COLON CANCER AND ITS MOLECULAR SUBSYSTEMS: NETWORK APPROACHES TO DISSECTING DRIVER GENE BIOLOGY

by

VISHAL N. PATEL

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Department of Genetics
CASE WESTERN RESERVE UNIVERSITY

August 2011
We hereby approve the dissertation* of Vishal N. Patel, candidate for

the degree of Doctor of Philosophy on July 6, 2011.

Committee Chair: Georgia Wiesner

Mark R. Chance

Sudha Iyengar

Mehmet Koyuturk

* We also certify that written approval has been obtained for any proprietary material contained therein
Table of Contents

I. List of Tables 6

II. List of Figures 7

III. List of Abbreviations 8

IV. Glossary 10

V. Abstract 14

VI. Colon Cancer and its Molecular Subsystems 15
   Construction, Interpretation, and Validation
   a. Colon Cancer
      i. Etiology
      ii. Development
      iii. The Pathway Paradigm
      iv. Cancer Subtypes and Therapies
   b. Molecular Subsystems
      i. Introduction
      ii. Construction
      iii. Interpretation
      iv. Validation
   c. Summary

VII. Prostaglandin dehydrogenase signaling 32
    One driver and an unknown path
    a. Introduction
       i. Colon Cancer
       ii. Mass spectrometry
b. Methods

c. Results

d. Discussion

e. Summary

VIII. *Apc* signaling with Candidate Drivers

*Two drivers, many paths*

a. Introduction

b. Methods

c. Results and Discussion

d. Summary

IX. *Apc-Cdkn1a* signaling

*Two drivers, one path*

a. Introduction

b. Methods

c. Results

   i. Driver Gene Network Prediction

   ii. Single Node Perturbations

      1. mRNA profiling

      2. Proteomic profiling

d. Discussion

e. Summary

X. Molecular Subsystems in Cancer

*The Present to the Future*

XI. Appendix I

*Prostaglandin dehydrogenase associated data*

XII. Appendix II

*Apc and Cdkn1a associated data*
XIII. Appendix III 128

*Ap* *c and *Cdkn1a* additional methods

XIV. References 132
List of Tables

1. Examples of molecular subsystems 27
2. Published evidence of interactions in the Apc-Cdkn1a 100
## List of Figures

1. Diagram of colorectal cancer development  
2. Pathways of Tp53  
3. Manifolds used in cancer biology  
4. Peptides to networks  
5. Bivariate t-distribution  
6. Mixing angle shrinkage  
7. Hpgd data processing flowchart  
8. Genomic width of proteomic data  
9. False discovery proportion for high-throughput validation  
10. False discovery proportion for individual validation  
11. Hsd11b2 and Selentr1: alternative splicing candidates  
12. Coexpression heatmap  
13. HNF4A binding to proteomic targets  
14. Mechanism by which Hpgd exerts its influence on both inflammation and neoplasia  
15. PETALS flowchart  
16. Blossom Algorithm  
17. Calculating bimodality of coexpression  
18. Apc blossom  
19. Top 3 petals highly coexpressed with Apc proteomic targets  
20. Hapl1 petal  
21. Flowchart for mining pathways from PPI data  
22. Apc-Cdkn1a pathway  
23. Differential expression of the Apc-Cdkn1a network  
24. Proteomic interactions of the Apc-Cdkn1a network  
25. Proteomic coexpression of the Apc-Cdkn1a network
<table>
<thead>
<tr>
<th>Acronym or Abbreviation</th>
<th>Full Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D-DIGE</td>
<td>Two Dimensional In Gel Electrophoresis</td>
</tr>
<tr>
<td>AP/MS</td>
<td>Affinity Purification-Mass Spectrometry</td>
</tr>
<tr>
<td>BioGRID</td>
<td>Biological General Repository for Interaction Datasets</td>
</tr>
<tr>
<td>CAN</td>
<td>Candidate cancer driver (gene)</td>
</tr>
<tr>
<td>CGH</td>
<td>Comparative Genomic Hybridization</td>
</tr>
<tr>
<td>ChIP-chip</td>
<td>Chromatin Immunoprecipitation chip</td>
</tr>
<tr>
<td>COX</td>
<td>Cyclooxygenase</td>
</tr>
<tr>
<td>DIGE</td>
<td>See “2D-DIGE”</td>
</tr>
<tr>
<td>DTT</td>
<td>Dithiothreitol</td>
</tr>
<tr>
<td>FAP</td>
<td>Familial Adenomatous Polyposis coli</td>
</tr>
<tr>
<td>FDP</td>
<td>False Discovery Proportion</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>FWER</td>
<td>Family-wise Error Rate</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>GWA</td>
<td>Genome-wide Association (study)</td>
</tr>
<tr>
<td>HPRD</td>
<td>Human Protein Reference Database</td>
</tr>
<tr>
<td>IBD</td>
<td>Inflammatory bowel disease</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>KO</td>
<td>Knock-out</td>
</tr>
<tr>
<td>KS</td>
<td>Kolmogorov-Smirnov</td>
</tr>
<tr>
<td>LC-MS/MS</td>
<td>Liquid chromatography coupled tandem Mass spectrometry</td>
</tr>
<tr>
<td>MIAME</td>
<td>Minimum Information About a Microarray Experiment</td>
</tr>
<tr>
<td>MIPS</td>
<td>Mammalian Protein Interaction Database</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MT</td>
<td>Mutant</td>
</tr>
<tr>
<td>OMIM</td>
<td>Online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>Pfam</td>
<td>Protein Family (database)</td>
</tr>
<tr>
<td>PGE2</td>
<td>Prostaglandin E2</td>
</tr>
<tr>
<td>PPI</td>
<td>Protein-protein interaction</td>
</tr>
<tr>
<td>PPIN</td>
<td>Protein-protein interaction network</td>
</tr>
<tr>
<td>PTM</td>
<td>Post-translational modification</td>
</tr>
<tr>
<td>SDS PAGE</td>
<td>Sodium dodecyl sulfate polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>siRNA</td>
<td>Small interfering RNA</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>TDP</td>
<td>True Discovery Proportion</td>
</tr>
<tr>
<td>UC</td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td>WT</td>
<td>Wild-type</td>
</tr>
</tbody>
</table>
# Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjacency matrix</td>
<td>Binary matrix; describes the node and edge set of a graph</td>
</tr>
<tr>
<td>Affinity Purification-Mass Spectrometry</td>
<td>Process of pulling down a target protein using antibodies, followed by subsequent analysis of the interacting proteins via MS</td>
</tr>
<tr>
<td>Alternative splicing</td>
<td>Alternative patterns of combining exons at the pre-mRNA level; performed by the spliceosome</td>
</tr>
<tr>
<td>Bimodality</td>
<td>Statistical parameter quantifying the density of extreme correlation values between two distributions</td>
</tr>
<tr>
<td>Binomial probability</td>
<td>The probability of a success in a sequence of independent experiments</td>
</tr>
<tr>
<td>Biological Unit</td>
<td>A peptide, exon, or protein</td>
</tr>
<tr>
<td>Bivariate t-distribution</td>
<td>A two-dimensional Student’s t distribution</td>
</tr>
<tr>
<td>Blossom</td>
<td>A conglomeration of petals linked via a central regulatory gene</td>
</tr>
<tr>
<td>Bootstrap (density)</td>
<td>Approach to approximating the null distribution; simulates new experiments by randomly sampling from the given samples</td>
</tr>
<tr>
<td>Chromatin Immunoprecipitation chip</td>
<td>Cross-linking of proteins to DNA, followed by antibody pull-down of the target protein, and hybridization of bound DNA to an array</td>
</tr>
<tr>
<td>Chromatogram</td>
<td>Time-dependent trace of peptides eluting from a chromatographic column</td>
</tr>
<tr>
<td>Clustering coefficient</td>
<td>A measure of the degree to which nodes in a graph cluster together; a reflection of edge density among nodes</td>
</tr>
<tr>
<td>Coexpression</td>
<td>Correlation (Pearson’s or Spearman’s) of mRNA expression measurements</td>
</tr>
<tr>
<td>Comparative Genomic Hybridization</td>
<td>Array-based approach measuring abundance of DNA across the genome for the analysis of copy number changes</td>
</tr>
<tr>
<td>Complement (probability)</td>
<td>The probability of an event not happening</td>
</tr>
<tr>
<td>Cooperativity</td>
<td>A principle describing the functional interaction between genes</td>
</tr>
<tr>
<td>Correlation</td>
<td>Covariance normalized by each variable’s variance such that the correlation ranges from -1 to 1</td>
</tr>
<tr>
<td>Covariance</td>
<td>A measure of how two variables change together</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>A measure of the amount of information available to estimate an unknown parameter</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Density</td>
<td>The distribution governing the likelihood of observing particular values of a random variable.</td>
</tr>
<tr>
<td>Differential expression</td>
<td>Change in mRNA or protein expression when comparing two experimental groups.</td>
</tr>
<tr>
<td>Differentially expressed</td>
<td>See “Differential expression”.</td>
</tr>
<tr>
<td>Driver (gene)</td>
<td>A gene that, when mutated, contributes to growth and development of a tumor.</td>
</tr>
<tr>
<td>Dysregulation</td>
<td>Patterns of mRNA/protein expression that deviate (e.g. in cancer) from “normal” or wild-type patterns.</td>
</tr>
<tr>
<td>Edge</td>
<td>A line in an interaction graph depicting a relationship between 2 nodes.</td>
</tr>
<tr>
<td>Eigenvalues</td>
<td>Scalar values associated with the eigenvectors of a matrix.</td>
</tr>
<tr>
<td>Eigenvector</td>
<td>Vector, $\mathbf{x}$, of a matrix, $\mathbf{A}$; when $\mathbf{x}$ is multiplied by $\mathbf{A}$, a scaled version of $\mathbf{x}$ is returned.</td>
</tr>
<tr>
<td>Empirical null distribution</td>
<td>Null density estimated from the data.</td>
</tr>
<tr>
<td>Empirical Type I Error Rate</td>
<td>Probability of a false positive; estimated from the data.</td>
</tr>
<tr>
<td>False Discovery Rate</td>
<td>Probability of a false positive among the set of tests rejecting the null hypothesis.</td>
</tr>
<tr>
<td>False negative</td>
<td>A situation where the null hypothesis should have been rejected, but was not.</td>
</tr>
<tr>
<td>False positive</td>
<td>An incorrectly rejected null hypothesis.</td>
</tr>
<tr>
<td>Family-wise Error Rate</td>
<td>A method of estimating the probability of making at least one false discovery among a family of tests (in multiple hypothesis testing).</td>
</tr>
<tr>
<td>Fisher’s method (statistics)</td>
<td>Method of combining p-values by taking the sum of their logarithm.</td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>A hierarchical vocabulary for describing genes/proteins.</td>
</tr>
<tr>
<td>Genome –wide Association</td>
<td>A study measuring thousands of polymorphisms across the genome to uncover genetic associations with disease.</td>
</tr>
<tr>
<td>Genomic width</td>
<td>The length of a biological unit (in base pairs).</td>
</tr>
<tr>
<td>Heterarchy</td>
<td>A principle that treats all genes as equal.</td>
</tr>
<tr>
<td>Hierarchy</td>
<td>A principle that gives importance to certain genes over others.</td>
</tr>
<tr>
<td>High dimensional</td>
<td>A situation in which the number of variables greatly exceeds the number of samples.</td>
</tr>
<tr>
<td>Homologues</td>
<td>Genes with sequences conserved throughout evolution.</td>
</tr>
<tr>
<td>Hotelling’s $T^2$</td>
<td>Multivariate statistical test; extension of the Student’s $t$ test.</td>
</tr>
<tr>
<td>Hub</td>
<td>A gene/protein with many interactions.</td>
</tr>
<tr>
<td>Hypomorph</td>
<td>Mutation that results in a gene with reduced function.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Ion suppression</td>
<td>Co-elution of multiple intense ions prevents the fragmentation and sequencing of less abundant species</td>
</tr>
<tr>
<td>Joint distribution</td>
<td>Statistical distribution governing the behavior of a set of variables</td>
</tr>
<tr>
<td>Knock-out</td>
<td>Homozygous deletion of a gene</td>
</tr>
<tr>
<td>Kuiper's test</td>
<td>Statistical test used to evaluate the similarity of two distributions</td>
</tr>
<tr>
<td>Liquid chromatography coupled tandem Mass spectrometry</td>
<td>Chromatographic separation of a peptide mixture, followed directly by mass spectrometric fragmentation; allows for reverse identification of peptides</td>
</tr>
<tr>
<td>Manifold</td>
<td>A map or structure that organizes genes/proteins</td>
</tr>
<tr>
<td>Marginal (distribution)</td>
<td>The density governing a single variable in a multivariate analysis</td>
</tr>
<tr>
<td>Maximum likelihood estimation</td>
<td>A method of estimating the parameters of a statistical model</td>
</tr>
<tr>
<td>Mixing angle</td>
<td>A parameter (related to Pearson's correlation) in the bivariate t-distribution that controls the correlation between the variables</td>
</tr>
<tr>
<td>Module</td>
<td>A group of genes organized according to similarity</td>
</tr>
<tr>
<td>Molecular Subsystem</td>
<td>A coupling of a manifold and a set of high-throughput molecular measurements</td>
</tr>
<tr>
<td>Multiple hypothesis testing</td>
<td>Significance testing of a large number of variables</td>
</tr>
<tr>
<td>Mutant</td>
<td>An organism, tissue, or cell possessing at least one (known) genetic aberration that differs from the “normal” or wild-type strain</td>
</tr>
<tr>
<td>Node</td>
<td>A vertex in an interaction graph; a protein in a PPI network</td>
</tr>
<tr>
<td>Null distribution</td>
<td>The distribution of a statistic under the null hypothesis</td>
</tr>
<tr>
<td>Parameterize</td>
<td>To obtain the parameters for</td>
</tr>
<tr>
<td>Pathway</td>
<td>A series of biochemical processes whereby a (molecular) stimulus is transduced into a functional output</td>
</tr>
<tr>
<td>Petal</td>
<td>A subnetwork in a “blossom”</td>
</tr>
<tr>
<td>p-value</td>
<td>A statistical measure interpreted as the probability of seeing a value as large as the one observed if the null hypothesis were true</td>
</tr>
<tr>
<td>Regularization, Regularized</td>
<td>Modification of a matrix to allow for the unique computation of its inverse, often by introducing sparsity</td>
</tr>
<tr>
<td>Shrinkage</td>
<td>See “Regularization”</td>
</tr>
<tr>
<td>Significance testing</td>
<td>Use of p-values to evaluate the null hypothesis</td>
</tr>
<tr>
<td>Smith-Waterman Algorithm</td>
<td>Computational algorithm for performing local sequence alignments</td>
</tr>
<tr>
<td>Somatic mutation</td>
<td>A DNA mutation arising in somatic cells; not found in the germline</td>
</tr>
<tr>
<td>Sparsity</td>
<td>To introduce zeros into a matrix</td>
</tr>
<tr>
<td>Spectral count</td>
<td>A measure of peptide intensity based on the number of MS/MS spectra collected for the parent ion</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>t-location-scale distribution</td>
<td>The Student's t distribution parameterized by four variables: mean, degrees of freedom, location, and scale</td>
</tr>
<tr>
<td>Transcription factor</td>
<td>A protein that regulates DNA transcription</td>
</tr>
<tr>
<td>t-statistic</td>
<td>A signal-to-noise measure between two experimental groups</td>
</tr>
<tr>
<td>Two Dimensional In Gel Electrophoresis</td>
<td>Procedure of separating proteins electrophoretically by isoelectric focusing and molecular weight</td>
</tr>
<tr>
<td>Type I Error Rate</td>
<td>Probability of a false positive; equivalent to the significance, or $\alpha$, level of a test</td>
</tr>
<tr>
<td>Western blot</td>
<td>Gel electrophoresis-based approach for protein expression analysis; relies upon antibody targeting of the protein of interest</td>
</tr>
<tr>
<td>XIC</td>
<td>Extracted ion chromatogram: ion current for a predefined set of precursor masses</td>
</tr>
</tbody>
</table>
Colon Cancer and its Molecular Subsystems:
Network Approaches to Dissecting Driver Gene Biology

Abstract
by
VISHAL N. PATEL

The progression of colorectal cancer is driven by the accumulation of mutations in a number of key genes, which synergize with each other to promote tumor growth. These co-regulatory genes are of particular importance as they provide alternative points for pharmacologic modulation of perturbed signaling pathways. In this work, we explored the signaling landscapes of three important regulatory genes – \textit{Hpgd}, \textit{Apc}, and \textit{Cdkn1a} – using genetically engineered mice as models of colon cancer. We assayed the intestinal epithelium using genomic (mRNA arrays) and proteomic (2D gels and label-free proteomics) modalities. To identify genes with a potential regulatory role, we employed protein interaction networks and mRNA coexpression networks to extrapolate from the proteomic measurements. In the process, we introduced novel bioinformatic approaches for testing network-based hypotheses, for applications where the signaling pathway is known a priori. In addition, we also introduced a novel statistical framework for interpreting label-free proteomic data in a genomic context. With the objective of identifying new regulatory targets, we found that the sparsity of proteomic data is ameliorated by relying on inferences from interaction networks and that these inferences are greatly improved by coupling proteomic data with genomic measurements.
Colorectal Cancer and its Molecular Subsystems: Construction, Interpretation, and Validation

Colon Cancer: Etiology

Colorectal cancer – a cancer of the large bowel – is the third leading cause of death among adult Americans, and it is estimated that 1 in 20 individuals will develop colorectal cancer in their lifetime. As the intestinal epithelium serves as a crucial medium for interaction with our environment (e.g. absorption of nutrients from food), it is not surprising that a variety of environmental triggers have been associated with an increased risk for developing colorectal cancer, such as smoking and saturated fat intake, among others. However, environmental insults can explain only a fraction of cancer risk, and much of the remaining story arises from an individual’s genetic background and his/her environment contribution to the initiation of colorectal cancer by inducing both somatic mutations in the epithelium and/or a chronic inflammatory response in the supportive tissue (called the lamina propria). The accumulation of genetic aberrations allows for clonal expansion of an aberrant precursor cell population, resulting in tumor formation.

Figure 1. Diagram of colorectal cancer development. An individual’s genetic background and his/her environment contribute to the initiation of colorectal cancer by inducing both somatic mutations in the epithelium and/or a chronic inflammatory response in the supportive tissue (called the lamina propria). The accumulation of genetic aberrations allows for clonal expansion of an aberrant precursor cell population, resulting in tumor formation.
background (see Figure 1). The most striking example of the genetic contribution to colorectal cancer can be found in the hereditary syndrome known as Familial Adenomatous Polyposis (FAP), in which numerous polyps grow throughout the large bowel at an early age, some of which invariably progress into adenocarcinomas. Known to run in an autosomal dominant fashion in families, the genetic basis for FAP was uncovered in the late 1980s as mapping uniquely to the gene \( Apc \). In demonstrating the genetic basis for colorectal cancer, studies of FAP patients paved the way for recent genome-wide experiments searching for genetic risk factors.

**Colon Cancer: Development**

A lifetime of environment insults, coupled with the natural aging process, results in the accumulation of damage to the DNA of colonic epithelial cells. The genetic character of non-hereditary colorectal cancer first became apparent in the early 1990s, when patterns of mutations were observed to be correlated with disease severity: it was observed that those aberrant crypt foci (ACFs) – the precursor epithelial lesion – that progress to neoplasia are often characterized by mutations in \( Apc \), and the growing tumor “builds” upon this first hit with mutations in additional genes, such as \( Kras \) and/or \( Tp53 \). By perturbing normal gene function, mutations can drastically alter the state of a cell, establishing new – and potentially uncontrollable – equilibria for cellular survival and proliferation. Cancer genes can be categorized as *tumor suppressors* if they suppress tumor growth or as *oncogenes* if they promote growth; consequently, mutated tumor suppressors (e.g. \( Tp53 \)) often have reduced or abrogated functions, while mutated oncogenes (e.g. \( Kras \)) have enhanced or new functions in the tumor. While mutations in such genes that actively sustain tumor development – referred to as “drivers” – are necessary for tumor progression, an individual tumor may be littered
with mutations in over 80 different genes\textsuperscript{13}, and these additional “passenger” mutations may have arisen simply due to the hyperproliferative and/or pro-mutagenic condition of the tumor.

While genetic aberrations are necessary for tumor initiation, promotion of tumor growth (see Figure 1) depends upon sustained injurious signals from the environment, which, in turn, induce further genetic damage. Though environmental insults can assault the epithelial cells directly, the supporting cells around the epithelium also mount a response to the insult by way of inflammation – a process observed in colorectal tumor specimens over a century ago\textsuperscript{14}. The chronic inflammatory response is characterized by an invasion of lymphocytes and other immune cells into the intestinal epithelium. As the inflammatory process is designed to repair tissue damage, the immune cells secrete a variety of communicatory molecules, called cytokines, which promote proliferation of the damaged epithelium, effectively allowing the cells to form a “wound” at the site of injury. By promoting cellular proliferation, inflammation creates an environment favorable to the accumulation of genetic insults. The inflammatory response can also damage DNA directly via the release of reactive oxygen and nitrogen species, free radical-containing molecules designed to damage invading bacteria and viruses\textsuperscript{15}. 

Figure 2. Pathways of \textit{Tp53} (shown above as p53). \textit{Tp53} controls the transcription of \textit{Cdkn1a} (shown above as p21), which, when mutated, also results in a predisposition to cancer. Tumorigenesis can also be triggered by mutations in \textit{Mdm2}. Protein interactions are indicated by blue diamonds; red arrows indicate transcriptional induction. Genes with red boxes have been found to be mutated in the germline; those in green boxes have been found to be mutated only somatically. Reprinted by permission from Macmillan Publishers Ltd: Nature Medicine ©

\textsuperscript{2}
The Pathway Paradigm

While the accumulation of somatic mutations is the hallmark of cancer, it has become increasingly clear that the mutant genes alone do not define the trajectory of the disease. Indeed, patients whose tumor’s mutational profiles are >85% distinct still result in similar histopathologic grading and patient survival (personal analysis of publicly available glioblastoma data). Given the apparent redundancy of gene function, the field of cancer biology relies upon a modular framework for molecular analysis, organizing genes according to signaling pathways: series of biochemical processes whereby a (molecular) stimulus is transduced into a functional output (see Figure 2 for an example). Though individual genes often appear to have their own identity (possessing unique functions and interaction partners), their redundancy becomes apparent when considered in the context of a signaling pathway. Mutated driver genes were recently found to co-exist within signaling pathways, supporting the hypothesis that multiple hits to a pathway or module are necessary to produce a functional effect. A more specific example involves the tumor suppressor $Tp53$, a gene frequently mutated in late-stage colorectal cancer. As $Tp53$ controls the transcription of a cell cycle inhibitor, $Cdkn1a$ (also known as $p21$), homozygous deletion of $Cdkn1a$ in mouse models also results in a tumorigenic predisposition; amplification of $Mdm2$ can also disrupt the $Tp53$ pathway (as in Figure 2).

In addition to functional redundancy, pathways also contain information on interactions. As shown in Figure 2, genes can affect each other through direct, physical interactions between their protein products, as well as via indirect regulation of gene transcription. Often, mutant genes in separate pathways (who, therefore, do not interact with each other via the aforementioned mechanisms) synergize with each other, and these long-
range genetic interactions are central to cancer biology, where dozens of mutations in myriad pathways are operating simultaneously. In fact, a mutant gene’s functional role is strongly affected by the presence of these synergistic, accompanying mutations. A lone mutation in $Kras$, for instance, leads only to self-limiting lesions and/or ACFs, while a $Kras$ mutation secondary to a hit in $Apc$ correlates with polyp formation and cancer $^{12}$.

