation of viral genome replication	BP
GO:0046676	6	100	-0.20063	0.703098	0.085714	0.919444	3.295659	4.841598	4.003979	5.289861	0.741874	0.943084	negative regulation of insulin secretion	BP
GO:0046873	6	59	0.252408	0.629428	-0.08571	0.919444	2.184069	3.566752	3.042768	3.427148	0.675546	0.943084	metal ion transmembrane transporter activity	MF
GO:0048854	6	34	0.441677	0.380566	0.085714	0.919444	1.811785	3.699449	1.817569	3.359042	0.446534	0.943084	brain morphogenesis	BP
GO:0050431	6	53	-0.09596	0.856506	0.085714	0.919444	1.863702	4.532962	1.571686	4.524894	0.87661	0.943084	transforming growth factor beta binding	MF
GO:0055088	6	78	-0.02256	0.966166	0.085714	0.919444	1.789608	4.553299	2.087711	4.304533	0.969787	0.943084	lipid homeostasis	BP
GO:0060135	6	70	0.383073	0.453497	-0.08571	0.919444	2.758274	4.889684	2.411515	4.980563	0.5162	0.943084	maternal process involved in female pregnancy	BP
GO:0060170	6	51	-0.18695	0.722835	0.085714	0.919444	1.162056	2.751843	1.350045	2.636462	0.760559	0.943084	cilium membrane	CC
GO:0071203	6	28	0.570887	0.236699	0.085714	0.919444	3.838014	5.654391	2.70468	5.767017	0.301297	0.943084	WASH complex	CC
GO:0071565	6	56	0.48133	0.333762	-0.08571	0.919444	5.245563	6.062814	5.4654	6.051746	0.400376	0.943084	nBAF complex	CC
GO:0004012	7	38	0.140815	0.763309	0.071429	0.906349	0.736742	3.318844	0.484792	2.80118	0.796884	0.939417	phospholipid-translocating ATPase activity	MF
GO:0004707	7	74	0.01469	0.975063	-0.07143	0.906349	1.839266	4.376119	2.680382	4.813743	0.977671	0.939417	MAP kinase activity	MF
GO:0043235	7	99	0.277375	0.547017	-0.07143	0.906349	0.485463	3.480122	0.047219	4.303502	0.600502	0.939417	receptor complex	CC
GO:0046326	7	108	-0.33657	0.460433	0.071429	0.906349	1.548996	4.143807	1.058894	4.34284	0.52227	0.939417	positive regulation of glucose import	BP
GO:0009165	8	46	-0.11582	0.78477	0.071429	0.881994	2.717129	4.526742	2.97321	4.088512	0.815204	0.916199	nucleotide biosynthetic process	BP
GO:0001947	9	112	-0.01805	0.963229	-0.06667	0.880093	0.803134	4.005524	0.406407	3.754132	0.967875	0.91473	heart looping	BP

GO:0008299	11	49	-0.10488	0.758924	-0.06364	0.860104	3.307927	4.780084	4.073393	5.415896	0.792748	0.89445	isoprenoid biosynthetic process	BP
GO:0045766	17	233	0.083013	0.751435	0.04902	0.853696	2.157506	4.099277	1.818579	4.665889	0.785801	0.88828	positive regulation of angiogenesis	BP
GO:0006783	12	71	-0.2553	0.423212	-0.06294	0.851682	3.521683	3.926829	3.905855	4.172798	0.487421	0.886676	heme biosynthetic process	BP
GO:0005763	18	51	0.232281	0.353661	0.048504	0.849928	6.48675	5.341391	6.488628	5.420436	0.421025	0.885341	mitochondrial small ribosomal subunit	CC
GO:0001938	17	202	0.038874	0.882241	-0.05147	0.846278	1.327589	4.295024	1.553511	4.234063	0.898047	0.88203	positive regulation of endothelial cell proliferation	BP
GO:0035019	9	68	-0.07151	0.854944	0.083333	0.843182	0.839581	4.531917	0.697827	4.770163	0.875489	0.879292	somatic stem cell maintenance	BP
GO:0060612	10	40	0.0752	0.836428	-0.07295	0.841271	2.690375	4.176714	2.098271	4.203965	0.860286	0.877787	adipose tissue development	BP
GO:0006013	7	18	0.029391	0.950126	-0.10714	0.839683	1.086868	3.995405	0.213404	4.088025	0.956246	0.876617	mannose metabolic process	BP
GO:0008324	11	66	0.161768	0.634645	0.072727	0.838825	4.242159	3.75972	4.297832	3.586884	0.680755	0.876617	cation transmembrane transporter activity	MF
GO:0033327	7	74	-0.04484	0.923949	0.107143	0.839683	4.694676	6.025466	5.80093	6.084058	0.932905	0.876617	Leydig cell differentiation	BP
GO:0034968	7	40	-0.12998	0.7812	0.107143	0.839683	0.538306	4.694806	-0.13488	4.902703	0.811945	0.876617	histone lysine methylation	BP
GO:0060716	7	37	-0.2336	0.614159	-0.10714	0.839683	2.55081	4.907479	1.325539	4.951885	0.660671	0.876617	labyrinthine layer blood vessel development	BP
GO:0070402	7	28	0.764417	0.045361	0.107143	0.839683	5.358128	2.465531	5.411195	3.045436	0.082494	0.876617	NADPH binding	MF
GO:0005761	22	75	-0.10212	0.651124	-0.04687	0.836419	6.164567	5.04572	6.386758	5.134603	0.694092	0.876137	mitochondrial ribosome	CC
GO:0070979	17	124	0.03779	0.885504	0.056373	0.83148	3.188915	4.949748	3.366466	5.308277	0.900391	0.871451	protein K11-linked ubiquitination	BP
GO:0005100	12	78	0.202693	0.52752	0.076923	0.817283	0.659569	3.934222	0.703818	4.289986	0.581823	0.85705	Rho GTPase activator activity	MF
GO:0000045	10	74	-0.00285	0.993775	-0.09091	0.811417	1.919718	4.576696	1.88459	5.014932	0.994305	0.851375	autophagic vacuole assembly	BP
GO:0001756	9	117	-0.05034	0.897675	0.1	0.809981	2.798997	4.278881	3.171368	4.761205	0.908823	0.850344	somitogenesis	BP

GO:0004177	23	96	0.247571	0.254729	-0.05435	0.805563	3.877978	4.07705	4.240732	4.739203	0.319689	0.84618	aminopeptidase activity	MF
GO:0005666	6	18	0.310437	0.549303	0.142857	0.802778	3.440333	4.227183	3.560238	4.314305	0.601954	0.843727	DNA-directed RNA polymerase III complex	CC
GO:0006241	6	34	0.346719	0.500762	0.142857	0.802778	3.541807	4.021962	3.662615	3.955451	0.558886	0.843727	CTP biosynthetic process	BP
GO:0006777	6	30	0.10283	0.846298	0.142857	0.802778	3.057003	3.81395	4.572423	3.746419	0.868091	0.843727	Mo-molybdopterin cofactor biosynthetic process	BP
GO:0007193	6	90	0.304023	0.558016	0.142857	0.802778	1.169721	2.890221	0.192976	3.460683	0.609011	0.843727	inhibition of adenylate cyclase activity by G-protein signaling pathway	BP
GO:0031648	6	46	0.537803	0.271071	0.142857	0.802778	3.352394	4.089137	3.324809	4.302306	0.335556	0.843727	protein destabilization	BP
GO:0034199	6	72	0.142421	0.787813	0.142857	0.802778	1.448272	4.059546	1.21921	4.357716	0.817913	0.843727	activation of protein kinase A activity	BP
GO:0042476	6	110	0.445107	0.376432	0.142857	0.802778	3.016215	4.404291	3.426106	4.292532	0.442514	0.843727	odontogenesis	BP
GO:0042640	6	36	-0.15693	0.766532	-0.14286	0.802778	2.500372	4.934465	3.117595	5.013499	0.798471	0.843727	anagen	BP
GO:0045749	6	54	0.440765	0.381667	0.142857	0.802778	0.672611	4.45561	0.025096	4.427356	0.447546	0.843727	negative regulation of S phase of mitotic cell cycle	BP
GO:0060444	6	54	0.181194	0.731183	0.142857	0.802778	3.603143	4.985588	4.803961	5.02928	0.76762	0.843727	branching involved in mammary gland duct morphogenesis	BP
GO:0006694	11	101	0.120571	0.723996	0.090909	0.796592	4.898655	4.704746	5.168903	4.415017	0.761353	0.842009	steroid biosynthetic process	BP
GO:0019894	11	64	0.758699	0.006788	-0.09091	0.796592	2.056325	4.893443	1.350706	4.413659	0.01858	0.842009	kinesin binding	MF
GO:0031418	16	84	0.069019	0.799509	0.070588	0.796653	2.76192	3.908979	2.546638	4.132171	0.827307	0.842009	L-ascorbic acid binding	MF
GO:0046965	8	68	0.564747	0.144707	0.119048	0.793006	1.37829	5.350354	0.686975	5.122601	0.201803	0.839574	retinoid X receptor binding	MF

GO:0000178	13	37	0.076387	0.804115	0.082418	0.792486	5.892467	5.002799	5.967425	4.811056	0.831613	0.839497	exosome (RNase complex)	CC
GO:0004602	7	46	0.204037	0.66078	0.142857	0.78254	6.029682	5.16464	6.107803	5.286291	0.702759	0.82943	glutathione peroxidase activity	MF
GO:0006390	7	26	0.194444	0.676106	0.142857	0.78254	4.216183	4.006701	4.643856	4.02589	0.717836	0.82943	transcription from mitochondrial promoter	BP
GO:0006476	7	41	0.30824	0.501216	0.142857	0.78254	3.139741	4.304743	2.961283	4.418351	0.55906	0.82943	protein deacetylation	BP
GO:0031293	7	47	-0.1856	0.690312	-0.14286	0.78254	2.911152	4.32924	2.432455	4.580642	0.730142	0.82943	membrane protein intracellular domain proteolysis	BP
GO:0042581	7	36	-0.03235	0.945117	-0.14286	0.78254	4.650235	4.587922	4.926256	4.426996	0.951715	0.82943	specific granule	CC
GO:0030520	9	93	0.116779	0.764787	0.116667	0.775628	3.468871	5.644236	3.451662	5.542371	0.797539	0.824435	estrogen receptor signaling pathway	BP
GO:0046329	9	92	-0.08187	0.834137	-0.11667	0.775628	1.597016	5.051806	1.532199	5.420283	0.858401	0.824435	negative regulation of JNK cascade	BP
GO:0048286	9	172	0.059568	0.879008	0.116667	0.775628	0.587715	3.206761	1.114035	2.525999	0.895241	0.824435	lung alveolus development	BP
GO:0007219	25	204	0.308103	0.134037	0.060769	0.772611	2.307316	4.787016	2.432455	4.935555	0.189675	0.822626	Notch signaling pathway	BP
GO:0022857	13	127	0.404651	0.170216	0.093407	0.764582	6.726652	6.253938	7.528755	6.669332	0.228785	0.81454	transmembrane transporter activity	MF
GO:0030518	13	35	0.166633	0.586377	-0.09341	0.764582	3.429866	4.350746	3.592326	4.21898	0.634794	0.81454	steroid hormone receptor signaling pathway	BP
GO:0008206	10	130	0.295908	0.406463	0.115152	0.758833	3.711967	3.706616	3.544795	4.078894	0.4719	0.809336	bile acid metabolic process	BP
GO:0046966	24	104	-0.09586	0.65591	-0.06696	0.755543	2.847296	4.651241	3.125091	4.6705	0.697975	0.806285	thyroid hormone receptor binding	MF
GO:0004143	8	88	0.018499	0.965322	-0.14286	0.752034	1.01886	3.688501	0.46234	3.838619	0.969459	0.802997	diacylglycerol kinase activity	MF
GO:0048661	8	195	0.273265	0.512565	0.142857	0.752034	1.522418	4.689901	1.454915	5.017571	0.568337	0.802997	positive regulation of smooth muscle cell proliferation	BP
GO:0050662	13	67	-0.00087	0.997755	-0.0989	0.75073	3.977718	4.184956	4.224317	4.558348	0.997755	0.80252	coenzyme binding	MF
GO:0006521	41	166	0.242658	0.126347	0.051916	0.746487	6.033826	5.555102	6.346333	5.801958	0.182001	0.798439	regulation of cellular amino acid metabolic process	BP
GO:0007220	9	51	-0.27669	0.47106	-0.13333	0.743541	2.659795	5.097063	3.273907	5.121863	0.532713	0.795741	Notch receptor processing	BP

GO:0032456	9	24	0.182973	0.637493	0.133333	0.743541	2.300023	4.180317	1.895395	3.682134	0.683028	0.795741	endocytic recycling	BP
GO:0006206	14	81	0.148233	0.613042	0.098901	0.738503	3.762449	4.379912	3.553859	4.178707	0.660226	0.791253	pyrimidine base metabolic process	BP
GO:0008277	11	180	0.205644	0.544101	0.118182	0.734252	1.267662	3.853783	-0.49876	3.818152	0.598351	0.787148	regulation of G-protein coupled receptor protein signaling pathway	BP
GO:0003678	12	48	0.245667	0.44151	0.111888	0.732775	2.188292	4.52689	1.616492	4.658547	0.504775	0.786134	DNA helicase activity	MF
GO:0004540	10	49	-0.04265	0.906883	-0.12727	0.732887	4.030075	3.566497	4.084876	3.563018	0.917155	0.786134	ribonuclease activity	MF
GO:0030414	17	285	-0.06094	0.816269	-0.09314	0.722578	4.263963	4.40822	4.267328	4.939667	0.841861	0.775965	peptidase inhibitor activity	MF
GO:0004867	21	304	0.104816	0.651151	0.085714	0.711371	4.513812	4.762634	5.28396	5.28103	0.694092	0.767073	serine-type endopeptidase inhibitor activity	MF
GO:0005663	6	28	0.509142	0.302278	0.2	0.713889	5.947455	5.695225	5.939335	5.559639	0.368751	0.767073	DNA replication factor C complex	CC
GO:0005721	6	18	0.35377	0.491483	0.2	0.713889	4.49101	4.995532	4.477097	4.785633	0.551155	0.767073	centromeric heterochromatin	CC
GO:0006734	7	27	0.489475	0.264897	0.178571	0.713095	6.78558	5.404291	6.803334	5.68167	0.329505	0.767073	NADH metabolic process	BP
GO:0007588	6	149	0.100109	0.850338	0.2	0.713889	2.40698	4.08734	2.655976	4.020161	0.871725	0.767073	excretion	BP
GO:0035257	6	58	0.454506	0.365186	0.2	0.713889	0.944933	4.914667	-0.11489	5.072559	0.431034	0.767073	nuclear hormone receptor binding	MF
GO:0045165	7	102	0.137289	0.769122	0.178571	0.713095	1.713083	3.820347	1.081614	3.919793	0.800724	0.767073	cell fate commitment	BP
GO:0070087	6	28	0.433538	0.390435	0.2	0.713889	4.561388	5.348476	3.953923	4.928984	0.455832	0.767073	chromo shadow domain binding	MF
GO:0070776	6	20	-0.13576	0.797611	-0.2	0.713889	0.355135	4.411923	0.604881	4.495243	0.826255	0.767073	MOZ/MORF histone acetyltransferase complex	CC
GO:0010388	9	32	-0.07711	0.843693	0.15	0.708069	4.678702	5.204871	4.782562	5.088497	0.866333	0.764763	cullin deneddylation	BP
GO:0010390	9	31	-0.28665	0.454569	0.15	0.708069	3.893637	5.335653	4.479685	5.445761	0.516555	0.764763	histone monoubiquitination	BP
GO:0001702	10	56	0.289063	0.417915	-0.13939	0.707204	1.42503	4.457764	0.836045	3.948033	0.482506	0.76471	gastrulation with	BP

													mouth forming second	
GO:0050661	14	130	-0.13122	0.65478	-0.11209	0.704285	4.261224	4.216618	4.542605	4.258537	0.697168	0.761994	NADP binding	MF
GO:0008630	13	125	0.406205	0.168424	0.120879	0.696085	2.788878	4.263779	2.55828	4.655165	0.226865	0.753556	DNA damage response, signal transduction resulting in induction of apoptosis	BP
GO:0007257	15	101	-0.0244	0.931206	-0.11071	0.695276	1.268355	4.48884	1.350706	4.622339	0.938846	0.753115	activation of JUN kinase activity	BP
GO:0015934	10	38	0.391117	0.263735	0.151515	0.681808	6.458516	7.04497	6.659613	6.21791	0.328792	0.738953	large ribosomal subunit	CC
GO:0019216	9	66	0.576222	0.104388	0.166667	0.677745	2.097659	3.832253	1.314733	2.720074	0.157338	0.734975	regulation of lipid metabolic process	BP
GO:0032007	9	40	0.175422	0.65167	0.166667	0.677745	0.71269	4.147382	1.19347	4.263584	0.694251	0.734975	negative regulation of TOR signaling cascade	BP
GO:0030914	13	42	0.213272	0.484179	0.131868	0.669269	2.734504	4.513679	2.500626	4.699925	0.54492	0.726624	STAGA complex	CC
GO:0006506	12	80	0.339022	0.281026	0.13986	0.667151	2.535495	3.841226	2.380258	3.881264	0.346432	0.724743	GPI anchor biosynthetic process	BP

Table S 12: Top 150 Highest Transcript-protein correlation by Gene Otology class in RWPE

Orange											
Annotation Cluster 1	Enrichment Score: 12.496483810424145										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0045449~regulation of transcription	280	20.05731	4.21E-17	936	1485	7682	1.547498	1.19E-13	1.19E-13	7.54E-14
GOTERM_MF_FAT	GO:0003677~DNA binding	247	17.69341	4.20E-15	892	1302	7328	1.558499	3.71E-12	3.71E-12	6.56E-12
GOTERM_MF_FAT	GO:0003700~transcription factor activity	119	8.524355	1.11E-13	892	500	7328	1.955229	9.81E-11	4.90E-11	1.73E-10

GOTERM_BP_FAT	GO:0006350~transcription	229	16.40401	2.47E-13	936	1216	7682	1.545612	7.00E-10	3.50E-10	4.42E-10
GOTERM_BP_FAT	GO:0006355~regulation of transcription, DNA-dependent	185	13.25215	4.33E-13	936	925	7682	1.641453	1.23E-09	4.09E-10	7.74E-10
GOTERM_BP_FAT	GO:0051252~regulation of RNA metabolic process	187	13.39542	2.23E-12	936	954	7682	1.608762	6.33E-09	1.58E-09	4.00E-09
GOTERM_MF_FAT	GO:0030528~transcription regulator activity	161	11.53295	7.09E-08	892	896	7328	1.476177	6.24E-05	2.08E-05	1.10E-04
Annotation Cluster 2	Enrichment Score: 4.737597178670993										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_MF_FAT	GO:0008270~zinc ion binding	222	15.90258	4.94E-07	892	1362	7328	1.339049	4.35E-04	1.09E-04	7.68E-04
GOTERM_MF_FAT	GO:0046914~transition metal ion binding	257	18.40974	2.67E-06	892	1653	7328	1.277265	0.002348	4.70E-04	0.004152
GOTERM_MF_FAT	GO:0046872~metal ion binding	339	24.28367	8.70E-05	892	2367	7328	1.176582	0.07373	0.006939	0.135178
GOTERM_MF_FAT	GO:0043169~cation binding	341	24.42693	1.17E-04	892	2391	7328	1.171643	0.09753	0.008515	0.181078
GOTERM_MF_FAT	GO:0043167~ion binding	343	24.5702	1.53E-04	892	2414	7328	1.167286	0.126019	0.010308	0.237612
Annotation Cluster 3	Enrichment Score: 3.480997610903036										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_CC_	GO:0031226~intrinsic	86	6.16045	1.33E-06	796	456	7021	1.663487	4.94E-04	1.65E-	0.0018

FAT	to plasma membrane		8							04	35
GOTERM_CC_FAT	GO:0005887~integral to plasma membrane	81	5.802292	1.16E-05	796	445	7021	1.605502	0.004319	6.18E-04	0.016054
GOTERM_CC_FAT	GO:0005886~plasma membrane	209	14.97135	0.003781	796	1568	7021	1.175672	0.755686	0.120251	5.093982
GOTERM_CC_FAT	GO:0044459~plasma membrane part	127	9.097421	0.203695	796	1041	7021	1.076066	1	0.961569	95.68649
Annotation Cluster 4	Enrichment Score: 3.3569652578172073										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_CC_FAT	GO:0034705~potassium channel complex	17	1.217765	2.85E-06	796	40	7021	3.748649	0.001061	2.65E-04	0.003937
GOTERM_CC_FAT	GO:0008076~voltage-gated potassium channel complex	17	1.217765	2.85E-06	796	40	7021	3.748649	0.001061	2.65E-04	0.003937
GOTERM_MF_FAT	GO:0005267~potassium channel activity	21	1.504298	3.12E-06	892	55	7328	3.136731	0.002744	4.58E-04	0.004852
GOTERM_MF_FAT	GO:0005249~voltage-gated potassium channel activity	18	1.289398	3.12E-06	892	42	7328	3.52082	0.002746	3.93E-04	0.004857
GOTERM_MF_FAT	GO:0022843~voltage-gated cation channel activity	20	1.432665	3.62E-06	892	51	7328	3.221665	0.003184	3.99E-04	0.005632
GOTERM_CC_FAT	GO:0034703~cation channel complex	19	1.361032	1.06E-05	796	53	7021	3.162013	0.003943	6.58E-04	0.014655
GOTERM_MF_FAT	GO:0022832~voltage-gated channel activity	22	1.575931	8.35E-05	892	72	7328	2.510214	0.070814	0.007318	0.129634
GOTERM_MF_FAT	GO:0005244~voltage-gated ion channel activity	22	1.575931	8.35E-05	892	72	7328	2.510214	0.070814	0.007318	0.129634

GOTERM_BP_FAT	GO:0006813~potassium ion transport	21	1.504298	1.10E-04	936	68	7682	2.534597	0.267176	0.027863	0.196186
GOTERM_CC_FAT	GO:0034702~ion channel complex	21	1.504298	1.44E-04	796	74	7021	2.503073	0.052093	0.005336	0.198276
GOTERM_MF_FAT	GO:0005261~cation channel activity	26	1.862464	3.34E-04	892	101	7328	2.114816	0.254618	0.020771	0.517655
GOTERM_MF_FAT	GO:0022836~gated channel activity	26	1.862464	8.42E-04	892	107	7328	1.996228	0.523313	0.04525	1.300012
GOTERM_BP_FAT	GO:0015672~monovalent inorganic cation transport	32	2.292264	9.71E-04	936	144	7682	1.823837	0.936146	0.108305	1.723026
GOTERM_MF_FAT	GO:0046873~metal ion transmembrane transporter activity	30	2.148997	9.76E-04	892	132	7328	1.867102	0.576576	0.046622	1.506342
GOTERM_MF_FAT	GO:0022838~substrate specific channel activity	30	2.148997	0.001417	892	135	7328	1.82561	0.712753	0.063545	2.179038
GOTERM_MF_FAT	GO:0022803~passive transmembrane transporter activity	31	2.22063	0.001752	892	143	7328	1.780928	0.786341	0.074266	2.689026
GOTERM_MF_FAT	GO:0015267~channel activity	31	2.22063	0.001752	892	143	7328	1.780928	0.786341	0.074266	2.689026
GOTERM_MF_FAT	GO:0005216~ion channel activity	29	2.077364	0.002611	892	134	7328	1.777927	0.899811	0.103769	3.981944
GOTERM_BP_FAT	GO:0030001~metal ion transport	38	2.722063	0.008446	936	205	7682	1.521347	1	0.477538	14.07967
GOTERM_MF_FAT	GO:0030955~potassium ion binding	13	0.931232	0.014601	892	50	7328	2.135964	0.999998	0.41685	20.4355
GOTERM_BP_FAT	GO:0006812~cation transport	44	3.151862	0.01766	936	257	7682	1.405135	1	0.67416	27.29686
GOTERM_MF_FAT	GO:0031420~alkali metal ion binding	18	1.289398	0.034327	892	88	7328	1.680391	1	0.64107	41.8937
GOTERM_BP_FAT	GO:0006811~ion transport	52	3.724928	0.06903	936	342	7682	1.247888	1	0.930426	72.18847
GOTERM_BP_	GO:0055085~transmem	39	2.79369	0.51671	936	311	7682	1.029207	1	0.99913	99.999

FAT	brane transport		6	2						5	78
Annotation Cluster 5	Enrichment Score: 3.298323733971374										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	73	5.229226	3.05E-05	936	372	7682	1.610565	0.082805	0.012272	0.05459
GOTERM_BP_FAT	GO:0051173~positive regulation of nitrogen compound metabolic process	74	5.30086	5.44E-05	936	385	7682	1.5775	0.142669	0.019057	0.097199
GOTERM_BP_FAT	GO:0010557~positive regulation of macromolecule biosynthetic process	71	5.08596	9.07E-05	936	371	7682	1.570663	0.226453	0.02535	0.162083
GOTERM_BP_FAT	GO:0009891~positive regulation of biosynthetic process	73	5.229226	1.36E-04	936	389	7682	1.540181	0.319436	0.031561	0.242823
GOTERM_BP_FAT	GO:0031328~positive regulation of cellular biosynthetic process	72	5.157593	1.69E-04	936	385	7682	1.534865	0.379852	0.036086	0.301393
GOTERM_BP_FAT	GO:0045941~positive regulation of transcription	63	4.512894	3.05E-04	936	332	7682	1.557403	0.578856	0.056022	0.544836
GOTERM_BP_FAT	GO:0010628~positive regulation of gene expression	63	4.512894	5.01E-04	936	338	7682	1.529756	0.758348	0.07587	0.893233
GOTERM_BP_FAT	GO:0045893~positive regulation of	53	3.796562	0.001347	936	283	7682	1.53705	0.978035	0.136583	2.38335

	transcription, DNA-dependent										
GOTERM_BP_FAT	GO:0051254~positive regulation of RNA metabolic process	53	3.796562	0.001853	936	287	7682	1.515627	0.994771	0.165693	3.264396
GOTERM_BP_FAT	GO:0006357~regulation of transcription from RNA polymerase II promoter	78	5.587393	0.001945	936	461	7682	1.388648	0.995971	0.167905	3.423664
GOTERM_BP_FAT	GO:0045944~positive regulation of transcription from RNA polymerase II promoter	39	2.793696	0.004895	936	205	7682	1.561382	0.999999	0.343701	8.405413
GOTERM_BP_FAT	GO:0010604~positive regulation of macromolecule metabolic process	80	5.730659	0.020966	936	522	7682	1.257818	1	0.706134	31.55188

Red											
Annotation Cluster 1	Enrichment Score: 4.9524705825017845										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0030216~keratinocyte differentiation	7	10	5.20E-08	54	31	7682	32.12306	2.61E-05	2.61E-05	7.49E-05
GOTERM_BP_FAT	GO:0009913~epidermal cell differentiation	7	10	1.12E-07	54	35	7682	28.45185	5.63E-05	2.81E-05	1.62E-04
GOTERM_BP_FAT	GO:0008544~epidermis development	9	12.85714	1.37E-07	54	89	7682	14.38577	6.85E-05	2.28E-05	1.97E-04
GOTERM_BP_FAT	GO:0007398~ectoderm development	9	12.85714	3.14E-07	54	99	7682	12.93266	1.57E-04	3.94E-05	4.53E-04

GOTERM_CC_FAT	GO:0001533~cornified envelope	5	7.142857	9.21E-07	45	13	7021	60.00855	1.23E-04	1.23E-04	0.001078
GOTERM_BP_FAT	GO:0031424~keratinization	5	7.142857	1.92E-06	54	14	7682	50.80688	9.62E-04	1.92E-04	0.002767
GOTERM_BP_FAT	GO:0030855~epithelial cell differentiation	7	10	8.90E-06	54	72	7682	13.83076	0.004447	7.43E-04	0.012814
GOTERM_BP_FAT	GO:0018149~peptide cross-linking	4	5.714286	1.33E-04	54	15	7682	37.9358	0.064505	0.00948	0.19153
GOTERM_BP_FAT	GO:0060429~epithelium development	7	10	1.93E-04	54	124	7682	8.030765	0.092363	0.012041	0.278245
GOTERM_CC_FAT	GO:0070161~anchoring junction	4	5.714286	0.042138	45	124	7021	5.032975	0.996877	0.617671	39.59242
GOTERM_CC_FAT	GO:0005856~cytoskeleton	10	14.28571	0.07782	45	851	7021	1.833399	0.999981	0.627274	61.2698
Annotation Cluster 2	Enrichment Score: 2.5683496481022687										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0016337~cell-cell adhesion	6	8.571429	6.77E-04	54	103	7682	8.286947	0.287718	0.036996	0.97074
GOTERM_BP_FAT	GO:0007155~cell adhesion	9	12.85714	0.001216	54	312	7682	4.103632	0.456305	0.059117	1.736756
GOTERM_BP_FAT	GO:0022610~biological adhesion	9	12.85714	0.001241	54	313	7682	4.090522	0.463162	0.054982	1.772609
GOTERM_CC_FAT	GO:0044459~plasma membrane part	12	17.14286	0.052183	45	1041	7021	1.798527	0.999239	0.549747	46.60683
Annotation Cluster 3	Enrichment Score: 1.9048898836370938										
Category	Term	Count	%	PValue	List Tota	Pop Hits	Pop Tota	Fold Enrichme	Bonferroni	Benjamini	FDR

					l		l	nt			
GOTERM_CC_FAT	GO:0005576~extracellular region	11	15.71429	0.003649	45	608	7021	2.82277	0.387289	0.217241	4.18988
GOTERM_CC_FAT	GO:0005615~extracellular space	6	8.571429	0.011176	45	218	7021	4.29419	0.778216	0.313749	12.33011
GOTERM_CC_FAT	GO:0044421~extracellular region part	6	8.571429	0.0473	45	318	7021	2.943816	0.998486	0.555862	43.29602
Annotation Cluster 4	Enrichment Score: 1.6428548903232476										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0016337~cell-cell adhesion	6	8.571429	6.77E-04	54	103	7682	8.286947	0.287718	0.036996	0.97074
GOTERM_CC_FAT	GO:0030054~cell junction	6	8.571429	0.034688	45	292	7021	3.205936	0.99118	0.611763	33.85671
GOTERM_CC_FAT	GO:0005911~cell-cell junction	4	5.714286	0.044719	45	127	7021	4.914086	0.997824	0.583462	41.47109
GOTERM_CC_FAT	GO:0043296~apical junction complex	3	4.285714	0.074464	45	72	7021	6.500926	0.999969	0.645454	59.58671
GOTERM_CC_FAT	GO:0016327~apicolateral plasma membrane	3	4.285714	7.81E-02	45	74	7021	6.325225	0.999981	0.596594	61.39719
Annotation Cluster 5	Enrichment Score: 1.6078873006578052										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_MF_FAT	GO:0005509~calcium ion binding	15	21.42857	3.64E-06	57	440	7328	4.382775	5.13E-04	5.13E-04	0.004297
GOTERM_MF_FAT	GO:0046872~metal ion binding	20	28.57143	0.446527	57	2367	7328	1.086281	1	0.996153	99.90774

GOTERM_MF_FAT	GO:0043169~cation binding	20	28.57143	0.467499	57	2391	7328	1.075378	1	0.996126	99.94154
GOTERM_MF_FAT	GO:0043167~ion binding	20	28.57143	0.487614	57	2414	7328	1.065132	1	0.996097	99.96291

Green											
Annotation Cluster 1	Enrichment Score: 4.39077234939865										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_CC_FAT	GO:0005739~mitochondrion	29	28.71287	2.23E-07	81	892	7021	2.81804	4.99E-05	4.99E-05	2.84E-04
GOTERM_CC_FAT	GO:0044429~mitochondrial part	19	18.81188	7.31E-06	81	492	7021	3.34736	0.001636	8.18E-04	0.009329
GOTERM_CC_FAT	GO:0005740~mitochondrial envelope	15	14.85149	2.01E-05	81	334	7021	3.89277	0.004482	0.001496	0.0256
GOTERM_CC_FAT	GO:0031967~organelle envelope	18	17.82178	2.98E-05	81	493	7021	3.164751	0.006662	0.00167	0.038089
GOTERM_CC_FAT	GO:0031975~envelope	18	17.82178	3.06E-05	81	494	7021	3.158345	0.006837	0.001371	0.039095
GOTERM_CC_FAT	GO:0031090~organelle membrane	22	21.78218	1.58E-04	81	796	7021	2.395651	0.034816	0.005889	0.201755
GOTERM_CC_FAT	GO:0019866~organelle inner membrane	12	11.88119	1.86E-04	81	265	7021	3.925087	0.040882	0.005945	0.237612
GOTERM_CC_FAT	GO:0031966~mitochondrial membrane	13	12.87129	1.90E-04	81	312	7021	3.611626	0.041747	0.005316	0.242741
GOTERM_CC_FAT	GO:0005743~mitochondrial inner membrane	10	9.90099	0.001817	81	246	7021	3.523537	0.334578	0.04425	2.294751
Annotation Cluster 2	Enrichment Score: 1.9562013451606715										

Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0015031~protein transport	16	15.84158	0.00256	86	610	7682	2.342966	0.916172	0.71047	3.952827
GOTERM_BP_FAT	GO:0045184~establishment of protein localization	16	15.84158	0.002685	86	613	7682	2.3315	0.925707	0.579613	4.141325
GOTERM_BP_FAT	GO:0008104~protein localization	16	15.84158	0.008949	86	698	7682	2.047578	0.999832	0.711131	13.18735
GOTERM_BP_FAT	GO:0046907~intracellular transport	9	8.910891	0.243341	86	538	7682	1.494294	1	0.98797	98.75595
Annotation Cluster 3	Enrichment Score: 1.4804953884813314										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0015909~long-chain fatty acid transport	3	2.970297	0.00514	86	10	7682	26.79767	0.993148	0.630886	7.787317
GOTERM_BP_FAT	GO:0015908~fatty acid transport	3	2.970297	0.01313	86	16	7682	16.74855	0.999997	0.797612	18.77337
GOTERM_BP_FAT	GO:0015718~monocarboxylic acid transport	3	2.970297	0.035513	86	27	7682	9.925065	1	0.841227	43.38239
GOTERM_BP_FAT	GO:0046942~carboxylic acid transport	4	3.960396	0.055001	86	78	7682	4.580799	1	0.887881	58.93421
GOTERM_BP_FAT	GO:0015849~organic acid transport	4	3.960396	0.055001	86	78	7682	4.580799	1	0.887881	58.93421
GOTERM_BP_FAT	GO:0006869~lipid transport	4	3.960396	0.07148	86	87	7682	4.106923	1	0.901078	68.86221
GOTERM_BP_FAT	GO:0010876~lipid localization	4	3.960396	0.083559	86	93	7682	3.84196	1	0.9164	74.65909
Annotation	Enrichment Score:										

Cluster 4	1.4152039870860111										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0034622~cellular macromolecular complex assembly	8	7.920792	0.008191	86	209	7682	3.419161	0.999648	0.734338	12.13691
GOTERM_BP_FAT	GO:0065003~macromolecular complex assembly	12	11.88119	0.015742	86	478	7682	2.242483	1	0.818198	22.09065
GOTERM_BP_FAT	GO:0034621~cellular macromolecular complex subunit organization	8	7.920792	1.73E-02	86	242	7682	2.952912	1	0.815595	24.0464
GOTERM_BP_FAT	GO:0043933~macromolecular complex subunit organization	12	11.88119	0.023639	86	508	7682	2.110053	1	0.808403	31.36395
GOTERM_BP_FAT	GO:0070271~protein complex biogenesis	7	6.930693	0.247154	86	382	7682	1.636856	1	0.98806	98.85099
GOTERM_BP_FAT	GO:0006461~protein complex assembly	7	6.930693	0.247154	86	382	7682	1.636856	1	0.98806	98.85099
Annotation Cluster 5	Enrichment Score: 1.396074843606603										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0007264~small GTPase mediated signal transduction	10	9.90099	5.24E-04	86	212	7682	4.213471	0.397328	0.397328	0.820448
GOTERM_MF_FAT	GO:0003924~GTPase activity	8	7.920792	5.92E-04	83	130	7328	5.433179	0.126376	0.126376	0.755763
GOTERM_MF_FAT	GO:0019001~guanyl nucleotide binding	10	9.90099	8.63E-04	83	225	7328	3.923963	0.178594	0.093686	1.098634

GOTERM_MF_FAT	GO:0005525~GTP binding	10	9.90099	8.63E-04	83	225	7328	3.923963	0.178594	0.093686	1.098634
GOTERM_MF_FAT	GO:0032561~guanyl ribonucleotide binding	10	9.90099	8.63E-04	83	225	7328	3.923963	0.178594	0.093686	1.098634
GOTERM_MF_FAT	GO:0000166~nucleotide binding	27	26.73267	0.013155	83	1524	7328	1.564178	0.951161	0.529899	15.59423
GOTERM_BP_FAT	GO:0006986~response to unfolded protein	4	3.960396	0.026072	86	58	7682	6.160385	1	0.817871	34.00569
GOTERM_BP_FAT	GO:0007242~intracellular signaling cascade	15	14.85149	0.037243	86	757	7682	1.769992	1	0.840397	44.95913
GOTERM_BP_FAT	GO:0051789~response to protein stimulus	4	3.960396	0.062074	86	82	7682	4.357345	1	0.89065	63.5119
GOTERM_MF_FAT	GO:0017076~purine nucleotide binding	20	19.80198	0.11122	83	1277	7328	1.382759	1	0.893564	77.89788
GOTERM_BP_FAT	GO:0010033~response to organic substance	9	8.910891	0.115808	86	443	7682	1.814741	1	0.952725	85.57665
GOTERM_MF_FAT	GO:0032553~ribonucleotide binding	19	18.81188	0.127172	83	1220	7328	1.374995	1	0.890861	82.47185
GOTERM_MF_FAT	GO:0032555~purine ribonucleotide binding	19	18.81188	0.127172	83	1220	7328	1.374995	1	0.890861	82.47185
GOTERM_CC_FAT	GO:0009898~internal side of plasma membrane	5	4.950495	0.231637	81	216	7021	2.006459	1	0.923173	96.53977
GOTERM_MF_FAT	GO:0001882~nucleoside binding	12	11.88119	0.700862	83	1097	7328	0.965788	1	0.99999	99.99998
GOTERM_MF_FAT	GO:0001883~purine nucleoside binding	11	10.89109	0.796195	83	1089	7328	0.89181	1	0.999998	100
GOTERM_MF_FAT	GO:0030554~adenyl nucleotide binding	10	9.90099	0.869946	83	1077	7328	0.819769	1	1	100
GOTERM_MF_FAT	GO:0005524~ATP binding	9	8.910891	0.894991	83	1011	7328	0.785957	1	1	100
GOTERM_MF_FAT	GO:0032559~adenyl ribonucleotide binding	9	8.910891	0.901677	83	1022	7328	0.777497	1	1	100

Blue											
Annotation Cluster 1	Enrichment Score: 6.548729456634398										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0045449~regulation of transcription	53	38.1295	1.46E-12	105	1485	7682	2.611166	1.29E-09	1.29E-09	2.27E-09
GOTERM_BP_FAT	GO:0006350~transcription	44	31.65468	3.29E-10	105	1216	7682	2.647306	2.92E-07	1.46E-07	5.12E-07
GOTERM_BP_FAT	GO:0006355~regulation of transcription, DNA-dependent	35	25.17986	2.09E-08	105	925	7682	2.768288	1.85E-05	6.18E-06	3.25E-05
GOTERM_BP_FAT	GO:0051252~regulation of RNA metabolic process	35	25.17986	4.56E-08	105	954	7682	2.684137	4.04E-05	1.01E-05	7.09E-05
GOTERM_MF_FAT	GO:0030528~transcription regulator activity	33	23.74101	3.75E-07	105	896	7328	2.570408	7.72E-05	7.72E-05	4.72E-04
GOTERM_MF_FAT	GO:0043565~sequence-specific DNA binding	14	10.07194	2.29E-04	105	292	7328	3.346119	0.046157	0.023351	0.288504
GOTERM_MF_FAT	GO:0003700~transcription factor activity	18	12.94964	6.08E-04	105	500	7328	2.512457	0.11768	0.024729	0.762543
GOTERM_MF_FAT	GO:0003677~DNA binding	32	23.02158	0.001706	105	1302	7328	1.715281	0.296496	0.056929	2.127192
Annotation Cluster 2	Enrichment Score: 3.9081490277452935										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0006355~regulation of transcription, DNA-dependent	35	25.17986	2.09E-08	105	925	7682	2.768288	1.85E-05	6.18E-06	3.25E-05

GOTERM_BP_FAT	GO:0051252~regulation of RNA metabolic process	35	25.17986	4.56E-08	105	954	7682	2.684137	4.04E-05	1.01E-05	7.09E-05
GOTERM_MF_FAT	GO:0030528~transcription regulator activity	33	23.74101	3.75E-07	105	896	7328	2.570408	7.72E-05	7.72E-05	4.72E-04
GOTERM_BP_FAT	GO:0006357~regulation of transcription from RNA polymerase II promoter	19	13.66906	3.94E-05	105	461	7682	3.01535	0.034326	0.006961	0.061307
GOTERM_BP_FAT	GO:0045893~positive regulation of transcription, DNA-dependent	14	10.07194	1.06E-04	105	283	7682	3.619317	0.089305	0.01547	0.164111
GOTERM_BP_FAT	GO:0051254~positive regulation of RNA metabolic process	14	10.07194	1.22E-04	105	287	7682	3.568873	0.102163	0.015277	0.189033
GOTERM_BP_FAT	GO:0045941~positive regulation of transcription	15	10.79137	1.40E-04	105	332	7682	3.305508	0.116986	0.015431	0.2182
GOTERM_BP_FAT	GO:0010628~positive regulation of gene expression	15	10.79137	1.70E-04	105	338	7682	3.24683	0.139474	0.016552	0.263384
GOTERM_BP_FAT	GO:0010557~positive regulation of macromolecule biosynthetic process	15	10.79137	4.42E-04	105	371	7682	2.958028	0.324041	0.038405	0.685231
GOTERM_BP_FAT	GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	15	10.79137	4.54E-04	105	372	7682	2.950077	0.331282	0.03592	0.704008
GOTERM_BP_FAT	GO:0031328~positive regulation of cellular biosynthetic process	15	10.79137	6.40E-04	105	385	7682	2.850464	0.4329	0.046169	0.990949
GOTERM_BP_FAT	GO:0051173~positive regulation of nitrogen compound metabolic process	15	10.79137	6.40E-04	105	385	7682	2.850464	0.4329	0.046169	0.990949

GOTERM_BP_FAT	GO:0009891~positive regulation of biosynthetic process	15	10.79137	7.09E-04	105	389	7682	2.821153	0.466495	0.047181	1.097049
GOTERM_BP_FAT	GO:0010604~positive regulation of macromolecule metabolic process	17	12.23022	0.001629	105	522	7682	2.382667	0.76416	0.098041	2.504455
GOTERM_BP_FAT	GO:0045944~positive regulation of transcription from RNA polymerase II promoter	10	7.194245	0.001768	105	205	7682	3.568873	0.791531	0.099253	2.715398
GOTERM_MF_FAT	GO:0008134~transcription factor binding	13	9.352518	0.011824	105	405	7328	2.240188	0.913728	0.263821	13.91248
GOTERM_MF_FAT	GO:0016563~transcription activator activity	9	6.47482	0.042633	105	277	7328	2.267561	0.999873	0.450279	42.23177
Annotation Cluster 3	Enrichment Score: 2.2605131805355603										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_MF_FAT	GO:0043565~sequence-specific DNA binding	14	10.07194	2.29E-04	105	292	7328	3.346119	0.046157	0.023351	0.288504
GOTERM_MF_FAT	GO:0003702~RNA polymerase II transcription factor activity	10	7.194245	5.75E-04	105	167	7328	4.17907	0.111753	0.029192	0.721911
GOTERM_MF_FAT	GO:0003700~transcription factor activity	18	12.94964	6.08E-04	105	500	7328	2.512457	0.11768	0.024729	0.762543
GOTERM_BP_FAT	GO:0045944~positive regulation of transcription from RNA polymerase II promoter	10	7.194245	0.001768	105	205	7682	3.568873	0.791531	0.099253	2.715398
GOTERM_MF_FAT	GO:0046983~protein dimerization activity	11	7.913669	0.021245	105	338	7328	2.271288	0.98801	0.357487	23.69707

GOTERM_MF_FAT	GO:0042803~protein homodimerization activity	6	4.316547	0.170568	105	207	7328	2.022912	1	0.723117	90.5141
GOTERM_MF_FAT	GO:0042802~identical protein binding	9	6.47482	0.292302	105	443	7328	1.417865	1	0.869308	98.71503
Annotation Cluster 4	Enrichment Score: 2.161855715020018										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_MF_FAT	GO:0008270~zinc ion binding	35	25.17986	3.78E-04	105	1362	7328	1.793441	0.074838	0.025595	0.474448
GOTERM_MF_FAT	GO:0046914~transition metal ion binding	37	26.61871	0.003187	105	1653	7328	1.562158	0.481906	0.089665	3.940734
GOTERM_MF_FAT	GO:0046872~metal ion binding	45	32.3741	0.019744	105	2367	7328	1.326815	0.983559	0.366468	22.21013
GOTERM_MF_FAT	GO:0043169~cation binding	45	32.3741	0.023546	105	2391	7328	1.313497	0.992617	0.359964	25.92582
GOTERM_MF_FAT	GO:0043167~ion binding	45	32.3741	0.027734	105	2414	7328	1.300982	0.996954	0.382958	29.82849