Cancer Subtypes and Therapies

In spite of the mutational heterogeneity found among colorectal cancer patients, it is clear that similarities between patients exist at more general phenotypic levels. Recently, various cancers have been clustered by their mRNA expression profiles $^{20,22}$ and proteomic signatures $^{23,24}$, pointing to the existence of molecular subtypes in cancer. Importantly, certain molecular subtypes show differences in outcome and response to therapy. For instance, it was recently found that a molecular therapy, cetuximab, targeting the epidermal growth factor receptor ($Egfr$) improved overall colorectal cancer survival, though resistance to therapy was widespread $^{25}$. Interestingly, $Egfr$’s mutational status alone is less helpful in understanding cetuximab efficacy than other molecular and genetic markers $^{26}$. In particular, oncogenic transformation of $Kras$ correlates with cetuximab resistance in patients, as these mutations activate $Egfr$-related pathways independent of $Egfr$’s mutational status $^{27}$. By understanding the genetic interactions and pathways linking $Kras$ to $Egfr$, prescribing practices for cetuximab therapy have been drastically altered, forcing physicians to consider a patient’s tumor genotype prior to administering treatment.

Molecular Subsystems: Introduction
The use of signaling pathways and other maps (e.g. the genome, interaction networks) to organize data has been a characteristic feature in the era of high-throughput data analysis of cancer. This approach has heralded a new subfield that uses theoretical frameworks to couple biological measurements together. In these recent approaches, the molecules of interest – e.g. driver genes – are studied in the context of an organizing manifold, e.g. a signaling network, and we refer to this joint structure between molecules and manifold as a molecular subsystem.

The manifold imposes a structure onto the system components, and, thus, a particular manifold represents one hypothesis for the architecture of driver gene cooperativity. As a subsystem is defined, in part, by its constituent molecules, subsystems can be constructed at individual levels of the molecular hierarchy to create networks of interacting genes, RNAs, or proteins. Alternatively, subsystems traversing multiple hierarchical levels exist, such as a regulatory network linking a transcription factor to the gene products whose expression it controls.

In the following, we first discuss approaches for measuring and mapping the components of molecular subsystems (“Construction”), and then discuss the biological and statistical meaning of the resultant subsystem (“Interpretation”). Since a molecular subsystem has two facets – the molecules and the manifold – that require experimental exploration, we conclude with a section on this important, but often overlooked, point (“Validation”).

**Molecular Subsystems: Construction**

**Measurements**
As molecular subsystems arise from the coordinated actions of many individual parts—be they nucleotides, amino acids, peptides, or proteins—we must be able to measure these numerous parts efficiently and en masse if we wish to study subsystems. However, our limited exploration of biological space prevents us from fully knowing all of a subsystem's components, and, consequently, our measurement tools must also be capable of revealing pieces unbeknownst to the observer.

Such tools have emerged in the past decade due, in no small part, to the sequencing of the genome. With the genome sequence, it became possible to create thorough compendia of genetic elements, and technologies soon followed for measuring these elements and their products. While some of these tools measure only pre-specified targets (e.g. microarrays), other tools—particularly in proteomics—take measurements stochastically from an experimental sample. Both types of technologies lend themselves to the study of molecular subsystems and have helped to usher in an era of “discovery science,” wherein large ‘omic datasets are explored to arrive at biological hypotheses ex post facto.

Array-based technologies for measuring mRNA transcripts were among the first to emerge in the mid 1990s; for a review of the technology see Allison et al. \(^{28}\) and Wheelan et al. \(^{29}\). The use of microarrays for measuring mRNA transcripts has become increasingly commonplace as their costs have declined, and large repositories for expression profiling experiments now exist (i.e. the Gene Expression Omnibus\(^{30}\)). Variants of expression profiling have been introduced for a variety of applications: arrays with probes spanning splice junctions for detecting alternative splicing\(^{31}\); probes targeting transposable elements\(^{32}\); and probes targeting microRNA (miRNA) species\(^{33}\).

For proteins, antibody arrays are the intuitive analog to microarrays. However, antibodies are notorious for their inconsistency and lack of specificity\(^{34}\). Moreover, with
current estimates of the occurrence of alternative splicing, the number of potential proteins is estimated to be well over 300,000, and, thus, specific antibody production becomes very costly. The alternative modality that has dominated the field is based on mass spectrometry (MS), wherein a sample is separated by liquid chromatography (LC) prior to ionization and sequencing by tandem rounds of MS. While DNA or RNA-based experiments can tile their probes on an array, LC-MS-based approaches sample semi-stochastically from a complex mixture of peptides. Though ripe for discovery, this critical step introduces a degree of uncertainty into the analysis that is still being addressed. Nonetheless, steady improvements in chromatographic reproducibility and spectral sensitivity have allowed for the development of quantitative LC-MS based methods where up to a few thousand proteins can be quantified in relative terms.

Figure 3. Three types of commonly used manifolds: (A) the genome, shown with peptide measurements made via LC-MS/MS; (B) a protein-protein interaction network, shown with mRNA measurements (data from Patel et al.); and (C) a coexpression network between proteomic targets and transcription factors, displayed as a heatmap.
Manifolds

While the aforementioned technologies provide point-measurements of subsystem components, an organizing principle, which we refer to as a manifold, is also necessary to describe the system. The genome has proven to be a surprisingly complex manifold due to the wealth of information hidden within it. The most obvious structure in this manifold is the gene itself. However, as our awareness of regulatory mechanisms grows – promoters, enhancers, miRNA, et al. - we push the boundaries of where a gene starts and where a gene stops. The hidden regulation of the genome has consistently created problems in the field of microarray analysis, where, for example, probes for a single gene often show inconsistent differential expression across the length of the gene, strongly suggestive of alternative splicing. Though this form of regulation was largely ignored early on, new array technologies with increased probe density across a gene are allowing for the improved detection of splice variants. Top-down, proteomic approaches also show promise for unbiased detection of splice isoforms by making use of the genome, as illustrated in Figure 1A.

Early subsystem analysis of microarray data focused on the use of ontologies to organize the measurements. Initial studies used the Gene Ontology (GO), the most widely used and well-curated ontology, to annotate individual molecular species, and the saturation of GO terms within a list of differentially expressed genes (or visa versa) remains a commonly used measure. While ontologies are used for finding a common annotation among a group of genes, their biological relevance is called into question as they are (1) biased towards well-studied genes and processes and (2) difficult to organize and interpret (e.g. higher terms in the GO hierarchy are less biologically informative).

An alternative manifold to represent more complex relationships is one based on networks, wherein a molecular unit may interact with any number of partners. A network is
constructed from a set of 1-dimensional elements, since only one dimension – the edge weight – is required to describe the relationship between any two points or nodes. The genome itself can be represented as a network in which the edges are measured by the inverse pairwise genomic distance between loci. Ontologies, as well, can be represented as networks, as was done for the set of gene-disease relationships found in OMIM (Online Mendelian Inheritance in Man) to create a disease network\textsuperscript{39}.

Transcriptional, or coexpression, networks are commonly used as they can be constructed from a (sufficiently large) set of microarray data, calculating pairwise correlations between all genes (see Figure 1C). Coexpression manifolds have proven useful because coexpressed genes (1) may have a common regulator or (2) may be regulating each other. With the genomic coverage of current microarrays, a coexpression network can provide a global picture of interactions among genes, as opposed to the more limited view provided by Protein-Protein Interaction (PPI) networks. However, coexpression networks have little meaning in the context of traditional pathways or interactions – two coexpressed genes may not physically interact or they may be in different cellular compartments – and, as such, they have been most useful in uncovering targets with strong transcriptional roles\textsuperscript{40}.

Protein-protein interaction networks are derived from databases compiled from specific experimental probes of physical interactions among proteins. These databases are largely composed of interactions measured via yeast two-hybrid (Y2H) or affinity purification-mass spectrometry (AP/MS). Some of the most widely used interaction databases are the Human Protein Reference Database (HPRD)\textsuperscript{41}, biological General Repository for Interaction Datasets (BioGRID)\textsuperscript{42}, Biomolecular Interaction Network Database (BIND)\textsuperscript{43}, Database of Interacting Proteins (DIP)\textsuperscript{44}, Online Predicted Human Interaction Database (OPHID)\textsuperscript{45}, Molecular Interaction database (MINT)\textsuperscript{46}, and IntAct\textsuperscript{47}.
AP/MS “pulls out” a complex of proteins interacting with a bait, and the set of interactions are modeled as either complete (with every protein interacting with every other protein) or as a “wheel-and-spoke” (with the bait alone interacting with every protein in the complex), though it is recognized that neither model is representative of most cases. The Y2H method focuses on pairwise interactions generated through genomic cloning of specific domains, and can be applied on a genome scale. These high-throughput screens for interactions have well known limitations that produce false positives (e.g. spurious interactions, or two proteins interacting in a Y2H may not interact in vivo) and false negatives (weak interactions may be missed), and several groups – both academic (e.g. HPRD) and commercial (e.g. MetaCore, Ingenuity) – have provided manual curation of the data.

While the genome is a “complete” manifold, PPI networks are certainly not, and our current knowledge of human PPIs is estimated to cover only 10% of the true number of interactions. While coupling multiple independent PPI networks together can serve to filter out the noise, multiplexing PPIs introduces additional biases, such as promoting well-studied proteins – which tend to be disease associated – and artificially inflating their value as “hubs”. However, as discussed previously, the construction of molecular subsystems hinges not on the contribution of any single molecule or interaction, but, rather, on the properties of the molecular ensemble. Consequently, subsystems based on PPIs can often tolerate a high degree of noise in the underlying manifold or in the measured molecules.

**Molecular Subsystems: Interpretation**

In designing and, more importantly, interpreting a molecular subsystem, we make use of two well-founded biological assumptions: hierarchy and heterarchy. **Molecular heterarchies** are systems in which there are no absolute governing molecules, but, rather, each
molecule is attributed a similar position of importance. The principle of heterarchy is implicit in the use of many manifolds, for the manifolds discussed here have no innate pecking order. Yet, the biological tools at our disposal are not amenable to studying heterarchies, as mouse models, siRNA transfections, et al. are designed to target a single (or a few) genes at a time. To use these tools efficiently, we are inadvertently biased towards looking for hierarchy: we may cluster nodes by degree or unearth regulatory “hub” proteins to discover hidden hierarchy in networks\textsuperscript{51, 52}.

The global hierarchy of molecular biology stems from the introduction of the central dogma, which gave primacy to “lower” levels (DNA) as controlling or storing information found at “higher” levels (RNA, protein). With the unceasing discovery of new forms of regulation (e.g. RNA-protein interactions, microRNA), our current understanding of reality is far more complex than imagined by the central dogma. Nonetheless, the original principles are still useful when integrating high-throughput data sources: mutations in the cancer genome, for example, that are in \textit{cis} with changes in mRNA expression may have a greater functional role than those not transcribed\textsuperscript{53}. Along these lines, considering how the protein product may be affected by a mutated driver gene (e.g. via the introduction of a stop codon or splicing alterations) can provide evidence as to the functional role of the mutation\textsuperscript{17, 54}.

Examples

In Table I, we have listed a few commonly studied molecular subsystems; some use two forms of high-throughput data to improve the power for biological inference. In the following section, we discuss the implications of using particular manifolds and provide examples from the literature.
Table I. Potential molecular subsystems and general questions that can be asked of them.

<table>
<thead>
<tr>
<th>Molecular species measured</th>
<th>Manifold</th>
<th>Pairs well with</th>
<th>Can be used to study</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA</td>
<td>Genome</td>
<td>Protein-DNA binding$^*$</td>
<td>Transcriptionally active euchromatin$^{55}$</td>
</tr>
<tr>
<td>mRNA</td>
<td>Genome</td>
<td>DNA dosage$^†$</td>
<td>Transcriptionally active copy number variation</td>
</tr>
<tr>
<td>Peptides$^‡$</td>
<td>Genome</td>
<td>-</td>
<td>Alternative splice variants that are translated</td>
</tr>
<tr>
<td>mRNA</td>
<td>Known pathway(s)</td>
<td>-</td>
<td>Transcriptional activity of pathway(s)</td>
</tr>
<tr>
<td>SNPs</td>
<td>Coexpression network</td>
<td>-</td>
<td>Transcriptional modules regulated by a SNP</td>
</tr>
<tr>
<td>mRNA</td>
<td>Coexpression network</td>
<td>miRNA binding (predictions)</td>
<td>Modules of genes regulated by a common miRNA species</td>
</tr>
<tr>
<td>SNPs</td>
<td>PPI network</td>
<td>mRNA data</td>
<td>Clustering of transcriptionally active SNPs$^{56}$</td>
</tr>
<tr>
<td>Proteins</td>
<td>PPI network</td>
<td>mRNA data</td>
<td>Dysregulated network neighborhoods$^{24}$</td>
</tr>
<tr>
<td>PTMs</td>
<td>PPI network</td>
<td>-</td>
<td>Modularity of PTM patterns</td>
</tr>
<tr>
<td>Mutations</td>
<td>PPI network</td>
<td>-</td>
<td>Genetic interactions in a pathway-context$^{57}$</td>
</tr>
<tr>
<td>Mutations</td>
<td>Disease network</td>
<td>-</td>
<td>Phenotypically related genes$^{30}$; new candidate disease genes$^{58}$</td>
</tr>
</tbody>
</table>

$^*$via Chip-seq; $^†$via CGH or SNP array; $^‡$digested from proteins in LC-MS/MS workflows;

The genome is a useful manifold as it can serve as a defensible point-of-causation in certain instances. To interpret mRNA expression changes in obese human populations, for example, Emilsson et al. correlated these expression changes with genetic polymorphisms by way of SNP arrays$^{59}$. This allowed expression changes with a genetic basis to be culled from those that arise by secondary or indirect mechanisms. Genomic location is also a valuable manifold in cancer studies, where genomic aberrations are frequent hallmarks of the disease$^{60}$. Relative amounts of DNA can be measured using comparative hybridization (CGH) arrays, as was done for the analysis of glioblastoma samples by Bredel et al., and, after mapping these measurements to the genome, were used to provide insight into correlated patterns of genomic gains and losses$^{61}$. 
Leveraging the modular nature of coexpression manifolds, Akavia et al. found that genes with correlated patterns of mRNA expression can serve to identify causal mutations in melanoma\(^6\). Similarly, Horvath et al. traversed coexpression modules in glioblastoma to uncover a gene, *Aspm*, with a previously unexamined regulatory role\(^4\).

Compiling PPIs allows one to extend the linear signaling pathway paradigm to create networks of physical interactions within which traditional pathways are embedded. Cerami et al. capitalized on this idea to identify clusters of physically connected driver genes in glioblastoma, providing support for the hypothesis that driver genes are mutated in different pathways\(^6\). Nibbe et al. (2010) found that groups of proteins proximal (as measured by protein-protein interactions) to multiple driver genes show higher levels of mRNA dysregulation than groups of proteins proximal to only a single driver gene, simultaneously illustrating that multiple levels of information in the molecular hierarchy can be integrated to identify hotspots in the signaling network\(^6\).

**Molecular Subsystems: Validation**

As discussed, subsystems are composed of two features – the molecular units and the organizing manifold – and both components require experimental validation. One approach to validation has been to demonstrate that analysis of a particular manifold can triangulate functionally relevant molecules. Nibbe et al. (2009) showed PPI networks are eminently useful in finding neighborhoods of dysregulated proteins in colorectal cancer\(^3\). Lim et al. used a PPI subnetwork for inherited ataxias to identify Purotrophrin-1, previously unknown to associate with these diseases. In addition, they validated their PPI manifold directly by testing randomly sampled interactions (which were compiled via Y2H) by coimmunoprecipitation\(^5\). The biological relevance of manifolds can also be supported by
demonstrating their evolutionary conservation, as was done for PPI networks\textsuperscript{65}, serving as a global validation of the manifold.

An alternative approach to validation is motivated statistically, using aggregate measures of the subsystem to assess its importance. The assumption in these approaches is that a manifold-molecule coupling should exhibit stronger coordinated changes (i.e. differential expression) than expected. Thus, when Nibbe et al. (2009) used exhaustive searches to identify transcriptional “hotspots” in PPI subnetworks, they tested the validity of these resultant hotspots statistically, permuting either the phenotype labels or the gene labels. These empirical null distributions represent two different null hypotheses: phenotype label permutation models a null distribution in which the measurements have no association with the known phenotype categories; gene name permutation creates a null distribution in which the interaction pattern among genes/proteins is abrogated.

In recently published work, we illustrated how proteomic data can be leveraged to statistically test the relevance of a candidate subnetwork\textsuperscript{66}. As proteomic data is sparse and measures targets downstream of oncogenes and tumor suppressors, a second map is required to associate proteomic targets with a hypothesized subsystem. A coexpression network can be particularly useful in this situation, as strong coexpression of subsystem genes with proteomic targets provides evidence for a regulatory role of the upstream, subsystem genes. This relationship can be directly tested by calculating the strength of coexpression between the subsystem and the proteomic targets, and then calculating significance against the background level of coexpression between the subsystem and all measured molecules\textsuperscript{67}.

Drawing from engineering systems theory, system structures can be tested by examining system properties upon failure of individual components, allowing us to gauge the
cascading effect of perturbations within the network architecture. This is analogous to the underpinnings of cancer biology, wherein we search for genes whose “failure” (i.e. tumor suppressors) critically impacts the functioning of the system. Using these principles, we demonstrated that the classic and most widely used model of gene failure – the knock-out mouse – can be re-envisioned in a subsystem context: the mutated gene is one component of a hypothesized molecular subsystem, and the knock-out mouse is a model perturbation of this subsystem. In our study, we tested a PPI network manifold by using mice mutated at two different subsystem genes, \( Apc \) and \( Cdkn1a \). Coupled with mRNA and proteomic measurements, we were able to test for (1) the differential expression of subsystem genes at the mRNA level and (2) the association – via PPIs or coexpression – of the subsystem with proteomic targets measured upon each of the two perturbations.

**Summary**

Herein, we have summarized a bioinformatic trend emerging in cancer biology, one in which ensembles of high-throughput molecular measurements are studied in the context of an organizing map, or manifold, and we dub this marriage of molecule and manifold a molecular subsystem. As studies of molecular subsystems often link disparate analytical tools and approaches, the field is becoming increasingly complex, especially for scientists unfamiliar with the past decade of bioinformatics research. For example, differential expression (of mRNA or protein) is a simple and frequent request of ‘omics experiments. However, the interpretation of these results is far from obvious. Traditional biology suggests that we simply cherry pick “targets” from this list for further experimental exploration, though at the expense of discarding valuable data. In this context, a basic understanding of manifolds – and the resultant molecular subsystems that can be studied with them –
prevents us from losing the forest for the trees, and we hope that our explanation of this concept proves useful both for scientists navigating an ‘omics experiment, as well as for experienced bioinformaticians developing new approaches for studying subsystems.
Introduction: Colon Cancer and Inflammation

While cancer in general is thought to arise from a cascade of genetic aberrations, colorectal cancer in particular has long been recognized as also being an inflammatory disease. Epidemiological evidence first suggested that patients who take aspirin regularly have lower rates of colonic polyps, implicating the cyclooxygenase (COX) pathway. The COX pathway converts arachidonic acid into thromboxanes, prostacyclins, and prostaglandins, with prostaglandin E2 (PGE$_2$) being the most abundant prostaglandin in gastrointestinal tumors. Non-selective inhibitors target both COX1 and COX2, reducing thromboxane and prostaglandin synthesis respectively, though inhibition of the constitutituve isozyme, COX1, thins the gastric mucosa and increases the likelihood of ulceration. Targeting COX2 specifically, e.g. with celecoxib, has ushered in a new line of treatment for FAP patients, though the concomitant increases in thromboxane are associated with cardiovascular side effects.

Consequently, alternative methods of pharmacologically modulating the COX pathway are needed. Apart from the synthesis of prostaglandins, an alternative mode of regulation lies in their degradation, which is catalyzed by 15-hydroxyprostaglandin dehydrogenase (protein name: HPGD, gene name: Hpgd). Hpgd, which specifically catabolizes PGE$_2$, is ubiquitously downregulated in colorectal cancer, supporting its role in
antagonizing the frequently upregulated COX2 inflammatory pathway\textsuperscript{73}. In fact, low levels of \textit{Hpgd} in human tumors have been correlated with a resistance to COX2 inhibitors\textsuperscript{74}. Decreased levels of \textit{Hpgd} are coupled, as expected, with increased levels of PGE\textsubscript{2}. Upon binding to its G-coupled receptor, EP2, PGE\textsubscript{2} can then activate several key oncogenic pathways: EGFR signaling\textsuperscript{75}; \(\beta\)-catenin/TCF signaling\textsuperscript{76,77}; PI3K/AKT signaling\textsuperscript{78}; and RAS-MAPK signaling\textsuperscript{79}. Though these signaling pathways provide insight into the tumor suppressive mechanism of \textit{Hpgd}, it is neither a comprehensive nor an unbiased list. In identifying alternative regulatory mechanisms of the COX pathway, we herein employ a proteomic approach to discovering perturbed pathways in mouse models of \textit{Hpgd}\textsuperscript{\text*/}. The goal is to provide an unbiased sampling of proteomic changes in colonic tissue as a function of the knockout. With these data in hand, we hope to infer specific pathways, in addition to those mentioned above, that could be therapeutically exploited for treatment. As seen in this chapter, we will develop models of the system that flow from peptide level information, to individual proteins, to networks of proteins that assist in reverse engineering the biology of \textit{Hpgd}\textsuperscript{\text*/} (Figure 1).

\textbf{Introduction: Mass Spectrometry}

In the past decade, liquid chromatography coupled mass spectrometry (LC-MS\textsuperscript{b}) has been increasingly used for proteomic profiling as it allows for a “bottom up” ‘omics approach, wherein a complex protein sample is enzymatically digested, the resulting mixture of peptides analyzed by LC-MS/MS (MS/MS indicating tandem mass spectrometry), and the component peptides finally identified by comparing the observed spectra to theoretical spectra. With the advent of modern, high mass accuracy MS instruments, case and control samples used for comparing protein expression can be analyzed in a label-free mode,
without resort to chemical derivitization of peptides that has been common in the past\textsuperscript{80}. Although this simplifies the biochemical preparation, the resultant data is very complex.

This multi-step workflow – and the massive amount of data generated – requires computational pipelines for data analysis and management, and many such packages have recently been developed (TransProteomic Pipeline\textsuperscript{81}, Rosetta Elucidator, Census\textsuperscript{83}). Programs like msInspect\textsuperscript{83} and the OpenMS Proteomics Pipeline\textsuperscript{84} provide essential routines, such as normalization, peak detection, and peak quantification. Commercial packages, e.g. Rosetta Elucidator, couple this functionality with database searching, allowing the user to annotate spectral features with peptides and protein assignments. After the spectra are acquired, the remainder of the workflow is largely statistical, and each step of the process has warranted the development of application-specific statistical routines: normalization of the spectra\textsuperscript{85, 86}; database matching\textsuperscript{87}; the normalization and interpretation of database matching scores\textsuperscript{88}; and the assignment of peptides to parent proteins\textsuperscript{35}.

One stage of the pipeline that has not been fully addressed, however, concerns the interpretation of the resulting intensity information per peptide and this relates to our specific goal of understanding protein dysregulation. Simply put, peptides are not proteins. While peptide-level intensity information – be it in the form of spectral counts, XIC, p-values, etc. – is technically informative and suggestive of protein level changes, it may be only minimally biologically informative, as numerous possibilities exist as to why a particular peptide exhibits differences in abundance between two biological conditions. Sophisticated tools, such as PeptideProphet\textsuperscript{88}, allow us to be more confident of a particular peptide being called “present” or “absent,” but they do not inform the scientist as to how the associated change in abundance should be interpreted biologically.
Thus, the question remains – and is not by any means solved in the field – as to how one should interpret a change in abundance at the peptide level. A particular peptide may be found to differ between two biological conditions for a number of technical reasons: (1) ion suppression in the mass spectrometer in one condition alone, (2) misalignment of the chromatograms, and (3) technical variation. With sufficient biological or technical replicates, and the increasing accuracy of both mass spectrometers and chromatographic separation systems, many of these issues can be overcome, leaving room for the more compelling biological explanations for differences in the abundance of a peptide: (1) differences in abundance of the parent protein between the two biological conditions (the result typically of interest), (2) the gain/loss of that peptide’s exon (alternative splicing) in one sample, or (3) post-translational modifications (PTMs) on the peptide that were not searched for explicitly in the spectrum-matching step.

As we are interested in making inferences about the behavior of the parent protein, peptide-level information is often summarized by one of two approaches: summing/averaging the intensities of all daughter peptides, or by selecting one or two “representative” peptides. Each of these approaches implicitly assumes that the regulation at hand is that of a change in whole-protein abundance, and they ignore the more subtle possibilities of alternative splicing and unsearched PTMs. Given the widespread occurrence of these two forms of regulation,
their consideration is critical to the proper analysis of label-free LC-MS/MS data. Thus, we are interested in both the global expression change, e.g. the overall change in protein expression, and the local changes that reflect the peptide-by-peptide behavior, and both may be biologically informative.

To incorporate these issues into the label-free quantification process, we introduce a statistical framework to quantify the dependency of a single peptide on the behavior of its sibling peptides, allowing one to prioritize “deviant” peptides, i.e. candidates for PTMs. In addition, we extend this framework in a genomic context to study the behavior of exons, thereby allowing one to identify deviant genomic regions, i.e. candidates for disease-specific alternative splicing.

**Methods**

**Sample preparation: 1D Gel Electrophoresis and Digestion**

The colonic mucosa of Hpgd⁻/⁻ mice (female, 2-3 months old) was harvested by scraping the intestinal surface with a glass slide; samples were collected from 5 Hpgd⁻/⁻ and 5 wild-type paired mice on the FVB background. To improve yields in proteomic discovery, the cell lysate (total of 20 µg protein) from each biological sample was separated via 1-D gel electrophoresis using a 4-20% gradient SDS PAGE gel system. Each of the 10 gel lanes was subfractionated into 6 bands, resulting in a total of 60 gel bands. All gel bands were rehydrated in a 1:1 acetonitrile:water mixture with mild shaking for 15 minutes, with the gel bands being crushed simultaneously. After treatment with 100% acetonitrile for 5 minutes, the crushed gel pieces were suspended in 50 mM ammonium bicarbonate buffer (pH 8) before being subjected to reduction (dithiothreitol 20mM, 56°C, 10 min) and alkylation (iodoacetmaide, 55mM, room temperature, 30 min), successively. For tryptic cleavage of the
denatured proteins, a total 500 ng of trypsin was added to 50 µL of 50 mM ammonium bicarbonate buffer (pH 8), and the digestion was left overnight at 37°C. 1% formic acid was added to stop the proteolytic processing and the resulting peptides were extracted from the gel pieces using a 50% acetonitrile-0.3% formic acid solvent; gel pieces were then dried and resolubilized in 0.1% formic acid/100% water. For all analyses, 5 µL of digest was used for LC-MS/MS processing.

**LC-MS/MS Analysis**

The digests were analyzed by LC-MS/MS using a UltiMate 3000 LC systems (Dionex, San Francisco, CA) that was interfaced to a LTQ-Orbitrap XL mass spectrometer (Thermo-Finnigan, Bremen, Germany). The platform was operated in the nano-LC mode using the standard nano-ESI API stack fitted with a picotip emitter (uncoated fitting, 10µm spray orifice, New Objective, Inc., Woburn, MA). The solvent flowrate through the column was maintained at 300 nL/min using a 1:1000 splitter system. The protein digests (5 µL) were injected into a reversed-phase C18 PepMap trapping column (0.3 x 5 mm, 5 µm particle size, Dionex Inc.) equilibrated with 0.1% Formic Acid (FA)/2% acetonitrile (v/v) and washed for 5 min with the equilibration solvent at a flow rate of 25 µL/min, using an isocratic loading pump operated through an autosampler. After the washing step, the trapping column was switched in-line with a reversed-phase C18 Acclaim PepMap 100 column (0.075 x 150 mm, Dionex Inc.) and the peptides were chromatographed using a linear gradient of acetonitrile from 5% to 50% in aqueous 0.1% formic acid over a period of 80 min at the above-mentioned flowrate such that the eluate was directly introduced to the mass spectrometer. A 100% acetonitrile elution step was subsequently performed for ten minutes prior to resetting the analytical column to the initial equilibration conditions for 10
more minutes at the end of the chromatographic run, accounting for a total of 90 minutes of LC-MS/MS time. The mass spectrometer was operated in a data-dependent MS to MS/MS switching mode, with the five most intense ions in each MS scan subjected to MS/MS analysis. The full scan was performed at 60000 resolution in the Orbitrap detector and the MS/MS fragmentation scans were performed in the ion trap detector (IT) CID mode such that the total scan cycle frequency was approximately 1 second. The threshold intensity for the MS/MS trigger was set at 1000 and the fragmentation was carried out using the CID mode using a normalized collision energy (NCE) of 35. The data were entirely collected in the profile mode for the full scan and centroid mode for the MS/MS scans. Dynamic exclusion function for previously selected precursor ions was enabled during the analysis such that the following parameters were applied: repeat count of 2, repeat duration of 45 seconds, exclusion duration of 60 seconds and exclusion size list of 150. Xcalibur software (version 2.0.7), Thermo-Finnigan Inc., San Jose, CA) was used for instrument control, data acquisition, and data processing.

Chromatographic reproducibility and full scan MS\(^1\) intensity robustness are critical to the success of this method. In order to monitor this efficiently and exhaustively, spiked external peptides from yeast enolase digest (400 fmoles on column-load) were used to keep track of the chromatographic performance of the LC-MS/MS system.

**Spectral Searching**

Rosetta Elucidator (Rosetta Inpharmatics) was used to normalize and process the raw LC-MS/MS data. Using Mascot, mass spectra were searched against the IPI database (with reverse sequences included for false discovery rate control), and matches were selected that had PeptideTeller probabilities >0.95. This resulted in 5701 peptides. After exporting
the results as a CSV file, the peptide intensities for each sample for log-transformed and mean-centered at a user-defined value, $\mu$, chosen such that all resulting intensities are greater than zero ($\mu=15$ for our data).

Genomic Mapping of Peptides

We used Ensembl (build 37) to provide a common genomic framework for mapping peptides annotated by Elucidator. Though the solution to assigning peptides to their parent proteins remains contested\textsuperscript{35, 93}, we chose to use the assignments generated by ProteinTeller (an incarnation of ProteinProphet) to illustrate the methods we develop. Peptides annotated with GI protein IDs were mapped to their associated gene-level IDs using the gene2accession file available at NCBI. For the subset of proteins identified in our data by Rosetta, corresponding exon sequences were downloaded from the Ensembl website.

Peptides within a protein were aligned to the protein’s exons using the Smith-Waterman algorithm and the BLOSUM80 matrix. Though most peptides were perfect alignments, this implementation allows for the possibility of polymorphisms or mutations that may result in slight changes in the coding sequence. The alignments were made in all three reading frames, and the alignment with the maximum score was chosen for each peptide.

Whole Protein-level Significance

To infer changes in protein abundance, we first assume that such regulation would result in unidirectional changes shared by the majority of peptides measured from a protein. To assess this, peptide-level t-statistics were first summarized at the exon level:
Where the t-statistic assigned to an exon, $t_e$, was defined as the average of the $k$ member peptides’ t-statistics. This prevents a single exon or domain from overshadowing undersampled regions of the protein. Then, the gene-level statistic, $t_g$, was defined as the sum of the $n_p$ exon-level t-statistics; $n_p$ only includes those exons that had daughter peptides measured in the LC-MS/MS experiment.

$$t_e = \frac{1}{k} \sum_{i=1}^{k} t_i$$

$$t_g = \sum_{p=1}^{n_p} t_p$$

We calculated the two-sided $p$-value of a gene being differentially expressed as $p(|t_g|)$ approximating the null distribution with the bootstrap density estimate (10,000 bootstrap samples). Since a low number of peptides from a parent protein prevent the discrimination of global (i.e. whole-gene) behavior from local (i.e. PTM or splicing) behavior, we limited significance testing to those proteins with $\geq 3$ peptides (regardless of peptide length).

**Peptide-Level Significance**

In analyzing differentially expressed peptides, we only consider peptides of length $\geq 9$ amino acids to reduce the frequency of irreproducible events. While current LC-MS/MS pipelines output peptide-level intensity, this information must be evaluated in a whole protein context to be biologically informative. At the individual peptide level, we are interested in cases when the behavior of a single peptide does not conform to the behavior of the rest of the protein, as such situations indicate local (as opposed to global) regulation of the protein, which may occur in the form of PTMs. Using t-statistics as our summary
measure of choice, we must calculate the conditional probability of observing a peptide \( i \) given the behavior of the protein’s remaining peptides:

\[
P(t_i \mid \tau, \Sigma) = \frac{P(t_i, \tau, \Sigma)}{P(\tau, \Sigma)}
\]

Where \( t_i \) is the t-statistic associated with peptide \( i \), and \( \tau \) is the statistic associated with the rest of the protein, defined as:

\[
\tau = \frac{1}{k - 1} \sum_{j \neq i} t_j
\]

\( \tau \) represents the average t-statistic of the sibling peptides (i.e. from the same parent protein) for peptide \( i \). The covariance between peptides, \( \Sigma \), must be included since the observed abundance of a particular peptide depends not only on the abundance of its sibling peptides, but also on the level of inter-peptide dependency. Thus, we are interested in calculating the probability of observing a certain behavior for peptide \( i \), given some observed behavior for the remainder of the protein. As seen above, this requires knowing (1) the joint distribution of \( t_i \) and \( \tau \), which is \( P(t_i, \tau, \Sigma) \), (2) the marginal distribution of \( \tau \), \( P(\tau, \Sigma) \), and (3) the covariance, \( \Sigma \), between all of the peptides.

The average (or sum) of t-statistics tends towards a normal distribution if the component distributions’ degrees of freedom are sufficiently large. Given the small sample size in our dataset – and most proteomic datasets – this tendency is not always observed. Furthermore, the patent lack of independence among peptides results in a sampling distribution that cannot be calculated analytically (i.e. from the convolution of the individual t-distributions). Thus, we approximate the distribution for \( \tau \) by generating 10,000 bootstrap samples and using maximum likelihood estimation to fit a t-location-scale distribution to the data (fixing the mean at zero); we call the resulting degrees of freedom \( \nu_\tau \) and the scale
parameter $\tau$. While the scale parameter can be estimated robustly, the degrees of freedom is usually overfit to the particular instance of the bootstrap density, resulting in unstable estimates for $\nu$. Thus, we conservatively fix the degrees of freedom, $\nu_\tau$, at 7.92, which serves as an empirical minimum bound for >95% of peptides in our data – regardless of any particular instance of the bootstrap density.

As bootstrap estimation preserves the correlation structure between peptides, this procedure has the added benefit of simplifying the probability as follows:

$$P(\tau_i | \Sigma_i) = \hat{P}(\tau_i)$$

Where $\hat{P}(\tau_i)$ refers to the bootstrap density estimate, and, thus, $\Sigma_i$ does not need to be explicitly calculated as it is inherent in the estimated density.

Similarly, we use bootstrap samples to parametrize the distribution of $\hat{P}(t_\tau, \tau_i)$. We choose to use the bivariate t-distribution as developed by Shaw and Lee as (1) it can be decomposed into the product of two t-distributions as the correlation between the marginals approaches zero (i.e. the property of statistical independence) and (2) it allows us to specify different degrees of freedom on the marginals\(^1\). The latter property is particularly important in many real-world applications, since assuming equidensity of the tails is a strong and often unrealistic assumption. The distribution is specified as follows:

$$P(t_\tau, \tau_i) = Ca^{\nu_\tau 1/2} \alpha_\tau^{1/2} \left( \sqrt{\alpha_\tau} \cdot t_\tau \right)^{\nu_\tau / 2} \cdot \Gamma \left( \frac{\nu_\tau + 1}{2} \right) \cdot \Gamma \left( \frac{\nu_\tau}{2} \right) \cdot \frac{\nu_\tau}{\nu_\tau + 1} \cdot \frac{\nu_\tau + 1}{\nu_\tau} \cdot \frac{1}{4\alpha_\tau} \cdot \gamma^2 + \ldots$$

$$\gamma \cdot \Gamma \left( \frac{\nu_\tau + 1}{2} \right) \cdot \Gamma \left( \frac{\nu_\tau}{2} + 1 \right) \cdot \frac{\nu_\tau}{\nu_\tau + 2} \cdot \frac{\nu_\tau}{\nu_\tau + 2} \cdot \frac{1}{4\alpha_\tau} \cdot \gamma^2 \right)$$
\[
\alpha_1 = 1 + \frac{t_1^2}{\nu_1 \cos^2 \theta}
\]

\[
\alpha_2 = 1 + \frac{t_2^2}{\nu_2 \cos^2 \theta}
\]

\[
\gamma = \frac{2 t_1 t_2 \sin \theta}{\sqrt{\nu_1 \cdot \nu_2 \cdot \cos^2 \theta}}
\]

\[
C = \frac{1}{\pi \cos \theta \cdot \sqrt{\nu_1 \cdot \nu_2}} \cdot \frac{\Gamma \left( \frac{\nu_1}{2} \right) \cdot \Gamma \left( \frac{\nu_2}{2} \right)}{\Gamma \left( \frac{\nu_1 - 1}{2} \right) \cdot \Gamma \left( \frac{\nu_2 - 1}{2} \right) \cdot \sqrt{\frac{\nu_1 - 1}{\nu_2 - 1}}}
\]

Where \( \Gamma(\cdot) \) is the gamma function, and \(_2F_1(\cdot)\) is the hypergeometric function. The “mixing angle,” \( \theta \), of the two marginals is related to Pearson’s correlation coefficient as:

\[
\rho = \sin \theta \cdot \frac{\Gamma \left( \frac{\nu_1 - 1}{2} \right) \cdot \Gamma \left( \frac{\nu_2 - 1}{2} \right)}{\Gamma \left( \frac{\nu_1}{2} \right) \cdot \Gamma \left( \frac{\nu_2}{2} \right) \cdot \sqrt{\frac{\nu_1 - 1}{\nu_2 - 1}}}
\]

\( \nu_1 \) and \( \nu_2 \) represent the degrees of freedom for the two statistics, \( t_1 \) and \( t_2 \), whose joint distribution we are modeling; the distribution is illustrated in Figure 2. For our data, these two statistics are \( t_1 \) and \( \tau_i \), where \( \tau_i \) must be rescaled by \( s_i \). The degrees of freedom for \( \tau_i \) was estimated above as \( \nu_\nu \), and the degrees of freedom for \( t_1 \) can be calculated analytically by the Welch-Satterthwaite equation:

\[
\nu_i = \frac{\left( \frac{s_i^2}{n_i} + \frac{s_\nu^2}{n_\nu} \right)^2}{\frac{1}{n_i - 1} \left( \frac{s_i^2}{n_i} \right)^2 + \frac{1}{n_\nu - 1} \left( \frac{s_\nu^2}{n_\nu} \right)^2}
\]

To solve for the “mixing angle”, \( \theta \), we calculate Pearson’s correlation, \( \rho \), between the \( t_i \) and \( \tau_i \) for each bootstrap sample (10,000 pairs). However, the small sample size in our dataset – as in most proteomic datasets – results in inflated values of \( \rho \) (and \( \theta \)) – a situation...
well-described in the statistics literature surrounding covariance estimation. To account for overestimation of the sample covariance, we construct a regularized estimate of \( \theta \), which we call \( \theta^* \). Regularization of the covariance is particularly important when specifying null bivariate distributions (Student’s \( t \) or normally distributed), as these distributions concentrate along the diagonal as \( \theta \) approaches \( \frac{\pi}{2} \), resulting in an underestimation of the real error – and misleading p-values – in tightly correlated data.

Current methods for covariance regularization introduce a degree of sparsity to the sample covariance matrix by shrinking the eigenvalues to a prespecified target – a process which can be distilled into a linear combination of two matrices:

\[
U^* = \lambda T + (1 - \lambda)U
\]

Where \( U \) is the sample covariance matrix, \( T \) is the target matrix, and \( \lambda \) is the shrinkage parameter. While Schafer and Strimmer outline several target matrices for different applications, the choice of a particular shrinkage estimator requires assumptions that are not always clear to the investigator. In addition, the estimation of \( \lambda \) is based on calculating variances of the sample (co)variances, which is not straightforward in our case, as our estimates are based on bootstrap samples.

To circumvent these issues, we first empirically determine the level of shrinkage optimal for our data, and then use this information to construct the envelope for \( \lambda \). We find the empirically optimal shrinkage as follows:

\[
\theta^*_i = \arg\min_\theta \left\{ \iint \left( f_i(t, \tau; \theta) - \hat{f}_i(t, \tau; \theta) \right)^2 d\tau dt \right\}
\]

Where \( f_i(t, \tau) \) is the analytical bivariate t-distribution of Shaw and Lee; \( \hat{f}_i(t, \tau) \) is the empirical distribution from the bootstrap samples for peptide \( i \); we use the kernel density
smoothed estimate of $\hat{f}(t, r)$ for computational efficiency\(^\text{37}\). Thus, for each peptide, we obtain a new mixing angle, $\theta^*$, that minimizes the squared error. We now pool information across peptides to make inferences about the true value of $\theta$, calculating the binned minimum bound of $l = 1 - \theta^* / \tilde{\theta}$ as follows:

$$\hat{l}(b_j) = \min\left(1 - \frac{\theta^*}{\tilde{\theta}}, b_j - \Delta x \leq \theta, \theta < b_j + \Delta x\right)$$

Where $b_j$ are the bin centers and $\Delta x$ their width; we use $\Delta x = 0.10$. The observed $\hat{l}(b)$ is used to construct an envelope for $\tilde{\lambda}(\theta)$ of the form:

$$\tilde{\lambda}(\theta) = \min\left(1, \frac{\alpha}{\tilde{\theta}^2} + \beta \theta + \kappa\right)$$

Which is suggested by Schafer and Strimmer for use in shrinking the sample covariance to zero identity, and they set $\beta = 0$\(^\text{96}\). While parameterizing with $\alpha = 0.1235$ and $\kappa = 0.1$ and $\beta = 0$ minimizes the squared error to our estimate of $\hat{l}(b)$, we found that this conservative form of shrinkage still results in overestimates of the observed estimate $\hat{l}(b)$, is shown by blue markers, which is fit to the Schafer-Strimmer form of the shrinkage estimate, $1 - \tilde{\lambda}(\theta)$ (dashed black line). We modified this shrinkage estimate by introducing a linear term to bound the estimate in the tails.

Figure 3. Top: peptide t-statistics versus the calculated p-value (based on the statistical framework herein). We see that poorly constrained mixing angles lead to false positive identifications, where a peptide with a small t-statistic is called significant. Bottom: coefficient of shrinkage, $1 - \tilde{\lambda}$, versus the mixing angle, $\theta$. The empirical minimum bound, $\hat{l}(b)$, is shown by blue markers, which is fit to the Schafer-Strimmer form of the shrinkage estimate, $1 - \tilde{\lambda}(\theta)$ (dashed black line). We modified this shrinkage estimate by introducing a linear term to bound the estimate in the tails.
correlation: overestimated correlation (i.e. mixing angle) creates an over-concentrated density along the diagonal, and then small values of \( t \) and \( \tau \) - slightly off-diagonal – result in exaggerated \( p \)-values and false positive identifications (Figure 3, top). Since inflated covariance estimates are problematic only at high levels of correlation, we introduce the linear term in calculating \( \hat{\lambda}(b) \) which imposes an upper bound on the mixing angle. We parameterize \( \hat{\lambda}(b) \) using \( \alpha = 0.025 \) and \( \kappa = 0.17 \) and \( \beta = 0.25 \) as this preserves the structure of the Schafer-Strimmer estimate for small mixing angles, and produces an increasing amount of shrinkage as \( \theta \) approaches 90°. Thus, we calculate the shrunken mixing angle, \( \hat{\theta} \), as:

\[
\hat{\theta} = \theta \left( 1 - \hat{\lambda}(\theta) \right)
\]

Where the target mixing angle is taken to be zero. Implicit in the use of a single function, \( \hat{\lambda}(\theta) \), to estimate the shrinkage for all peptides is the assumption that certain statistical properties, such as the \( \text{Var}(\theta) \), are universally shared. The estimated envelope, \( \hat{\lambda}(\theta) \), is shown in Figure 3 (bottom), along with \( \hat{i}(b) \).

As the null densities \( \hat{P}(t, \tau) \) and \( \hat{P}(\tau) \) are estimated from bootstrap samples – a process that scrambles phenotype membership – the resultant probability of observing a particular peptide, \( \hat{P}(t|\tau) \), is aptly suited to hypothesis testing, with the null hypothesis being “no association with phenotype.” We calculate the \( p \)-value for a peptide as \( \hat{P}(t \geq t_i|\tau) \), i.e. the probability of observing a t-statistic more extreme than \( t_i \), given the observed value for \( \tau \).

**Exon-level Significance**

The framework for evaluating the significance of a single peptide can be extended to study the behavior of exons, provided they have been proteomically probed. To provide a
better consensus estimate for the behavior of particular exon, we limited our analysis to only those with \( \geq 2 \) peptides (regardless of peptide length). As before, the statistic used to represent an individual exon is the average of its \( k \) constituent peptides:

\[
t_e = \frac{1}{k} \sum_{i=1}^{k} t_i
\]

The dependence between peptides prevents the distribution of \( t_e \) from being calculated analytically, and we determine its associated scale, \( \sigma_e \), by fitting 10,000 bootstrap estimates to a \( t \)-location-scale distribution via maximum likelihood. As before, the sensitivity of the degrees of freedom, \( \nu_e \), to the particular instance of the bootstrap distribution results in unreliable estimates, and, thus, we fix \( \nu_e \) at 7.45 to capture >95\% of all exon’s degrees of freedom. We are interested in the probability:

\[
P(t_e | \tau, \Sigma) = \hat{P}(t_e | \tau)
\]

Where \( \tau \) is defined as:

\[
\tau = \frac{1}{n_e - 1} \sum_{i} \tilde{t}_i
\]

\( \tau \) is the average of the \( \tilde{t}_i \) for the \( n_e - 1 \) exons that have daughter peptides measured in the experiment. The specification of the distribution for \( \hat{P}(\tilde{t}_i | \tau) \) follows as above for the peptide-level analysis.

**Statistics of False Discoveries and Multiple Testing**

To evaluate our \( p \)-values, we must calculate rates of false discoveries or errors, i.e. situations where the null hypothesis is incorrectly rejected. An issue unique to LC-MS based analyses is the issue of sampling: peptides are measured semi-stochastically from the complex mixture injected into the mass spectrometer, and, based on the data observed, the
existence of the parent protein is inferred. Additionally, within the mass spectrometer, a peptide’s signal can be suppressed by other, more intense peptides that co-elute. Thus, a false discovery in an LC-MS/MS pipeline may arise due to misassignment of a peptide, or to suppression by other peptides, or from other unknown issues. For a biological unit, \( U \) – representing a protein, exon, or peptide – we can estimate the rate of false positives by modeling “non-biological” units that have a scrambled biological meaning. We synthesize “non-biological units” by assigning random peptides to the unit, \( U \), being tested; these simulated units are called non-biological because the substitution of a random peptide produces a unit that has no clear biological interpretation. To calculate the error rate of a real unit, \( U \), being called significant by the stochastic segregation of signal/noise alone, we find the proportion of “non-biological units” that results in \( p \)-values significant at the pre-specified testing level, \( \alpha \).