Table S 13: DAVID GO clustering analysis results in VCaP

Name	ConceptType	#Genes	Coeff	OddsRatio	P-Value	FDR	Direction
mitochondrial part	GO Cellular Component	459	0.477479055	19.44024111	1.55E-34	5.03E-32	up
mitochondrial membrane	GO Cellular Component	292	0.483099941	20.13131986	5.13E-27	8.31E-25	up
mitochondrial envelope	GO Cellular Component	309	0.462403553	17.70163243	7.79E-26	8.41E-24	up
organelle inner membrane	GO Cellular Component	236	0.494350298	21.58920383	9.29E-25	7.52E-23	up

mitochondrial inner membrane	GO Cellular Component	215	0.505933338	23.20058325	2.01E-24	1.30E-22	up
organelle envelope	GO Cellular Component	458	0.386747746	11.06166663	5.23E-24	2.82E-22	up
envelope	GO Cellular Component	460	0.385486555	10.97530627	6.17E-24	2.86E-22	up
microtubule cytoskeleton	GO Cellular Component	292	-0.349829787	0.113715561	3.83E-19	1.55E-17	down
endoplasmic reticulum part	GO Cellular Component	353	0.350464048	8.828602302	6.27E-17	2.22E-15	up
subsynaptic reticulum	GO Cellular Component	363	0.346436222	8.610353299	6.84E-17	2.22E-15	up
cytoskeletal part	GO Cellular Component	385	-0.293816416	0.161063936	2.67E-16	7.85E-15	down
respiratory chain	GO Cellular Component	61	0.595390491	40.45210038	1.49E-15	4.02E-14	up
nuclear membrane-endoplasmic reticulum network	GO Cellular Component	320	0.345020918	8.534952503	2.26E-15	5.64E-14	up
mitochondrial lumen	GO Cellular Component	189	0.415265891	13.20657444	2.73E-15	5.78E-14	up
mitochondrial matrix	GO Cellular Component	189	0.415265891	13.20657444	2.73E-15	5.78E-14	up
endoplasmic reticulum membrane	GO Cellular Component	313	0.346637401	8.621125111	2.85E-15	5.78E-14	up
mitochondrial membrane part	GO Cellular	103	0.489332535	20.92636677	3.02E-14	5.75E-13	up

		Component						
	microtubule	GO Cellular Component	129	-0.403025137	0.081704739	3.26E-14	5.86E-13	down
	Oxidative phosphorylation	KEGG Pathway	90	0.602292584	42.22499604	1.05E-14	1.75E-12	up
	GTPase regulator activity	GO Molecular Function	169	-0.372373458	0.098849666	7.20E-15	2.98E-12	down
	nucleoside-triphosphatase regulator activity	GO Molecular Function	174	-0.362892132	0.104849176	2.25E-14	4.64E-12	down
	oxidoreductase activity	GO Molecular Function	366	0.317626291	7.198861712	4.21E-14	5.80E-12	up
	mitochondrial respiratory chain	GO Cellular Component	56	0.563677263	33.21605306	5.01E-13	8.54E-12	up
	GTPase activator activity	GO Molecular Function	95	-0.436687419	0.066281718	1.13E-13	9.36E-12	down
	small GTPase regulator activity	GO Molecular Function	123	-0.399892894	0.083310756	1.13E-13	9.36E-12	down
	GTP catabolic process	GO Biological Process	59	-0.52379397	0.038574099	2.74E-14	1.59E-11	down
	regulation of GTP catabolic process	GO Biological Process	59	-0.52379397	0.038574099	2.74E-14	1.59E-11	down
	regulation of GTPase activity	GO Biological Process	59	-0.52379397	0.038574099	2.74E-14	1.59E-11	down
	transmembrane transporter activity	GO Molecular Function	227	0.368367218	9.867612143	3.01E-13	1.97E-11	up
	enzyme regulator activity	GO Molecular Function	344	-0.273181859	0.183100848	3.34E-13	1.97E-11	down

	regulation of small GTPase mediated signal transduction	GO Biological Process	102	-0.429668498	0.0692369	9.09E-14	3.24E-11	down
	regulation of signaling process	GO Biological Process	286	-0.297871525	0.157055695	1.68E-13	3.24E-11	down
	regulation of nucleotide catabolic process	GO Biological Process	61	-0.505452116	0.04323146	1.69E-13	3.24E-11	down
	regulation of purine nucleotide catabolic process	GO Biological Process	61	-0.505452116	0.04323146	1.69E-13	3.24E-11	down
	regulation of Ras GTPase activity	GO Biological Process	53	-0.529230312	0.037292652	1.74E-13	3.24E-11	down
	cell cycle phase	GO Biological Process	272	-0.302901267	0.1522224	1.80E-13	3.24E-11	down
	regulation of signal transduction	GO Biological Process	282	-0.298581529	0.15636423	1.98E-13	3.24E-11	down
	nucleoside triphosphate catabolic process	GO Biological Process	67	-0.488350952	0.048079	2.16E-13	3.24E-11	down
	purine ribonucleoside triphosphate catabolic process	GO Biological Process	64	-0.495003951	0.046131671	2.42E-13	3.24E-11	down
	ribonucleoside triphosphate catabolic process	GO Biological Process	64	-0.495003951	0.046131671	2.42E-13	3.24E-11	down
	GTP metabolic process	GO Biological Process	64	-0.489600194	0.047707182	5.21E-13	6.24E-11	down
	purine nucleoside triphosphate catabolic process	GO Biological Process	66	-0.484413545	0.049269978	5.38E-13	6.24E-11	down

	purine ribonucleotide catabolic process	GO Biological Process	67	-0.477711696	0.051365375	9.89E-13	1.08E-10	down
	ligase activity, forming carbon-nitrogen bonds	GO Molecular Function	147	-0.358528029	0.107731723	2.48E-12	1.28E-10	down
	hydrogen ion transmembrane transporter activity	GO Molecular Function	51	0.576402535	35.94952463	3.29E-12	1.43E-10	up
	acid-amino acid ligase activity	GO Molecular Function	125	-0.377453275	0.095777817	3.45E-12	1.43E-10	down
	ribonucleotide catabolic process	GO Biological Process	68	-0.472818853	0.052951236	1.42E-12	1.45E-10	down
	Parkinson's disease	KEGG Pathway	89	0.541592459	28.95621559	2.74E-12	2.29E-10	up
	protein amino acid phosphorylation	GO Biological Process	271	-0.288307935	0.166673107	3.69E-12	3.57E-10	down
	regulation of nucleotide metabolic process	GO Biological Process	74	-0.451761696	0.060354372	4.38E-12	4.01E-10	down
	enzyme activator activity	GO Molecular Function	141	-0.354127785	0.110718382	1.24E-11	4.64E-10	down
	M phase	GO Biological Process	221	-0.308028648	0.14744835	5.52E-12	4.81E-10	down
	regulation of Ras protein signal transduction	GO Biological Process	91	-0.419266346	0.073860579	5.98E-12	4.96E-10	down
	oxidative phosphorylation	GO Biological Process	72	0.511959016	24.08585285	6.96E-12	5.51E-10	up

cell cycle process	GO Biological Process	358	-0.253583389	0.206816907	1.60E-11	1.21E-09	down
microtubule-based process	GO Biological Process	137	-0.357812537	0.108211819	1.69E-11	1.23E-09	down
protein serine/threonine kinase activity	GO Molecular Function	169	-0.325062384	0.132637515	3.96E-11	1.36E-09	down
cellular respiration	GO Biological Process	78	0.489921163	21.00305761	2.19E-11	1.53E-09	up
Rab GTPase activator activity	GO Molecular Function	20	-0.670649118	0.01548596	5.10E-11	1.62E-09	down
oxidation reduction	GO Biological Process	348	0.285605888	5.899860264	3.58E-11	2.25E-09	up
purine nucleotide catabolic process	GO Biological Process	72	-0.440325131	0.0648001	3.65E-11	2.25E-09	down
regulation of Rab GTPase activity	GO Biological Process	20	-0.675894483	0.01498929	3.75E-11	2.25E-09	down
regulation of Rab protein signal transduction	GO Biological Process	20	-0.675894483	0.01498929	3.75E-11	2.25E-09	down
Ras GTPase activator activity	GO Molecular Function	44	-0.510410866	0.041919528	8.12E-11	2.39E-09	down
generation of precursor metabolites and energy	GO Biological Process	204	0.349053446	8.751546	4.51E-11	2.62E-09	up
regulation of catabolic process	GO Biological Process	126	-0.361490638	0.105766373	4.90E-11	2.75E-09	down
microtubule cytoskeleton organization	GO Biological Process	100	-0.391007041	0.088040748	5.42E-11	2.89E-09	down
electron transport chain	GO Biological	88	0.464098929	17.88912456	5.48E-11	2.89E-09	up

	Process						
small conjugating protein ligase activity	GO Molecular Function	111	-0.369719088	0.100493801	1.22E-10	3.36E-09	down
inorganic cation transmembrane transporter activity	GO Molecular Function	79	0.469958332	18.55254277	2.12E-10	5.47E-09	up
Cardiac muscle contraction	KEGG Pathway	26	0.714765181	84.94342799	1.29E-10	7.18E-09	up
ubiquitin-protein ligase activity	GO Molecular Function	97	-0.379607509	0.094504115	3.64E-10	8.84E-09	down
structural constituent of ribosome	GO Molecular Function	139	0.382435647	10.76917267	4.51E-10	1.03E-08	up
respiratory electron transport chain	GO Biological Process	52	0.528599666	26.71004638	2.74E-10	1.40E-08	up
ATP synthesis coupled electron transport	GO Biological Process	45	0.550088025	30.52608282	3.04E-10	1.47E-08	up
mitochondrial ATP synthesis coupled electron transport	GO Biological Process	45	0.550088025	30.52608282	3.04E-10	1.47E-08	up
cation transmembrane transporter activity	GO Molecular Function	122	0.396251033	11.7346368	7.39E-10	1.61E-08	up
nucleotide catabolic process	GO Biological Process	83	-0.401009218	0.082734786	4.36E-10	2.05E-08	down
monovalent inorganic cation transmembrane transporter activity	GO Molecular Function	64	0.486460732	20.55620427	1.05E-09	2.07E-08	up
cytochrome-c oxidase activity	GO Molecular Function	15	0.728307806	92.40190465	1.20E-09	2.07E-08	up
heme-copper terminal oxidase activity	GO Molecular Function	15	0.728307806	92.40190465	1.20E-09	2.07E-08	up
oxidoreductase activity, acting on heme group of donors	GO Molecular Function	15	0.728307806	92.40190465	1.20E-09	2.07E-08	up
oxidoreductase activity, acting on heme group of donors, oxygen as acceptor	GO Molecular Function	15	0.728307806	92.40190465	1.20E-09	2.07E-08	up

regulation of cellular catabolic process	GO Biological Process	109	-0.363829352	0.104240261	4.98E-10	2.28E-08	down
nucleobase, nucleoside and nucleotide catabolic process	GO Biological Process	85	-0.395882137	0.085413396	5.54E-10	2.41E-08	down
nucleobase, nucleoside, nucleotide and nucleic acid catabolic process	GO Biological Process	85	-0.395882137	0.085413396	5.54E-10	2.41E-08	down
intrinsic to organelle membrane	GO Cellular Component	75	0.448708225	16.25736067	1.77E-09	2.87E-08	up
ion transmembrane transporter activity	GO Molecular Function	160	0.349481587	8.774862524	2.43E-09	4.02E-08	up
cytoskeleton organization	GO Biological Process	243	-0.266571986	0.190778841	1.30E-09	5.53E-08	down
ribosome	GO Cellular Component	173	0.330800272	7.813039332	3.61E-09	5.57E-08	up
mitosis	GO Biological Process	173	-0.300278462	0.154723912	1.72E-09	6.96E-08	down
nuclear division	GO Biological Process	173	-0.300278462	0.154723912	1.72E-09	6.96E-08	down
Huntington's disease	KEGG Pathway	126	0.411594744	12.90868102	2.61E-09	1.09E-07	up
microtubule organizing center	GO Cellular Component	134	-0.318728933	0.137962219	8.04E-09	1.18E-07	down
integral to organelle membrane	GO Cellular Component	68	0.44592569	15.97864914	9.98E-09	1.41E-07	up
Alzheimer's disease	KEGG Pathway	106	0.427681417	14.2659046	5.12E-09	1.71E-07	up
M phase of mitotic cell cycle	GO Biological Process	180	-0.289175575	0.165776817	4.65E-09	1.84E-07	down

Chemokine signaling pathway	KEGG Pathway	56	-0.48185232	0.050060484	7.39E-09	1.95E-07	down
Glioma	KEGG Pathway	25	-0.628358948	0.020140862	8.17E-09	1.95E-07	down
phosphotransferase activity, alcohol group as acceptor	GO Molecular Function	285	-0.23531019	0.231688701	1.28E-08	2.04E-07	down
organelle fission	GO Biological Process	181	-0.28460635	0.170551678	8.22E-09	3.18E-07	down
substrate-specific transporter activity	GO Molecular Function	278	0.261541996	5.080354329	3.30E-08	5.05E-07	up
heterocycle catabolic process	GO Biological Process	88	-0.364598728	0.10374304	1.44E-08	5.45E-07	down
cellular nitrogen compound catabolic process	GO Biological Process	90	-0.357092693	0.108696994	2.45E-08	9.09E-07	down
microtubule associated complex	GO Cellular Component	53	-0.415907096	0.075418731	1.02E-07	1.38E-06	down
Chronic myeloid leukemia	KEGG Pathway	40	-0.509927007	0.042045769	6.87E-08	1.44E-06	down
RNA processing	GO Biological Process	455	0.212295057	3.740916286	3.96E-08	1.44E-06	up
mitotic cell cycle	GO Biological Process	296	-0.224680625	0.247510559	5.34E-08	1.90E-06	down
T cell receptor signaling pathway	KEGG Pathway	38	-0.511077754	0.041746154	1.20E-07	2.23E-06	down

	kinase activity	GO Molecular Function	314	-0.210972892	0.269519687	1.52E-07	2.23E-06	down
	Focal adhesion	KEGG Pathway	82	-0.391785229	0.087615999	1.43E-07	2.39E-06	down
	condensed chromosome	GO Cellular Component	92	-0.336172546	0.123788523	2.03E-07	2.64E-06	down
	cell division	GO Biological Process	202	-0.257183514	0.202241096	7.79E-08	2.71E-06	down
	ubiquitin thiolesterase activity	GO Molecular Function	38	-0.454385848	0.05937809	2.11E-07	3.00E-06	down
	Natural killer cell mediated cytotoxicity	KEGG Pathway	29	-0.549119959	0.032956547	2.34E-07	3.55E-06	down
	ribonucleoprotein complex	GO Cellular Component	428	0.199459193	3.454095793	3.42E-07	4.27E-06	up
	ribosomal subunit	GO Cellular Component	110	0.344368105	8.500396537	3.71E-07	4.45E-06	up
	transmembrane transport	GO Biological Process	282	0.249219641	4.70583111	1.36E-07	4.63E-06	up
	structural molecule activity	GO Molecular Function	264	0.2474829	4.655313402	3.41E-07	4.69E-06	up
	Renal cell carcinoma	KEGG Pathway	34	-0.513899017	0.041020596	3.73E-07	5.08E-06	down
	Ubiquitin mediated proteolysis	KEGG Pathway	88	-0.370610161	0.099938837	3.96E-07	5.08E-06	down
	antigen processing and presentation of peptide antigen via MHC class I	GO Biological Process	11	0.699498506	77.25480223	2.34E-07	7.83E-06	up

RNA binding	GO Molecular Function	515	0.180583101	3.071767718	6.87E-07	9.15E-06	up
ligase activity	GO Molecular Function	240	-0.221787667	0.252000695	8.08E-07	1.04E-05	down
glycoprotein biosynthetic process	GO Biological Process	67	0.415930914	13.26126826	3.36E-07	1.09E-05	up
antigen processing and presentation of peptide antigen	GO Biological Process	12	0.677755597	67.49040381	3.38E-07	1.09E-05	up
NADH dehydrogenase (quinone) activity	GO Molecular Function	37	0.486434934	20.55290878	1.16E-06	1.37E-05	up
NADH dehydrogenase (ubiquinone) activity	GO Molecular Function	37	0.486434934	20.55290878	1.16E-06	1.37E-05	up
NADH dehydrogenase activity	GO Molecular Function	37	0.486434934	20.55290878	1.16E-06	1.37E-05	up
Neurotrophin signaling pathway	KEGG Pathway	62	-0.404310752	0.081054552	1.18E-06	1.40E-05	down
mitochondrial respiratory chain complex I	GO Cellular Component	39	0.465952811	18.09641975	1.56E-06	1.68E-05	up
NADH dehydrogenase complex	GO Cellular Component	39	0.465952811	18.09641975	1.56E-06	1.68E-05	up
respiratory chain complex I	GO Cellular Component	39	0.465952811	18.09641975	1.56E-06	1.68E-05	up
Pathways in cancer	KEGG Pathway	128	-0.308548938	0.146972361	1.69E-06	1.88E-05	down
guanyl-nucleotide exchange factor activity	GO Molecular Function	52	-0.384242314	0.091820895	1.68E-06	1.92E-05	down
electron carrier activity	GO Molecular Function	70	0.39119475	11.37163462	1.77E-06	1.97E-05	up
Insulin signaling pathway	KEGG Pathway	61	-0.398605701	0.083979864	2.12E-06	2.21E-05	down
protein K48-linked ubiquitination	GO Biological	14	-	0.02278734	9.73E-07	3.08E-05	down

	Process		0.608493748				
oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor	GO Molecular Function	42	0.455624435	16.97136145	2.84E-06	3.09E-05	up
mitochondrial electron transport, NADH to ubiquinone	GO Biological Process	36	0.488526668	20.82182654	1.07E-06	3.32E-05	up
Non-small cell lung cancer	KEGG Pathway	25	-0.531450936	0.036781536	3.46E-06	3.40E-05	down
Regulation of actin cytoskeleton	KEGG Pathway	83	-0.34846926	0.114681118	4.26E-06	3.95E-05	down
regulation of hydrolase activity	GO Biological Process	141	-0.268804465	0.188150255	1.46E-06	4.45E-05	down
glycoprotein metabolic process	GO Biological Process	86	0.363685679	9.584660569	1.61E-06	4.85E-05	up
Vascular smooth muscle contraction	KEGG Pathway	32	-0.476484247	0.051758694	7.27E-06	6.39E-05	down
Apoptosis	KEGG Pathway	42	-0.432748095	0.06792441	7.91E-06	6.61E-05	down
early endosome	GO Cellular Component	80	-0.315628786	0.140645997	6.90E-06	7.21E-05	down
spindle	GO Cellular Component	107	-0.281277446	0.174116774	8.02E-06	8.12E-05	down
Ras protein signal transduction	GO Biological Process	102	-0.296168644	0.1587266	3.00E-06	8.85E-05	down
cysteine-type peptidase activity	GO Molecular Function	67	-0.331873174	0.1271406	8.39E-06	8.89E-05	down
ErbB signaling pathway	KEGG Pathway	37	-0.445484903	0.062755184	1.17E-05	9.27E-05	down
Toll-like receptor signaling pathway	KEGG Pathway	31	-0.472146465	0.053172962	1.24E-05	9.43E-05	down
intrinsic to endoplasmic reticulum membrane	GO Cellular Component	38	0.442570275	15.64890296	1.03E-05	1.01E-04	up

energy derivation by oxidation of organic compounds	GO Biological Process	101	0.332166114	7.879640035	3.96E-06	1.15E-04	up
microtubule organizing center organization	GO Biological Process	24	-0.486575198	0.04861252	4.35E-06	1.24E-04	down
T cell receptor signaling pathway	GO Biological Process	16	-0.556086564	0.031560149	4.42E-06	1.24E-04	down
centrosome organization	GO Biological Process	23	-0.49298288	0.046714746	4.50E-06	1.24E-04	down
regulation of catalytic activity	GO Biological Process	366	-0.176721333	0.333452841	4.71E-06	1.28E-04	down
RIG-I-like receptor signaling pathway	KEGG Pathway	29	-0.475282523	0.052146687	1.88E-05	1.37E-04	down
endoplasmic reticulum lumen	GO Cellular Component	47	0.406296713	12.49058101	1.69E-05	1.61E-04	up
Fc epsilon RI signaling pathway	KEGG Pathway	21	-0.523926298	0.03854239	2.54E-05	1.77E-04	down
oxidoreductase activity, acting on NADH or NADPH	GO Molecular Function	60	0.379870293	10.59884465	1.75E-05	1.80E-04	up
purine ribonucleotide metabolic process	GO Biological Process	132	-0.260378127	0.198265543	6.80E-06	1.82E-04	down
B cell receptor signaling pathway	KEGG Pathway	28	-0.473128364	0.052849482	2.84E-05	1.89E-04	down
intracellular protein kinase cascade	GO Biological Process	201	-0.220243159	0.254431173	7.32E-06	1.90E-04	down
signal transmission via phosphorylation event	GO Biological Process	201	-0.220243159	0.254431173	7.32E-06	1.90E-04	down

transferase activity, transferring phosphorus-containing groups	GO Molecular Function	411	-0.155962248	0.379369638	2.16E-05	2.17E-04	down
ribonucleotide metabolic process	GO Biological Process	139	-0.252526087	0.208180317	8.86E-06	2.27E-04	down
activation of immune response	GO Biological Process	25	-0.467916556	0.054589265	9.35E-06	2.36E-04	down
integral to endoplasmic reticulum membrane	GO Cellular Component	34	0.442417737	15.63407543	2.74E-05	2.54E-04	up
RNA splicing	GO Biological Process	252	0.219993949	3.924253634	1.05E-05	2.61E-04	up
soluble fraction	GO Cellular Component	148	-0.234466879	0.232906134	2.93E-05	2.63E-04	down
Endocytosis	KEGG Pathway	111	-0.282876124	0.172395461	4.41E-05	2.83E-04	down
antigen receptor-mediated signaling pathway	GO Biological Process	18	-0.516058714	0.04047371	1.29E-05	3.08E-04	down
immune response-activating cell surface receptor signaling pathway	GO Biological Process	18	-0.516058714	0.04047371	1.29E-05	3.08E-04	down
immune response-regulating cell surface receptor signaling pathway	GO Biological Process	18	-0.516058714	0.04047371	1.29E-05	3.08E-04	down
immune response-activating signal transduction	GO Biological Process	23	-0.474059045	0.052544693	1.43E-05	3.31E-04	down
immune response-regulating signaling pathway	GO Biological Process	23	-0.474059045	0.052544693	1.43E-05	3.31E-04	down
Shigellosis	KEGG Pathway	39	-0.41021859	0.07813261	5.43E-05	3.36E-04	down
protein modification by small protein conjugation or removal	GO Biological Process	220	-0.205942626	0.278078255	1.50E-05	3.44E-04	down
ubiquitin-dependent protein catabolic process	GO Biological Process	190	-0.217855725	0.258234303	1.58E-05	3.57E-04	down

mRNA processing	GO Biological Process	251	0.21574066	3.821884593	1.63E-05	3.63E-04	up
Prostate cancer	KEGG Pathway	42	-0.397381798	0.084621057	6.14E-05	3.66E-04	down
antigen processing and presentation	GO Biological Process	19	0.53735829	28.20420779	1.82E-05	4.02E-04	up
thiolester hydrolase activity	GO Molecular Function	50	-0.345418133	0.116876399	4.39E-05	4.32E-04	down
mTOR signaling pathway	KEGG Pathway	24	-0.477774822	0.051345228	7.67E-05	4.42E-04	down
regulation of cell cycle	GO Biological Process	220	-0.202995188	0.283218799	2.06E-05	4.47E-04	down
nucleoside triphosphate metabolic process	GO Biological Process	127	-0.251983937	0.208882912	2.12E-05	4.57E-04	down
regulation of protein metabolic process	GO Biological Process	326	-0.171939241	0.343511392	2.39E-05	5.08E-04	down
regulation of cell cycle process	GO Biological Process	77	-0.303109262	0.152025764	2.46E-05	5.16E-04	down
ribonucleoside triphosphate metabolic process	GO Biological Process	118	-0.257458887	0.201895289	2.49E-05	5.16E-04	down
glycosylation	GO Biological Process	50	0.396026329	11.71826147	2.60E-05	5.20E-04	up
macromolecule glycosylation	GO Biological Process	50	0.396026329	11.71826147	2.60E-05	5.20E-04	up
protein amino acid glycosylation	GO Biological Process	50	0.396026329	11.71826147	2.60E-05	5.20E-04	up
nucleobase, nucleoside and nucleotide metabolic process	GO Biological Process	248	-0.191151173	0.304851767	2.64E-05	5.22E-04	down
Basal transcription factors	KEGG Pathway	36	0.423595329	13.90820513	9.39E-05	5.23E-04	up
Small cell lung cancer	KEGG Pathway	42	-	0.089627512	1.00E-04	5.41E-04	down