For a significant protein (\( p \)-value< \( \alpha \)) with \( n \) peptides in \( e \) exons, we simulate a non-biological protein by randomly collecting sets of peptides (of size \( n \)), summarizing them into \( e \) exons (as before), and calculating significance as was done for the “real” protein. When the unit of interest is a differentially expressed exon, only the peptides within the exon of interest are randomized, keeping \( \tau_e \) constant and requiring recalculation of \( t_e \) and \( \hat{\Theta} \). For differentially expressed peptides, as well, only the peptide of interest is randomized; \( \tau_i \) remains fixed, and \( \hat{\Theta} \) - representing the covariance – is recalculated. For each tested unit, \( U \), we can calculate the proportion, \( \hat{\alpha}_U \), that may arise due to non-biological assignment of peptides and stochastic segregation of signal:

\[
\hat{\alpha}_U = \frac{\# \{ p_{\text{rand},i} \leq \frac{\alpha}{2} \}}{n_{\text{rand}}} \cdot 2
\]
Where $n_{\text{rand}}$ is the number of non-biological units generated (for accuracy, we set $n_{\text{rand}} \geq 1000$); and $p_{\text{rand},i}$ is the $p$-value for the $i^{th}$ random unit. We use $\alpha/2$ since we are using two-sided hypothesis testing here. $\hat{\alpha}_U$ represents the probability of a unit – protein, exon, or peptide – being marked as “significant” based on random segregation of signal in the measured data, and this randomization process takes into account the incorrect assignment of peptides to their parent protein. Though the $\alpha$-level itself is typically used to provide a false positive rate for an individual test, $\hat{\alpha}_U$ serves as a more general false positive estimate as it takes into account the sampling issues unique to LC-MS workflows. Thus, the proportion $\hat{\alpha}_U$ represents the empirical type I error rate (the “hat” signifies that the parameter is estimated empirically). Use of $\hat{\alpha}_U$ also accounts for any misspecifications in the underlying probability model, which is important as the true form of this distribution cannot be known exactly.

Each significant molecular unit, $U$ – a protein, exon, or peptide – will have a unique proportion of false positives, $\hat{\alpha}_U$, as the random $p$-values depend on (1) the $\alpha$-level of the unit being tested, (2) the number of daughter peptides, and, for exons and peptides, (3) the intensities of the sibling peptides.

In multiple hypothesis testing, the probability of making a false discovery increases with an increasing number of comparisons. However, there is no absolute way to quantify or control this rate. While family-wise error rates (FWER) have been decried as being too conservative, their calculation is intuitive and straightforward. False discovery rate (FDR) control, on the other hand, has garnered much attention in the past 15 years for its practical motivation, but its actual calculation is riddled with ambiguity, as the proportion of true null hypotheses must be estimated from the distribution of test statistics or $p$-values. No single
approach is appropriate in all settings, and we demonstrate how error rates must be considered with the validation strategy in mind.

The FWER calculates the probability of making a false discovery among a set, or family, of tests. If the probability of a single test being false is $\alpha$, and $n$ identical comparisons are made (where tests are identical if they are tested at the same level, $\alpha$), then we are interested in knowing the probability that the null hypothesis, $H_o$, is false ($H_o=0$). The probability of all null hypotheses being incorrectly rejected is:

$$P(H^*_1 = 0, H^*_2 = 0, \ldots, H^*_n = 0) = \alpha^n$$

The probability that a null hypothesis is true is $(1-\alpha)$, so the probability of all null hypotheses being correctly rejected, i.e. the probability of no false rejections is $(1-\alpha)^n = P(H^*_i = 1, \forall i)$.

Thus, the probability of there being a false rejection in $n$ comparisons is:

$$P(H^*_i = 0, \text{for at least one } i) = 1 - P(H^*_i = 1, \forall i) = 1 - (1-\alpha)^n$$

A number of alternative strategies consider the rate of false discoveries only among $r$ rejected hypotheses, which has been referred to as a false discovery proportion, or the tail probability for the proportion of false positives. The false discovery rate, FDR, attempts to control the average proportion of false discoveries. The FWER has also been extended to allow for $k$ false positives – $k=1$ in the classical form of the FWER. The deluge of approaches for calculating an experimentally useful and simple parameter – the proportion of false positives – is both overwhelming and misleading, as different approaches can provide vastly different estimates of false discoveries (such as, in our experience, the FDR of Benjamini and Hochberg versus the pFDR or $q$-value of Storey). We argue that the interpretation of false discoveries, and the process by which they ought to be calculated, must depend on the validation strategy employed.
In the analysis of our $Hpgd^+$ data, we are interested in three separate validation strategies: high-throughput validation (targeted LC-MS/MS of differentially expressed peptides and proteins); network analysis (constructing a molecular subsystem from the set of candidate proteins); and one-by-one validation (PCR of alternative splicing candidates). Our high-throughput validation strategy of choice is targeted LC-MS/MS, and – like many high-throughput platforms – does not exhibit fluctuations in either (1) time or (2) cost upon the addition of another peptide/target. In other words, the “price” of making a single false positive claim in the discovery phase is negligible, and calculating the FWER – the probability of making at least one error – serves only as a statistical exercise with no practical value.

In a similar vein, when we construct molecular subsystems from lists of proteins, the resulting subsystem is often robust to perturbations in any individual unit. In network analysis, subnetwork constructions can also be robust against the inclusion or exclusion of any individual node in the seed set. Therefore, the probability of making at least one incorrect claim is an uninformative calculation, as we are, instead, interested in the probability that the majority of the targets or seeds are correct. If we assume a uniform false positive rate for each test/protein, then this latter probability can be restated simply as the number of successes in $r$ rejections, i.e. a cumulative binomial probability:

$$\text{FDP}_H(x, r) = P(\leq x \text{ successes in } r \text{ rejections}) = \sum_{i=0}^{r} \binom{r}{i} p^i (1-p)^{r-i}, \text{ where } p = 1 - \max_{i=1}^{r} \hat{\alpha},$$

We will refer to this as a false discovery probability, $\text{FDP}_H$, where the subscript “H” indicates “high-throughput” validation. In validation approaches – high-throughput technologies or network analysis – where “success” depends on the bulk of the targets being correct, we suggest that statements of the form above are more useful than those made by
FWERs or FDRs. One downside of the above approach is that it relies on a subjective
definition of \( x \). We choose to define \( x \) as \( x = r - \omega \), and use the maximum value of \( \omega \) for
which \( \frac{\omega}{r} \) exceeds a pre-specified proportion (e.g. 0.70, so that 70% of the hypotheses are
correctly rejected). The probability of success for a given set of hypotheses is set as the
maximum empirical type I error rate for the group, \( \max \alpha_{i} \). The complement of this
probability is more easily interpreted:

\[
\text{TDP}_{n} = 1 - \text{FDP}_{n}
\]

Which is a “true discovery probability.” The TDP\(_{n}\) is the probability that at least \( x \) of our \( r \)
targets are correct.

To tune the \( \text{FDP}_{n}(x,r) \), all \( r \) rejections are first sorted in increasing order by their \( p \)-
values. Then, we calculate \( \text{FDP}_{n}(g - \omega, g) \), for increasing \( g \in [\omega, r] \) with fixed \( \omega \). We fix the
probability of success for each nested family of tests as \( \alpha_{i} = 1 - \max \hat{\alpha}_{i} \). Plotting \( \text{FDP}_{n}(g - \omega, g) \)
versus \( p \)-values allows us to choose a comfortable level of \( \text{FDP}_{n} \), and targets with \( p \)-values
outside this range can be removed from the analysis.

The third validation strategy that we will also pursue is the “one-by-one” approach.
For the proteins that are candidates for alternative splicing, we design PCR primers specific
to the spliced exon and look for multiple species via gel electrophoresis. Like other one-by-
one approaches (e.g. Western blots), this strategy is costly and labor intensive, yet necessary.
In these situations, it is critical that we are confident about every candidate we choose to
pursue: we can only pursue a small number of candidates, and a large proportion of false
positives among this small subset would lead to disappointing results that may only reflect
the rate of false positives in the discovery phase, rather than true biological non-existence.
Therefore, we calculate the *false discovery probability for individual validation*, FDP$_I$, using the empirical type I error rate for each exon, $\hat{\alpha}_i$, on ordered $p$-values, $g \leq r$:

\[
FDP_I = P(\text{all } g \text{ hypotheses are not true}) = P(\geq 1 \text{ among } g \text{ is a false positive}) =
\]

\[
P(H'_i = 0 \text{ for at least one } i, \; i \in \{1, \ldots, g\}) = 1 - \prod_{i=1}^{g} (1 - \hat{\alpha}_i)
\]

Where the subscript “I” indicates that validation will be performed “individually,” or one-by-one. This formulation of the FDP$_I$ uses empirical estimates of the local (individual) false positive rates, $\hat{\alpha}_i$, which accounts for any misspecifications of the probability model.

The FDR literature focuses on allowing the investigator to choose a cut-off probability, i.e. “controlling” the level of FDR, the targets below which can be pursued for validation. In practice, however, investigators know *a priori* what a feasible number of targets for validation may be. With one-by-one strategies, the investigator may know that pursuing 5 proteins via Western blotting, for example, is economically feasible. Consequently, he or she is only interested in knowing the probability of an error among those 5 targets, and the above formulation for FDP$_I$ is a straightforward measure in such applications.
Coexpression Network Construction

We hypothesized that the proteomic targets regulated by changes in abundance may be controlled transcriptionally. To uncover transcription factors that may be regulating these targets, we constructed a coexpression network using publicly available microarray data. We first downloaded a diverse set of experiments specific to the mouse intestinal epithelium from the Gene Expression Omnibus\textsuperscript{30}; we limited our search to experiments performed on the Affymetrix Mouse 430\_2 platform. We downloaded a set of 20 experiments spanning a range of experimental perturbations and genotypes, for a total of 206 arrays. Each experiment was normalized independently using dChip, as it has been shown to preserve correlation structure throughout its normalization process\textsuperscript{104, 105}. Outlier arrays were treated as missing data.

Next, the 206 arrays were used to build a bipartite coexpression graph between the proteomic targets and known transcription factors (TFs). We used the curated set of transcription factors found at TfDb\textsuperscript{106}. For genes covered by multiple probes, the average correlation profile of the probes was used as the representative,

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{coexpression_network.png}
\caption{Workflow for the analysis of \textit{Hpgd}\textsuperscript{-/-} vs wild-type (WT) FVB mice. Proteins, exons, and peptides were considered in a hierarchical fashion (Venn diagram, bottom).}
\end{figure}
taking the average across the rows (proteomic targets) first, and then the columns (transcription factors).

To test if a transcription factor has a specific association with the proteomic targets, we used the one-sided Kolmogorov-Smirnov (KS) test. For a given transcription factor, we compared the distribution of the proteomic targets to the background distribution (i.e. between the transcription factor and the rest of the probes on the array). We chose to examine positive correlations only – hence, using the one-sided KS test – as these are more likely to be indicative of the proteomic targets being directly regulated by a transcription factor.

To further test the hypothesis that our set of proteomic targets shares a common regulator, we simulated a null distribution for a transcription factor’s KS statistic by calculating the KS statistic for stochastically generated sets of proteomic targets (of equal size); we calculated \( p \)-values by counting the number of random sets whose KS statistic exceeded the real KS statistic. \( p \)-values – from both the KS test itself and from the empirical null – were corrected for multiple hypothesis testing using the Bonferroni approach.

Figure 5. Genomic width of proteomic data. Peptides measured in our shotgun proteomic data were binned by their genomic width (number of base pairs); the length was also calculated for exons covered by \( \geq 2 \) peptides and genes covered by \( \geq 3 \) peptides (as assigned by Rosetta Elucidator). Each group was normalized by the total number of events observed in that category. Approximating normal distributions are shown as shadows. A curve of theoretical exponentially decaying noise is also plotted.
Results

Genomic Mapping and Differential Expression

10 mice with 6 gel fractions each resulted in 60 RAW files from the LC-MS/MS analysis. In Rosetta, searching the resultant spectra against the IPI mouse database mapped 5701 peptides into 1210 proteins. Mapping these peptides onto their genomic coordinates resulted in the identification of 3813 exons spread across the 1210 proteins. The workflow is illustrated in Figure 4.

We then calculated the significance of differential expression for the 3 classes of events – proteins, exons, and peptides – which correspond to candidates that may be regulated by changes in abundance, alternative splicing, or PTMs, respectively. 42 proteins showed evidence for changes in abundance – i.e. differential expression of the majority of daughter peptides – and a larger set of 81 showed weaker evidence (p-value<0.05). As a differentially expressed protein must contain differentially expressed exons, we focused on those exons whose parent protein did not show evidence for differential expression as candidates for alternative splicing. Hence, the 12 exons (p-value<0.05) do not intersect with the set of 81 proteins with evidence for differential expression (Venn diagram at the bottom of Figure 4). In a similar hierarchical fashion, exons contain differentially expressed peptides, and we focus on the subset of 48 peptides not subsumed by such exons; we used the set of 76 exons with weaker evidence for differential expression (p-value<0.10) for exclusion.

Additionally, we organized the three events by their “genomic width,” i.e. their nucleotide length. The frequency of each event was normalized by the total number of that event observed. When plotted against the logarithm of the number of nucleotides, the three events form three roughly normal distributions (Figure 5). We also plotted a theoretical noise level (not based on real data) which is important in assessing the reproducibility of the three
events. The distribution of peptides, many of which have very short genomic widths, is strongly convolved with both biologic and technical (i.e. the LC-MS/MS pipeline) noise. As measurements are integrated across larger and larger spans of the genome, the noise level should decrease and the corresponding signal increase, allowing us to detect events at larger genomic widths.

Proteomic Targets

Tables I-III (Appendix I) list the proteomic targets discovered for each type of event. As the candidate proteins and exons will be used in network analysis, we wanted to ensure that the majority of these targets are not false positives; similarly for peptides, as they are slated for validation via MassMatrix\textsuperscript{107}. For this application, we use the false discovery probability for high-throughput validation, $\text{FDP}_{\text{ut}}$, tuning the proportion of true discoveries to be $>90\%$ for proteins and peptides ($\omega=4$), and $>80\%$ for exons ($\omega=2$); different numbers of acceptable false positives were required due to the differences in size of the datasets. Based on Figure 6, we removed targets for which the $\text{FDP}_{\text{ut}}$ falls below 30%. For proteins, this led to the removal of 9 proteins with the highest $\hat{\alpha}$, in other words, there is a 70\% probability that more than 29 of the remaining 33 proteins are true. Four exons were removed, and there is a 70\% probability that at least 6 of the remaining 8 exons are true discoveries. 26 peptides were removed, and there is a similarly high (70\%) probability that at least 18 out of the remaining 22 are correct.
Since candidates for alternative splicing are slated for validation via PCR, we calculate the *false discovery probability for individual validation*, \( \text{FDP}_I \), for the set of candidate exons (Figure 7). When we want to be confident that all our targets are correct – we are worried that we have at least one error among the group - then this probability grows quickly with the size of the group, and there is a very high probability (57%) of at least one false discovery being made if all 12 exons are pursued. Thus, we control the \( \text{FDP}_I \) at 10%, which limits our choice of candidates to *Selenbp1* and *Hsd11b2*. The interpretation of a 10% \( \text{FDP}_I \) is that there is a 90% probability that no false discovery has been made among these two choices. The genomic mappings of the proteomic evidence for *Selenbp1* and *Hsd11b2* is shown in Figure 8.

**Network Mining**

Setting a comfortable probability for \( \text{FDP}_H \) (30%) resulted in...
33 proteins with evidence for changes in abundance (i.e. candidates for transcriptional/translational regulation) and 8 alternatively spliced candidates (i.e. 8 proteins each with a differentially expressed exon). We note that while transcriptional changes are one important form of regulation that affects protein abundance, observed changes in protein level can also be attributed to regulation of protein degradation and/or post-transcriptional processing. Given \textit{Hpgd}'s intersection with Wnt signaling, however, we hypothesized that \textit{Hpgd} will have notable transcriptional effects and we chose to focus on this form of regulation alone. We hypothesized that the set of proteomic targets regulated by changes in abundance may be under the control of common transcription factors, and we used the topology of a coexpression network to examine this. To detect common transcription factors directly regulating the group, we limited analysis to only those proteomic targets with unidirectional changes, focusing on the subset of 25 upregulated targets. A murine intestine-specific coexpression network was constructed, spanning 20 different experimental conditions and 206 arrays, yielding gene-gene correlations that are robust to a wide range of experimental perturbations and genotypes. To examine whether a transcription factor (from TfDb) specifically controls our proteomic targets, we used a one-tailed Kolmogorov-Smirnov test between the proteomic-specific correlations and the background correlations (i.e. the correlations between the transcription factor and all the probes on the array). 966 transcription factors were tested at the $\alpha = 0.001/966 = 0.000001$ level, resulting in 46 significant hits with the Bonferroni correction for multiple testing. The 46-by-25 matrix of proteins was hierarchically clustered (complete linkage; euclidean distance), allowing us to identify a subset of 16 transcription factors and 18 proteomic targets with robust correlations (Figure 9). The subset of 8 downregulated proteins did not exhibit significant correlations with any transcription factors, likely due to the small size of the set.
Discussion

Genomic Mapping and Differential Expression

While the proteomics literature has focused on the detection of molecular events with either large or small genomic widths (proteins and peptides, respectively), we demonstrate that LC-MS/MS based workflows offer the ability to probe an unexplored intermediary region: exons. This is illustrated in Figure 5, where exons with ≥2 peptides form a consolidated and notable intermediary distribution between proteins and peptides.

Significance Testing and Error Rates

Calculating the significance of differential expression for the three categories of events – peptides, proteins, and peptides.

Figure 8. Genomic mappings of proteomic evidence for *Selenbp1* (top) and *Hsd11b2* (bottom) alternative splicing. WT peptides shown in red; *Hpgd* knock-out (KO) peptides in blue. The arrow indicates the exon with ≥2 peptides and evidence for local differential expression. The genomic coordinates are for the positive strand, 5’ oriented to the left.
exons, and proteins – yielded sets of 48, 12, and 42 targets, respectively. A less conservative type I error threshold ($\alpha = 0.10$) was required for the analysis of peptides and exons since, as illustrated by Figure 5, these events are more highly convolved with noise in the data. This potentially higher rate of false positives can be addressed, however, by calculating empirical type I error rates, $\hat{\alpha}_u$, which take into account the sampling biases of LC-MS/MS workflows and misspecification of our probability estimates. For all peptides with $p$-values $\leq 0.05$ (two sided test, $\alpha = 0.10$), the average $\hat{\alpha}_u$ was 0.104, with a standard deviation of 0.036. For all exons with $p$-values $\leq 0.05$ (two sided test, $\alpha = 0.10$), the average $\hat{\alpha}_u$ was 0.077, with a standard deviation of 0.033. For the 42 proteins (two sided test, $\alpha = 0.05$), the average $\hat{\alpha}_u$ was 0.059, with a standard deviation of 0.015. Thus, the average values of $\hat{\alpha}_u$ are all close to their theoretical values, indicating that the underlying bootstrap distributions employed are accurate representations of a unique null case: the case when the event observed is due to the random segregation of peptides in the data.

![Figure 9. Patterns of coexpression between 25 upregulated proteomic targets and the most significant transcription factors ($\alpha = 0.001/966$).](image)

Hierarchical clustering reveals a highly coexpressed cluster of proteomic targets (CHCHD3 through RAB1B) driving the association with a cluster of transcription factors (FEM1C through TNFAIP3); clusters are indicated by blue on the dendrogram.
The trend has been to apply “FDR control” to the set of \( p \)-values (or, in our case, the \( \hat{\alpha}_c \)) for the enticing idea of limiting the number of false positives in a high-throughput dataset. However, the actual interpretation of such measures is ambiguous in the laboratory, where a “false positive” has an experimental, rather than statistical, meaning: false positives are those discoveries that would not validate in either (A) a reproduction of the original experiment or (B) an orthogonal experiment to measure the targets. We argue that the blanket application of FDR or FWER control is not appropriate in all settings, as the number of false positives depends on the particular validation strategy. Consequently, we use two measures of the probability of false discoveries among the set of rejected null hypotheses: the FDP for high-throughput validation (FDP\(_H\)), a binomial cumulative probability that informs us about the “truth” of the majority of our targets; and the FDP for individual validation (FDP\(_I\)), which focuses on a single error being made.

**Validation Strategies**

As mentioned, the three sets of proteomic targets – proteins, exons, and peptides – require unique validation strategies. In this work, we pursue differentially expressed proteins using coexpression network analysis (discussed below), an approach aptly suited to uncovering transcriptional regulators of these targets. Proteins with differentially expressed
exons are posited to be candidates for alternative splicing, as local differential expression provides evidence as to the skipping or retention of that particular exon during mRNA processing. After accounting for multiple hypothesis testing, we have slated *Selenbp1* and *Hsd11b2* for further study via PCR. Selecting primers to amplify select portions of these genes around the candidate exons is expected to result in multiple transcripts if alternative splicing exists, and these transcripts can be distinguished via gel electrophoresis.

Finally, differentially expressed peptides are candidates for demonstrating a wide variety of regulation, such as PTMs, as well as noise: incomplete tryptic digestion, partially modified cysteines and/or methionines, ion suppression, etc. Modified peptides are expected to show differences in intensity, as quantitation is typically based on the unmodified species alone for computational efficiency. Searching the candidate peptides against common PTMs in MassMatrix, however, confirmed that these peptide targets are buried deeply in the noise inherent in LC-MS/MS workflows (data not shown). This issue may be particular to our disease model, as PTMs are often transient signal transducers, while the *Hpgd<sup>−/−</sup>* model represents a chronic inflammatory state at equilibrium (2-3 months of age). We believe that modified peptide populations are more likely to be detected in short-term studies where changes in the gene expression profile have yet to be hardwired, such as an experiment searching for acutely (<1 week) perturbed pathways following administration of a tyrosine kinase inhibitor. For experiments designed with this particular target set in mind, an assortment of strong inhibitors must be used to ensure stability of PTMs. Additionally, differential expression of a lone peptide may arise due to misassignment of that peptide, and, as such, experiments designed to target this peptide population ought to focus on unique hits alone (avoiding the use of the Parsimony Principle in peptide assignment).
Signaling Pathways

Intracellular PGE2 must exit the cell through either passive diffusion or through ATP-dependent transport by way of MRP4\textsuperscript{108}. MRP4 levels have been found to be increased in a variety of human colonic tumors, as well as in polyps from \textit{Apc}^\textsubscript{Min/+} mice, with much higher levels of MRP4 expression than PGT expression\textsuperscript{109}. This suggests that the mechanisms for prostaglandin uptake and removal attempt to compensate for upstream changes in prostaglandin synthesis (i.e. \textit{Cox1/2}) or degradation (i.e. \textit{Hpgd}). As PGE2 removal is ATP-dependent, we expect to see an increase in the glycolytic capacity of the tissue in response to decreased levels of \textit{Hpgd}. Indeed, our proteomic results are saturated with such targets: among alternatively spliced candidates, we find LONP (lon peptidase 1, involved in mitochondrial DNA replication), PGD (6-phosphogluconate, of the pentose phosphate pathway shunt), and IDH (isocitrate dehydrogenase, of the citric acid cycle). Among candidates regulated by abundance, ALDOB (Fructose-1,6-bisphosphate aldolase) stands out as metabolic enzyme, and ATP5F1 and ATP5H are members of ATP synthase, both of which are upregulated (1.5 fold change) upon knock-out of \textit{Hpgd}.

After PGE\textsubscript{2} exits the cell, it can then bind to G-protein coupled receptors, EP2 and EP4, on neighboring cells. The subsequent increase in cAMP then triggers a number of signaling cascades, and it was recently shown that PGE\textsubscript{2} can activate RAP1A (i.e. increase the proportion of GTP bound Rap1) in an EP2-dependent manner\textsuperscript{110}. RAP1A operates in a variety of structural pathways: cell-cell adhesion\textsuperscript{111}, migration, and invasion, suggesting that it may play a role in metastasis\textsuperscript{112}.

In addition to RAP1A, we identified 6 additional members of the Ras superfamily: RAB1B, RAB6, RAB11B, RAB5A, RAB14, and RAB18, all of which were upregulated (1.7 average fold change for the group) upon knock-out of \textit{Hpgd}. The RAB family of proteins
play roles in regulating intracellular membrane traffic, with RAB1B\textsuperscript{113} and RAB6\textsuperscript{114} controlling exocytic vesicular transport from the Golgi, and RAB11B, RAB5A, RAB18, and RAB14 playing roles in endosomal trafficking (though these roles are not necessarily distinct)\textsuperscript{115, 116}. These changes suggest the cell compensates for Hpgd gene loss by up-regulation of these Ras superfamily members.

**Pathway Discovery**

The discovery of cohorts of proteomic targets with related function suggested that the group of targets may (1) be under the control of a common regulator and/or (2) operate in related pathways. To uncover common regulators, we used a coexpression network specific to mouse intestinal epithelium to identify transcription factors that may control the expression of our proteomic targets. Though the activity of many transcription factors is transiently regulated post-translationally, use of a coexpression network allows us to uncover transcription factors that show changes in abundance at the mRNA level – a more stable, long-term form of regulation likely to be operative in these three month old mice. Using a Kolmogorov-Smirnov test, we compared the correlations between each transcription factor and the set of measured proteomic targets. Implicit in this approach is the assumed independence of the transcription factors, which is a tenuous biological assumption, and, consequently, we refrain from making group-wise inferences about the resulting set of significant transcription factors, which are expected to be correlated with each other; rather, we focus on identifying individual transcription factors supported by multiple lines of evidence.

We identified a set of significantly correlated transcription factors, which contained a subcluster of 16 targets with higher coexpression connected to 18 of the upregulated
proteomic targets (Figure 9). Then, using DNA-protein interactions (in MetaCore), this set of 18 proteomic candidates was mined for known associations with transcription factors. MetaCore’s proprietary analysis returned a number of candidate subnetworks, with *Hnf4a* showing the most interactions (based on data from ChIP-chip experiments) with our proteomic targets than any other transcription factor. Given *Hnf4a’s* strong correlation with the upregulated proteomic targets, we infer that *Hnf4a* may be upregulated upon knock-out of *Hpgd*, as well.

*Hnf4a* (hepatocyte nuclear factor receptor 4α) is a transcription factor that plays critical roles in intestinal cell differentiation and homeostasis\textsuperscript{117}. *Hnf4a* may form a link between diet and intestinal health, as it was also recently discovered that *Hnf4a* can reversibly bind fatty acids, being neutrally affected by omega-6 fatty acids (i.e. linoleic acid)\textsuperscript{118}, inhibited by omega-3 fatty acids, and activated by saturated fatty acids\textsuperscript{119}. In a landmark connection with inflammation, a genome-wide association study in individuals of European ancestry recently revealed that the most significant polymorphism associated with ulcerative colitis, a common form of inflammatory bowel disease (IBD), lies in the 3’ untranslated region of *Hnf4a*\textsuperscript{120}.

*Hnf4a* appears to plays different roles in inflamed versus neoplastic epithelium. While loss of *Hnf4a* alone increases both cellular proliferation and apoptosis in mouse intestinal epithelium\textsuperscript{121}, concomitant loss of *Apc* leads to a lower polyp burden than observed with wild-type *Hnf4a*\textsuperscript{122}. In colorectal tumor biopsies from 40 patients, *Hnf4a* mRNA and protein levels were both increased on average\textsuperscript{122}. Similarly, in the HT29 human colon cancer cell line, knocking down *Hnf4a* with siRNA or fatty acyl analogues also inhibits tumor growth both in vitro and in xenografts\textsuperscript{119}. These results indicate that *Hnf4a* likely serves as a tumor suppressor in the context of a genetic background that is prone to tumorigenic insults.
While HNF4A is upregulated in tumors, perhaps damping tumor growth via its pro-apoptotic effects, inflamed tissue appears to downregulate HNF4A. In biopsies from ulcerative colitis patients, HNF4A mRNA levels were reduced relative to normal patients (though the number of normal patients analyzed was much smaller than ulcerative colitis patients)\textsuperscript{123}. In a conditional villin-Cre system, Ahn et al. found that inflammatory stimuli reduce the expression of HNF4A and, in turn, mice mutated at Hnf4a have intestinal epithelia more susceptible to damage by such stimuli. In fact, loss of Hnf4a induces an inflammatory state, with elevated levels of a variety of cytokines (e.g. IL-1β, TNF-α, IFNγ)\textsuperscript{123}. As Hpgd\textsuperscript{-/-} is highly inflammatory, HNF4A upregulation in our model may be a compensatory mechanism to offset this state, illustrated in Figure 11.

As mentioned, the Hpgd\textsuperscript{-/-} mouse has an increased susceptibility to colonic neoplasia induced by mutations in Apc, implicating the Wnt signaling pathway in its pathogenesis\textsuperscript{124}. The increased EP2 receptor activation by PGE\textsubscript{2} was shown to result in increased association of the G\textsubscript{α} subunit of the EP2 receptor with axin, releasing GSK-3β and allowing it to become inactivated\textsuperscript{76}. The inactivation of GSK-3β leads to a concomitant freeing of β-catenin (CTNNB1), which can then translocate to the nucleus to activate transcription of the Wnt target genes. In the Hpgd\textsuperscript{-/-} mouse, the loss of Apc is necessary to observe neoplasia, and one hypothesis is that Apc serves, in part, to sequester free CTNNB1 from the activating effects of prostaglandins\textsuperscript{76}. From the coexpression network analysis, CTNNB1 was found to be significantly and strongly coexpressed alongside HNF4A, implying that CTNNB1 is also upregulated upon knock-out of Hpgd. The upregulation of CTNNB1 is likely to occur independently of HNF4A, as loss of HNF4A in conditional mouse knock-outs is followed by an increase in nuclear CTNNB1 in the small intestinal crypts and villi, indicating that HNF4A opposes the transcriptional effects of CTNNB1 (see Figure 11)\textsuperscript{121}. 
The *Hpgd* mouse was first envisioned as a model for colorectal cancer, showing evidence of colonic neoplasia upon crossing with *Apc*<sup>Min/+</sup>, and the discovery of *Hnf4a* underscores the link between colorectal cancer and inflammation. It has long been known that patients with chronic inflammatory bowel diseases, IBDs (Crohn's disease, CD, and ulcerative colitis, UC), have an increased risk of developing colorectal cancer, though the pathogenesis differs from spontaneous colorectal cancer. While *Apc* is considered the “gatekeeper” for sporadic colorectal cancer, *Apc* mutations are infrequent in neoplasias arising from UC patients<sup>125</sup>, and *Tp53* mutations, instead, are correlated with histological progression from predysplastic to neoplastic epithelium<sup>126, 127</sup>. In light of our data, loss of *Apc* in UC may simply be a redundant mutation, given the upregulation of *Ctnnb1* we hypothesize occurs in mouse pro-inflammatory epithelium.

**Summary**

Herein, we have introduced a novel approach to deconvoluting proteomic data, making use of the genome as a scaffold on which peptide-level information should be interpreted. In the analysis of label-free proteomic data from the intestine of the *Hpgd*<sup>−/−</sup> mouse, we identified subsets of proteins regulated by changes in abundance and by changes in splicing. Conservatively limiting the probability of making a single error among the latter category, we present two candidates for alternative splicing – *Selenbp1* and *Hsd11b2* – to be pursued using PCR. For the set of proteins regulated by abundance, we control the false discovery probability of the entire group to distill our targets down to a set of 33 candidates. To uncover common regulators of this set of targets, we constructed a coexpression network specific to the mouse intestinal epithelium, thus revealing a tightly correlated cluster.
of transcription factors. Among these, we found CTNNB1, a well-known player in cancer previously implicated in the pathogenesis of perturbed prostaglandin signaling; and HNF4A, a regulator of many cellular processes whose connections to colorectal cancer and inflammation are still being explored.

Given that HNF4A is a transcriptional controller of several of our upregulated proteomic targets (with tightly correlated mRNA expression patterns), our data suggest that HNF4A itself is upregulated in response to knockout of Hpgd. Though a hypermorphic mouse mutant has not yet been constructed, we hypothesize that constitutively active or overexpressed HNF4A would result in a hyperproliferative and/or less apoptotic epithelium. Additionally, it is known that Hpgd signaling relies, in part, on CTNNB1, and this was confirmed in our unbiased proteomic screen, as CTNNB1 was highly coexpressed with a large cluster of our proteomic targets. Our data suggest that both HNF4A and CTNNB1 are upregulated, and, based on previous mouse models, the upregulation of CTNNB1 likely occurs independently of HNF4A. In fact, we hypothesize that loss of Hpgd leads to a neoplastic predisposition due to the upregulation of CTNNB1, and the increase in HNF4A serves to counterbalance these effects.

As this study was motivated by the hope of uncovering novel regulatory mechanisms – and drug targets – of prostaglandin signaling, Hnf4a is a promising target given its known ability to bind a variety of fatty acids.

Figure 11. Hypothesized mechanism by which 15-prostaglandin dehydrogenase (Hpgd) exerts its influence on both inflammation and neoplasia. A pro-inflammatory state (induced by loss of Hpgd) results in decreased inhibition of both Hnf4a and Ctnnb1. Tcf indicates the T-cell family of transcription factors that induce transcription of Wnt target genes.
If *Hnf4a* truly does play a tumor suppressive role, the difficulty, then, is in engineering a linoleic acid capable of activating the protein. Additionally, designing linoleic acids that are not absorbed beyond the intestinal epithelium offers the hope of creating a tissue specific molecular therapy. Such a therapy would be of immense benefit to UC patients and colorectal cancer patients alike, as it would provide a much needed alternative to the current mainstays of therapy: NSAIDs and coxibs.
Apc Signaling with Candidate Drivers: Two Drivers, Many Paths

Introduction

It is clear that sporadic colorectal cancer – as well as other cancers – is largely the product of acquired somatic mutations. While many of these mutations are functionally relevant to the tumor (“driver” genes), the most well-studied cancer driver gene remains Apc (adenomatous polyposis coli), thought to be the first hit in the majority of nonhereditary colon cancers. While Apc is commonly known as an antagonist to β-catenin and WNT signaling, a growing body of evidence points to the importance of Apc in a variety of cellular contexts – from microtubule polymerization to cell migration. Apc also plays important roles in chromosome segregation and stability, localizing to spindles, kinetochores, and centrosomes in mitosis.

The myriad aspects of Apc signaling may not be relevant in all cellular contexts, however, as signaling depends upon the background gene expression program and, in cancer biology, is often the result of multiple mutations. In fact, mouse models mutated at two driver genes simultaneously have shown a synergistic (i.e. non-additive) increase in tumor burden, such as in Pten-Apc, Kras-Tgfβ, and Apc-Trp53. Such genetic synergy suggests that the pathways emanating from the two genes intersect downstream, supporting the idea that only a subset of all possible pathways are involved in a tumor harboring a mutation in Apc. We hypothesize that these mutations have distinct synergistic effects on the cancer.
phenotype, such that the activities of these networks are greatly associated with the measured downstream changes in the proteome of the intestine. We argue that these measured molecular changes can be leveraged to elucidate which pathways are most relevant to the disease model at hand.

To prioritize the various pathways associated with a cancer driver gene, we have developed a computational framework to predict the set of pathways functionally related to Apc signaling in mouse models. Our algorithm mines chains of proteins (simple paths) from a protein-protein interaction network; these paths are then filtered by tissue-specific mRNA coexpression and Gene Ontology (GO) annotation rule mining\(^\text{138}\). To identify biologically relevant paths, we constrain our search space to pathways connected to previously identified cancer driver (CAN) genes\(^\text{17}\), as many of these pairings are expected to be simultaneously mutated. The set of paths linking Apc to each CAN-gene comprises a subnetwork, which we refer to as a petal in the Apc blossom. As each petal is based on in silico predictions, we then use publicly available functional genomic and proteomic data from the intestine of the Apc\(^{1638N+/-}\) mouse to assess the biological relevance of each petal. As proteins themselves are the mediators of cellular functions, we mapped proteome-level measurements (from 2D-DIGE) to each petal, using mRNA-level coexpression to quantify the strength of the relationship. Though transcriptional activity (i.e. mRNA level) does not strictly correlate with translational activity (i.e. protein level), coexpression information can still be helpful in uncovering regulatory hot spots in protein networks\(^\text{40}\). Testing each petal against such functional data correlates gene and protein expression readouts with specific driver gene relationships, thereby allowing the experimenter to identify the petal most likely to be operative in this particular mouse model. The paper is organized as follows: (1) we introduce the computational framework used to predict signaling pathways between Apc and other CAN
genes, resulting in the Apc blossom; (2) we introduce the concept of testing networks against functional data, specifically proteomic measurements; (3) we then discuss the expected bimodality of mRNA coexpression patterns, and present a novel parameter to measure this property; (4) we conclude with a discussion of the discovered Apc-Hapln1 petal.

Methods

Blooming the Blossom

The Apc blossom is built using the Blossom algorithm and the PathFinder architecture\(^\text{139}\); additional details can be found in Appendix III. A recent study compared various frameworks for detecting signaling pathways, and PathFinder was found to have the best recall rate\(^\text{140}\). In the Blossom algorithm, networks (i.e. pathways) connecting proteins of interest are built by integrating and mining multiple datasets. First, the network of publicly available interactions\(^\text{41, 42, 47, 141}\)
 (>80,000 interactions) is filtered to remove less reliable interactions, i.e. likely false positives, and, then, new interactions are added to enrich the network to account for missing interactions, i.e. false negatives. To remove false positives, we use a logistic regression model that incorporates (i) the number of times a PPI is observed, (ii) coexpression measurements for the corresponding genes, (iii) the protein’s small world clustering coefficient, and (iv) the protein subcellular localization data of interacting partners\textsuperscript{142}.

Coexpression values (Pearson’s correlation coefficient) are calculated from mRNA expression profiles of the laser-capture microdissected epithelium from the $A_{pc}^{Min/+}$ mouse (series GSE422\textsuperscript{143}), providing coregulatory information specific to our tissue and organism of interest. The logistic regression model is trained on positive (1000 PPIs from the MIPS database of interactions\textsuperscript{144}) and negative training data sets (1000 randomly selected PPIs not in MIPS). After accounting for false positives, we impute false negatives based on the observation that proteins with similar sequences share similar interaction partners in the same organism\textsuperscript{145}. Thus, an interaction is inferred between two proteins if no record of interaction exists and there exists at least one interaction between the protein families (based on sequence information) of these two proteins (Pfam release 23.0 used\textsuperscript{146}). Repeating these trials 100 times, an optimized cut off point for the probability of a true interaction is set, and a network of reliable interactions is formed (~ 30K PPIs).

These steps resulted in a filtered network with predicted edges within which we searched for pathways linking APC and CAN-genes. GO Biological Process annotations are used to generate functional association rules from known pathways\textsuperscript{139}. Association rules are tuples representing a noteworthy relationship, in this case functional relationships, between two interacting proteins. For each protein, leaf terms on the GO term graph are used. In addition, the average absolute coexpression is calculated for each path, and paths are then
filtered according to a set threshold ($\gamma=0.6$). The $p$-value is calculated for a given path with the null hypothesis being that *every* path has a large number of association rules associated with it, but the average number of rules on these paths is uniform. Significant ($p_{threshold}<0.05$) paths are merged into a subnetwork, thus representing a *petal* in the *Apc blossom*. An empty set is returned when there is not a significant path.

Formally, let $G(V,E)$ denote the PPI network gathered from publicly available interactions. Also, let $G'$ and $G''$ be networks built on the same set of nodes, $V$, using the procedures outlined above, where false positive interactions, $F$, are removed, $E'=(E\cdot F)$, to obtain $G'(V,E')$, and a set of additional interactions, $H$, are imputed (based on sequence information) to obtain $E''$ and $G''(V,E'')$.

The objective of the proposed framework is to find a petal for a given protein $c_a$ (e.g. *Apc*) and each protein $c_i$ in the candidate set of proteins $c_i \subset V$. To reduce the search space, the Blossom algorithm employs a network diameter heuristic. Namely, for each node pair ($c_a$ and $c_i$), let $d_{ij}$ denote the shortest path between them in $G(V,E)$, i.e. the graph without inferred edges. We then search $G'(V,E')$ for every path that connects $c_a$ and $c_i$ with path length smaller than $d_{ij}$. This guarantees at least one path for consideration if the two nodes are connected.

The paths on the network are discovered using all paths depth first search (*AllPathDFS*), where every path connecting $c_a$ and $c_i$ that is less than $d_{ij}$, $\Phi$, is identified. In the final step of the algorithm, these paths are compared against the null distribution for significance. For the shortest path calculation, a single-source shortest path solution is used (i.e. Dijkstra’s algorithm). The Blossom algorithm’s run time is the same as the all-paths depth first search: $O(n^{d_{max}})$ where $d_{max} = \max_{c_i \in C} d_{ij}$. The algorithm is shown in Figure 2.
Figure 2. The Blossom algorithm that returns the blossom network for protein $c_a$.

Plucking Petals

For a particular petal, a single node perturbation (e.g. a mutation at $Apc$) within the petal itself will perturb pathways that are expected to associate with the given petal more strongly than others, assuming that the network predictions were accurate. To identify the “best” petal in the $Apc$ blossom, we employed a mouse mutant, $Apc^{1638N+/-}$, representing a perturbation at the stamen. As proteins are the ultimate mediators of function, targets from proteomic experiments – such as label-free, or, in our case, 2D DIGE – represent an ideal dataset for assessing the downstream effects of such perturbations. However, proteomic technologies often sample the most abundant quartile of proteins\textsuperscript{147}, while cancer network predictions – such as those in the $Apc$ blossom – often focus on low-abundance signaling.
proteins. To make inferences about these petals, a relational map must be used to connect the proteomic targets to the petal of interest. Coexpression networks are currently the most informative and accessible mapping available, as proteins correlated at the mRNA-level are hypothesized to be coregulated. Thus, for a hypothesized petal, $P$, mRNA coexpression (Pearson’s correlation coefficient) was calculated between the nodes, $p \in P$, and the 2D-DIGE targets, $d \in D$ (where $D \subseteq S$, and $S$ is the set of all genes on the array) measured in the $Apc^{1638N+/}$ mouse intestinal epithelium. The 2D-DIGE targets’ Mascot DAT files are available through the Proteomics Identifications Database, accession number 10638. The corresponding microarray data have been made publically available through the Gene Expression Omnibus, accession number GSE19338. Two fractions, representing crypts and villi, were available with 4 samples in each group (8 samples each, wild-type and $Apc^{1638N+/}$). The proteins identified within each fraction were pooled to arrive at a set of 31 2D-DIGE targets. For calculating mRNA differential expression, Robust Multiarray Averaging was used to normalize the samples; differential expression was calculated between the 8 mutant samples versus the 8 wild-type samples. For coexpression, the wild-type and $Apc^{1638N+/}$ microarray data were normalized by dChip to avoid artificially inflating coexpression values.

Additionally, mRNA coexpression is more informative for nodes that are known to be differentially expressed, as these nodes are regulated differently between wild-type (WT) and mutant (MT) tissue; a node with low differential expression may have many coexpression linkages simply due to its uniform expression profile, which is shared by the majority of genes (as most genes are not differentially regulated). To focus on genes with strong levels of both coexpression and differential expression, we compute the active coexpression as follows:
\[ \vec{r}_i' = \alpha_i \cdot \vec{r}_i \]

Where \( \vec{r}_i \) is the vector of coexpression between node \( i \) (in petal \( P \)) and all other genes on the array; \( \alpha_i \) is the activity of node \( i \), defined as the scaled, absolute differential expression:

\[ t_i = \frac{\mu_{Apc,i} - \mu_{WT,i}}{\sqrt{\frac{\sigma_{Apc,i}^2}{n_{Apc,i}} + \frac{\sigma_{WT,i}^2}{n_{WT,i}}}} \]

\[ \alpha_i = \frac{|t_i|}{\max|t|} \]

Where \( \mu_{MT,i} \) is the average expression of a gene, \( i \), across the samples in the mutant, \( MT \) (in our case, \( Apc^{1638N/+} \)), and \( \sigma^2 \) is the associated variance; these parameters are defined respectively for the wild-type (WT) samples. The active coexpression matrix, \( R'(P,D) \), between a given petal, \( P \), and the 2D-DIGE targets, \( D \), is then vectorized, \( \text{vec}(R'(P,D)) \). The distribution of \( \text{vec}(R'(P,D)) \) is expected to be leptokurtic (i.e. higher peak, fatter tails), as it is a product of a normal and a folded normal distribution.

Figure 3. (A) Active coexpression, \( \vec{r}_i' \), values for the expected and sample distributions for a petal. Note the extreme peakedness of the data around the origin. (B) The cumulative distribution functions (CDFs) of the expected and sample values are shown. This particular petal has more negative and positive active correlations to the 2D-DIGE targets than to the rest of the genes on the array. To assess the bimodality, \( \beta \), of the data, we examine the area between the two CDFs, as well as the center of mass of this area.
distribution (see Figure 3B). With coexpression measurements, we are particularly interested in the tails of the distribution, as these are expected to exhibit two modes – one positive and one negative – if subgroups of coexpressed 2D-DIGE targets exist. Thus, we developed a measure of bimodality, \( \beta \), borrowing concepts from engineering physics:

\[
\Delta F_p(x) = F_{p,D}(x) - F_{p,S}(x)
\]

\[
\beta_p = l_{<0} \int \Delta F_p(x) dx + l_{\geq 0} \int \Delta F_p(x) dx
\]

\( F_{p,D} \) is the empirical cumulative distribution function (CDF) for the vectorized \( R'(P, D) \) over the range of active coexpression values, \( x \); \( F_{p,S} \) is the empirical CDF for the vectorized matrix \( R'(P, S) \), i.e. the expected active coexpression to all genes on the array; and the sample deviation, \( \Delta F_p \), is simply the difference of the two CDFs. \( l_{<0} \) is the moment arm of the distribution defined classically as:

\[
l_{<0} = \frac{\sum_{x<0} \Delta F_p(x) \cdot x}{\sum_{x<0} \Delta F_p(x)} = \frac{\int_{-\infty}^{0} \Delta F_p(x) \cdot x \, dx}{\int_{-\infty}^{0} \Delta F_p(x) \, dx}
\]

And \( l_{\geq 0} \) is defined similarly for \( x \geq 0 \). Thus, \( l_{<0} \) and \( l_{\geq 0} \) represent the centers of mass for the negative and positive active coexpression values’ deviation from the expected distribution (Figure 3). The bimodality, \( \beta_p \), then, is simply the torque of the distribution, \( \Delta F_p(x) \), around the center: negative values of \( \beta_p \) indicate a clockwise skewing of the tails, with greater mass distributed at extreme (high and low) values of \( r' \) than the background; positive values of \( \beta_p \) indicate a counterclockwise skew, where the sample distribution is more leptokurtic than the
background, and, hence, possesses less correlation than expected. Further insight can be gained by noting that the denominator of the center of mass, $l_o$, cancels out, leaving:

$$
\beta_p = \int_{-\infty}^{\infty} \Delta F_p(x) x \, dx + \int_{0}^{\infty} \Delta F_p(x) x \, dx = \int_{-\infty}^{\infty} x \left( F_{p,0}(x) - F_{p,S}(x) \right) \, dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{p,D}(y) - f_{p,S}(y) \, dy \, dx
$$

Changing the order of integration allows us to formulate $\beta_p$ in terms of the probability density functions (PDFs) of our targets, $f_{p,D}(x)$, and the background, $f_{p,S}(x)$:

$$
\beta_p = -\frac{1}{2} x \left( f_{p,0}(x) - f_{p,S}(x) \right) \, dx = -\frac{1}{2} \left( E(x_{p,0}^2) - E(x_{p,S}^2) \right)
$$

Where $E(\cdot)$ indicates the expectation. Thus, we see that $\beta_p$ is the difference between the second moments of the two distributions (or the difference of their variances, if both distributions are centered at zero). While this ultimate formulation of $\beta_p$ is statistically simple, we present the initial formulation – in terms of the center of mass and torque – to provide an intuitive understanding of its motivation and meaning. As mentioned, we use the empirical CDF/PDF to calculate $\beta_p$. We calculated the significance, $p$, of $\beta_p$ for a network-petal, $P$, as follows:

$$
p = \frac{\# \beta_{\text{rand}} < \beta_p}{\# \beta_{\text{rand}}}
$$

With $\beta_{\text{rand}}$ being the bimodality for a randomly selected set of candidate 2D-DIGE targets; 10000 such sets (of cardinality equal to that of $P$) were generated. The $p$-value, then, represents the probability of attaining a value of at least $|\beta_p|$ via stochastic generation of 2D-DIGE targets.
Results and Discussion