			0.388132754				
purine ribonucleoside triphosphate metabolic process	GO Biological Process	117	-0.256644727	0.202919405	2.87E-05	5.62E-04	down
regulation of glucose import	GO Biological Process	11	-0.588596264	0.025786734	3.02E-05	5.84E-04	down
NOD-like receptor signaling pathway	KEGG Pathway	26	-0.456553889	0.058583424	1.13E-04	5.90E-04	down
polyol metabolic process	GO Biological Process	22	-0.465531634	0.055404378	3.45E-05	6.60E-04	down
regulation of ARF protein signal transduction	GO Biological Process	21	-0.472179768	0.053161958	3.58E-05	6.75E-04	down
transcription initiation from RNA polymerase II promoter	GO Biological Process	54	0.38064986	10.65031748	3.60E-05	6.75E-04	up
Colorectal cancer	KEGG Pathway	30	-0.430536986	0.068864213	1.36E-04	6.89E-04	down
small ribosomal subunit	GO Cellular Component	55	0.360991984	9.425546435	7.97E-05	6.98E-04	up
protein autoubiquitination	GO Biological Process	11	-0.583037222	0.026693162	3.86E-05	7.16E-04	down
protein amino acid N-linked glycosylation	GO Biological Process	26	0.477670924	19.46343535	4.13E-05	7.57E-04	up
Melanoma	KEGG Pathway	21	-0.483442306	0.049568265	1.58E-04	7.74E-04	down
meiosis I	GO Biological Process	15	-0.522881225	0.038793526	4.58E-05	8.22E-04	down
purine nucleoside triphosphate metabolic process	GO Biological Process	122	-0.247046969	0.215391044	4.58E-05	8.22E-04	down
spindle pole	GO Cellular Component	37	-0.373231004	0.098324267	1.03E-04	8.60E-04	down
endosome	GO Cellular Component	203	-0.193209793	0.300976483	1.05E-04	8.60E-04	down

condensed chromosome kinetochore	GO Cellular Component	55	-0.322745349	0.134561241	1.06E-04	8.60E-04	down
regulation of microtubule cytoskeleton organization	GO Biological Process	26	-0.433621415	0.067556759	4.96E-05	8.65E-04	down
regulation of phosphate metabolic process	GO Biological Process	199	-0.202459318	0.284163553	4.97E-05	8.65E-04	down
regulation of phosphorus metabolic process	GO Biological Process	199	-0.202459318	0.284163553	4.97E-05	8.65E-04	down
organic alcohol transport	GO Biological Process	11	0.599380332	41.46766076	5.23E-05	9.01E-04	up
transcription initiation	GO Biological Process	66	0.347908657	8.68950486	5.56E-05	9.48E-04	up
hydrogen ion transporting ATP synthase activity, rotational mechanism	GO Molecular Function	10	0.604292786	42.75314852	1.02E-04	9.80E-04	up
spindle organization	GO Biological Process	42	-0.364330347	0.103916215	5.85E-05	9.89E-04	down
cellular lipid metabolic process	GO Biological Process	266	0.19621317	3.385115262	6.00E-05	0.001004985	up
glucose import	GO Biological Process	12	-0.555928564	0.031591153	6.21E-05	0.00102842	down
translational elongation	GO Biological Process	90	0.308007338	6.781137762	6.26E-05	0.00102842	up
integral to peroxisomal membrane	GO Cellular Component	10	0.587162719	38.43567978	1.42E-04	0.001073986	up
intrinsic to peroxisomal membrane	GO Cellular Component	10	0.587162719	38.43567978	1.42E-04	0.001073986	up
Golgi membrane	GO Cellular Component	227	0.19750164	3.412329818	1.45E-04	0.001073986	up
membrane fraction	GO Cellular Component	317	0.170136491	2.878679312	1.46E-04	0.001073986	up
MAPK signaling pathway	KEGG Pathway	83	-	0.166166711	2.29E-04	0.001093514	down

			0.288797569				
SH3 domain binding	GO Molecular Function	57	-0.3144631	0.141668575	1.17E-04	0.001097754	down
purine nucleotide metabolic process	GO Biological Process	154	-0.220906555	0.253384378	6.77E-05	0.001102235	down
regulation of glucose transport	GO Biological Process	12	-0.553282118	0.032115017	6.97E-05	0.001124189	down
regulation of organelle organization	GO Biological Process	120	-0.243086252	0.220758533	7.24E-05	0.001156323	down
enzyme binding	GO Molecular Function	350	-0.151399344	0.390281272	1.26E-04	0.001156949	down
protein catabolic process	GO Biological Process	245	-0.181944549	0.322802672	7.58E-05	0.001194557	down
RNA elongation from RNA polymerase II promoter	GO Biological Process	41	0.40434698	12.34014805	7.62E-05	0.001194557	up
protein modification by small protein conjugation	GO Biological Process	188	-0.202229208	0.28457021	7.88E-05	0.001225363	down
protein K63-linked ubiquitination	GO Biological Process	10	-0.584288973	0.026486318	7.97E-05	0.001228372	down
regulation of microtubule-based process	GO Biological Process	28	-0.413580938	0.076516916	8.29E-05	0.001258504	down
nucleoside phosphate metabolic process	GO Biological Process	232	-0.185018666	0.316694245	8.39E-05	0.001258504	down
nucleotide metabolic process	GO Biological Process	232	-0.185018666	0.316694245	8.39E-05	0.001258504	down
mitotic cell cycle checkpoint	GO Biological Process	31	-0.398567649	0.083999726	8.62E-05	0.001282051	down
focal adhesion	GO Cellular Component	51	-0.322365766	0.134879041	1.89E-04	0.001361519	down

proteoglycan metabolic process	GO Biological Process	14	0.551244629	30.74629	9.24E-05	0.001363458	up
Hepatitis C	KEGG Pathway	59	-0.323030956	0.134322615	3.03E-04	0.001407154	down
regulation of blood pressure	GO Biological Process	20	0.498669574	22.17656223	9.70E-05	0.001418696	up
mitochondrial ribosome	GO Cellular Component	48	0.360878701	9.418913091	2.13E-04	0.001447391	up
organellar ribosome	GO Cellular Component	48	0.360878701	9.418913091	2.13E-04	0.001447391	up
microbody	GO Cellular Component	69	0.315623055	7.109796314	2.19E-04	0.001447391	up
peroxisome	GO Cellular Component	69	0.315623055	7.109796314	2.19E-04	0.001447391	up
condensed chromosome, centromeric region	GO Cellular Component	58	-0.30446684	0.150748546	2.24E-04	0.001452739	down
protein polyubiquitination	GO Biological Process	21	-0.452150039	0.060208888	1.02E-04	0.00147309	down
Acute myeloid leukemia	KEGG Pathway	25	-0.43738848	0.065993568	3.41E-04	0.001504817	down
N-Glycan biosynthesis	KEGG Pathway	34	0.403365275	12.26509109	3.42E-04	0.001504817	up
DNA integrity checkpoint	GO Biological Process	35	-0.37703747	0.096025633	1.14E-04	0.001644407	down
holo TFIIH complex	GO Cellular Component	10	0.573352975	35.27463126	2.60E-04	0.001650439	up
S phase	GO Biological Process	16	-0.491643952	0.047105077	1.20E-04	0.001709661	down
regulation of mitotic cell cycle	GO Biological Process	85	-0.270814396	0.1858147	1.21E-04	0.001714299	down
HOPS complex	GO Cellular Component	12	-0.521954797	0.03901752	2.77E-04	0.00172688	down
neuron projection	GO Cellular Component	133	-0.21628796	0.260762588	2.95E-04	0.001799696	down

perinuclear region of cytoplasm	GO Cellular Component	170	-0.195337121	0.297023609	3.00E-04	0.001799696	down
ubiquitin ligase complex	GO Cellular Component	81	-0.262494752	0.195674637	3.19E-04	0.0018767	down
microtubule binding	GO Molecular Function	45	-0.332662543	0.126518425	2.16E-04	0.001938799	down
cell-substrate adherens junction	GO Cellular Component	52	-0.310095004	0.145566984	3.37E-04	0.001951241	down
ubiquitin-specific protease activity	GO Molecular Function	25	-0.409532701	0.078466363	2.26E-04	0.001954432	down
UDP-glycosyltransferase activity	GO Molecular Function	37	0.399721935	11.99050606	2.27E-04	0.001954432	up
Endometrial cancer	KEGG Pathway	27	-0.418371624	0.074272413	4.65E-04	0.001989047	down
ARF GTPase activator activity	GO Molecular Function	17	-0.465651006	0.055363292	2.43E-04	0.002046694	down
Leukocyte transendothelial migration	KEGG Pathway	41	-0.358356585	0.107846567	5.05E-04	0.002107272	down
transferase activity, transferring hexosyl groups	GO Molecular Function	66	0.322802853	7.434216735	2.55E-04	0.002109285	up
modification-dependent macromolecule catabolic process	GO Biological Process	195	-0.191984623	0.30327685	1.53E-04	0.00212976	down
modification-dependent protein catabolic process	GO Biological Process	195	-0.191984623	0.30327685	1.53E-04	0.00212976	down
response to retinoic acid	GO Biological Process	16	0.520226531	25.35571148	1.60E-04	0.002214663	up
DNA damage checkpoint	GO Biological Process	34	-0.372805274	0.098584752	1.79E-04	0.002447738	down
tubulin binding	GO Molecular Function	65	-0.283981433	0.171215323	3.09E-04	0.002503695	down
motor activity	GO Molecular Function	52	-0.308556106	0.146965814	3.19E-04	0.002535659	down
microsome	GO Cellular	105	0.255175521	4.88327381	4.62E-04	0.002623334	up

	Component						
serine hydrolase activity	GO Molecular Function	45	0.365157801	9.672749735	3.50E-04	0.002726686	up
protein ubiquitination	GO Biological Process	171	-0.198174934	0.291831251	2.21E-04	0.003006955	down
RNA elongation	GO Biological Process	45	0.372118008	10.10032483	2.25E-04	0.003030346	up
fatty acid metabolic process	GO Biological Process	111	0.263667661	5.147911941	2.29E-04	0.003066249	up
cell-substrate junction	GO Cellular Component	57	-0.290866992	0.164043381	5.51E-04	0.003080693	down
TGF-beta signaling pathway	KEGG Pathway	22	-0.436646404	0.066298614	7.57E-04	0.00308307	down
monocarboxylic acid metabolic process	GO Biological Process	160	0.224953078	4.047078335	2.44E-04	0.003224476	up
lipid metabolic process	GO Biological Process	381	0.152907983	2.586390286	2.44E-04	0.003224476	up
transferase activity, transferring glycosyl groups	GO Molecular Function	97	0.267989742	5.288058781	4.32E-04	0.00330557	up
regulation of S phase	GO Biological Process	12	-0.522227102	0.038951548	2.55E-04	0.003339529	down
ribosome biogenesis	GO Biological Process	116	0.257003079	4.939052034	2.58E-04	0.003346249	up
large ribosomal subunit	GO Cellular Component	57	0.318413905	7.234184448	6.34E-04	0.003483706	up
Peroxisome	KEGG Pathway	57	0.320330477	7.320864168	8.77E-04	0.003486558	up
Pancreatic cancer	KEGG Pathway	33	-0.373696291	0.098040365	9.06E-04	0.003517184	down
kinetochore	GO Cellular Component	72	-0.262596663	0.195550749	6.63E-04	0.003579866	down
Ribosome	KEGG Pathway	78	0.281863061	5.764212025	0.001045216	0.003967069	up
Rho protein signal transduction	GO Biological	47	-	0.135159603	3.10E-04	0.004002945	down

	Process		0.322031401				
negative regulation of microtubule depolymerization	GO Biological Process	12	-0.51630267	0.040412394	3.22E-04	0.004093776	down
regulation of microtubule depolymerization	GO Biological Process	12	-0.51630267	0.040412394	3.22E-04	0.004093776	down
unsaturated fatty acid metabolic process	GO Biological Process	17	0.495801905	21.78484483	3.29E-04	0.004151514	up
regulation of ARF GTPase activity	GO Biological Process	16	-0.468466404	0.054403047	3.33E-04	0.004165179	down
Bladder cancer	KEGG Pathway	19	-0.44829168	0.061670036	0.00112336	0.004168915	down
small GTPase mediated signal transduction	GO Biological Process	194	-0.182930403	0.320831003	3.46E-04	0.004274106	down
regulation of cytoskeleton organization	GO Biological Process	66	-0.281620006	0.173746495	3.46E-04	0.004274106	down
kinase regulator activity	GO Molecular Function	43	-0.319566322	0.137246122	5.90E-04	0.0044297	down
cellular protein catabolic process	GO Biological Process	219	-0.173365994	0.340479041	3.64E-04	0.004442442	down
heterocycle metabolic process	GO Biological Process	238	-0.167453273	0.353222727	3.65E-04	0.004442442	down
aerobic respiration	GO Biological Process	29	0.419527321	13.56099808	3.74E-04	0.004522376	up
proteolysis involved in cellular protein catabolic process	GO Biological Process	218	-0.172576208	0.34215429	4.01E-04	0.004815538	down
protein processing	GO Biological Process	31	0.408733347	12.68116179	4.08E-04	0.004870251	up
positive regulation of NF-kappaB transcription factor activity	GO Biological Process	21	-0.421611108	0.072792103	4.28E-04	0.005069808	down
cytoskeleton-dependent intracellular transport	GO Biological Process	25	-0.395126664	0.085815352	4.65E-04	0.00545265	down
ncRNA metabolic process	GO Biological	188	0.200462272	3.475695004	4.67E-04	0.00545265	up

	Process						
positive regulation of immune response	GO Biological Process	34	-0.353184995	0.111368993	4.91E-04	0.005696691	down
carboxylesterase activity	GO Molecular Function	36	0.376955309	10.40857009	7.78E-04	0.005736629	up
lipid biosynthetic process	GO Biological Process	185	0.20068286	3.480462979	5.07E-04	0.005842725	up
protein oligomerization	GO Biological Process	98	-0.234980613	0.23216373	5.28E-04	0.006049633	down
ribonucleoprotein complex biogenesis	GO Biological Process	171	0.206606597	3.610979367	5.48E-04	0.006233614	up
ion transport	GO Biological Process	191	0.196592892	3.393112955	5.52E-04	0.006244182	up
microtubule-based movement	GO Biological Process	42	-0.324568025	0.133045637	5.56E-04	0.006246603	down
organelle outer membrane	GO Cellular Component	80	0.266556533	5.241168065	0.001203952	0.006394764	up
Fructose and mannose metabolism	KEGG Pathway	27	-0.383721197	0.092118742	0.001767733	0.006417638	down
regulation of phosphorylation	GO Biological Process	184	-0.180407767	0.325900365	5.85E-04	0.006532278	down
cell surface	GO Cellular Component	95	0.247362741	4.651838389	0.001260348	0.006586336	up
protein maturation by peptide bond cleavage	GO Biological Process	21	0.45274342	16.67020382	5.96E-04	0.00661227	up
positive regulation of response to stimulus	GO Biological Process	71	-0.263907662	0.193964002	6.27E-04	0.006906491	down
regulation of microtubule polymerization or depolymerization	GO Biological Process	18	-0.434909507	0.067018127	6.35E-04	0.006951232	down
outer membrane	GO Cellular Component	81	0.262095058	5.097845849	0.001403039	0.007130644	up
endosomal part	GO Cellular	120	-	0.284860904	0.00143053	0.007130644	down

		Component		0.202064918				
	endosome membrane	GO Cellular Component	120	-0.202064918	0.284860904	0.00143053	0.007130644	down
	mitochondrial outer membrane	GO Cellular Component	69	0.278659624	5.650592218	0.001474548	0.007238688	up
	cell cycle checkpoint	GO Biological Process	65	-0.271830542	0.184644987	6.74E-04	0.007334121	down
	negative regulation of microtubule polymerization or depolymerization	GO Biological Process	14	-0.470050263	0.053870182	7.22E-04	0.007804757	down
	Role of BRCA1, BRCA2 and ATR in Cancer Susceptibility	Biocarta Pathway	13	-0.610444766	0.022512716	8.01E-05	0.00784613	down
	negative regulation of organelle organization	GO Biological Process	53	-0.292250986	0.162638498	7.38E-04	0.007932008	down
	endosome transport	GO Biological Process	54	-0.289868087	0.165064898	7.50E-04	0.008014594	down
	unsaturated fatty acid biosynthetic process	GO Biological Process	11	0.537399543	28.21143959	7.64E-04	0.008096048	up
	coenzyme metabolic process	GO Biological Process	109	0.244917423	4.581680253	7.67E-04	0.008096048	up
	cellular ketone metabolic process	GO Biological Process	325	0.15038785	2.546198729	7.86E-04	0.008244956	up
	MAPKKK cascade	GO Biological Process	88	-0.238478429	0.227171515	8.03E-04	0.008353961	down
	lipid catabolic process	GO Biological Process	78	0.280223946	5.70579317	8.06E-04	0.008353961	up
	mitochondrial nucleoid	GO Cellular Component	27	0.388249952	11.16541771	0.001782086	0.008491117	up
	nucleoid	GO Cellular Component	27	0.388249952	11.16541771	0.001782086	0.008491117	up
	Drug metabolism - other enzymes	KEGG Pathway	17	-0.441657775	0.064265651	0.002448132	0.008678625	down
	Wnt signaling pathway	KEGG Pathway	49	-	0.155749532	0.002494455	0.008678625	down

			0.299215348				
protein maturation	GO Biological Process	38	0.366128023	9.731248242	8.44E-04	0.008689925	up
proton-transporting ATP synthase complex	GO Cellular Component	13	0.484948921	20.36397665	0.001891073	0.008879822	up
Fc gamma R-mediated phagocytosis	KEGG Pathway	40	-0.321931081	0.135243895	0.002622952	0.008939447	down
proton-transporting ATPase activity, rotational mechanism	GO Molecular Function	13	0.504438736	22.98608545	0.001249499	0.009053385	up
insoluble fraction	GO Cellular Component	332	0.13581346	2.325714789	0.002004879	0.009279724	up
Rho GTPase binding	GO Molecular Function	17	-0.42356	0.071915792	0.001336237	0.009458331	down
GTPase binding	GO Molecular Function	60	-0.266185703	0.191237374	0.00135119	0.009458331	down
interphase	GO Biological Process	71	-0.256964419	0.202516653	9.32E-04	0.009541961	down
spliceosomal complex	GO Cellular Component	125	0.210636577	3.702557418	0.002093221	0.009552162	up
regulation of stress-activated protein kinase signaling cascade	GO Biological Process	32	-0.346833219	0.11585307	9.51E-04	0.009679938	down
macromolecule catabolic process	GO Biological Process	342	-0.133954001	0.434972869	9.68E-04	0.009796853	down
Spliceosome	KEGG Pathway	117	0.219744625	3.918177908	0.002947315	0.009844032	up
centrosome cycle	GO Biological Process	14	-0.461565953	0.056786792	9.89E-04	0.009953848	down
replication fork	GO Cellular Component	29	-0.33982061	0.121013651	0.002314731	0.010408615	down
dynein complex	GO Cellular Component	13	-0.449905508	0.061054619	0.002345151	0.010408615	down
ATP synthesis coupled proton	GO Biological	23	0.426500498	14.16159116	0.001059718	0.010542677	up

	transport	Process						
	energy coupled proton transport, down electrochemical gradient	GO Biological Process	23	0.426500498	14.16159116	0.001059718	0.010542677	up
	microtubule depolymerization	GO Biological Process	14	-0.458542665	0.05786382	0.001103867	0.010919502	down
	negative regulation of cell cycle process	GO Biological Process	21	-0.398817868	0.083869207	0.001116215	0.010979267	down
	mRNA metabolic process	GO Biological Process	294	0.152696053	2.582986083	0.00112734	0.011026402	up
	transition metal ion transport	GO Biological Process	35	0.369625026	9.945047421	0.00114007	0.011088617	up
	dephosphorylation	GO Biological Process	79	-0.242661597	0.221341897	0.001150814	0.011130926	down
	serine-type peptidase activity	GO Molecular Function	44	0.333611657	7.950745538	0.001641662	0.011300105	up
	rRNA processing	GO Biological Process	86	0.261760256	5.087249995	0.001183127	0.011380243	up
	cellular hormone metabolic process	GO Biological Process	22	0.429053915	14.38810664	0.001213383	0.011607138	up
	I-kappaB kinase/NF-kappaB cascade	GO Biological Process	76	-0.244419896	0.21893643	0.001286291	0.012226158	down
	rRNA metabolic process	GO Biological Process	87	0.258730818	4.992369453	0.001292138	0.012226158	up
	Thyroid cancer	KEGG Pathway	14	-0.456608017	0.05856372	0.003839757	0.012573321	down
	inositol or phosphatidylinositol phosphatase activity	GO Molecular Function	17	-0.414335558	0.076158916	0.001868558	0.012651054	down
	Jak-STAT signaling pathway	KEGG Pathway	25	-0.370201011	0.100193276	0.003994383	0.012828114	down
	oxidoreduction coenzyme metabolic process	GO Biological Process	38	0.354849912	9.072547856	0.001365291	0.012848495	up
	Phosphatidylinositol signaling system	KEGG Pathway	35	-0.325968099	0.13189304	0.00409926	0.012916536	down