Blossoming \textit{Apc}

To study CAN-gene pathways operative in the \textit{Apc}^{1638N+/-} mouse model, we used the Blossom algorithm to identify pathways connecting \textit{Apc} to 68 other CAN genes. First, likely false positives from the underlying PPI network are filtered out. Next, using this filtered PPI network, we were able to find paths linking \textit{Apc} to 42 of the CAN-genes, forming subnetworks, which we refer to as petals. After imputing interaction edges using sequence homology, this number was increased to 65. However, filtering out paths whose (i) average mRNA coexpression was low (r < 0.6), a significance threshold validated in similar studies\textsuperscript{149} and (ii) support of GO annotation association rules based on known signaling pathways and functional annotations \textsuperscript{[11]} was weak (p-value>0.05), the number of \textit{Apc}-CAN-gene petals was reduced to 24 (Figure 4). The petals identified vary in the number of nodes (from 3-35) and edges (from 2-80) they

Figure 4. The \textit{Apc} blossom. Using the Blossom algorithm, we search for paths in the filtered and imputed PPI network that connect \textit{Apc} to other CAN-genes. For the CAN-genes that possess at least one path to \textit{Apc}, this resulted in 24 petals (p < 0.05) – one petal for each CAN-gene.
contain, with some nodes being shared among multiple petals. A blossom can be constructed for a wide variety of genes, with the stipulation that corresponding microarray data is available. In our case study of Apc, we employ mRNA expression data from intestinal tumors harvested from Apc\textsuperscript{Min/+} mice. As multiple mutations are present in these samples, coexpression measurements calculated for this dataset are representative of the tumor microenvironment; as such, both Apc signaling, as well as additional CAN-gene signaling, are likely to be active simultaneously. While the presence of these multiple, active pathways increases the signal associated with cross-talk within in each petal, it does not allow us to determine which pathways are most strongly associated with Apc signaling alone. To answer this question, as outlined in the next section, we used mice with a particular heterozygous mutation in Apc - 1638N - that results in a mild intestinal cancer phenotype, thereby minimizing the noise arising from the many pathways activated in a full-blown tumor. Since we are interested in assessing the systems-level effects of such mutations, we focus on measuring the downstream effects of these genes via ‘omic experiments.
Plucking Petals

The \textit{Apc}^{1638N+/-} mouse model represents a perturbation of the stamen in the \textit{Apc} blossom, and such a perturbation is expected to have far-reaching molecular effects. This was supported by the 2D-DIGE proteomic experiments that identified 31 proteins with a variety of cellular functions from the intestinal epithelium compared to wild-type. We hypothesized that if one of the petals in the blossom truly reveals signaling of this mutation in \textit{Apc}, then the nodes of this petal are more likely to associate with the 2D-DIGE targets than a random group of proteins. To gauge this association, we used a map of coexpression compiled from the corresponding \textit{Apc}^{1638N+/-} intestinal epithelium mRNA-expression profile. Assuming that the signaling molecules in a petal are upstream of the 2D-DIGE targets, strong coexpression between a petal and the proteomic targets can help to identify the

![Figure 5. Top three significant \textit{Apc} petals. The top three petals that are found to be significant after coexpression correlation with proteomic targets are shown. The \textit{Apc-Hapln1} (left, \(p = 0.0068\)), \textit{Apc-Kras} (middle, \(p = 0.0157\)), and \textit{Apc-Prkdl1} (right, \(p = 0.0167\)) petals are shown. The darker nodes represent CAN-gene proteins. When searching for paths, CAN-genes were not differentiated from other proteins on the network, hence multiple CAN-genes exist in some of the petals. The dashed edges represent novel interactions predicted to exist on the network, whereas solid edges are known interactions.](image-url)
causative signaling events that led to these measured changes in the abundance of the proteome. Since coexpression is most informative when it relates to differentially expressed nodes (i.e. those that differ between the mutant and wild-type mice), we modulated the coexpression values associated with the nodes in each petal by their respective levels of differential expression. This allows for the identification of groups of nodes where any individual node may have a low level of expression, but the collective level of expression across nodes may be high; this is seen in Figure 6, where the most significantly \( p=0.0068 \) coexpressed network is shown. We further posited that, if a group of proteins truly is coregulated, then we expect to see deviations in the tails of the coexpression distribution within when compared to the expected (background) distribution. To gauge this deviation, we introduced the bimodality, \( \beta \), of coexpression: a measure based on the mass (i.e. area under the curve) of the cumulative distribution functions’ and the distance of the mass from the origin; this is illustrated by data from a sample petal in Figure 3. In Figure 6, the 31 2D-DIGE targets are shown on the periphery of the petal, ranked by their degree (i.e. sum) of absolute coexpression. This representation also facilitates the prioritization of 2D-DIGE targets, placing emphasis on those targets whose regulation is supported by multiple elements of the candidate subnetwork. Much of the coregulation can be explained by a few key signaling intermediates – notably, TGFB1, which has both a high level of differential expression, as well as strong coexpression links. Signaling molecules like TGFB1 are hypothesized to lie upstream of ‘omic measurements, and, thus, the petal at the heart of Figure 5 represents a potential set of intermediaries by which the signal arising from a mutation in \( \text{Apc} \) blossoms into proteome-level manifestations (i.e. the 2D-DIGE targets).
Summary

To understand how a mutation affects information flow in a tumor, one must consider both the proximal and distal signaling effects. Proximally, a mutation in a gene may result in a truncated protein product that affects physical interactions, or it may result in a hyperphosphorylated and active state. These small, upstream effects are then amplified and result in distal changes in signaling, affecting mRNA and protein levels of tens to hundreds of seemingly unrelated nodes. While the field of cell signaling is adept at dissecting the proximal effects of a mutation – mechanistically mapping out perturbed pathways – it has not yet developed the tools to fully understand the distal effects and, more importantly, their...
connection with more proximal signaling. Indeed, currently available commercial software suites for network analysis can only associate these distal effects amongst themselves, with no regard to the upstream causative mutation. In this study, we present a method by which the distal effects measured in two ‘omics experiments – microarray and proteomics – can be simultaneously leveraged to test network-based hypotheses. After testing the hypotheses (i.e. petals) against proteomic evidence, the refined petal subnetworks we present not only reveal the relationship between upstream genetic interference and downstream proteomic effects, but they also allow us to prioritize other cancer-driver genes that are likely to act cooperatively with $Apc$ to drive tumorigenesis. This new approach – linking in silico predictions with experimental measurements – provides a way forward in mining context-specific pathways that may prove useful in identifying pathways active in individual cancer patients.
**Apc-Cdkn1a Signaling:**
*Two Drivers, a Single Path*

**Introduction**

Recent large-scale sequencing efforts revealed that any given colorectal tumor includes (on average) 80 mutations, with as many as 15 lying in frequently mutated “driver” genes. In support of the hypothesis that these key genes function cooperatively in driving tumorigenesis, mouse models mutated at two driver genes simultaneously have shown a synergistic increase in tumor burden, including: *Pten-Apc*\(^{135}\), *Kras-Tgfβ*\(^{136}\), and *Apc-Trp53*\(^{137}\). The evidence of synergistic, i.e. non-additive, increases in tumor burden suggest that the signaling pathways of two mutated genes may intersect downstream, and, thus, predicting and interrogating these points of intersection – *as a biological network* – is of significant interest. To trace the connections between genes, a variety of high-throughput datasets – e.g. protein-protein interactions (PPIs), gene coexpression, and transcription factor relationships – have been employed to infer functional associations that lend themselves to analysis as networks, in which each gene or protein is represented as a node and an interaction as an edge. Furthermore, network-based analyses can be used to identify biomarkers\(^{24}\), to predict tumor progression\(^{150}\), or to reveal the molecular alterations underlying disease\(^{151}\).

However, our current knowledge of biological networks is far from complete. The coverage of current interactome databases is estimated to be less than 10% of the total number of interactions\(^{152}\). Thus, when interpolating the connections between driver genes,
network-based analyses that rely solely upon confirmed interactions may lack essential connections. As one goal of our research is to predict and analyze the functional paths between driver genes, a critical step is to develop a predictive framework to infer and evaluate novel connections between genes. The framework proposed here (modeled on Pathfinder\textsuperscript{139}) infers missing edges using predictions from protein family relationships and filters these paths based on known association rules. On the other hand, since a cancer gene participates in multiple signaling pathways, there may be dozens – if not, hundreds – of paths by which two proteins functionally interact. Thus, a computational approach is required to limit the network space to the specific biological context of interest. To extract functionally relevant subnetworks, the framework detects highly probable signaling pathways based on gene-gene mRNA coexpression and Gene Ontology\textsuperscript{138} association rules mined from published pathways.

We used this computational method to elucidate the connections between a well-known driver gene of intestinal cancer, \textit{Apc} (\textit{adenomatous polyposis coli}), to another gene also involved in cancer, \textit{Cdkn1a} (previously known as \textit{p21}). Though \textit{Cdkn1a} was not found to be mutated in populations of human colorectal cancers studied to date\textsuperscript{17}, its expression level correlates with neoplastic progression and has a prognostic value greater than even that of \textit{Trp53}\textsuperscript{153}. Further supporting its importance in neoplasia, the double mutant mouse, \textit{Apc}\textsuperscript{1638N+/; Cdkn1a/−}, exhibits a synergistic increase in its tumor burden\textsuperscript{19}. After predicting the network linking \textit{Apc} and \textit{Cdkn1a}, we evaluated the relevance of these predictions by manipulating the underlying system: generating \textit{in vivo} network perturbations in two mouse models, followed by systems-level 'omic measurements from the small intestinal epithelium. The 'omic measurements – both proteomic and genomic – of the perturbed system were
used for the statistical testing of the predicted network, thus introducing the concept of evaluating *in silico* predictions against context-specific biological data.

**Methods**

**Network Analysis Framework**

The network analysis framework (illustrated in Figure 1, and explained in Appendix III) employs the PathFinder architecture outlined previously. The raw network of publicly available physical interactions is first pruned of false positives using a logistic regression model that incorporates (i) the number of times a PPI is observed, (ii) the Pearson correlation of expression measurements for the corresponding genes, (iii) the proteins’ small world clustering coefficient, and (iv) the protein subcellular localization data of interacting partners. Positive (1000 PPIs from the MIPS database of interactions) and negative training data sets (1000 randomly selected PPIs that are not in MIPS) are used in 1000 cross-validation trials to acquire the parameters that maximize the likelihood of a true interaction.

False negative interactions are inferred using sequence homology relationships. It was observed that proteins with similar sequences share similar interaction partners in the same organism, and, thus, proteins from the same family are also likely to have similar interaction patterns. The Pfam database, utilizing multiple sequence alignments and hidden Markov models (HMMs), uses sequence similarity to formulate protein family classifications and serves as a useful tool for exploiting these relationships. Hence, we inferred an interaction edge if (i) two proteins do not interact with each other in the PPI network, and (ii) there exists at least one interaction between the families of these two proteins.
To identify those paths relevant to our model system of interest, coexpression data based on microarray experiments from the Apc\textsuperscript{Min/+} mouse small intestinal epithelium were obtained from the Gene Expression Omnibus (series GSE422145); this study used laser-capture microdissection to sample the crypts of adenomas, carcinomas, and normal epithelium. In our implementation, we used Pfam release 23.0\textsuperscript{146} and the Gene Ontology release in August 2008\textsuperscript{138}. The search algorithm was extended to find pathways up to 6 nodes in length, and the threshold for the average coexpression of pathways was $|\rho| = 0.70$. 

Figure 1. Framework for prediction of driver gene networks. The process begins with a two-step filtering process to account for false positives and false negatives in interaction databases. After selecting the driver genes of interest, pathways are predicted and then pruned using both GO term association rules and gene-gene coexpression values. Finally, the significant pathway segments are merged to arrive at a network connecting the two driver genes. The framework incorporates tissue-specific mRNA coexpression at two levels: in the pairwise filtering of false positives; and in the filtering of paths by average coexpression. The logistic regression model is trained on gold-standard interactome databases (see Appendix III for additional details).
Mouse Intestinal Epithelium Isolation

All animals were handled in strict accordance with good animal practice as defined by the relevant national and/or local animal welfare bodies, and all animal work was approved by the Institutional Animal Care and Use Committee (IACUC) of Albert Einstein College of Medicine (permit number 20070805). \(Apc^{1638N/+}\) and \(Cdkn1a^{-/-}\) C57BL6/J mice were generated as described previously \(^{19}\) and tissue samples were harvested using the method outlined by Weiser et al., resulting in crypt and villus populations of cells from the small intestine of \(Apc^{1638N/+}\), \(Cdkn1a^{-/-}\), and wild-type mice \(^{155}\).

2D Differential In Gel Electrophoresis

2D Differential In Gel Electrophoresis (2D-DIGE) was performed as previously described \(^{147}\). Differentially expressed proteins from crypt and villus fractions were identified in the mutant mice (\(Apc^{1638N/+}\) and \(Cdkn1a^{-/-}\)) relative to the respective fractions from wild-type mice (4 replicates each). Univariate t-tests (unequal variances and equal sample sizes) and multivariate linear regression (coded in the R package LIMMA \(^{156}\)) were performed. Gel spots were selected for LC-MS/MS identification based on these two t-statistics at the 0.05 level of significance.

Gel spots were excised, trypsin digested, and the peptides were subsequently analyzed by tandem LC-MS/MS on a LC Packings/Dionex Ultimate 3000 HPLC-Orbitrap XL (Finnigan, San Jose, CA) system \(^{147}\). For interpretation of the MS/MS spectra, the MASCOT software package was used to search the SwissProt database; a null database of reversed peptide sequences was searched simultaneously to account for false positives.
Identified proteins are listed in Appendix II. Mascot DAT files have been made publicly available through the Proteomics Identifications Database [148], accession number 10638.

**Gene Expression Profiling**

Microarray studies for crypt and villus populations from \( Ap\text{-}1638N^{+/-} \), \( Cdkn1a^{+/-} \), and wild-type mice (4 replicates each) were conducted on Affymetrix Mouse Genome 2.0 chips according to published procedures [157]. All data is MIAME compliant and the raw data have been made publicly available through the MIAME compliant database, the Gene Expression Omnibus [30], accession number GSE19338.

**Network mRNA Analysis**

Raw .CEL files were processed in MATLAB using the Robust Multiarray Averaging procedure [158]. To deal with multiple probes capturing different aspects of a gene product’s behavior, we used all probes to represent a gene. Thus, in the following analysis, each \( Ap\text{-}Cdkn1a \) network node, \( i \), was represented by \( k_i \) probes on the array, resulting in a matrix of size \( q \times n \), where \( q = \sum_{i=1}^{p} k_i \) and \( p = 20 \). To determine whether the \( Ap\text{-}Cdkn1a \) network nodes were collectively differentially expressed in a tissue compartment (crypts or villi), we extended Hotelling’s \( T^2 \) statistic – a classical approach useful for testing gene groups [159] – to incorporate multiple experiments, as follows:

\[
V^2 = \left( \overline{x}_{Ap} - \overline{x}_{WT} \right) S_{Ap}^{-1/2} S_{Cdkn1a}^{-1/2} \left( \overline{x}_{Cdkn1a} - \overline{x}_{WT} \right)'
\]

Where \( \overline{x}_G \) is the vector of mean mRNA intensity for all the \( q \) probes for a genetic background, \( G \), where \( G \in \{ Ap\text{-}1638N^{+/-}, Cdkn1a^{+/-} \} \).
Cdkn1a<sup>−/−</sup>; and WT indicating wild-type C57BL6/J). S is the absolute value of the unbiased pooled sample covariance matrix for each mutant:

\[
S = \frac{1}{n_{\text{Mutant}} + n_{\text{WT}} - 2} \left[ \sum_{i=1}^{n_{\text{Mutant}}} (x_{i,\text{Mutant}} - \overline{x}_{\text{Mutant}})(x_{i,\text{Mutant}} - \overline{x}_{\text{Mutant}}) + \sum_{i=1}^{n_{\text{WT}}} (x_{i,\text{WT}} - \overline{x}_{\text{WT}})(x_{i,\text{WT}} - \overline{x}_{\text{WT}}) \right]
\]

Where Mutant can refer to either Apc<sup>1638N+/−</sup> or Cdkn1a<sup>−/−</sup>, and the absolute value in S is used to avoid imaginary components when taking the inverse root of S in \(V^2\). It should be noted that probes corresponding to Apc and Cdkn1a themselves were excluded, as these are expected to have extremely low intensity values (in the respective mutants) that would skew the perceived aggregate network effect. In \(V^2\), the difference of means, \((\overline{x}_{\text{Mutant}} - \overline{x}_{\text{WT}})\), for each mutant may be positive or negative for a probe \(i\), so, unlike \(T^2\), \(V^2\) can be either positive or negative.

Given that \(p > n\), sample covariance estimates are not positive definite, and hence, the inverse is singular. To circumvent this issue, we set all covariances to zero for initial calculation of \(V^2\) and then calculate the significance of \(V^2\) using a permutation test (i.e. stochastically generating new “mutant” and “wild-type” phenotype labels), thus preserving the underlying covariance structure in the null distribution. Setting the off-diagonal elements of S to zero simplifies \(V^2\) to:

\[
V^2 = \left\{ t_{\text{Apc}}, t_{\text{Cdkn1a}} \right\}
\]

Thus, \(V^2\) is simply the sum of the product of scaled t-statistics calculated for each probe, in each of the two experimental perturbations. As the number of samples was small (\(n = 4\) for mutant and wild-type, each), random \(N(0, \sigma_q)\) noise was added to each permutated matrix to obtain an interpolated and smoothed empirical null distribution; the standard deviation, \(\sigma_{q,G}\), of the noise for each probe, \(q\), in the genetic background, \(G\), was estimated by the
sample standard deviation of each probe. 10000 such permutations were calculated to obtain the null distributions, which, as expected, resemble F-distributions. Since Apc and Cdkn1a are both tumor suppressors and hypothesized to affect our network of interest in a similar fashion, we expect the t-statistics to vary in the same direction if the null hypothesis (of no joint effect) is to be rejected. Hence, we compute the \( p \)-value of \( V^2 \) as the number of null observations greater than our observed value of \( V^2 \). Calculating the \( p \)-value for the negative tail of the distribution would be useful if the perturbations were expected to have opposite molecular effects (e.g. Apc+/− paired with a Stat3+/− hypomorph).

While we present an analysis for a 2-node perturbation of a network, this analysis is extensible to \( k \) experimental perturbations by computing pairwise \( V^2 \) statistics, resulting in a matrix:

\[
V^2 \text{ Matrix} = \begin{bmatrix}
V^2_{11} & V^2_{12} & \cdots & V^2_{1k} \\
V^2_{21} & \ddots & \vdots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
V^2_{k1} & \cdots & \cdots & V^2_{kk}
\end{bmatrix}
\]

Where \( V^2_{jk} \) represents the \( V^2 \) statistic between perturbations \( j \) and \( k \); as shown, the diagonal reduces to a scaled version of Hotelling’s \( T^2 \) statistic for each experiment. As the statistics are each of a different scale, they cannot be compared directly, and, therefore, the significance of each matrix element should be calculated (as above) via a permutation test. Then, for the matrix of \( p \)-values, the diagonal elements provide information about the significance of individual experiments, while the off-diagonal values provide information about pairwise experimental significance. The total experimental support for network perturbations can then be calculated by aggregating off-diagonal \( p \)-values, e.g. by Fisher’s method. We
recommend this approach for dealing with $k \geq 3$ perturbations; for $k = 2$ perturbations, as in our case, the $p$-values can be interpreted directly.

**Analysis of Proteomic Targets**

To assess the importance of physical proximity, the topological distance between $Apc-Cdkn1a$ network nodes and the respective proteomic targets was calculated. Physical PPI networks were assembled from BioGRID \(^{42}\), the Human Protein Reference Database (HPRD) \(^{41}\), and IntAct \(^47\). Each network node was tested independently for the number of 2-hop paths connecting it to a set of $n$ experimentally measured proteins, expressed as follows:

$$\sum_{k=1}^{n} d_{ik}$$

Where $a_{ij}$ is the entry at row $i$ and column $j$ in the adjacency matrix, $A$, of the PPI network; $i$ is a protein in the $Apc-Cdkn1a$ network; $j$ is an intermediate protein; and $k$ is an experimentally measured protein. In this case, the experimental proteins were the proteomic targets from either $Apc^{1638N+/-}$ or $Cdkn1a^{1/-}$ mice. If there is at least one intermediate protein, $j$, for which a two-hop path exists between nodes $i$ and $k$, then the 2-hop distance, $d_{ik}$, is 1; the total connectivity, $\eta_i$, of protein $i$ to the set of 2D-DIGE targets is simply the sum of the $d_{ik}$.

Significance was calculated against an empirical null formulated from 10000 randomly generated sets of proteins also of size $n$.

To assess patterns of coregulation, mRNA coexpression values (Spearman’s correlation coefficient) were calculated from the corresponding set of normalized microarray experiments, spanning wild-type, $Apc^{1638N+/-}$, and $Cdkn1a^{1/-}$ crypts and villi; the probe with
maximum intensity was used as the representative for a gene. To test the significance of mRNA-level correlations, a modified Kuiper’s test statistic, $K$, was calculated between the group correlations (i.e. all probes on the array) and sample correlations (i.e. set of 2D-DIGE targets) for each node in the network independently; it is calculated as the sum of the maximal and minimal deviations of the sample, $F_s$, and control (i.e. entire array), $F$, cumulative distribution functions:

$$K = D^* + D^- = \max(F_s - F) + \min(F_s - F)$$

As per the suggestions of Subramanian et al.\textsuperscript{37}, the Kuiper’s statistic, $K$, was modified to improve its ability to detect bimodal shifts in location of the sample distribution (as one would expect coexpressed groups of proteins to show both positive and negative correlations):

$$F_s = P_s(r \leq r_s) = \frac{1}{N_s} \sum_{r \leq r_s} |r| \quad \text{where} \quad N_s = \sum_{r \leq r_s} |r|$$

Where $S$ is the set of proteins being tested (either the $\text{Apc}^{638NN+/}$ or $\text{Cdkn1a}^{-/-}$ 2D-DIGE targets); $r$ is the ordered vector of correlation coefficients between the respective 2D-DIGE targets and a single network node; and $N_s$ normalizes $F_s$ to have sum 1. Significance testing was performed using a normal approximation of the empirical null: the empirical null was assembled from the modified $K$ calculated for 500 randomly selected protein sets, each of size $n = |S|$, and maximum likelihood estimation was used to fit a normal distribution. For exploring and illustrating the connections of significant ($\alpha = 0.05$) network nodes, we examine the subset of correlations, $r_y$, where $y \leq z$ such that $D^* = P_s(r \leq r_y) - F$ and $F < P_s(r \leq r_y)$; and the subset of correlations, $r_z$, where $p \geq q$ such that $D^- = F - P_s(r \leq r_q)$ and $F > P_s(r \leq r_q)$ (analogous to the “leading edge” subset of GSEA\textsuperscript{37}). To identify differentially expressed nodes, we chose those nodes where the $t$-statistic (unequal variance)
of the maximum intensity probe was such that \( |t| \geq \Phi^{-1}(0.95) \) in either the crypt or the villus compartment, where \( \Phi^{-1} \) is the normal inverse cumulative distribution function.

Testing each node in the \( Apc-Cdkn1a \) network independently resulted in a \( p \)-value for each of the \( H_{o,i} \) null hypotheses, where \( i \in n \), and each hypothesis, \( H_{o,i} \), assumes that there is no relationship (physically-based or coexpression-based) between the \( Apc-Cdkn1a \) network node, \( i \), and the 2D-DIGE targets. To test the group null hypothesis that all \( H_{o,i} \) are simultaneously true, \( p \)-values were aggregated into a statistic, \( \tau \), suggested by Fisher; significance was assessed against a \( \chi^2 \) distribution with \( 2n \) degrees of freedom\(^{160} \). The mutated node (\( Apc \) in \( Apc^{1638N/+} \) or \( Cdkn1a \) in \( Cdkn1a^{-/-} \)) was excluded from the respective analyses, as their extreme expression patterns skew the group-wise results.

Figure 2. The \( Apc-Cdkn1a \) network. Solid edges represent previously known interactions; dashed edges represent predicted interactions; and edges marked with a “\( v \)” represent predicted interactions that have been validated recently in the published literature.
Results

Driver Gene Network Predictions

The double mutant \textit{Apc}\textsuperscript{1638N+/Cdkn1a\textsuperscript{−/−}} mouse was previously shown to exhibit a synergistic increase in its tumor burden when compared with the single mutants\textsuperscript{19}. To identify the potential connections between \textit{Apc} and \textit{Cdkn1a}, we constructed a predictive framework that, first, learns the annotation patterns characteristic of known signaling pathways (e.g. those found in KEGG\textsuperscript{162} and others) and, then, couples these patterns with tissue specific coexpression data to extract the most likely chains of interacting proteins involved in \textit{Apc-Cdkn1a} signaling (illustrated in Figure 1). To identify only high-confidence pathways, a two-phase filtering process was first applied to the global PPI network. In the first phase, edges – compiled from mammalian interactions in BioGRID\textsuperscript{42} and HPRD\textsuperscript{41} – were pruned from the network if they did not resemble likely interactions (as defined by a logistic regression model), with the goal of reducing false positives among the reported interactions. To account for false negatives (Phase 2), interactions were added to the network by inferring relationships that are preceded in model organisms based on protein family relationships. After applying these measures to generate a synthetic network, we searched for likely connections between \textit{Apc} and \textit{Cdkn1a} using both gene coexpression data and Gene Ontology association rules.

To emphasize nodes and edges relevant to our biological system, we introduced a tissue-specific bias in our search for \textit{Apc-Cdkn1a} connections by using gene expression data from the intestinal epithelium of \textit{Apc}\textsuperscript{Min/+} mice. From these data, we calculated the mRNA-level coexpression value for individual edges via the gene-gene Pearson correlation coefficient. Next, all paths in the synthetic network linking the gene products of \textit{Apc} and \textit{Cdkn1a} were queried, and the predicted paths were filtered based on (i) the support of
association rules for GO annotations and (ii) the average coexpression along a path; the result (at a significance level of \( \alpha=0.01 \)) is shown in Figure 2. The \( \text{Apc-Cdkn1a} \) network includes a number of previously known interactions (solid lines), as well as predicted interactions (dashed lines) based on: (i) protein family relationships, (ii) strength of GO association rules, and (iii) microarray coexpression along the specific path connecting \( \text{Apc} \) to \( \text{Cdkn1a} \). As genetic interactions were included in the original interaction databases, the predicted network includes both physical and functional relationships.

At a systems-level, the proposed \( \text{Apc-Cdkn1a} \) network bears the statistically unlikely property of being saturated with oncogenes: 8 of the 20 proteins are annotated as oncogenes in OMIM (\( p\text{-value} < 5 \times 10^{-10} \) by Fisher’s exact test), and many of the remaining genes have been experimentally shown to act as oncogenes (e.g. \( \text{Erbb}^{3,172} \), \( \text{Shc1}^{174} \), \( \text{Map2k1}^{175} \)). Although the \( \text{Apc-Cdkn1a} \) network contains many well-studied proteins, the node degree (i.e. number of interactions) within the subnetwork does not strictly correlate with the node degree in the unfiltered interaction database (Pearson’s correlation=0.51). For instance, while \( \text{AKT1} \) has many known interactions, its commonly studied biological partners – namely, \( \text{GSK3B} \) and \( \text{PTEN} \) (both of which are associated with \( \text{Apc}^{135} \) and \( \text{Cdkn1a}^{176} \) signaling) – do not appear in the network. Other known interactions, such as that between \( \text{SHC1} \) and \( \text{SRC}^{177} \), are also absent from the network. Since our algorithm predicts connections biased by the biology of the system under study (through the use of gene expression data from \( \text{Apc}^{\text{Min/+}} \) mouse intestinal tissue), a particular protein or edge may not appear in the network if the pathway (i.e. chain of proteins) on which it resides does not meet the gene coexpression and/or GO association rule thresholds.
Table 1. Published evidence validating interactions predicted in the *Apc-Cdkn1a* network.

<table>
<thead>
<tr>
<th>Protein A</th>
<th>Protein B</th>
<th>Interaction Type</th>
<th>System</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPK8</td>
<td>CAV2</td>
<td>Functional</td>
<td>Human, fetal fibroblasts; Human, lung tissue</td>
<td>CAV1 forms hetero-oligomers with CAV2, and CAV1 inhibits TGF-beta or IL-6 induced phosphorylation of Mapk8 in fibroblasts(^{163})</td>
</tr>
<tr>
<td>SRC</td>
<td>APC</td>
<td>Functional</td>
<td>Mouse, colon epithelial cell line</td>
<td>Stable expression of SRC resulted in increased proliferation of <em>Apc(^{Min/+})</em> cells versus <em>Apc(^{+/+})</em> cells(^{165})</td>
</tr>
<tr>
<td>SRC</td>
<td>CCND1</td>
<td>Functional</td>
<td>Human, breast cancer cell line</td>
<td>SRC transfection leads to CCND1-CDK4-p27 complex formation(^{166})</td>
</tr>
<tr>
<td>SRC</td>
<td>CDK4</td>
<td>Phosphorylation</td>
<td>Human, colon cancer cell line</td>
<td>SRC phosphorylates CDK4(^{169})</td>
</tr>
<tr>
<td>SRC</td>
<td>PCNA</td>
<td>Functional</td>
<td>Human, ovarian cancer xenograft</td>
<td>Administration of a small molecule inhibitor of SRC results in decreased staining for PCNA in mouse carrying the xenograft(^{170})</td>
</tr>
<tr>
<td>CDK4</td>
<td>CAV2</td>
<td>Functional</td>
<td>Mouse, ES cells</td>
<td>Expression of CDK4 decreased upon knock-down of Caveolin-1(^{171})</td>
</tr>
</tbody>
</table>

Though CAV2 was discovered in the subnetwork, CAV1 and CAV2 are located adjacent to each other on chromosome 7 and express co-localizing proteins that form a stable complex.

Conversely, the *Apc-Cdkn1a* network includes novel associations: those not contained within the source databases (dashed edges in Figure 2). Several of these interactions have recently been validated in focused studies (see Table 1), providing confidence that the framework is useful. In addition, the *Apc-Cdkn1a* network also suggests that certain interactions previously associated with other cancer models – such as the SRC-
CCND1 functional association found in prostate cancer\textsuperscript{167}, or the phosphorylation of CDK4 by SRC in a cell line\textsuperscript{169} – are relevant in this model of colon cancer.

**Single Node Perturbations: mRNA Profiling**

As the \textit{Apc-Cdkn1a} network represents the intersection of signaling pathways emanating from \textit{Apc} and from \textit{Cdkn1a}, we expect to observe functional changes in network-associated proteins in response to perturbations at either \textit{Apc} or \textit{Cdkn1a}. Single-node perturbations were developed in mouse models with mutations in either \textit{Apc} (namely, \textit{Apc}\textsuperscript{1638N+/+}) or \textit{Cdkn1a} (\textit{Cdkn1a}+/−). While the \textit{Apc-Cdkn1a} network was generated using tumor-specific \textit{Apc}\textsuperscript{Min/+} data – a model harboring a number of background genetic lesions\textsuperscript{178} – the intestinal tissue obtained from the \textit{Apc}\textsuperscript{1638N+/+} and \textit{Cdkn1a}+/− mice at 3 months of age is relatively polyp free, thus allowing us to gauge the effect of a single genetic perturbation on the pre-neoplastic epithelium. Although this removes potential bias that is introduced by subsequent mutations of neoplastic tissue, this approach may also attenuate the flow of information between the two genes.

Since we are using the two perturbations to determine how well the \textit{Apc-Cdkn1a} network can capture biological phenomena, we introduced a multivariate statistic, $V^2$ to test if differences in mean mRNA abundance exist jointly between the \textit{Apc}\textsuperscript{1638N+/+} and \textit{Cdkn1a}+/− models. By using $V^2$, as illustrated in Figure 3, genes with mild differential expression in the two individual mutants can contribute to the overall support of the network, as $V^2$ rewards those genes where each of the two independent t-statistics are both greater than 1. Statistical significance of $V^2$ was tested against a permutation null, and, as our perturbations involved two tumor suppressors expected to have molecular effects in the same direction, we used the positive tail of the distribution. Knowing that many molecules “switch” expression (i.e. high
to low, or vice versa) in the transition from crypts to villi\textsuperscript{147}, the microarray datasets for these two biological compartments were tested separately. We found that the \textit{Apc-Cdkn1a} network was strongly supported (\textit{p}-value=0.002) by the joint mRNA differential expression in the two mutants’ crypt compartment. Network coherence was weaker (\textit{p}-value=0.060) in the villus compartment, and the network as a whole was not differentially expressed in the villi of either mutant, noted in the two \textit{V}^2 matrices’ \textit{p}-values:

\begin{align*}
\text{\textit{p}-value}(V^2_{\text{crypt}}) &= \text{Apc}^{1638N+/\textendash} \quad \text{Cdkn1a}^{-/\textendash} \\
&= \begin{bmatrix} 0.871 & 0.002 \\ 0.002 & 0.009 \end{bmatrix} \\
\text{\textit{p}-value}(V^2_{\text{villus}}) &= \text{Apc}^{1638N+/\textendash} \quad \text{Cdkn1a}^{-/\textendash} \\
&= \begin{bmatrix} 0.645 & 0.060 \\ 0.060 & 0.247 \end{bmatrix}
\end{align*}

Where, as mentioned, the diagonal elements indicate the significance of differential expression \textit{within} a mutant (as per Hotelling’s \textit{T}^2), and the off-diagonal elements indicate significance of joint differential expression \textit{across} mutants (as per \textit{V}^2). In the crypts, the network was differentially expressed in \textit{Cdkn1a}^{-/\textendash} (\textit{p}-value=0.009), but not in \textit{Apc}^{1638N+/\textendash} (\textit{p}-value=0.871), and, yet, was jointly supported by differential expression across both mouse models (\textit{p}-value=0.002). This illustrates that small mRNA-level changes that are shared between multiple perturbations – on a gene-by-gene basis – provide joint support for the network hypothesis, while any individual perturbation may fail to demonstrate the claim.
To illustrate how the joint consideration of gene-wise behavior operates, each network node has color-coded bubbles for the t-statistics of both $A_{pc}^{1638N+/-}$ and $Cdkn1a^{-/-}$ in Figure 3; the sum is shown at the intersection of each gene’s bubbles. Though $V^2$ employs products of t-statistics, the sum is better suited for visually demonstrating the principle that small mRNA effects can have a significant impact when considered together. We observe that several nodes that are differentially expressed in the crypts – ERBB3, JAK2, MAPK8, et al. – are no longer differentially expressed in the villus. In addition, some genes – e.g. CCNE1, CAV2, FGFR1, EGFR – switch their direction of expression.

Figure 3. Differential expression of the $A_{pc}$-$Cdkn1a$ network in villi (top) and crypts (bottom) for the $A_{pc}^{1638N+/-}$ and $Cdkn1a^{-/-}$ mouse models. Each network gene is represented by two overlapping bubbles colored according to the t-statistics (unequal variance) in the two mutants: the lower left bubble of a gene corresponds to the t-statistic for $A_{pc}^{1638N+/-}$, and the upper left bubble to the t-statistic for $Cdkn1a^{-/-}$. The intersection of the two bubbles corresponds to the sum of the t-statistics, illustrating how the significance of small effects can be strengthened when considered jointly. Nodes downregulated in the mutant are colored pink, those upregulated in the mutant are yellow, and neutral t-statistics are grey. While $V^2$ is calculated using all probes for each gene, we use only the probe with maximum intensity to calculate the t-statistics for visualization.
between the crypts and villi.

**Single Node Perturbations: Proteomic Profiling**

The 2D-DIGE analysis reported 12 proteins differentially expressed for the Cdkn1a\(^{-/-}\) intestinal epithelium (crypts and villi combined) versus wild type, and 31 proteins differentially expressed in the epithelium of the Apc\(^{1638N+/-}\) mice versus wild type (Appendix II, Table I). To test our network-based hypothesis, we first assumed that the set of regulatory molecules in the Apc-Cdkn1a network are independent. Then, the one and two-hop physical interactions were assessed for each network node (see Figure 4). While directly interacting neighbors (one hop) are typically useful in mapping signaling pathways, they did not associate much of the proteomic data with the network. Also, the few direct connections were not statistically significant; EGFR, for example, tends to have many interactions, and, thus, EGFR’s direct connections to the 2D-DIGE targets were not more likely than expected by chance. However, analysis of indirect interactions (two hops) from network nodes captured the relationship to the majority of the 2D-DIGE targets (individual node’s \(p\)-value < 0.005), as illustrated in Figure 4B. Considering the network as a whole and aggregating the \(p\)-values (aggregate statistic, \(\tau\)), the network was significantly (\(p\)-value of \(\tau < 1x10^{-15}\)) physically associated with either the Apc\(^{1638N+/-}\) or Cdkn1a\(^{-/-}\) 2D-DIGE targets, suggesting that the proteome-level effects are at most 4-hops away from the causative mutations.
Based on the physical proximity of the 2D-DIGE targets to the Apc-Cdkn1a network, we hypothesized that these network proteins might be controlling the expression of the 2D-DIGE targets. To examine this relationship further, we studied the pattern of mRNA-level coexpression between network nodes and the 2D-DIGE targets. As before, the network nodes were assumed to be independent, and the pattern of coexpression was assessed for each node individually using a modified Kuiper’s test statistic, $K$; nodes identified as (i) being differentially expressed ($|k|\geq \Phi^{-1}(0.95)$) in either crypts or villi and (ii)
having significant (α=0.05) coexpression with the 2D-DIGE targets are highlighted in Figure 5. Fifteen nodes in the Apc-Cdkn1a network had significant mRNA-level correlations to the Apc1638N+/ 2D-DIGE targets, and four of these were also differentially expressed. On the other hand, eight nodes had significant correlations to the Cdkn1a+/ 2D-DIGE targets, and four of these were individually differentially expressed. Considering coexpression relationships from the Apc-Cdkn1a network as a whole, the p-value of τ for coexpression between Apc-Cdkn1a network nodes and Apc1638N+/ 2D-DIGE targets was strongly significant (all nodes excluding Apc and Cdkn1a, p-value<1x10⁻²⁰; differentially expressed nodes, p-value=1.4x10⁻⁵). Given the magnitude of these group-wise statistics, however, the evidence for Apc-Cdkn1a network coexpression with the Cdkn1a+/ 2D-DIGE targets was not as well-supported (all nodes, p-value of τ =3.1x10⁻⁸; differentially expressed nodes, p-value of τ =1.6x10⁻⁵). Given that τ can be influenced by a few small p-values, we also calculated the probability of observing k p-values less than α=0.05, which, as a binomial distribution, is more sensitive to larger p-values. This also indicated that the Cdkn1a+/ 2D-DIGE targets were least supported by coexpression with the Apc-Cdkn1a network, with the p-values separated by two orders of magnitude again (p-value for coexpression of Cdkn1a+/ targets with differentially expressed network nodes was 3.3x10⁻⁵; p-value for coexpression of Apc1638N+/ targets was 3.1x10⁻⁵).

Finally, knowing that the use of a single probe per gene is often misleading, the above calculations for proteomic coexpression were also performed using all microarray probes for the network nodes and the proteomic targets. By this approach, the network as a whole was strongly coexpressed in either mutant (for differentially expressed probes, excluding those belonging to Apc or Cdkn1a, p-value of τ <1x10⁻²⁰). However, for ease of
interpretability and visualization, we discuss the results of the analysis using only the maximum intensity probe per gene.

**Discussion**

As underscored by Wood et al.\textsuperscript{17}, colorectal cancer is the product of mutations in multiple genes operating simultaneously. Though tumors differ at the genetic level, their phenotypes intersect at histopathologic levels and, in our view, at the molecular level, as well, implying that the connections between unique sets of mutations may often merge at downstream signalling hubs. Thus, the reduction of genetic heterogeneity into clinically meaningful biological networks could have an impact in the context of personalized medicine. As a step towards this goal, we developed a computational framework capable of predicting functional connections between genes mutated in cancer, and we applied our methodology to define a network between \textit{Apc} and \textit{Cdkn1a}.

Networks, as abstractions of underlying molecular phenomena, offer the hope of distilling system-level structure from biological complexity. Given the many degrees of freedom in the underlying datasets (PPIs, microarrays, et al.), however, numerous network structures are possible for a particular biological context, and candidate networks are often evaluated solely based on topological significance (e.g. the G-score of MetaCore \textsuperscript{179}). Yet, the value of a network may not be reflected topologically; for example, a highly-connected hub may not be highly active in a particular model system. If a network model truly reflects the underlying biology, then – from an engineering standpoint – perturbations of the underlying system should be manifest in and around the predicted network. Specifically, since a network connecting two cancer genes represents cross-connections between signaling pathways, one
would expect that perturbations along a pathway would result in altered regulation of network-associated nodes.

A systems-based approach required to biologically evaluate network coherence, however, is not immediately amenable to the tools of classical molecular biology, which are designed to target single molecules or, at most, a few at a time. As an alternative, we outline the use of in vivo single node perturbations – by way of mouse models with targeted inactivation of specific loci – followed by gene and protein expression analysis to gauge systems-level effects. For further investigation, numerous mouse models for cancer biology are already available, and these resources can be productively mined to expand our understanding of cancer gene networks. Though we have demonstrated the value of biologically testing network predictions using 2D-DIGE and microarray data, many other types of screening tools could also be employed to test the functional coherence of predicted networks.

Due to the differences in coverage of proteomic and gene expression data, different approaches were required to probe the potential functional coherence of the network: $V^2$ – a multivariate statistic – was used to gauge the effect of single node perturbations on mRNA-levels of the Apc-Cdkn1a network, while, due to reduced coverage, relational maps – physical and coexpression – were required to assess the effect of driver gene mutations on the proteome. We found that the Apc-Cdkn1a network was supported by the joint differential expression of mRNA in two different network perturbations, with stronger differential expression being observed in the crypts (Figure 3). The $V^2$ statistic is presented in a framework that is extensible to multiple network perturbations – a feature that proves necessary in evaluating the biological coherence of networks, as small mRNA-level effects of an individual perturbation may fail to lend adequate support for a predicted network;
coupling multiple perturbations together via the $V^2$ matrix allows the integrity of the network to be assessed via a biologically multidimensional approach. It should be noted that, in testing the mRNA-level support for an individual network, a “self-contained” hypothesis is necessary, embodied by null distributions – namely, the permutation null used here – modeling the population from which the samples (mice) were drawn; gene randomization methods, on the other hand, compare network expression patterns to stochastically chosen gene groups, which are bound to have a different and/or reduced covariance structure.

Figure 5. Coexpression between the 2D-DIGE targets and the differentially expressed $Apc^{-/}$ and $Cdkn1a^{-/}$ network nodes. To examine the network-based hypothesis, each network node was tested independently for significant correlation with the $Apc^{1638N+/}$ or $Cdkn1a^{-/}$ 2D-DIGE targets using a modified Kuiper’s test statistic. The 2D-DIGE targets are ordered by the amount of second-degree physical interaction, per Figure 4; node size is proportional to the number of coexpression interactions with differentially expressed signaling proteins. $Apc^{1638N+/}$. 2D-DIGE targets marked with a “*” were also found in the $Cdkn1a^{-/}$ intestinal epithelium.
(especially without incorporating network structure to generate the null gene sets), leading to overinflated significance values.

After applying $V^2$ to the mRNA data, we found that the network was better able to capture joint differential expression in the crypts than in the villi, suggesting that oncogenic transformations are initiated in the crypts by the network genes and then transduced to downstream targets in the villi. This is reflected in Figure 3, where the t-statistic of individual nodes is colored for both mouse models, and more nodes are seen to be brightly colored (i.e. highly differentially expressed) in the crypts. Interestingly, a large contingent of oncogenes – ERBB3, JAK2, MAP2K1, MAPK8 – are clearly downregulated in the crypts, while their expression levels diminish considerably in the villi, indicating that these genes turn “off” during the crypt-to-villus transition – a well-known feature of this biological compartment.

Though the downregulation of oncogenes may appear counterintuitive, it is to be expected in these particular mouse models, where the tumorigenic phenotype is mild and the tissue has been harvested in the pre-neoplastic regime. Before the onset of tumors, the downregulation of oncogenes represents a homeostatic reflex of the tissue to the genetic perturbations, i.e. protective downregulation of oncogenes to compensate for the loss of $Cdkn1a$ or $Apc$. In addition to genes turning “off” in the crypt-to-villus transition, several genes appear to switch their pattern of expression entirely. In particular, CDK3, CAV2, EGFR, and FGFR1 exhibit this behavior, which suggests that they play two different roles in the two compartments.

For visualization, we show only the probe with maximum intensity across all samples for each gene. Given the extent of alternative splicing and array manufacturing variation, however, a single probe can be misleading. Hence, for calculation of statistical significance, we use all probes to model a single gene – a more robust approach that is amenable to
matrix-based calculations (such as $V^2$). We concede that, since each gene is represented by a different number of probes, genes with many probes contribute proportionally more weight to the final $V^2$ statistic. This is a useful feature, however, as we have more confidence about the true behavior of these well-probed genes – many of which are well-studied and important in cancer, such as EGFR (7 probes) and MAPK8 (5 probes) – and, thus, they deserve greater weight than the highly variable, single-probe genes.

Proteomic data, however, requires different analytical considerations, as the protein levels of the network nodes may not be directly measured in a given proteomic experiment. To make inferences about our network-based hypothesis, we used two different mappings: one based on physical interactions, and another based on mRNA correlations. From Figure 4, it is clear that the proteins measured in the 2D-DIGE experiment are not merely a random sampling from the proteome. Rather, they are physically close to (i) the hypothesized network as a whole ($p$-value of $\tau < 1 \times 10^{-15}$) and (ii) several individual signaling molecules ($p$-values $< 0.005$). Specifically, the 2D-DIGE targets from the $Apc^{1638N+/-}$ and $Cdkn1a^{-/-}$ experiments have significant physical proximity to CTNNB1, FGFR1, ERBB3, CAV2, and CDKN1A itself. The tight physical proximity of the predicted network nodes to experimentally measured targets suggested that the signaling molecules more proximal to the mutations may regulate the proteomic targets, and we used mRNA-level coexpression to examine this relationship further. While coexpression relationships abound between the $Apc^{1638N+/-}$ 2D-DIGE targets and the $Apc-Cdkn1a$ network nodes, this is less significant for the $Cdkn1a^{-/-}$ targets ($p$-value of $\tau = 3.3 \times 10^{-3}$). Taken together with the results of the mRNA analysis, this suggests that the hypothesized network nodes more effectively capture $Cdkn1a^{-/-}$ signaling at the mRNA level rather than at a proteomic level, whereas the opposite is true of $Apc$ signaling.
While the network as a whole showed differences in the level of proteomic coexpression, two differentially expressed nodes – MAP2K1 and AKT1 – were significantly coexpressed with the measured proteome in both network perturbations (Figure 5). Interestingly, AKT1 was also found to be closely physically associated with the proteomic targets in *Cdkn1a*+/−, while MAP2K1 was physically associated with *Apc*1638N+/+. 2D-DIGE targets (Figure 4). As we know that mRNA coexpression can provide evidence regarding the regulatory role of proteins⁴⁰, the mutual discovery of MAP2K1 and AKT1 in the two network perturbations – via both coexpression and physical connectivity to the perturbed proteome – suggests that these two proteins may serve as intersection points of *Apc* and *Cdkn1a* signaling. Also of interest is the observation that coexpression connections and physical connections tend to associate different subsets of the proteomic targets, as the more physically distant proteomic targets (e.g. SULT1B1, OTC, KRT19) are also the ones that tend to be coexpressed with multiple network nodes. Not only does this illustrate that physical and coexpression maps capture different dimensions of biological function, but it also illustrates the necessity of using both maps to provide complementary information in evaluating the molecular context of network hypotheses.