P-P-bond-hydrolysis-driven transmembrane transporter activity	GO Molecular Function	58	0.29547339	6.272978752	0.001992053	0.013059015	up
primary active transmembrane transporter activity	GO Molecular Function	58	0.29547339	6.272978752	0.001992053	0.013059015	up
DNA-directed RNA polymerase II, holoenzyme	GO Cellular Component	65	0.268169087	5.293955945	0.00314129	0.013753756	up
channel activity	GO Molecular Function	47	0.318506338	7.238341236	0.002175918	0.01382545	up
passive transmembrane transporter activity	GO Molecular Function	47	0.318506338	7.238341236	0.002175918	0.01382545	up
microtubule-based transport	GO Biological Process	17	-0.421233514	0.072963117	0.001491855	0.013914077	down
fatty acid catabolic process	GO Biological Process	31	0.378597254	10.51532338	0.001494504	0.013914077	up
microtubule polymerization or depolymerization	GO Biological Process	20	-0.397298932	0.084664647	0.001542426	0.01424858	down
stress-activated protein kinase signaling cascade	GO Biological Process	40	-0.309094871	0.146474564	0.001546802	0.01424858	down
Fatty acid metabolism	KEGG Pathway	27	0.36431364	9.622138114	0.004609852	0.014256395	up
regulation of epithelial cell differentiation	GO Biological Process	11	0.517993334	25.00624442	0.001559743	0.014292171	up
Long-term potentiation	KEGG Pathway	26	-0.359618964	0.107003801	0.004728902	0.014358665	down
Axon guidance	KEGG Pathway	43	-0.29617918	0.158716208	0.004947719	0.014754804	down
vesicular fraction	GO Cellular Component	109	0.213698137	3.773678184	0.003415668	0.014755687	up
fat-soluble vitamin metabolic process	GO Biological Process	10	0.527949233	26.60229739	0.001767632	0.016099492	up
glucose transport	GO Biological Process	16	-0.424915763	0.071312408	0.00179397	0.016099492	down
hexose transport	GO Biological Process	16	-0.424915763	0.071312408	0.00179397	0.016099492	down

monosaccharide transport	GO Biological Process	16	-0.424915763	0.071312408	0.00179397	0.016099492	down
spindle assembly	GO Biological Process	13	-0.455549796	0.05895013	0.001821114	0.016259279	down
cytosolic ribosome	GO Cellular Component	66	0.261452589	5.077532304	0.003873385	0.016375724	up
cell projection	GO Cellular Component	264	-0.131950943	0.44042135	0.003891762	0.016375724	down
active transmembrane transporter activity	GO Molecular Function	106	0.223831373	4.018964417	0.002634631	0.016486405	up
RNA polymerase	KEGG Pathway	25	0.36609578	9.729298538	0.005841034	0.017113206	up
cellular response to lipopolysaccharide	GO Biological Process	10	-0.495050031	0.046118462	0.001953311	0.017262509	down
cellular response to molecule of bacterial origin	GO Biological Process	10	-0.495050031	0.046118462	0.001953311	0.017262509	down
DNA damage response, signal transduction	GO Biological Process	62	-0.256436937	0.203181611	0.001971602	0.017336157	down
small conjugating protein-specific protease activity	GO Molecular Function	28	-0.336364694	0.123640793	0.002834207	0.017470562	down
Antigen processing and presentation	KEGG Pathway	28	0.35139719	8.879949099	0.00608908	0.017532351	up
small GTPase binding	GO Molecular Function	54	-0.26031903	0.198338373	0.003037775	0.018450014	down
cofactor catabolic process	GO Biological Process	24	0.402720988	12.21607996	0.002126701	0.018605958	up
regulation of binding	GO Biological Process	97	-0.212252835	0.267384338	0.002205817	0.019201635	down
Valine, leucine and isoleucine degradation	KEGG Pathway	36	0.319059015	7.26324531	0.006790926	0.019221774	up
JNK cascade	GO Biological Process	34	-0.319613073	0.137206253	0.00224224	0.01942159	down

G-protein coupled receptor protein signaling pathway	GO Biological Process	67	-0.245777362	0.217097225	0.002283204	0.019628127	down
regulation of cellular component organization	GO Biological Process	283	-0.134809712	0.432665863	0.002288633	0.019628127	down
positive regulation of stress-activated protein kinase signaling cascade	GO Biological Process	10	-0.489516496	0.047732003	0.00231344	0.019743621	down
response to oxidative stress	GO Biological Process	96	0.236010616	4.334964774	0.002357593	0.019948728	up
cellular lipid catabolic process	GO Biological Process	50	0.30781881	6.77319742	0.00236639	0.019948728	up
cofactor metabolic process	GO Biological Process	142	0.198661235	3.437009363	0.002371848	0.019948728	up
amino acid binding	GO Molecular Function	20	-0.375130163	0.097170613	0.003347502	0.0200365	down
translation	GO Biological Process	327	0.135576797	2.322296719	0.002401492	0.020100947	up
negative regulation of translation	GO Biological Process	23	-0.366296562	0.102654163	0.002437267	0.020302786	down
Cell cycle	KEGG Pathway	76	-0.226453002	0.244799282	0.007465867	0.020779997	down
RNA polymerase II transcription factor activity	GO Molecular Function	117	0.208033064	3.643132765	0.003566588	0.021042867	up
transport vesicle membrane	GO Cellular Component	24	-0.34161728	0.119669977	0.005142626	0.021361678	down
amine binding	GO Molecular Function	24	-0.348643494	0.114557009	0.003678974	0.021400227	down
vesicle-mediated transport	GO Biological Process	387	-0.116597091	0.484515806	0.002605804	0.021540581	down

negative regulation of cytoskeleton organization	GO Biological Process	31	-0.326939586	0.131099147	0.002610605	0.021540581	down
RNA splicing, via transesterification reactions	GO Biological Process	94	0.235856193	4.330806614	0.00262446	0.021552762	up
icosanoid metabolic process	GO Biological Process	14	0.469066363	18.44998618	0.002639013	0.021570528	up
regulation of JNK cascade	GO Biological Process	27	-0.3423452	0.119129845	0.002774551	0.022507001	down
cation transport	GO Biological Process	143	0.195065396	3.361055261	0.00277944	0.022507001	up
double-strand break repair	GO Biological Process	41	-0.291993924	0.162898527	0.00289029	0.023296272	down
vacuolar transport	GO Biological Process	25	-0.350654191	0.113134447	0.002922587	0.02344804	down
regulation of kinase activity	GO Biological Process	149	-0.172703732	0.341883237	0.003006405	0.024009868	down
RNA polymerase II carboxy-terminal domain kinase activity	GO Molecular Function	10	0.505089111	23.07917919	0.004196079	0.024069177	up
Adipocytokine signaling pathway	KEGG Pathway	30	-0.32115875	0.13589459	0.008796593	0.024082477	down
protein domain specific binding	GO Molecular Function	196	-0.147225308	0.400537617	0.004351933	0.024621207	down
cation-transporting ATPase activity	GO Molecular Function	17	0.429670482	14.44334361	0.004464289	0.024915557	up
interphase of mitotic cell cycle	GO Biological Process	67	-0.239287576	0.226032042	0.003136774	0.024936639	down
nucleolus	GO Cellular Component	510	0.099078228	1.851011706	0.00608092	0.024939469	up
proton-transporting two-sector ATPase complex	GO Cellular Component	26	0.356634204	9.173710237	0.006396231	0.025592245	up
Golgi apparatus part	GO Cellular Component	280	0.129608262	2.237735807	0.0064628	0.025592245	up
coated membrane	GO Cellular	51	-	0.211127866	0.006556038	0.025592245	down

	Component		0.250263782				
membrane coat	GO Cellular Component	51	-0.250263782	0.211127866	0.006556038	0.025592245	down
clathrin coat of trans-Golgi network vesicle	GO Cellular Component	10	-0.451843839	0.06032357	0.007072613	0.026423775	down
trans-Golgi network transport vesicle membrane	GO Cellular Component	10	-0.451843839	0.06032357	0.007072613	0.026423775	down
external side of plasma membrane	GO Cellular Component	31	0.331423925	7.843379556	0.007140297	0.026423775	up
mitochondrial small ribosomal subunit	GO Cellular Component	18	0.39995633	12.007985	0.007176828	0.026423775	up
organellar small ribosomal subunit	GO Cellular Component	18	0.39995633	12.007985	0.007176828	0.026423775	up
regulation of MAPKKK cascade	GO Biological Process	53	-0.260696904	0.197873154	0.00341371	0.02701486	down
positive regulation of kinase activity	GO Biological Process	83	-0.21745818	0.258873082	0.003469348	0.027289124	down
nuclear mRNA splicing, via spliceosome	GO Biological Process	85	0.239736569	4.436513525	0.00349539	0.027289124	up
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	GO Biological Process	85	0.239736569	4.436513525	0.00349539	0.027289124	up
response to tumor necrosis factor	GO Biological Process	10	-0.47516848	0.052183659	0.003531451	0.02737711	down
ncRNA processing	GO Biological Process	153	0.184708215	3.151533535	0.00353811	0.02737711	up
nucleoplasm part	GO Cellular Component	375	0.110890819	1.992007619	0.007628221	0.027770154	up
chromosome	GO Cellular Component	300	-0.115616078	0.487478732	0.007774671	0.027988817	down
Arachidonic acid metabolism	KEGG Pathway	14	0.416236386	13.28646723	0.01043069	0.028095568	up

Integrin Signaling Pathway	Biocarta Pathway	26	-0.402630857	0.081905185	0.001089562	0.028665447	down
Growth Hormone Signaling Pathway	Biocarta Pathway	13	-0.503574009	0.043739001	0.001420145	0.028665447	down
PTEN dependent cell cycle arrest and apoptosis	Biocarta Pathway	15	-0.47758933	0.051404451	0.001459957	0.028665447	down
CXCR4 Signaling Pathway	Biocarta Pathway	15	-0.47752301	0.051425642	0.001462523	0.028665447	down
Citrate cycle (TCA cycle)	KEGG Pathway	26	0.339765932	8.260722967	0.011188265	0.029657782	up
microbody part	GO Cellular Component	40	0.295792381	6.285426653	0.008489582	0.029898093	up
peroxisomal part	GO Cellular Component	40	0.295792381	6.285426653	0.008489582	0.029898093	up
cofactor binding	GO Molecular Function	154	0.175498018	2.976212155	0.005431735	0.029910754	up
cytosolic small ribosomal subunit	GO Cellular Component	32	0.321498657	7.374205447	0.008645429	0.030119559	up
Glycosaminoglycan biosynthesis - chondroitin sulfate	KEGG Pathway	11	0.439445571	15.34795142	0.012386826	0.031171105	up
Chagas disease (American trypanosomiasis)	KEGG Pathway	35	-0.291231739	0.163671956	0.012466934	0.031171105	down
Toxoplasmosis	KEGG Pathway	48	-0.256969647	0.202510073	0.01249133	0.031171105	down
Tryptophan metabolism	KEGG Pathway	18	0.378952841	10.53858615	0.012505773	0.031171105	up
lysosomal membrane	GO Cellular Component	44	0.282905222	5.801665812	0.009066186	0.031249407	up
magnesium ion binding	GO Molecular Function	74	-0.216088865	0.261085429	0.005758851	0.031294808	down
sulfur amino acid metabolic process	GO Biological Process	15	-0.409274195	0.078592522	0.004105485	0.031626766	down
macromolecule transmembrane transporter activity	GO Molecular Function	11	0.479188288	19.64784009	0.006051723	0.032043098	up

	protein transmembrane transporter activity	GO Molecular Function	11	0.479188288	19.64784009	0.006051723	0.032043098	up
	mitochondrial proton-transporting ATP synthase complex	GO Cellular Component	11	0.455626283	16.97155639	0.00947075	0.032300241	up
	chromosomal part	GO Cellular Component	247	-0.122686222	0.466523544	0.009605474	0.032418474	down
	regulation of localization	GO Biological Process	239	-0.136507831	0.428123887	0.004232109	0.032458602	down
	negative regulation of cellular component organization	GO Biological Process	86	-0.210123888	0.270945491	0.004288517	0.032746964	down
	calmodulin binding	GO Molecular Function	56	-0.239958194	0.225091986	0.00628525	0.032858331	down
	dendrite	GO Cellular Component	60	-0.224148743	0.248330044	0.009951436	0.033239847	down
	M phase of meiotic cell cycle	GO Biological Process	34	-0.302602039	0.152505734	0.004406419	0.033354675	down
	meiosis	GO Biological Process	34	-0.302602039	0.152505734	0.004406419	0.033354675	down
	regulation of transferase activity	GO Biological Process	156	-0.163102499	0.362903576	0.004436979	0.03344061	down
	phosphoric ester hydrolase activity	GO Molecular Function	142	-0.161941842	0.365530673	0.006558939	0.033501766	down
	ubiquitin protein ligase binding	GO Molecular Function	43	-0.265414207	0.192156472	0.006570564	0.033501766	down
	chromosome, centromeric region	GO Cellular Component	99	-0.181114569	0.32447199	0.010255117	0.033904671	down
	copper ion transport	GO Biological Process	11	0.48535155	20.41499487	0.004527686	0.033977159	up
	ion transmembrane transport	GO Biological Process	38	0.323069046	7.446525235	0.00459049	0.034300612	up

organic acid metabolic process	GO Biological Process	316	0.128494483	2.222300354	0.004620277	0.034373715	up
regulation of protein kinase activity	GO Biological Process	141	-0.169503861	0.348749957	0.00463976	0.034373715	down
nuclear membrane	GO Cellular Component	83	0.212576059	3.74745484	0.010659411	0.034885346	up
microbody membrane	GO Cellular Component	33	0.310085251	6.869273231	0.010942802	0.035103644	up
peroxisomal membrane	GO Cellular Component	33	0.310085251	6.869273231	0.010942802	0.035103644	up
lipoprotein binding	GO Molecular Function	13	0.451142423	16.50516486	0.006971909	0.035114616	up
Cell Cycle: G1/S Check Point	Biocarta Pathway	15	-0.462385899	0.056498163	0.002171196	0.035462874	down
aminoglycan metabolic process	GO Biological Process	20	0.40266071	12.21150461	0.004831434	0.035642062	up
rRNA binding	GO Molecular Function	21	0.385679996	10.98850832	0.007213898	0.035895661	up
Long-term depression	KEGG Pathway	23	-0.335806092	0.124070756	0.014625504	0.035918518	down
regulation of cellular response to stress	GO Biological Process	59	-0.241769711	0.22257214	0.004906243	0.03604122	down
substrate-specific channel activity	GO Molecular Function	41	0.301195927	6.500081067	0.007368167	0.036226821	up
acetyl-CoA metabolic process	GO Biological Process	32	0.341556632	8.353165837	0.00500899	0.036641396	up
cellular macromolecule catabolic process	GO Biological Process	304	-0.120738193	0.472205715	0.00507753	0.036987363	down

establishment of protein localization	GO Biological Process	552	-0.093559123	0.55909645	0.005182427	0.037594189	down
regulation of cellular protein metabolic process	GO Biological Process	286	-0.123544418	0.464042034	0.005232892	0.037802759	down
cellular glucan metabolic process	GO Biological Process	18	-0.375936845	0.096684695	0.005312857	0.037908539	down
glucan metabolic process	GO Biological Process	18	-0.375936845	0.096684695	0.005312857	0.037908539	down
glycogen metabolic process	GO Biological Process	18	-0.375936845	0.096684695	0.005312857	0.037908539	down
regulation of translation	GO Biological Process	95	-0.197061386	0.293857806	0.005413883	0.038471713	down
oxidoreductase activity, acting on peroxide as acceptor	GO Molecular Function	17	0.409594724	12.74922761	0.008094494	0.038872397	up
peroxidase activity	GO Molecular Function	17	0.409594724	12.74922761	0.008094494	0.038872397	up
nuclear envelope	GO Cellular Component	153	0.158103788	2.6712675	0.012283717	0.039018866	up
carboxylic acid metabolic process	GO Biological Process	313	0.126388399	2.193403329	0.005539605	0.039036971	up
oxoacid metabolic process	GO Biological Process	313	0.126388399	2.193403329	0.005539605	0.039036971	up
cellular response to biotic stimulus	GO Biological Process	32	-0.303264841	0.151878847	0.005560694	0.039036971	down
protein kinase regulator activity	GO Molecular Function	39	-0.269170229	0.187723061	0.008359777	0.03968492	down
glutamine metabolic process	GO Biological Process	14	-0.407963926	0.079235098	0.005736832	0.040066687	down
regulation of transcription factor activity	GO Biological Process	57	-0.241303916	0.223217361	0.005776415	0.040066687	down

	regulation of transcription regulator activity	GO Biological Process	57	-0.241303916	0.223217361	0.005776415	0.040066687	down
	oligosaccharide metabolic process	GO Biological Process	17	0.418081618	13.43970529	0.005816909	0.040187457	up
	negative regulation of translational initiation	GO Biological Process	12	-0.4289135	0.069562524	0.005949484	0.040940919	down
	embryonic epithelial tube formation	GO Biological Process	17	-0.379469005	0.094585494	0.006034577	0.041200777	down
	neural tube formation	GO Biological Process	17	-0.379469005	0.094585494	0.006034577	0.041200777	down
	endosome organization	GO Biological Process	14	-0.405640664	0.080387406	0.006137286	0.041738342	down
	hydrolase activity, acting on ester bonds	GO Molecular Function	308	-0.111451733	0.500259238	0.009014588	0.042307102	down
	hydrogen transport	GO Biological Process	29	0.346772131	8.62834654	0.006297979	0.042499153	up
	proton transport	GO Biological Process	29	0.346772131	8.62834654	0.006297979	0.042499153	up
	actin filament-based process	GO Biological Process	132	-0.168762053	0.350361422	0.006324978	0.042516554	down
	viral genome expression	GO Biological Process	10	0.486698853	20.58664637	0.006391588	0.042598867	up
	viral transcription	GO Biological Process	10	0.486698853	20.58664637	0.006391588	0.042598867	up
	plasma membrane organization	GO Biological Process	11	-0.439026987	0.065324987	0.006420649	0.042598867	down
	mitochondrial ATP synthesis coupled proton transport	GO Biological Process	10	0.48645779	20.55582841	0.006435096	0.042598867	up
	positive regulation of epithelial cell proliferation	GO Biological Process	11	0.473024323	18.90943092	0.006492884	0.042818601	up
	recycling endosome	GO Cellular	19	-0.33836552	0.122112918	0.013735672	0.043207357	down

	Component						
ER-Golgi intermediate compartment	GO Cellular Component	45	0.266852623	5.250821149	0.013912689	0.043343377	up
centrosome	GO Cellular Component	81	-0.189573926	0.307854609	0.014132105	0.043607637	down
channel regulator activity	GO Molecular Function	17	-0.363164326	0.104671965	0.009521572	0.043899691	down
ATPase activity, coupled to transmembrane movement of substances	GO Molecular Function	51	0.267752332	5.280262498	0.009672813	0.043899691	up
hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	GO Molecular Function	51	0.267752332	5.280262498	0.009672813	0.043899691	up
pore complex	GO Cellular Component	64	0.229041291	4.151218325	0.014438723	0.044133456	up
metal cluster binding	GO Molecular Function	41	0.291823389	6.132288415	0.009919243	0.044528774	up
positive regulation of binding	GO Biological Process	52	-0.246244772	0.216467522	0.006813838	0.044765633	down
regulation of mitotic metaphase/anaphase transition	GO Biological Process	16	-0.383357566	0.09232715	0.006862175	0.044913713	down
glutathione transferase activity	GO Molecular Function	10	0.473052033	18.91268749	0.010193074	0.045253335	up
calmodulin-dependent protein kinase activity	GO Molecular Function	10	-0.434315739	0.067265883	0.01029979	0.045253335	down
mammary gland morphogenesis	GO Biological Process	12	0.458784574	17.30795685	0.006942186	0.045267215	up
Bacterial invasion of epithelial cells	KEGG Pathway	40	-0.262493442	0.195676231	0.018787613	0.045471469	down
negative regulation of neuron apoptosis	GO Biological Process	26	-0.320781968	0.136213167	0.007028641	0.045659942	down

	protein deubiquitination	GO Biological Process	33	-0.293097196	0.161785449	0.007056155	0.045668274	down
	coenzyme binding	GO Molecular Function	124	0.179005414	3.041797111	0.010517375	0.045722902	up
	p53 signaling pathway	KEGG Pathway	34	-0.27923113	0.176345165	0.019189143	0.045779813	down
	morphogenesis of embryonic epithelium	GO Biological Process	26	-0.320281826	0.136637201	0.0071456	0.046075888	down
	Links between Pyk2 and Map Kinases	Biocarta Pathway	18	-0.417519256	0.074666888	0.003319565	0.046473911	down
	positive regulation of transcription factor activity	GO Biological Process	34	-0.288851834	0.166110683	0.00728896	0.046654704	down
	positive regulation of transcription regulator activity	GO Biological Process	34	-0.288851834	0.166110683	0.00728896	0.046654704	down
	ATPase activity, coupled to movement of substances	GO Molecular Function	52	0.261750067	5.086927854	0.010961311	0.047156475	up
	Inositol phosphate metabolism	KEGG Pathway	31	-0.287628928	0.167377914	0.020190085	0.047489356	down
	mitochondrial large ribosomal subunit	GO Cellular Component	17	0.378101774	10.48299425	0.015911035	0.047733104	up
	organellar large ribosomal subunit	GO Cellular Component	17	0.378101774	10.48299425	0.015911035	0.047733104	up
	ATM Signaling Pathway	Biocarta Pathway	12	-0.467581848	0.054702934	0.004670063	0.047753208	down
	Signaling of Hepatocyte Growth Factor Receptor	Biocarta Pathway	24	-0.364264299	0.103958878	0.004698167	0.047753208	down
	PKC-catalyzed phosphorylation of inhibitory phosphoprotein of myosin phosphatase	Biocarta Pathway	15	-0.427820448	0.070036663	0.005115393	0.047753208	down
	Fc Epsilon Receptor I Signaling in Mast Cells	Biocarta Pathway	12	-0.461399003	0.056845741	0.005360054	0.047753208	down

oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	GO Molecular Function	53	0.258591261	4.988041498	0.011348081	0.047927597	up
Ras GTPase binding	GO Molecular Function	50	-0.235698966	0.231129597	0.01137265	0.047927597	down
negative regulation of protein complex disassembly	GO Biological Process	25	-0.323343119	0.134062286	0.007536747	0.048026735	down
positive regulation of transferase activity	GO Biological Process	87	-0.197372231	0.293290684	0.007558487	0.048026735	down
ER-Golgi intermediate compartment membrane	GO Cellular Component	18	0.37018306	9.97959633	0.016163327	0.04804512	up
regulation of neuron apoptosis	GO Biological Process	38	-0.274560032	0.181539322	0.007882094	0.04990082	down

Table S 14: Discordance-based LRPath Gene Ontology and pathway analysis results with FDR <= 0.05

Name	ConceptType	#Genes	Coeff	OddsRatio	P-Value	FDR	Direction
ectoderm development	GO Biological Process	70	-0.29244408	0.162443448	1.63E-15	1.61E-12	down
epidermis development	GO Biological Process	65	-0.296634434	0.158267798	1.85E-15	1.61E-12	down
extracellular matrix	GO Cellular Component	58	-0.294481258	0.160399835	6.26E-14	2.03E-11	down
cell adhesion	GO Biological Process	220	-0.211939819	0.26790498	5.75E-13	3.34E-10	down

	biological adhesion	GO Biological Process	221	-0.210783743	0.269836689	7.72E-13	3.36E-10	down
	wound healing	GO Biological Process	57	-0.28251543	0.172782331	1.52E-12	5.29E-10	down
	cell surface	GO Cellular Component	95	-0.247915139	0.214232066	9.28E-12	1.50E-09	down
	tissue development	GO Biological Process	243	-0.195736299	0.296287684	1.91E-11	5.54E-09	down
	serine-type endopeptidase inhibitor activity	GO Molecular Function	15	-0.37131091	0.099504562	2.35E-11	6.16E-09	down
	peptidase regulator activity	GO Molecular Function	53	-0.278352324	0.177310897	2.98E-11	6.16E-09	down
	response to wounding	GO Biological Process	135	-0.223254447	0.24971403	2.99E-11	7.43E-09	down
	basal plasma membrane	GO Cellular Component	11	-0.391509148	0.087766454	1.59E-10	1.72E-08	down
	extracellular matrix part	GO Cellular Component	25	-0.320368741	0.136563418	2.41E-10	1.95E-08	down
	basal part of cell	GO Cellular Component	12	-0.375101077	0.097188179	6.24E-10	3.52E-08	down

	extracellular region part	GO Cellular Component	156	-0.205784902	0.27835096	6.53E-10	3.52E-08	down
	keratinocyte differentiation	GO Biological Process	25	-0.315359679	0.140881409	3.64E-10	7.91E-08	down
	DNA damage response, signal transduction by p53 class mediator	GO Biological Process	25	-0.312278924	0.143604666	7.13E-10	1.33E-07	down
	epidermal cell differentiation	GO Biological Process	29	-0.301697976	0.153364983	7.62E-10	1.33E-07	down
	blood vessel morphogenesis	GO Biological Process	81	-0.237879632	0.22801846	9.41E-10	1.49E-07	down
	ECM-receptor interaction	KEGG Pathway	26	-0.343438742	0.118322991	2.59E-09	4.32E-07	down
	basement membrane	GO Cellular Component	20	-0.318250885	0.138372697	1.07E-08	4.94E-07	down
	blood vessel development	GO Biological Process	97	-0.221791577	0.251994572	4.35E-09	6.31E-07	down
	vasculature development	GO Biological Process	99	-0.215963913	0.261288248	1.33E-08	1.79E-06	down
	regulation of epithelial cell differentiation	GO Biological Process	11	-0.358507654	0.107745365	1.96E-08	2.44E-06	down
	cell-substrate junction assembly	GO Biological Process	24	-0.295608502	0.1592801	3.72E-08	4.15E-06	down
	cell junction assembly	GO Biological Process	37	-0.267516448	0.189662352	3.82E-08	4.15E-06	down
	angiogenesis	GO Biological Process	67	-0.232053253	0.236425998	4.26E-08	4.36E-06	down

Arrhythmogenic right ventricular cardiomyopathy (ARVC)	KEGG Pathway	23	-0.328449091	0.129875058	7.59E-08	6.34E-06	down
proteinaceous extracellular matrix	GO Cellular Component	46	-0.249836113	0.211689748	1.59E-07	6.43E-06	down
negative regulation of cell adhesion	GO Biological Process	24	-0.291911263	0.162982231	7.43E-08	6.60E-06	down
chemotaxis	GO Biological Process	33	-0.271174514	0.185399313	7.58E-08	6.60E-06	down
taxis	GO Biological Process	33	-0.271174514	0.185399313	7.58E-08	6.60E-06	down
caveola	GO Cellular Component	25	-0.286440078	0.168619121	2.18E-07	7.85E-06	down
cell junction organization	GO Biological Process	45	-0.250747094	0.210494676	1.02E-07	8.46E-06	down
receptor complex	GO Cellular Component	24	-0.285495831	0.169611508	4.11E-07	1.33E-05	down
Dilated cardiomyopathy	KEGG Pathway	19	-0.332625915	0.126547227	2.67E-07	1.49E-05	down
basolateral plasma membrane	GO Cellular Component	104	-0.195700138	0.296354276	8.43E-07	2.48E-05	down
cell-substrate adhesion	GO Biological Process	56	-0.22824575	0.242087056	6.79E-07	5.20E-05	down
regulation of chemotaxis	GO Biological Process	11	-0.33468572	0.124937635	6.87E-07	5.20E-05	down
positive regulation of multicellular organismal process	GO Biological Process	60	-0.223701605	0.24902106	7.71E-07	5.59E-05	down
tissue regeneration	GO Biological Process	12	-0.323258922	0.134132453	1.33E-06	9.27E-05	down
regeneration	GO Biological Process	29	-0.26273504	0.195382656	1.45E-06	9.70E-05	down
regulation of behavior	GO Biological Process	12	-0.321363379	0.135721884	1.73E-06	1.12E-04	down

response to biotic stimulus	GO Biological Process	152	-0.170901273	0.345734398	1.94E-06	1.21E-04	down
response to other organism	GO Biological Process	102	-0.187192077	0.312445447	3.56E-06	2.13E-04	down
Cytokine-cytokine receptor interaction	KEGG Pathway	11	-0.347073593	0.115680134	6.07E-06	2.54E-04	down
endothelial cell proliferation	GO Biological Process	18	-0.28620369	0.168867014	4.37E-06	2.54E-04	down
Vascular smooth muscle contraction	KEGG Pathway	32	0.322811683	7.434624688	8.95E-06	2.99E-04	up
regulation of blood coagulation	GO Biological Process	12	-0.312526508	0.14338388	5.69E-06	3.09E-04	down
regulation of wound healing	GO Biological Process	12	-0.312526508	0.14338388	5.69E-06	3.09E-04	down
lipid catabolic process	GO Biological Process	78	0.209742252	3.682036133	8.74E-06	4.61E-04	up
insulin-like growth factor receptor signaling pathway	GO Biological Process	11	0.308974893	6.822035414	1.12E-05	5.75E-04	up
ossification	GO Biological Process	41	-0.227791574	0.242771317	1.38E-05	5.91E-04	down
regulation of apoptosis	GO Biological Process	370	-0.122280002	0.46770277	1.39E-05	5.91E-04	down
regulation of programmed cell death	GO Biological Process	371	-0.122096346	0.468236889	1.41E-05	5.91E-04	down
regulation of cell death	GO Biological Process	375	-0.121630342	0.469594882	1.42E-05	5.91E-04	down
cytokine-mediated signaling pathway	GO Biological Process	28	-0.249952931	0.211536121	1.43E-05	5.91E-04	down
initiation of signal transduction	GO Biological Process	28	-0.249952931	0.211536121	1.43E-05	5.91E-04	down

signal initiation by diffusible mediator	GO Biological Process	28	-0.249952931	0.211536121	1.43E-05	5.91E-04	down
signal initiation by protein/peptide mediator	GO Biological Process	28	-0.249952931	0.211536121	1.43E-05	5.91E-04	down
response to external stimulus	GO Biological Process	171	-0.154074793	0.383845766	1.53E-05	6.21E-04	down
Hypertrophic cardiomyopathy (HCM)	KEGG Pathway	20	-0.291193426	0.16371093	2.37E-05	6.61E-04	down
amine transmembrane transporter activity	GO Molecular Function	19	-0.281705811	0.17365387	7.28E-06	6.82E-04	down
amino acid transmembrane transporter activity	GO Molecular Function	17	-0.287002779	0.168030496	9.69E-06	6.82E-04	down
carboxylic acid transmembrane transporter activity	GO Molecular Function	23	-0.267218941	0.19001334	9.90E-06	6.82E-04	down
organic acid transmembrane transporter activity	GO Molecular Function	23	-0.267218941	0.19001334	9.90E-06	6.82E-04	down
regulation of response to external stimulus	GO Biological Process	39	-0.229046077	0.240885972	1.75E-05	6.94E-04	down
regulation of steroid hormone receptor signaling pathway	GO Biological Process	14	0.291146016	6.106528144	1.96E-05	7.60E-04	up
nucleolus	GO Cellular Component	510	-0.107006742	0.514270976	2.84E-05	7.66E-04	down
regulation of coagulation	GO Biological Process	15	-0.286795228	0.168247369	2.09E-05	7.90E-04	down
epithelium development	GO Biological Process	120	-0.167490521	0.353140973	2.57E-05	9.52E-04	down
synaptic vesicle	GO Cellular Component	21	0.283182191	5.811660546	4.04E-05	0.001005769	up

small molecule catabolic process	GO Biological Process	202	0.155892047	2.634801838	2.82E-05	0.001023778	up
epithelial cell differentiation	GO Biological Process	62	-0.199703165	0.289072746	3.06E-05	0.001087322	down
digestive tract development	GO Biological Process	13	-0.292960953	0.161922491	3.22E-05	0.001121726	down
positive regulation of response to external stimulus	GO Biological Process	15	-0.283062869	0.172195504	3.32E-05	0.001121726	down
DNA damage response, signal transduction resulting in induction of apoptosis	GO Biological Process	20	-0.26436069	0.193418685	3.35E-05	0.001121726	down
cellular component movement	GO Biological Process	207	-0.140701293	0.417110795	3.53E-05	0.001143973	down
response to insulin stimulus	GO Biological Process	58	0.213854624	3.777349886	3.61E-05	0.001143973	up
DNA damage response, signal transduction	GO Biological Process	62	-0.198563134	0.291128054	3.61E-05	0.001143973	down
odontogenesis	GO Biological Process	19	-0.266709467	0.190615911	3.75E-05	0.00116622	down
regulation of leukocyte migration	GO Biological Process	10	-0.309724692	0.14590237	3.94E-05	0.00120277	down
integrin-mediated signaling pathway	GO Biological Process	22	-0.256809824	0.202711314	4.11E-05	0.001227622	down
response to abiotic stimulus	GO Biological Process	150	-0.153998574	0.384027625	4.16E-05	0.001227622	down
endopeptidase inhibitor activity	GO Molecular Function	39	-0.228793316	0.241264654	2.53E-05	0.001307481	down
peptidase inhibitor activity	GO Molecular Function	39	-0.228793316	0.241264654	2.53E-05	0.001307481	down

calcium ion binding	GO Molecular Function	173	-0.151123984	0.390949714	2.86E-05	0.001311718	down
endopeptidase regulator activity	GO Molecular Function	42	-0.22223254	0.251304949	3.51E-05	0.001449737	down
muscle tissue development	GO Biological Process	54	-0.203123103	0.282993747	5.37E-05	0.001559506	down
myeloid leukocyte differentiation	GO Biological Process	23	-0.25145429	0.209571592	5.72E-05	0.001632447	down
ligase activity, forming carbon-sulfur bonds	GO Molecular Function	15	0.277862369	5.622664874	4.62E-05	0.001654697	up
GTPase regulator activity	GO Molecular Function	169	0.15729416	2.657860718	4.81E-05	0.001654697	up
cytokine receptor binding	GO Molecular Function	30	-0.238433926	0.227234352	5.42E-05	0.001721341	down
positive regulation of immune system process	GO Biological Process	65	-0.192335809	0.302615676	6.16E-05	0.001728755	down
serine-type endopeptidase activity	GO Molecular Function	30	0.239390463	4.426981223	6.17E-05	0.001820832	up
locomotion	GO Biological Process	174	-0.14375074	0.409280511	7.27E-05	0.002010322	down
response to glucocorticoid stimulus	GO Biological Process	30	-0.232485534	0.235791701	8.48E-05	0.002306027	down
cell-cell adhesion	GO Biological Process	69	-0.186692991	0.313416039	8.75E-05	0.002344295	down
membrane raft	GO Cellular Component	57	-0.197707637	0.292679981	1.03E-04	0.002379765	down
response to mechanical stimulus	GO Biological Process	26	-0.239653314	0.225518874	9.85E-05	0.002566604	down

	defense response	GO Biological Process	138	-0.152128305	0.388517217	9.88E-05	0.002566604	down
	cell-matrix adhesion	GO Biological Process	47	-0.205802014	0.27832136	1.02E-04	0.002615411	down
	blood coagulation	GO Biological Process	30	-0.230855721	0.238192089	1.04E-04	0.002615411	down
	negative regulation of developmental process	GO Biological Process	73	-0.182385385	0.321919524	1.06E-04	0.002641753	down
	RNA processing	GO Biological Process	455	-0.103244797	0.526435769	1.13E-04	0.002772715	down
	regulation of immune system process	GO Biological Process	113	-0.160335848	0.369197163	1.16E-04	0.002774582	down
	inflammatory response	GO Biological Process	65	-0.187648003	0.311561416	1.16E-04	0.002774582	down
	response to corticosteroid stimulus	GO Biological Process	32	-0.225994567	0.245497709	1.19E-04	0.002804658	down
	skin development	GO Biological Process	14	-0.276225418	0.17967013	1.23E-04	0.002825307	down
	regulation of cell adhesion	GO Biological Process	60	-0.191335208	0.304503307	1.24E-04	0.002825307	down
	digestive system development	GO Biological Process	15	-0.271577554	0.184935518	1.25E-04	0.002825307	down
	nucleoside-triphosphatase regulator activity	GO Molecular Function	174	0.15041931	2.546696583	1.04E-04	0.002873394	up
	negative regulation of cell differentiation	GO Biological Process	62	-0.189245266	0.308484043	1.30E-04	0.00290213	down
	coagulation	GO Biological Process	32	-0.224010223	0.248543911	1.51E-04	0.003335753	down

regulation of cell proliferation	GO Biological Process	268	-0.120567254	0.472707615	1.61E-04	0.003467759	down
cell fate specification	GO Biological Process	10	0.290446389	6.080035175	1.61E-04	0.003467759	up
regulation of small GTPase mediated signal transduction	GO Biological Process	102	0.173443575	2.938454747	1.86E-04	0.00392772	up
regulation of cellular catabolic process	GO Biological Process	109	0.170243293	2.880590618	1.87E-04	0.00392772	up
anchoring junction	GO Cellular Component	92	-0.16850933	0.350912124	1.85E-04	0.003993539	down
gland morphogenesis	GO Biological Process	30	-0.225274655	0.246598519	2.01E-04	0.004164271	down
locomotory behavior	GO Biological Process	59	-0.188210772	0.310473668	2.06E-04	0.004188142	down
multi-organism process	GO Biological Process	332	-0.110778619	0.502356274	2.07E-04	0.004188142	down
Focal adhesion	KEGG Pathway	82	-0.190324191	0.306422547	1.93E-04	0.004262825	down
Long-term depression	KEGG Pathway	23	0.308261348	6.791850705	2.04E-04	0.004262825	up
extracellular space	GO Cellular Component	117	-0.155485338	0.380495686	2.22E-04	0.004486302	down
GTP metabolic process	GO Biological Process	64	0.194051775	3.339949648	2.25E-04	0.00449432	up
protein maturation by peptide bond cleavage	GO Biological Process	21	-0.244764412	0.218468181	2.36E-04	0.004561255	down
purine ribonucleotide metabolic process	GO Biological Process	132	0.159348988	2.692019117	2.37E-04	0.004561255	up
response to drug	GO Biological Process	93	-0.164217179	0.360398321	2.40E-04	0.004561255	down

leukocyte differentiation	GO Biological Process	58	-0.187826432	0.311216126	2.40E-04	0.004561255	down
hemostasis	GO Biological Process	32	-0.220011486	0.254797756	2.41E-04	0.004561255	down
melanosome	GO Cellular Component	72	-0.1783283	0.330139332	2.54E-04	0.004578307	down
pigment granule	GO Cellular Component	72	-0.1783283	0.330139332	2.54E-04	0.004578307	down
negative regulation of cell death	GO Biological Process	179	-0.13400212	0.434842813	2.54E-04	0.004755984	down
regulation of ion homeostasis	GO Biological Process	11	-0.284610128	0.170547674	2.68E-04	0.004891163	down
GTP catabolic process	GO Biological Process	59	0.196263537	3.386175012	2.73E-04	0.004891163	up
regulation of GTP catabolic process	GO Biological Process	59	0.196263537	3.386175012	2.73E-04	0.004891163	up
regulation of GTPase activity	GO Biological Process	59	0.196263537	3.386175012	2.73E-04	0.004891163	up
Insulin signaling pathway	KEGG Pathway	61	0.23908232	4.41851171	2.80E-04	0.005190971	up
Notch signaling pathway	GO Biological Process	22	-0.239761341	0.225367524	2.99E-04	0.005320454	down
purine ribonucleoside triphosphate catabolic process	GO Biological Process	64	0.191144351	3.280143701	3.12E-04	0.005426972	up
ribonucleoside triphosphate catabolic process	GO Biological Process	64	0.191144351	3.280143701	3.12E-04	0.005426972	up
positive regulation of cytokine production	GO Biological Process	28	-0.225183969	0.246737535	3.16E-04	0.005443982	down
purine ribonucleotide catabolic process	GO Biological Process	67	0.188589806	3.228480961	3.24E-04	0.005536562	up
negative regulation of apoptosis	GO Biological Process	175	-0.133106628	0.437269518	3.28E-04	0.005545708	down
negative regulation of programmed cell death	GO Biological Process	176	-0.13266593	0.438468738	3.37E-04	0.005642173	down

nucleoside triphosphate catabolic process	GO Biological Process	67	0.188038907	3.217446762	3.45E-04	0.005716934	up
Amoebiasis	KEGG Pathway	33	-0.233609039	0.234151105	3.47E-04	0.005800083	down
regulation of nucleotide catabolic process	GO Biological Process	61	0.192187594	3.3014791	3.58E-04	0.005830017	up
regulation of purine nucleotide catabolic process	GO Biological Process	61	0.192187594	3.3014791	3.58E-04	0.005830017	up
purine ribonucleoside triphosphate metabolic process	GO Biological Process	117	0.161163599	2.722549085	3.69E-04	0.005941629	up
ribonucleoside triphosphate metabolic process	GO Biological Process	118	0.16062216	2.713403575	3.75E-04	0.005987296	up
desmosome	GO Cellular Component	15	-0.264810595	0.192878645	3.59E-04	0.0061145	down
striated muscle tissue development	GO Biological Process	50	-0.191102022	0.3049449	4.06E-04	0.006387462	down
purine nucleoside triphosphate catabolic process	GO Biological Process	66	0.187239846	3.20150901	4.07E-04	0.006387462	up
response to organic cyclic substance	GO Biological Process	53	-0.187846717	0.311176896	4.16E-04	0.006465618	down
ribonucleotide catabolic process	GO Biological Process	68	0.185490701	3.166896265	4.22E-04	0.006498914	up
ribonucleotide metabolic process	GO Biological Process	139	0.151798327	2.568615696	4.35E-04	0.006648819	up
COPI-coated vesicle	GO Cellular Component	14	0.285507605	5.89625776	4.28E-04	0.006926958	up
coated vesicle	GO Cellular Component	94	0.176494276	2.994696106	4.59E-04	0.007079173	up
small GTPase regulator activity	GO Molecular Function	123	0.15871416	2.681419465	2.75E-04	0.007092422	up
response to bacterium	GO Biological Process	43	-0.197664387	0.292758659	4.76E-04	0.007208594	down
regulation of developmental growth	GO Biological Process	17	0.25218934	4.793485997	4.83E-04	0.007244663	up

carboxypeptidase activity	GO Molecular Function	10	0.282350899	5.781714038	3.14E-04	0.007267494	up
GTPase activator activity	GO Molecular Function	95	0.16952289	2.867722972	3.17E-04	0.007267494	up
muscle organ development	GO Biological Process	69	-0.172533253	0.34224564	5.16E-04	0.007618296	down
tissue remodeling	GO Biological Process	18	-0.246465886	0.216170271	5.16E-04	0.007618296	down
muscle structure development	GO Biological Process	85	-0.162089478	0.365195452	5.30E-04	0.007756459	down
positive regulation of protein transport	GO Biological Process	23	-0.23115602	0.237747981	5.55E-04	0.008052488	down
response to steroid hormone stimulus	GO Biological Process	70	-0.170902867	0.345730972	5.73E-04	0.00818727	down
behavior	GO Biological Process	99	-0.154182708	0.383588426	5.74E-04	0.00818727	down
purine nucleoside triphosphate metabolic process	GO Biological Process	122	0.155208975	2.623640759	5.79E-04	0.008197591	up
leukocyte chemotaxis	GO Biological Process	10	-0.28236786	0.172940861	5.94E-04	0.008336561	down
response to inorganic substance	GO Biological Process	113	-0.147521786	0.399800309	6.12E-04	0.008364431	down
purine nucleotide catabolic process	GO Biological Process	72	0.179262513	3.046661086	6.14E-04	0.008364431	up
estrogen receptor signaling pathway	GO Biological Process	12	0.267488972	5.271627465	6.14E-04	0.008364431	up
nucleoside triphosphate metabolic process	GO Biological Process	127	0.152828879	2.585119126	6.15E-04	0.008364431	up
lipid transporter activity	GO Molecular Function	26	0.23074997	4.195534028	3.87E-04	0.008409404	up
anatomical structure formation involved in morphogenesis	GO Biological Process	143	-0.136763949	0.427442994	6.31E-04	0.008514637	down
regulation of calcium ion transport	GO Biological Process	16	-	0.209992437	6.57E-04	0.008798721	down

				0.251131485				
	regulation of Ras protein signal transduction	GO Biological Process	91	0.167454938	2.831104212	6.67E-04	0.008866121	up
	induction of apoptosis by intracellular signals	GO Biological Process	34	-0.207138359	0.276019511	6.72E-04	0.008866121	down
	protein processing	GO Biological Process	31	-0.211916542	0.267943737	6.95E-04	0.009096946	down
	regulation of peptidase activity	GO Biological Process	51	-0.184989955	0.316750756	7.10E-04	0.009226172	down
	di-, tri-valent inorganic cation transport	GO Biological Process	55	-0.1806911	0.325327025	7.42E-04	0.009573316	down
	negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	GO Biological Process	14	-0.257784735	0.201486861	7.63E-04	0.009765055	down
	phosphoric ester hydrolase activity	GO Molecular Function	142	0.14749272	2.500796935	4.82E-04	0.009951015	up
	heterocycle catabolic process	GO Biological Process	88	0.167451178	2.831038048	7.87E-04	0.010000106	up
	carboxylic acid catabolic process	GO Biological Process	65	0.181232724	3.084193989	8.26E-04	0.010350354	up
	organic acid catabolic process	GO Biological Process	65	0.181232724	3.084193989	8.26E-04	0.010350354	up
	female sex differentiation	GO Biological Process	26	-0.219894603	0.254982904	8.39E-04	0.010383843	down
	cell migration	GO Biological Process	148	-0.13291975	0.437777646	8.41E-04	0.010383843	down
	phospholipase activity	GO Molecular Function	25	0.22953199	4.163896782	5.40E-04	0.010579832	up
	extracellular matrix binding	GO Molecular Function	11	-0.279372493	0.17619031	5.64E-04	0.010579832	down
	negative regulation of cell growth	GO Biological Process	32	-0.207894907	0.27472481	8.71E-04	0.01067961	down
	regulation of catabolic process	GO Biological Process	126	0.149540341	2.532823318	9.11E-04	0.011091051	up
	organelle fusion	GO Biological Process	15	-	0.20964806	9.21E-04	0.011136855	down

			0.251395588				
generation of a signal involved in cell-cell signaling	GO Biological Process	40	0.203099588	3.533130635	9.49E-04	0.011312938	up
signal release	GO Biological Process	40	0.203099588	3.533130635	9.49E-04	0.011312938	up
hormone secretion	GO Biological Process	24	0.227770765	4.118570102	9.74E-04	0.011530315	up
ion transport	GO Biological Process	191	-0.120788702	0.472057517	0.001008941	0.011868685	down
response to lipopolysaccharide	GO Biological Process	32	-0.205452504	0.27892655	0.001104318	0.012854296	down
growth	GO Biological Process	188	-0.120632341	0.472516447	0.001112017	0.012854296	down
hair cycle process	GO Biological Process	13	-0.257936278	0.201297194	0.001129642	0.012854296	down
hair follicle development	GO Biological Process	13	-0.257936278	0.201297194	0.001129642	0.012854296	down
molting cycle process	GO Biological Process	13	-0.257936278	0.201297194	0.001129642	0.012854296	down
alcohol metabolic process	GO Biological Process	208	0.125067508	2.175471777	0.001190351	0.013431032	up
regulation of endothelial cell proliferation	GO Biological Process	15	-0.24841073	0.213573267	0.001195755	0.013431032	down
regulation of nitric oxide biosynthetic process	GO Biological Process	10	-0.273398843	0.182854109	0.001261061	0.01407376	down
skeletal system development	GO Biological Process	58	-0.172735782	0.341815148	0.001283446	0.014232348	down
hormone transport	GO Biological Process	25	0.222625204	3.988951428	0.001304397	0.014373135	up
regulation of cell activation	GO Biological Process	48	-0.181878497	0.322935205	0.001344249	0.014553961	down
hair cycle	GO Biological Process	14	-0.251257539	0.209827998	0.001345886	0.014553961	down
molting cycle	GO Biological Process	14	-0.251257539	0.209827998	0.001345886	0.014553961	down

renal system development	GO Biological Process	29	-0.208409785	0.273847161	0.001390925	0.0148361	down
cell motility	GO Biological Process	152	-0.127455467	0.452899162	0.001397542	0.0148361	down
localization of cell	GO Biological Process	152	-0.127455467	0.452899162	0.001397542	0.0148361	down
response to molecule of bacterial origin	GO Biological Process	34	-0.199526379	0.289390512	0.001415648	0.01493723	down
microtubule	GO Cellular Component	129	0.152174671	2.574630275	0.00106478	0.015681301	up
negative regulation of epithelial cell proliferation	GO Biological Process	10	0.267032585	5.256696922	0.001497269	0.015703281	up
response to radiation	GO Biological Process	85	-0.152452939	0.387734182	0.001534944	0.016002026	down
acid-thiol ligase activity	GO Molecular Function	10	0.271317325	5.398552411	9.07E-04	0.01629198	up
Salivary secretion	KEGG Pathway	23	0.284590194	5.862736799	0.001073633	0.016299695	up
purine nucleotide metabolic process	GO Biological Process	154	0.135076234	2.315083735	0.001615919	0.016745923	up
positive regulation of intracellular protein transport	GO Biological Process	15	-0.244756346	0.218479134	0.00162774	0.016768617	down
regulation of cytokine-mediated signaling pathway	GO Biological Process	10	-0.269967521	0.18679522	0.001653697	0.016935806	down
regulation of cytokine production	GO Biological Process	46	-0.181722141	0.323249151	0.001684987	0.017155331	down
protein binding, bridging	GO Molecular Function	39	-0.197357559	0.293317427	0.001054561	0.018147243	down
positive regulation of lymphocyte activation	GO Biological Process	28	-0.207041807	0.276185182	0.001871676	0.018945283	down
cholesterol transport	GO Biological Process	14	-0.247048603	0.215388856	0.001901266	0.01902359	down
sterol transport	GO Biological Process	14	-	0.215388856	0.001901266	0.01902359	down

			0.247048603				
receptor metabolic process	GO Biological Process	27	-0.208721052	0.273317941	0.001926667	0.019167582	down
envelope	GO Cellular Component	460	-0.087280621	0.581342697	0.001373952	0.019354801	down
regulation of fatty acid oxidation	GO Biological Process	15	0.243297235	4.535779569	0.001963146	0.019419533	up
fructose metabolic process	GO Biological Process	13	0.250414741	4.740911784	0.001974654	0.019423012	up
positive regulation of cell communication	GO Biological Process	134	-0.129526735	0.447106751	0.002005435	0.019614952	down
mesenchymal cell development	GO Biological Process	13	-0.250632568	0.210644544	0.002051702	0.019844517	down
mesenchymal cell differentiation	GO Biological Process	13	-0.250632568	0.210644544	0.002051702	0.019844517	down
nitric oxide biosynthetic process	GO Biological Process	14	-0.245884106	0.216953257	0.002086134	0.019955822	down
nitric oxide metabolic process	GO Biological Process	14	-0.245884106	0.216953257	0.002086134	0.019955822	down
Arachidonic acid metabolism	KEGG Pathway	14	0.314441683	7.057789017	0.001441625	0.020062617	up
regulation of Ras GTPase activity	GO Biological Process	53	0.180864552	3.077145275	0.002153499	0.020487662	up
actin filament bundle assembly	GO Biological Process	19	0.230340501	4.184871277	0.002176892	0.020597657	up
long-chain fatty acid transport	GO Biological Process	12	0.252983484	4.817201736	0.002233125	0.021015514	up
homophilic cell adhesion	GO Biological Process	18	-0.229582545	0.24008421	0.002317868	0.021695741	down
positive regulation of intracellular transport	GO Biological Process	17	-0.232827874	0.235290586	0.002337932	0.021766522	down
Chemokine signaling pathway	KEGG Pathway	56	0.219585713	3.914310323	0.001715845	0.022042015	up
cholesterol homeostasis	GO Biological Process	13	-0.248538785	0.213403371	0.002413839	0.022207997	down
sterol homeostasis	GO Biological Process	13	-0.248538785	0.213403371	0.002413839	0.022207997	down
anti-apoptosis	GO Biological Process	107	-0.137486204	0.425528702	0.002423618	0.022207997	down

regulation of hormone levels	GO Biological Process	57	0.175819119	2.982157163	0.002485668	0.022620983	up
response to organic substance	GO Biological Process	337	-0.092488427	0.562829056	0.002494675	0.022620983	down
regulation of nucleotide metabolic process	GO Biological Process	74	0.163481589	2.762052347	0.002519233	0.022725309	up
ligase activity	GO Molecular Function	240	0.115915133	2.055187584	0.001378775	0.022777363	up
cytokine production	GO Biological Process	56	-0.167246651	0.353676583	0.00257132	0.02307561	down
response to calcium ion	GO Biological Process	30	-0.199466925	0.289497457	0.002597399	0.023190115	down
leukocyte migration	GO Biological Process	21	-0.218966367	0.256458055	0.002664192	0.02366509	down
response to oxidative stress	GO Biological Process	96	-0.141286551	0.415596457	0.002689527	0.023768867	down
lipase activity	GO Molecular Function	31	0.208376195	3.650909782	0.001510264	0.023989965	up
elevation of cytosolic calcium ion concentration	GO Biological Process	14	-0.242254678	0.221902345	0.002764687	0.024309701	down
regulation of cell-cell adhesion	GO Biological Process	14	-0.241906454	0.222383078	0.00283871	0.024835149	down
endothelial cell migration	GO Biological Process	19	-0.223466535	0.249385114	0.002937875	0.025574199	down
organelle envelope	GO Cellular Component	458	-0.085068372	0.589390342	0.001907457	0.025750675	down
cytosolic calcium ion homeostasis	GO Biological Process	16	-0.233127709	0.234852564	0.003019634	0.026155142	down
ligase activity, forming carbon-nitrogen bonds	GO Molecular Function	147	0.134451104	2.306107229	0.001715634	0.026242844	up

Malaria	KEGG Pathway	10	-0.284673177	0.170480862	0.002213367	0.026402305	down
di-, tri-valent inorganic cation homeostasis	GO Biological Process	62	-0.160316073	0.369242539	0.003069952	0.026459342	down
iron ion transport	GO Biological Process	16	-0.232727693	0.235437121	0.003113109	0.026699126	down
peptide hormone secretion	GO Biological Process	21	0.220958091	3.947837422	0.003155455	0.026836121	up
skeletal muscle tissue development	GO Biological Process	32	-0.193650676	0.300152962	0.003159911	0.026836121	down
double-stranded DNA binding	GO Molecular Function	55	-0.17312878	0.340981342	0.001921888	0.027372498	down
transmembrane receptor protein kinase activity	GO Molecular Function	17	-0.237201127	0.228981963	0.002012815	0.027372498	down
acid-amino acid ligase activity	GO Molecular Function	125	0.139734002	2.383075935	0.002059718	0.027372498	up
serine-type peptidase activity	GO Molecular Function	44	0.188009308	3.216854989	0.002061547	0.027372498	up
enzyme inhibitor activity	GO Molecular Function	95	-0.145442514	0.404999995	0.002160473	0.027372498	down
structure-specific DNA binding	GO Molecular Function	86	-0.149955367	0.39379932	0.002187149	0.027372498	down
Gap junction	KEGG Pathway	27	0.260466941	5.046525375	0.002485372	0.027670475	up
cell chemotaxis	GO Biological Process	12	-0.24918138	0.212552849	0.00331875	0.028048268	down
fat cell differentiation	GO Biological Process	30	-0.196297256	0.295256587	0.003377264	0.028404911	down
regulation of metal ion transport	GO Biological Process	20	-0.2184978	0.257205939	0.003440627	0.028798711	down
response to gamma radiation	GO Biological Process	10	-0.259482419	0.199372261	0.003577765	0.029803295	down
serine hydrolase activity	GO Molecular Function	45	0.184940349	3.156083285	0.002466671	0.029962792	up
positive regulation of cellular component movement	GO Biological Process	46	-0.173055333	0.341137017	0.003659367	0.030337892	down

positive regulation of anti-apoptosis	GO Biological Process	16	-0.230271993	0.239057733	0.003742316	0.030740412	down
positive regulation of cell migration	GO Biological Process	41	-0.178635464	0.329509728	0.003743232	0.030740412	down
response to DNA damage stimulus	GO Biological Process	239	-0.100717813	0.534768287	0.003837123	0.031363529	down
myoblast differentiation	GO Biological Process	10	-0.257718861	0.201569363	0.004040665	0.03287289	down
regulation of homeostatic process	GO Biological Process	29	-0.195714911	0.296327069	0.004091968	0.032880685	down
positive regulation of cell activation	GO Biological Process	31	-0.192127763	0.303007189	0.004098282	0.032880685	down
positive regulation of leukocyte activation	GO Biological Process	31	-0.192127763	0.303007189	0.004098282	0.032880685	down
lipid particle	GO Cellular Component	13	-0.251064316	0.210080113	0.00256098	0.033190298	down
regulation of canonical Wnt receptor signaling pathway	GO Biological Process	12	-0.245973845	0.216832297	0.00416906	0.033295106	down
peptide secretion	GO Biological Process	22	0.215110076	3.806936582	0.004196337	0.033338881	up
regulation of cellular component movement	GO Biological Process	80	-0.144773426	0.406687538	0.00421284	0.033338881	down
extracellular matrix organization	GO Biological Process	30	-0.193452547	0.300522767	0.004237396	0.033381475	down
small molecule biosynthetic process	GO Biological Process	232	0.107946409	1.955888621	0.004366662	0.034224402	up
monocarboxylic acid metabolic process	GO Biological Process	160	0.123028707	2.148081669	0.004383711	0.034224402	up
response to peptide hormone stimulus	GO Biological Process	82	0.152223879	2.57541774	0.004441274	0.034519007	up
microtubule cytoskeleton	GO Cellular Component	292	0.105542594	1.92688722	0.00278381	0.034690558	up
regulation of membrane potential	GO Biological Process	41	-	0.334181259	0.004505507	0.034862608	down

			0.176370211				
negative regulation of cell size	GO Biological Process	36	-0.182986511	0.320719153	0.004533912	0.034927169	down
vesicle localization	GO Biological Process	25	0.207840872	3.638784023	0.004558708	0.034963481	up
appendage morphogenesis	GO Biological Process	29	-0.194082482	0.299348579	0.0046449	0.035313411	down
limb morphogenesis	GO Biological Process	29	-0.194082482	0.299348579	0.0046449	0.035313411	down
regulation of T cell proliferation	GO Biological Process	19	-0.217313977	0.25910518	0.004678078	0.035411013	down
cellular membrane fusion	GO Biological Process	33	-0.186606389	0.313584763	0.004874391	0.036737295	down
metal ion transport	GO Biological Process	104	-0.13127505	0.442275191	0.004948066	0.037131825	down
nucleobase, nucleoside and nucleotide metabolic process	GO Biological Process	248	0.103944965	1.907850516	0.005000387	0.037363405	up
regulation of blood pressure	GO Biological Process	20	0.217027807	3.852578896	0.005172007	0.038480613	up
gliogenesis	GO Biological Process	26	-0.198472261	0.291292511	0.005199928	0.03852372	down
cell-substrate junction	GO Cellular Component	57	-0.165586493	0.357344434	0.003301974	0.03856594	down
synapse part	GO Cellular Component	63	0.17617752	2.988806797	0.003332859	0.03856594	up
response to oxygen levels	GO Biological Process	59	-0.156428734	0.378271428	0.00527663	0.038926329	down
regulation of anti-apoptosis	GO Biological Process	22	-0.20722594	0.275869319	0.005355827	0.03934386	down
regulation of leukocyte proliferation	GO Biological Process	23	-0.204440563	0.280686194	0.005485694	0.039572681	down
regulation of lymphocyte proliferation	GO Biological Process	23	-0.204440563	0.280686194	0.005485694	0.039572681	down
regulation of mononuclear cell proliferation	GO Biological Process	23	-0.204440563	0.280686194	0.005485694	0.039572681	down

	regulation of ARF protein signal transduction	GO Biological Process	21	0.213860958	3.777498591	0.0054882	0.039572681	up
	ncRNA metabolic process	GO Biological Process	188	-0.106124374	0.517098762	0.005500625	0.039572681	down
	mesenchyme development	GO Biological Process	15	-0.228278809	0.242037323	0.005641974	0.040422539	down
	cellular response to stress	GO Biological Process	361	-0.083273175	0.596002659	0.00576057	0.041103082	down
	protein maturation	GO Biological Process	38	-0.17693479	0.333010792	0.005845726	0.041540443	down
	nucleoside phosphate metabolic process	GO Biological Process	232	0.104697547	1.916794425	0.005932026	0.041812375	up
	nucleotide metabolic process	GO Biological Process	232	0.104697547	1.916794425	0.005932026	0.041812375	up
	ureteric bud development	GO Biological Process	13	-0.235588165	0.231288803	0.006088609	0.042743014	down
	Adipocytokine signaling pathway	KEGG Pathway	30	0.245044885	4.585310975	0.004096861	0.042760989	up
	response to vitamin	GO Biological Process	33	-0.18357227	0.319553774	0.006137749	0.042914944	down
	extrinsic to plasma membrane	GO Cellular Component	22	-0.214320436	0.263970608	0.00389503	0.043297889	down
	apical part of cell	GO Cellular Component	66	-0.155863809	0.379601793	0.004110883	0.043297889	down
	coated vesicle membrane	GO Cellular Component	55	0.180577211	3.071655288	0.004142699	0.043297889	up
	fatty acid metabolic process	GO Biological Process	111	0.134514823	2.307020596	0.006284535	0.043661646	up
	negative regulation of growth	GO Biological Process	38	-0.175955513	0.335043616	0.006297196	0.043661646	down
	amine catabolic process	GO Biological Process	36	0.186074829	3.1784134	0.006319779	0.043661646	up
	activation of caspase activity	GO Biological Process	25	-0.197727526	0.292643807	0.006409542	0.044106768	down
	regulation of ossification	GO Biological Process	20	-0.209894139	0.271332624	0.006456167	0.044219874	down
	nucleotide catabolic process	GO Biological Process	83	0.147031129	2.493633404	0.006476777	0.044219874	up

	nucleobase, nucleoside and nucleotide catabolic process	GO Biological Process	85	0.145681413	2.472804391	0.006621545	0.044856456	up
	nucleobase, nucleoside, nucleotide and nucleic acid catabolic process	GO Biological Process	85	0.145681413	2.472804391	0.006621545	0.044856456	up
	hormone binding	GO Molecular Function	10	0.253621001	4.836324963	0.00399207	0.045170486	up
	protein complex binding	GO Molecular Function	103	-0.135214414	0.431579048	0.004055189	0.045170486	down
	kinase activity	GO Molecular Function	314	0.095596095	1.811385945	0.004142546	0.045170486	up
	microtubule motor activity	GO Molecular Function	27	0.203409882	3.539950336	0.004156122	0.045170486	up
	response to estrogen stimulus	GO Biological Process	37	-0.176226852	0.33447912	0.006832184	0.046104	down
	regulation of lymphocyte activation	GO Biological Process	41	-0.17092953	0.34567369	0.00688017	0.046248554	down
	regulation of ion transport	GO Biological Process	24	-0.198461931	0.291311213	0.007114239	0.047474742	down
	cellular response to insulin stimulus	GO Biological Process	45	0.173954952	2.947808031	0.007117121	0.047474742	up
	regulation of T cell activation	GO Biological Process	37	-0.175603679	0.335776995	0.007153968	0.047538389	down
	peptidyl-serine phosphorylation	GO Biological Process	21	-0.205599799	0.278671343	0.007200673	0.047564109	down
	development of primary female sexual characteristics	GO Biological Process	22	-0.202972895	0.28325804	0.007232602	0.047564109	down
	peptide transport	GO Biological Process	25	0.20156463	3.499587773	0.0072834	0.047564109	up
	mRNA processing	GO Biological Process	251	-0.092907929	0.561363647	0.007303132	0.047564109	down
	appendage development	GO Biological Process	30	-0.186139204	0.314496539	0.007321758	0.047564109	down
	limb development	GO Biological Process	30	-0.186139204	0.314496539	0.007321758	0.047564109	down

positive regulation of signaling pathway	GO Biological Process	125	-0.11900698	0.477313514	0.007402823	0.047911953	down
heterocycle metabolic process	GO Biological Process	238	0.101201214	1.875594895	0.007465836	0.048140816	up
regulation of protein kinase B signaling cascade	GO Biological Process	17	-0.216558714	0.26032419	0.007639796	0.049080757	down

Table S 15: Concordance-based LRPath Gene Ontology and pathway analysis results with FDR <= 0.05

References:

1. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
2. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
3. Rothberg, J.M., et al., *An integrated semiconductor device enabling non-optical genome sequencing.* Nature, 2011. **475**(7356): p. 348-52.
4. Levene, M.J., et al., *Zero-mode waveguides for single-molecule analysis at high concentrations.* Science, 2003. **299**(5607): p. 682-6.
5. Korlach, J., et al., *Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures.* Proc Natl Acad Sci U S A, 2008. **105**(4): p. 1176-81.
6. Parkinson, N.J., et al., *Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA.* Genome Res, 2012. **22**(1): p. 125-33.
7. Yamamoto, M., et al., *Use of serial analysis of gene expression (SAGE) technology.* J Immunol Methods, 2001. **250**(1-2): p. 45-66.
8. Brenner, S., et al., *Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.* Nat Biotechnol, 2000. **18**(6): p. 630-4.
9. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.
10. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
11. Xu, X., et al., *Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets.* BMC Bioinformatics, 2013. **14 Suppl 9**: p. S1.
12. Pushkarev, D., N.F. Neff, and S.R. Quake, *Single-molecule sequencing of an individual human genome.* Nat Biotechnol, 2009. **27**(9): p. 847-50.
13. Ideker, T., et al., *Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.* Science, 2001. **292**(5518): p. 929-34.
14. Rives, A.W. and T. Galitski, *Modular organization of cellular networks.* Proc Natl Acad Sci U S A, 2003. **100**(3): p. 1128-33.
15. Petti, A.A. and G.M. Church, *A network of transcriptionally coordinated functional modules in Saccharomyces cerevisiae.* Genome Res, 2005. **15**(9): p. 1298-306.
16. Sam, L., et al., *Discovery of protein interaction networks shared by diseases.* Pac Symp Biocomput, 2007: p. 76-87.
17. Lage, K., et al., *A human phenome-interactome network of protein complexes implicated in genetic disorders.* Nat Biotechnol, 2007. **25**(3): p. 309-16.

18. Xu, J. and Y. Li, *Discovering disease-genes by topological features in human protein-protein interaction network.* Bioinformatics, 2006. **22**(22): p. 2800-5.
19. Goh, K.I., et al., *The human disease network.* Proc Natl Acad Sci U S A, 2007. **104**(21): p. 8685-90.
20. Krauthammer, M., et al., *Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease.* Proc Natl Acad Sci U S A, 2004. **101**(42): p. 15148-53.
21. He, X. and J. Zhang, *Why do hubs tend to be essential in protein networks?* PLoS Genet, 2006. **2**(6): p. e88.
22. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-2.
23. Yu, H., et al., *The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.* PLoS Comput Biol, 2007. **3**(4): p. e59.
24. Said, M.R., et al., *Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae.* Proc Natl Acad Sci U S A, 2004. **101**(52): p. 18006-11.
25. Tuck, D.P., H.M. Kluger, and Y. Kluger, *Characterizing disease states from topological properties of transcriptional regulatory networks.* BMC Bioinformatics, 2006. **7**: p. 236.
26. Dijkstra, E.W., *A note on two problems in connection with graphs.* Numerische Mathematik, 1959(1): p. 83–89.
27. Feldman, I., A. Rzhetsky, and D. Vitkup, *Network properties of genes harboring inherited disease mutations.* Proc Natl Acad Sci U S A, 2008. **105**(11): p. 4323-8.
28. Semon, M., D. Mouchiroud, and L. Duret, *Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance.* Hum Mol Genet, 2005. **14**(3): p. 421-7.
29. Berneburg, M. and A.R. Lehmann, *Xeroderma pigmentosum and related disorders: defects in DNA repair and transcription.* Adv Genet, 2001. **43**: p. 71-102.
30. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.
31. Sunohara, T., et al., *Ribosome stalling during translation elongation induces cleavage of mRNA being translated in Escherichia coli.* J Biol Chem, 2004. **279**(15): p. 15368-75.
32. Baker, K.E. and R. Parker, *Nonsense-mediated mRNA decay: terminating erroneous gene expression.* Curr Opin Cell Biol, 2004. **16**(3): p. 293-9.
33. Zeng, Y., R. Yi, and B.R. Cullen, *MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms.* Proc Natl Acad Sci U S A, 2003. **100**(17): p. 9779-84.
34. Wilkinson, K.D., *Ubiquitination and deubiquitination: targeting of proteins for degradation by the proteasome.* Semin Cell Dev Biol, 2000. **11**(3): p. 141-8.
35. Anderson, L. and J. Seilhamer, *A comparison of selected mRNA and protein abundances in human liver.* Electrophoresis, 1997. **18**(3-4): p. 533-7.
36. Lichtinghagen, R., et al., *Different mRNA and protein expression of matrix metalloproteinases 2 and 9 and tissue inhibitor of metalloproteinases 1 in benign and malignant prostate tissue.* Eur Urol, 2002. **42**(4): p. 398-406.
37. Bruce, C., et al., *Proteomics and the analysis of proteomic data: 2013 overview of current protein-profiling technologies.* Curr Protoc Bioinformatics, 2013. **Chapter 13**: p. Unit 13 21.

38. Matthiesen, R. and A.S. Carvalho, *Methods and algorithms for quantitative proteomics by mass spectrometry.* Methods Mol Biol, 2013. **1007**: p. 183-217.

39. Mirza, S.P. and M. Olivier, *Methods and approaches for the comprehensive characterization and quantification of cellular proteomes using mass spectrometry.* Physiol Genomics, 2008. **33**(1): p. 3-11.

40. Nagaraj, N., et al., *Deep proteome and transcriptome mapping of a human cancer cell line.* Mol Syst Biol, 2011. **7**: p. 548.

41. Chen, G., et al., *Discordant protein and mRNA expression in lung adenocarcinomas.* Mol Cell Proteomics, 2002. **1**(4): p. 304-13.

42. Cox, B., T. Kislinger, and A. Emili, *Integrating gene and protein expression data: pattern analysis and profile mining.* Methods, 2005. **35**(3): p. 303-14.

43. Gry, M., et al., *Correlations between RNA and protein expression profiles in 23 human cell lines.* BMC Genomics, 2009. **10**: p. 365.

44. Shankavaram, U.T., et al., *Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study.* Mol Cancer Ther, 2007. **6**(3): p. 820-32.

45. Ghazalpour, A., et al., *Comparative analysis of proteome and transcriptome variation in mouse.* PLoS Genet, 2011. **7**(6): p. e1001393.

46. Akan, P., et al., *Comprehensive analysis of the genome transcriptome and proteome landscapes of three tumor cell lines.* Genome Med, 2012. **4**(11): p. 86.

47. Waters, K.M., J.G. Pounds, and B.D. Thrall, *Data merging for integrated microarray and proteomic analysis.* Brief Funct Genomic Proteomic, 2006. **5**(4): p. 261-72.

48. Nie, L., et al., *Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications.* Crit Rev Biotechnol, 2007. **27**(2): p. 63-75.

49. Hegde, P.S., I.R. White, and C. Debouck, *Interplay of transcriptomics and proteomics.* Curr Opin Biotechnol, 2003. **14**(6): p. 647-51.

50. Kvam, V.M., P. Liu, and Y. Si, *A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data.* Am J Bot, 2012. **99**(2): p. 248-56.

51. Garber, M., et al., *Computational methods for transcriptome annotation and quantification using RNA-seq.* Nat Methods, 2011. **8**(6): p. 469-77.

52. Dillies, M.A., et al., *A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.* Brief Bioinform, 2013. **14**(6): p. 671-83.

53. Nahnsen, S., et al., *Tools for label-free peptide quantification.* Mol Cell Proteomics, 2013. **12**(3): p. 549-56.

54. DeSouza, L.V. and K.W. Siu, *Mass spectrometry-based quantification.* Clin Biochem, 2013. **46**(6): p. 421-31.

55. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.

56. Pedrioli, P.G., *Trans-proteomic pipeline: a pipeline for proteomic analysis.* Methods Mol Biol, 2010. **604**: p. 213-38.

57. Fermin, D., et al., *Abacus: a computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis.* Proteomics, 2011. **11**(7): p. 1340-5.

58. Vogel, C. and E.M. Marcotte, *Insights into the regulation of protein abundance from proteomic and transcriptomic analyses.* Nat Rev Genet, 2012. **13**(4): p. 227-32.

59. Vogel, C., et al., *Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line.* Mol Syst Biol, 2010. **6**: p. 400.

60. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-42.

61. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.

62. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.

63. Agoston, A.T., et al., *Increased protein stability causes DNA methyltransferase 1 dysregulation in breast cancer.* J Biol Chem, 2005. **280**(18): p. 18302-10.

64. Lu, X. and Y. Kang, *Hypoxia and hypoxia-inducible factors: master regulators of metastasis.* Clin Cancer Res, 2010. **16**(24): p. 5928-35.

65. Wan, M., et al., *Yin Yang 1 plays an essential role in breast cancer and negatively regulates p27.* Am J Pathol, 2012. **180**(5): p. 2120-33.

66. Wu, S., et al., *Transcription factor YY1 contributes to tumor growth by stabilizing hypoxia factor HIF-1alpha in a p53-independent manner.* Cancer Res, 2013. **73**(6): p. 1787-99.

67. Curtis, D.J. and M.P. McCormack, *The molecular basis of Lmo2-induced T-cell acute lymphoblastic leukemia.* Clin Cancer Res, 2010. **16**(23): p. 5618-23.

68. Aplan, P.D., et al., *Involvement of the putative hematopoietic transcription factor SCL in T-cell acute lymphoblastic leukemia.* Blood, 1992. **79**(5): p. 1327-33.

69. Lecuyer, E., et al., *Protein stability and transcription factor complex assembly determined by the SCL-LMO2 interaction.* J Biol Chem, 2007. **282**(46): p. 33649-58.

70. Maher, C.A., et al., *Transcriptome sequencing to detect gene fusions in cancer.* Nature, 2009. **458**(7234): p. 97-101.

71. Kim, J.H., et al., *Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer.* Genome Res, 2011. **21**(7): p. 1028-41.

72. Grasso, C.S., et al., *The mutational landscape of lethal castration-resistant prostate cancer.* Nature, 2012. **487**(7406): p. 239-43.

73. Roychowdhury, S., et al., *Personalized oncology through integrative high-throughput sequencing: a pilot study.* Sci Transl Med, 2011. **3**(111): p. 111ra121.

74. Sultan, M., et al., *A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.* Science, 2008. **321**(5891): p. 956-60.

75. Li, H., et al., *Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model.* Proc Natl Acad Sci U S A, 2008. **105**(51): p. 20179-84.

76. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-5.

77. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

78. Au, K.F., et al., *Detection of splice junctions from paired-end RNA-seq data by SpliceMap.* Nucleic Acids Res, 2010.

79. Bryant, D.W., Jr., et al., *Supersplat--spliced RNA-seq alignment.* Bioinformatics, 2010. **26**(12): p. 1500-5.

80. Morin, R., et al., *Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing.* Biotechniques, 2008. **45**(1): p. 81-94.

81. Shah, S.P., et al., *Mutation of FOXL2 in granulosa-cell tumors of the ovary.* N Engl J Med, 2009. **360**(26): p. 2719-29.

82. Berger, M.F., et al., *Integrative analysis of the melanoma transcriptome.* Genome Res.

83. Tuch, B.B., et al., *Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations.* PLoS One. **5**(2): p. e9317.

84. Palacios, G., et al., *A new arenavirus in a cluster of fatal transplant-associated diseases.* N Engl J Med, 2008. **358**(10): p. 991-8.

85. Briese, T., et al., *Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa.* PLoS Pathog, 2009. **5**(5): p. e1000455.

86. Nakamura, S., et al., *Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach.* PLoS One, 2009. **4**(1): p. e4219.

87. Maher, C.A., et al., *Chimeric transcript discovery by paired-end transcriptome sequencing.* Proc Natl Acad Sci U S A, 2009. **106**(30): p. 12353-8.

88. Palanisamy, N., et al., *Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma.* Nat Med, 2010. **16**(7): p. 793-8.

89. Ozsolak, F., et al., *Direct RNA sequencing.* Nature, 2009. **461**(7265): p. 814-8.

90. Kozarewa, I., et al., *Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes.* Nat Methods, 2009. **6**(4): p. 291-5.

91. Dohm, J.C., et al., *Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.* Nucleic Acids Res, 2008. **36**(16): p. e105.

92. Campbell, P.J., et al., *Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.* Nat Genet, 2008. **40**(6): p. 722-9.

93. Ng, S.B., et al., *Targeted capture and massively parallel sequencing of 12 human exomes.* Nature, 2009. **461**(7261): p. 272-6.

94. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

95. Lipson, D., et al., *Quantification of the yeast transcriptome by single-molecule sequencing.* Nat Biotechnol, 2009. **27**(7): p. 652-8.

96. Ramskold, D., et al., *An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data.* PLoS Comput Biol, 2009. **5**(12): p. e1000598.

97. Bruford, E.A., et al., *The HGNC Database in 2008: a resource for the human genome.* Nucleic Acids Res, 2008. **36**(Database issue): p. D445-8.

98. Mamanova, L., et al., *FRT-seq: amplification-free, strand-specific transcriptome sequencing.* Nat Methods. **7**(2): p. 130-2.

99. Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. **310**(5748): p. 644-8.

100. Korenchuk, S., et al., *VCaP, a cell-based model system of human prostate cancer.* In Vivo, 2001. **15**(2): p. 163-8.

101. R Development Core Team, *R: A Language and Environment for Statistical Computing*. 2009, R Foundation for Statistical Computing: Vienna, Austria.

102. Carmona-Saez, P., et al., *GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists.* Genome Biol, 2007. **8**(1): p. R3.
103. Vogel, C. and E.M. Marcotte, *Insights into the regulation of protein abundance from proteomic and transcriptomic analyses.* Nat Rev Genet. **13**(4): p. 227-32.
104. Maier, T., M. Guell, and L. Serrano, *Correlation of mRNA and protein in complex biological samples.* FEBS Lett, 2009. **583**(24): p. 3966-73.
105. Bello, D., et al., *Androgen responsive adult human prostatic epithelial cell lines immortalized by human papillomavirus 18.* Carcinogenesis, 1997. **18**(6): p. 1215-23.
106. Fagan, A., A.C. Culhane, and D.G. Higgins, *A multivariate analysis approach to the integration of proteomic and gene expression data.* Proteomics, 2007. **7**(13): p. 2162-71.
107. Greenbaum, D., et al., *Comparing protein abundance and mRNA expression levels on a genomic scale.* Genome Biol, 2003. **4**(9): p. 117.
108. Wu, L., et al., *Global survey of human T leukemic cells by integrating proteomics and transcriptomics profiling.* Mol Cell Proteomics, 2007. **6**(8): p. 1343-53.
109. Guo, Y., et al., *How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes.* Acta Biochim Biophys Sin (Shanghai), 2008. **40**(5): p. 426-36.
110. Lundberg, E., et al., *Defining the transcriptome and proteome in three functionally different human cell lines.* Mol Syst Biol, 2010. **6**: p. 450.
111. Moghaddas Gholami, A., et al., *Global Proteome Analysis of the NCI-60 Cell Line Panel.* Cell Rep, 2013. **4**(3): p. 609-20.
112. Keller, A., et al., *A uniform proteomics MS/MS analysis platform utilizing open XML file formats.* Mol Syst Biol, 2005. **1**: p. 2005 0017.
113. Hebenstreit, D., et al., *RNA sequencing reveals two major classes of gene expression levels in metazoan cells.* Mol Syst Biol, 2011. **7**: p. 497.
114. Tomlins, S.A., et al., *Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer.* Nature, 2007. **448**(7153): p. 595-9.
115. Yocum, A.K., et al., *Development of selected reaction monitoring-MS methodology to measure peptide biomarkers in prostate cancer.* Proteomics, 2010. **10**(19): p. 3506-14.
116. Kessner, D., et al., *ProteoWizard: open source software for rapid proteomics tools development.* Bioinformatics, 2008. **24**(21): p. 2534-6.
117. Keller, A., et al., *A uniform proteomics MS/MS analysis platform utilizing open XML file formats.* Molecular systems biology, 2005. **1**: p. 2005 0017.
118. Fenyo, D. and R.C. Beavis, *A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes.* Anal Chem, 2003. **75**(4): p. 768-74.
119. Nesvizhskii, A.I., et al., *A statistical model for identifying proteins by tandem mass spectrometry.* Anal Chem, 2003. **75**(17): p. 4646-58.
120. Keller, A., et al., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.* Anal Chem, 2002. **74**(20): p. 5383-92.
121. Zhang, Y., et al., *Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins.* Anal Chem, 2010. **82**(6): p. 2272-81.

122. Siegel, R., D. Naishadham, and A. Jemal, *Cancer statistics, 2012.* CA Cancer J Clin, 2012. **62**(1): p. 10-29.

123. Pascal, L.E., et al., *Correlation of mRNA and protein levels: cell type-specific gene expression of cluster designation antigens in the prostate.* BMC Genomics, 2008. **9**: p. 246.

124. Nordengren, J., et al., *Discordant expression of mRNA and protein for urokinase and tissue plasminogen activators (u-PA, t-PA) in endometrial carcinoma.* Int J Cancer, 1998. **79**(2): p. 195-201.

125. Suvannasankha, A., et al., *Breast cancer resistance protein (BCRP/MXR/ABCG2) in acute myeloid leukemia: discordance between expression and function.* Leukemia, 2004. **18**(7): p. 1252-7.

126. Ropponen, K.M., et al., *Expression of transcription factor AP-2 in colorectal adenomas and adenocarcinomas; comparison of immunohistochemistry and in situ hybridisation.* J Clin Pathol, 2001. **54**(7): p. 533-8.

127. Keller, A., et al., *Experimental protein mixture for validating tandem mass spectral analysis.* OMICS, 2002. **6**(2): p. 207-12.

128. Cordwell, S.J. and T.E. Thingholm, *Technologies for plasma membrane proteomics.* Proteomics, 2010. **10**(4): p. 611-27.

129. Vuckovic, D., et al., *Membrane proteomics by high performance liquid chromatography-tandem mass spectrometry: Analytical approaches and challenges.* Proteomics, 2013. **13**(3-4): p. 404-23.

130. Beck, M., et al., *The quantitative proteome of a human cell line.* Mol Syst Biol, 2011. **7**: p. 549.

131. Gray, N.K. and M.W. Hentze, *Regulation of protein synthesis by mRNA structure.* Mol Biol Rep, 1994. **19**(3): p. 195-200.

132. Gedeon, T. and P. Bokes, *Delayed protein synthesis reduces the correlation between mRNA and protein fluctuations.* Biophys J, 2012. **103**(3): p. 377-85.

133. Lapenna, S. and A. Giordano, *Cell cycle kinases as therapeutic targets for cancer.* Nat Rev Drug Discov, 2009. **8**(7): p. 547-66.

134. Kolfschoten, I.G., et al., *A genetic screen identifies PITX1 as a suppressor of RAS activity and tumorigenicity.* Cell, 2005. **121**(6): p. 849-58.

135. Belandia, B., et al., *Hey1, a mediator of notch signaling, is an androgen receptor corepressor.* Mol Cell Biol, 2005. **25**(4): p. 1425-36.

136. Ding, Z., et al., *SMAD4-dependent barrier constrains prostate cancer growth and metastatic progression.* Nature, 2011. **470**(7333): p. 269-73.

137. Wisniewski, J.R., et al., *Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma.* Mol Syst Biol, 2012. **8**: p. 611.

138. Ostlund, G. and E.L. Sonnhammer, *Quality criteria for finding genes with high mRNA-protein expression correlation and coexpression correlation.* Gene, 2012. **497**(2): p. 228-36.

139. Sartor, M.A., G.D. Leikauf, and M. Medvedovic, *LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data.* Bioinformatics, 2009. **25**(2): p. 211-7.

140. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.

141. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.

142. Fraser, S.P., J.A. Grimes, and M.B. Djamgoz, *Effects of voltage-gated ion channel modulators on rat prostatic cancer cell proliferation: comparison of strongly and weakly metastatic cell lines.* Prostate, 2000. **44**(1): p. 61-76.

143. Abdul, M. and N. Hoosein, *Expression and activity of potassium ion channels in human prostate cancer.* Cancer Lett, 2002. **186**(1): p. 99-105.

144. Sikes, R.A., et al., *Therapeutic approaches targeting prostate cancer progression using novel voltage-gated ion channel blockers.* Clin Prostate Cancer, 2003. **2**(3): p. 181-7.

145. Sheng, T., et al., *Activation of the hedgehog pathway in advanced prostate cancer.* Mol Cancer, 2004. **3**: p. 29.

146. Karhadkar, S.S., et al., *Hedgehog signalling in prostate regeneration, neoplasia and metastasis.* Nature, 2004. **431**(7009): p. 707-12.

147. Karlou, M., et al., *Hedgehog signaling inhibition by the small molecule smoothened inhibitor GDC-0449 in the bone forming prostate cancer xenograft MDA PCa 118b.* Prostate, 2012. **72**(15): p. 1638-47.

148. Kon, S., et al., *Smap1 deficiency perturbs receptor trafficking and predisposes mice to myelodysplasia.* J Clin Invest, 2013. **123**(3): p. 1123-37.

149. Thome, M., *Multifunctional roles for MALT1 in T-cell activation.* Nat Rev Immunol, 2008. **8**(7): p. 495-500.

150. Noh, K.H., et al., *Activation of Akt as a mechanism for tumor immune evasion.* Mol Ther, 2009. **17**(3): p. 439-47.

151. Quilliam, L.A., J.F. Rebhun, and A.F. Castro, *A growing family of guanine nucleotide exchange factors is responsible for activation of Ras-family GTPases.* Prog Nucleic Acid Res Mol Biol, 2002. **71**: p. 391-444.

152. Aksamitiene, E., A. Kiyatkin, and B.N. Kholodenko, *Cross-talk between mitogenic Ras/MAPK and survival PI3K/Akt pathways: a fine balance.* Biochem Soc Trans, 2012. **40**(1): p. 139-46.

153. Musicco, M., et al., *Inverse occurrence of cancer and Alzheimer disease: a population-based incidence study.* Neurology, 2013. **81**(4): p. 322-8.

154. Hebron, M.L., I. Lonskaya, and C.E. Moussa, *Nilotinib reverses loss of dopamine neurons and improves motor behavior via autophagic degradation of alpha-synuclein in Parkinson's disease models.* Hum Mol Genet, 2013. **22**(16): p. 3315-28.

155. Reimers, M. and V.J. Carey, *Bioconductor: an open source framework for bioinformatics and computational biology.* Methods Enzymol, 2006. **411**: p. 119-34.

156. Jiang, H. and W.H. Wong, *Statistical inferences for isoform expression in RNA-Seq.* Bioinformatics, 2009. **25**(8): p. 1026-32.

157. Wang, L., et al., *DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.* Bioinformatics, 2010. **26**(1): p. 136-8.

158. Xia, Z., et al., *NSMAP: a method for spliced isoforms identification and quantification from RNA-Seq.* BMC Bioinformatics, 2011. **12**: p. 162.

159. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
160. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.
161. van de Vijver, M.J., et al., *A gene-expression signature as a predictor of survival in breast cancer.* N Engl J Med, 2002. **347**(25): p. 1999-2009.
162. Chen, H.Y., et al., *A five-gene signature and clinical outcome in non-small-cell lung cancer.* N Engl J Med, 2007. **356**(1): p. 11-20.
163. Sotiriou, C. and L. Pusztai, *Gene-expression signatures in breast cancer.* N Engl J Med, 2009. **360**(8): p. 790-800.
164. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-6.
165. Carlson, J.J. and J.A. Roth, *The impact of the Oncotype Dx breast cancer assay in clinical practice: a systematic review and meta-analysis.* Breast Cancer Res Treat, 2013. **141**(1): p. 13-22.
166. Lee, U., et al., *A prognostic gene signature for metastasis-free survival of triple negative breast cancer patients.* PLoS One, 2013. **8**(12): p. e82125.
167. Chen, R., et al., *Personal omics profiling reveals dynamic molecular and medical phenotypes.* Cell, 2012. **148**(6): p. 1293-307.
168. Jayapandian, M., et al., *Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together.* Nucleic Acids Res, 2007. **35**(Database issue): p. D566-71.
169. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2002. **30**(1): p. 52-5.
170. Lussier, Y., et al., *PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing.* Pac Symp Biocomput, 2006: p. 64-75.
171. Eyre, T.A., et al., *The HUGO Gene Nomenclature Database, 2006 updates.* Nucleic Acids Res, 2006. **34**(Database issue): p. D319-21.
172. National Library of Medicine. *Medical Subject Headings (MeSH®) Fact Sheet*. 1999 27 May 2005; Available from: http://www.nlm.nih.gov/pubs/factsheets/mesh.html.
173. National Library of Medicine. *Unified Medical Language System® Fact Sheet*. 2006 23 March 2006; Available from: http://www.nlm.nih.gov/pubs/factsheets/umls.html.
174. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
175. Kasprzyk, A., et al., *EnsMart: a generic system for fast and flexible access to biological data.* Genome Res, 2004. **14**(1): p. 160-9.
176. Witten, I.H. and E. Frank, *Data Mining: Practical machine learning tools and techniques*. 2nd ed. 2005, San Francisco: Morgan Kaufmann.
177. Hammer, Ø., D.A.T. Harper, and P.D. Ryan, *PAST: Paleontological Statistics Software Package for Education and Data Analysis.* Palaeontologia Electronica, 2001. **4**(1): p. 9.