While the 2D-DIGE studies revealed many differentially expressed proteins, annexin A2 (ANXA2) was among the most highly ranked in its physical proximity to the hypothesized network. At the protein level, ANXA2 was upregulated in both mouse models of colon cancer (see Appendix II, Tables I and II). From studies of prostate cancer⁴¹, ANXA2 upregulation is expected since *in vitro* experiments indicate that overexpression promotes a more invasive, proliferative cell phenotype. Though ANXA2 had high mRNA expression in one population of human colorectal tumors⁴², it is downregulated in some populations of human prostate tumors⁴¹ and colorectal cancer cell lines⁴³. Since activation
of either Apc or Cdkn1a signaling leads to upregulation of ANXA2 in our studies, activation of alternative or repressive pathways may lead to such unexpected downregulation of ANXA2 in some tumors.

In addition to ANXA2, the mutant mice exhibited other protein-level alterations that can potentially contribute to tumorigenesis. Elongation factor E2 (EEF2), for example, was found to be upregulated in both mouse models, and its tumorigenic potential in gastrointestinal\textsuperscript{184} and breast\textsuperscript{185} cancers is well known. Though drugs inhibiting the EEF2 pathway (via its kinase) exist\textsuperscript{186}, and molecular chemotherapy targeting ANXA2 can also be envisioned, our network-based hypothesis suggests these changes may be controlled by specific upstream signaling molecules that integrate the information from mutated genes. Thus, in patients where levels of ANXA2 or EEF2 are elevated, molecular therapy targeting the proposed upstream network targets – such as MAP2K1 or AKT1 – may be more effective.

In conclusion, we outline a novel method for identifying networks that connect signaling pathways associated with cancer driver genes. The first step towards statistically analyzing these novel subnetworks was pursued using single node perturbations of the system in vivo, followed by network interrogation via high-throughput -omics experiments. Together, the various lines of evidence – mRNA differential expression, 2D-DIGE-target physical proximity, and 2D-DIGE-target coexpression – strengthen the hypothesis that the Apc-Cdkn1a network helps to mediate both Apc and Cdkn1a signaling. Thus, we show that using ‘omic data to test a network-based hypothesis not only allows one to assess the biological validity of in silico predictions, it also allows one to prune the hypothesis to identify molecular targets (e.g. SRC and EGFR) that are likely to integrate the various signaling pathways perturbed in cancer.
The Present

Our technologies continue to evolve, allowing us to measure a growing number of molecules with increasing proficiency. While at first glance these measurements appear unambiguous, providing information about mRNA, peptides, and proteins, the picture becomes much more complex when we begin to consider the interactions among these players. The study of protein interaction patterns, for example, surfaced as a field in its own right owing to its complexity (culminating in the construction of large publicly available PPI datasets\(^4\)), and these resulting data can now be leveraged in our analyses of high-throughput molecular measurements. A variety of both manifolds and molecular measurements exist, and my work lives at the interface of the two, focusing on the study of molecular subsystems in colorectal cancer.

Herein, we began with a proteomic study of prostaglandin dehydrogenase (\(Hpgd\)), where we sought to discover previously known and unknown pathways involved in \(Hpgd\) signaling. While proteomics is often asked to address simple questions – “which proteins are differentially expressed?” – we illustrated that mapping peptide-level information to the genome, a well-defined manifold, can allow us to probe more nuanced forms of biological regulation, such as alternative splicing. In viewing LC-MS/MS measurements as part of a
molecular subsystem, we were able to provide a highly informative set of molecular targets where the follow-up experiments are self-evident: proteins regulated by abundance, with many peptides changing across the entire span of the gene, can be pursued by Western blots; while candidates for alternative splicing require PCR to probe for different transcripts arising from the spliced portion of the gene. Additionally, we elucidated the biological context of the differentially expressed proteins by examining patterns of coexpression with known transcription factors, demonstrating the utility of coexpression manifolds in “reverse engineering” regulatory pathways.

In contrast, our studies of Apc involved predicting pathways first and experimentally testing their biological merit after the fact (Chapter 3). When a gene, like Apc, is mutated in a tumor, many pathways are likely to be perturbed, and the question is, then, in identifying the pathway, or subset of pathways, most involved in disease progression. We approached the problem computationally, first generating a set of candidate pathways based on PPIs with other driver genes, and then parsing the resulting search space of possible signaling pathways. The predicted PPI manifolds were evaluated by the strength of their coexpression with proteomic targets, in effect leveraging one manifold (the coexpression network) to evaluate another (the signaling pathway).

These studies of Apc signaling hinge on a critical assumption: though a genetic mouse model may have a single mutation, the mutated gene resides in a larger subsystem of genes. Rather than viewing the mutated gene as the paramount feature of the system, we flipped the paradigm on its head, viewing the mutation as simply one piece of a larger puzzle. As our knowledge of the molecular “wiring diagram” improves, we expect that more experiments will be designed with this principle in mind: a scientist can begin with the
biological circuitry (i.e. the signaling pathway) known a priori, and then methodically perturb key nodes in the circuit to affect measurable outputs.

When such an experiment is designed, it is of interest to determine if and how various perturbations affect the system of interest. If different perturbations affect the system in similar ways, this phenomenon serves as one form of validation for the molecular subsystem of interest. In Chapter 4, we demonstrated this principle using genetic perturbations of a PPI network. We observed similar changes in mRNA expression in the crypts of the small intestine upon mutation of either \textit{Apc} or \textit{Cdkn1a}, indicating that our manifold is “coherent”: genetic perturbations at multiple points in the manifold generate similar molecular and phenotypic responses. A finding of incoherence, where perturbations produce inconsistent effects, would suggest that either the molecular subsystem does not represent a true biological unit, or that the phenotype measured is inappropriate for the perturbation induced. For instance, molecular effects may be manifest at the protein level more so than the mRNA level, and, in this case, a proteomic experiment would be more appropriate for gauging system coherence.

**Caveats**

While molecular subsystems hold much promise for organizing and interpreting high-throughput data, they suffer from shortcomings in both high-throughput measurement technologies and manifolds. High-throughput arrays for mRNA expression have progressed significantly since their introduction in the 1990s, with current whole exon array providing robust measurements across the expanse of a gene. A problem not fully addressed in the literature is how these exon-level measurements should be summarized to make gene-level inferences, a problem we dealt with in our proteomic studies (Chapter 2). Once such a
summary measure can be produced, there remains the question of how to assess its significance – no small feat when dealing with high-dimensional data. Current approaches hinge upon intricate statistical procedures to make inferences about the true (i.e. population) variance for a gene’s measurements, as the sample size is often too small to simply use the sample variance. No absolute solution has yet been discovered for these issues, and the adoption of an approach rests upon its ability to validate.

Manifolds began to be assembled shortly after the sequencing of the genome, and our current knowledge of manifolds remains limited. Protein-protein interaction (PPI) networks, while a direct generalization of the signaling pathway paradigm, remain far from complete. Yet, high-throughput efforts to pull down PPIs carry the cost of increased false positives. The genome, perhaps the most intuitive manifold to use, has a well-known structure, though much of its functional relevance remains unknown. In the past decade alone, scientists have rapidly progressed from initially referring to much of the genome as “junk DNA” and “dark matter”, to recognizing that there is a host of regulatory machinery – non-coding RNAs, in particular – of which we are not fully aware. Any particular manifold can capture only a limited form of regulation, e.g. transcriptional regulation by coexpression networks, or physical interactions by a PPI network, which, in turn, limits the kind of inferences we can make of our data. While the genome contains a wide variety of regulatory structures embedded within it, one must be aware how high-throughput data is being mapped onto the genome, for, inevitably, much regulation will be ignored.

The Future
With the advent of many high-throughput molecular technologies in the past decade, we have been eagerly anticipating the coming era of “personalized medicine.” However, we have also come to realize that measuring mutations, mRNA, or proteins is not enough to tell a biological story; in fact, our measurement prowess only adds to the complexity by showering us with more information to process. With the work that I have presented here, I argue that molecular biology is in need of paradigm shifts to change the way it views molecular measurements. A set of polymorphisms identified by a genome-wide association study (GWAS), for example, are of limited utility when considered alone, but can be part of a compelling story when evaluated in a framework or protein or genetic interactions; in fact, with the small effect size uncovered by many GWA studies, considering subsystems of SNPs may be necessary to identify coherent molecular groups with population-scale relevance. Genomic sequencing of tumors is subject to similar pitfalls, as tumors are riddled with dozens of mutations, many of which may not be relevant to pathogenesis. Rather than generating encyclopedic lists of mutations, we must consider the various interactions amongst mutations to arrive at systems-level hypotheses suitable for testing. In this case, the types of interactions we ought to consider are broad and span many domains – from genes to cells to tissues; indeed, the interaction between a tumor and its microenvironment is a critical factor that affects the disease progression (discussed in Chapter 1). At the present, our molecular therapies are designed to target individual nodes in a molecular subsystem, and, consequently, the field of systems biology is oriented to uncovering new targets for drug development. However, one can anticipate that new therapies or therapeutic regimens will ultimately arise that take into consideration subsystems and pathways, and we have witnessed this begin to happen with cetuximab therapy in colon cancer patients.
Appendix I

Prostaglandin dehydrogenase signaling:

Associated Data
Table I. \(p\)-values and empirical type I error rates for proteins. Proteins that were excluded from further analysis are highlighted in orange.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Ensemble ID</th>
<th># Daughter Peptides</th>
<th>Fold Change</th>
<th>(p)-value</th>
<th>Empirical Type I Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagln2</td>
<td>ENSMUSG00000026547</td>
<td>7</td>
<td>1.76</td>
<td>0.001</td>
<td>0.052</td>
</tr>
<tr>
<td>Rab6</td>
<td>ENSMUSG00000030704</td>
<td>4</td>
<td>1.69</td>
<td>0.001</td>
<td>0.072</td>
</tr>
<tr>
<td>Cldn3</td>
<td>ENSMUSG00000070473</td>
<td>3</td>
<td>1.93</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>Lypla1</td>
<td>ENSMUSG00000025903</td>
<td>3</td>
<td>1.83</td>
<td>0.002</td>
<td>0.062</td>
</tr>
<tr>
<td>Mettl7b</td>
<td>ENSMUSG00000025347</td>
<td>3</td>
<td>2.41</td>
<td>0.002</td>
<td>0.070</td>
</tr>
<tr>
<td>Gsta1</td>
<td>ENSMUSG00000074183</td>
<td>3</td>
<td>1.82</td>
<td>0.004</td>
<td>0.070</td>
</tr>
<tr>
<td>Rab11b</td>
<td>ENSMUSG00000077450</td>
<td>9</td>
<td>1.72</td>
<td>0.005</td>
<td>0.062</td>
</tr>
<tr>
<td>Arf4</td>
<td>ENSMUSG00000021877</td>
<td>4</td>
<td>1.64</td>
<td>0.005</td>
<td>0.080</td>
</tr>
<tr>
<td>Arpc3</td>
<td>ENSMUSG00000029465</td>
<td>4</td>
<td>1.96</td>
<td>0.006</td>
<td>0.046</td>
</tr>
<tr>
<td>Krt14</td>
<td>ENSMUSG00000045545</td>
<td>9</td>
<td>0.66</td>
<td>0.007</td>
<td>0.042</td>
</tr>
<tr>
<td>Eif6</td>
<td>ENSMUSG00000027613</td>
<td>3</td>
<td>1.67</td>
<td>0.009</td>
<td>0.058</td>
</tr>
<tr>
<td>Tmed10</td>
<td>ENSMUSG00000021248</td>
<td>3</td>
<td>1.52</td>
<td>0.010</td>
<td>0.052</td>
</tr>
<tr>
<td>Rab5a</td>
<td>ENSMUSG00000017831</td>
<td>3</td>
<td>1.47</td>
<td>0.010</td>
<td>0.060</td>
</tr>
<tr>
<td>Eno3</td>
<td>ENSMUSG000000660600</td>
<td>6</td>
<td>0.44</td>
<td>0.011</td>
<td>0.044</td>
</tr>
<tr>
<td>Rap1a</td>
<td>ENSMUSG00000068798</td>
<td>3</td>
<td>2.15</td>
<td>0.012</td>
<td>0.082</td>
</tr>
<tr>
<td>Chcd3</td>
<td>ENSMUSG00000053768</td>
<td>3</td>
<td>1.63</td>
<td>0.012</td>
<td>0.084</td>
</tr>
<tr>
<td>Krt6a</td>
<td>ENSMUSG00000058354</td>
<td>5</td>
<td>0.44</td>
<td>0.013</td>
<td>0.050</td>
</tr>
<tr>
<td>Hexb</td>
<td>ENSMUSG00000021665</td>
<td>4</td>
<td>0.69</td>
<td>0.014</td>
<td>0.070</td>
</tr>
<tr>
<td>Ethc1</td>
<td>ENSMUSG00000064254</td>
<td>3</td>
<td>1.61</td>
<td>0.014</td>
<td>0.084</td>
</tr>
<tr>
<td>Prdx1</td>
<td>ENSMUSG00000028691</td>
<td>10</td>
<td>1.38</td>
<td>0.015</td>
<td>0.056</td>
</tr>
<tr>
<td>Rpll0</td>
<td>ENSMUSG00000008682</td>
<td>3</td>
<td>1.92</td>
<td>0.015</td>
<td>0.056</td>
</tr>
<tr>
<td>Slc9a3r1</td>
<td>ENSMUSG00000020733</td>
<td>6</td>
<td>0.73</td>
<td>0.015</td>
<td>0.056</td>
</tr>
<tr>
<td>Rab14</td>
<td>ENSMUSG00000026878</td>
<td>5</td>
<td>1.75</td>
<td>0.015</td>
<td>0.060</td>
</tr>
<tr>
<td>Atp5f1</td>
<td>ENSMUSG0000000563</td>
<td>4</td>
<td>1.51</td>
<td>0.016</td>
<td>0.064</td>
</tr>
<tr>
<td>Gene</td>
<td>ENSMUSG0000000000000000000</td>
<td>q-value</td>
<td>False Discovery Rate</td>
<td>False Non-Discovery Rate</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------</td>
<td>---------</td>
<td>----------------------</td>
<td>--------------------------</td>
<td></td>
</tr>
<tr>
<td>Rab18</td>
<td>ENSMUSG00000073639</td>
<td>3</td>
<td>1.65</td>
<td>0.016</td>
<td>0.084</td>
</tr>
<tr>
<td>Akr1c14</td>
<td>ENSMUSG00000033715</td>
<td>3</td>
<td>0.65</td>
<td>0.017</td>
<td>0.074</td>
</tr>
<tr>
<td>Sri</td>
<td>ENSMUSG00000035161</td>
<td>3</td>
<td>1.51</td>
<td>0.018</td>
<td>0.052</td>
</tr>
<tr>
<td>Krt5</td>
<td>ENSMUSG00000061527</td>
<td>6</td>
<td>0.89</td>
<td>0.018</td>
<td>0.052</td>
</tr>
<tr>
<td>Gsta4</td>
<td>ENSMUSG00000032248</td>
<td>7</td>
<td>1.63</td>
<td>0.018</td>
<td>0.056</td>
</tr>
<tr>
<td>Arp5h</td>
<td>ENSMUSG00000034566</td>
<td>6</td>
<td>1.55</td>
<td>0.018</td>
<td>0.060</td>
</tr>
<tr>
<td>Arf5</td>
<td>ENSMUSG00000020440</td>
<td>3</td>
<td>2.01</td>
<td>0.018</td>
<td>0.072</td>
</tr>
<tr>
<td>Rab1b</td>
<td>ENSMUSG00000244870</td>
<td>6</td>
<td>1.46</td>
<td>0.019</td>
<td>0.050</td>
</tr>
<tr>
<td>Pgam1</td>
<td>ENSMUSG00000011752</td>
<td>6</td>
<td>1.65</td>
<td>0.019</td>
<td>0.054</td>
</tr>
<tr>
<td>Aldob</td>
<td>ENSMUSG00000028307</td>
<td>4</td>
<td>0.51</td>
<td>0.019</td>
<td>0.058</td>
</tr>
<tr>
<td>Krt19</td>
<td>ENSMUSG00000020911</td>
<td>29</td>
<td>0.65</td>
<td>0.020</td>
<td>0.032</td>
</tr>
<tr>
<td>Uqcrfs1</td>
<td>ENSMUSG00000038462</td>
<td>4</td>
<td>1.78</td>
<td>0.020</td>
<td>0.066</td>
</tr>
<tr>
<td>Gli1</td>
<td>ENSMUSG00000024026</td>
<td>4</td>
<td>1.47</td>
<td>0.021</td>
<td>0.052</td>
</tr>
<tr>
<td>Myl12b</td>
<td>ENSMUSG0000034868</td>
<td>4</td>
<td>1.27</td>
<td>0.021</td>
<td>0.052</td>
</tr>
<tr>
<td>Rps23</td>
<td>ENSMUSG00000049517</td>
<td>5</td>
<td>1.88</td>
<td>0.022</td>
<td>0.052</td>
</tr>
<tr>
<td>Spr</td>
<td>ENSMUSG00000033735</td>
<td>6</td>
<td>1.62</td>
<td>0.022</td>
<td>0.058</td>
</tr>
<tr>
<td>Copz1</td>
<td>ENSMUSG00000060992</td>
<td>3</td>
<td>1.54</td>
<td>0.022</td>
<td>0.062</td>
</tr>
<tr>
<td>Knt42</td>
<td>ENSMUSG00000053654</td>
<td>5</td>
<td>1.04</td>
<td>0.023</td>
<td>0.072</td>
</tr>
</tbody>
</table>
Table II. *p*-values and empirical type I error rates for exons. Exons that were excluded from further high-throughput analysis are highlighted in orange. *Selenbp1* and *Hsd11b2* are slated for validation via PCR.

<table>
<thead>
<tr>
<th>Parent Gene</th>
<th>Ensemble ID</th>
<th>Exon Start</th>
<th>Exon Stop</th>
<th>Exon Fold Change</th>
<th><em>p</em>-value</th>
<th># of Peptides In This Exon</th>
<th>Empirical Type I Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selenbp1</td>
<td>ENSMUSG00000068874</td>
<td>94746386</td>
<td>94746464</td>
<td>0.92</td>
<td>0.014</td>
<td>2</td>
<td>0.08</td>
</tr>
<tr>
<td>Hsd11b2</td>
<td>ENSMUSG00000031891</td>
<td>108042646</td>
<td>108043029</td>
<td>0.74</td>
<td>0.016</td>
<td>2</td>
<td>0.020</td>
</tr>
<tr>
<td>Pgd</td>
<td>ENSMUSG00000028961</td>
<td>148539159</td>
<td>148539338</td>
<td>0.71</td>
<td>0.023</td>
<td>2</td>
<td>0.066</td>
</tr>
<tr>
<td>Lonp1</td>
<td>ENSMUSG00000041168</td>
<td>56753724</td>
<td>56753986</td>
<td>1.69</td>
<td>0.036</td>
<td>2</td>
<td>0.066</td>
</tr>
<tr>
<td>Idh1</td>
<td>ENSMUSG00000025950</td>
<td>65215073</td>
<td>65215178</td>
<td>0.72</td>
<td>0.038</td>
<td>2</td>
<td>0.098</td>
</tr>
<tr>
<td>Ahcy</td>
<td>ENSMUSG00000027597</td>
<td>154889516</td>
<td>154889633</td>
<td>0.93</td>
<td>0.040</td>
<td>2</td>
<td>0.132</td>
</tr>
<tr>
<td>D10Jhu81e</td>
<td>ENSMUSG000000053329</td>
<td>77626142</td>
<td>77626298</td>
<td>1.43</td>
<td>0.041</td>
<td>2</td>
<td>0.100</td>
</tr>
<tr>
<td>Krt17</td>
<td>ENSMUSG00000035557</td>
<td>100119723</td>
<td>100119884</td>
<td>0.61</td>
<td>0.044</td>
<td>2</td>
<td>0.058</td>
</tr>
<tr>
<td>Atp5a1</td>
<td>ENSMUSG00000025428</td>
<td>78017471</td>
<td>78017637</td>
<td>1.01</td>
<td>0.044</td>
<td>3</td>
<td>0.076</td>
</tr>
<tr>
<td>Bsg</td>
<td>ENSMUSG00000023175</td>
<td>79172803</td>
<td>79172885</td>
<td>0.92</td>
<td>0.045</td>
<td>2</td>
<td>0.052</td>
</tr>
<tr>
<td>Anxa2</td>
<td>ENSMUSG00000032231</td>
<td>69337459</td>
<td>69337581</td>
<td>1.23</td>
<td>0.048</td>
<td>2</td>
<td>0.046</td>
</tr>
<tr>
<td>Hsd17bl0</td>
<td>ENSMUSG00000025260</td>
<td>148437940</td>
<td>148438513</td>
<td>1.37</td>
<td>0.049</td>
<td>2</td>
<td>0.020</td>
</tr>
</tbody>
</table>
Table III. *p*-values and empirical type I error rates for peptides. Peptides that were excluded from further analysis are highlighted in orange.

<table>
<thead>
<tr>
<th>Gene</th>
<th>EnsembleGeneID</th>
<th>Peptide Sequence</th>
<th>Fold Change</th>
<th>p-value</th>
<th># of Sibling Peptides</th>
<th>Empirical Type I Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hint2</td>
<td>ENSMUSG000000028470</td>
<td>MAAAYLLAVGLR</td>
<td>2.88</td>
<td>0.007</td>
<td>4</td>
<td>0.012</td>
</tr>
<tr>
<td>Nme1</td>
<td>ENSMUSG000000037601</td>
<td>EISLWFQPEELVEYK</td>
<td>0.23</td>
<td>0.008</td>
<td>4</td>
<td>0.002</td>
</tr>
<tr>
<td>Acadm</td>
<td>ENSMUSG000000062908</td>
<td>AAVEVDSGR</td>
<td>0.44</td>
<td>0.010</td>
<td>7</td>
<td>0.01</td>
</tr>
<tr>
<td>Copg</td>
<td>ENSMUSG000000030058</td>
<td>SIATLAITTLK</td>
<td>1.89</td>
<td>0.015</td>
<td>7</td>
<td>0.046</td>
</tr>
<tr>
<td>Hspa5</td>
<td>ENSMUSG000000026864</td>
<td>IDTRNELESYASLKL</td>
<td>0.37</td>
<td>0.016</td>
<td>26</td>
<td>0.042</td>
</tr>
<tr>
<td>Krt10</td>
<td>ENSMUSG000000019761</td>
<td>QSLEASLAETEGR</td>
<td>0.51</td>
<td>0.017</td>
<td>6</td>
<td>0.028</td>
</tr>
<tr>
<td>Atp5b</td>
<td>ENSMUSG000000025393</td>
<td>VLDSGAPIK</td>
<td>0.52</td>
<td>0.019</td>
<td>25</td>
<td>0.014</td>
</tr>
<tr>
<td>Pgm1</td>
<td>ENSMUSG000000029171</td>
<td>ELDELYGAIIEHFQFPQK</td>
<td>4.41</td>
<td>0.019</td>
<td>6</td>
<td>0.042</td>
</tr>
<tr>
<td>Ctc</td>
<td>ENSMUSG000000047126</td>
<td>VGYTPDVIFLLR</td>
<td>4.24</td>
<td>0.023</td>
<td>29</td>
<td>0.034</td>
</tr>
<tr>
<td>Krt8</td>
<td>ENSMUSG000000049382</td>
<td>QLEALGQEQK</td>
<td>0.59</td>
<td>0.024</td>
<td>28</td>
<td>0.038</td>
</tr>
<tr>
<td>Cs</td>
<td>ENSMUSG00000005683</td>
<td>EGSSIGAIDSR</td>
<td>0.46</td>
<td>0.025</td>
<td>9</td>
<td>0.022</td>
</tr>
<tr>
<td>Gm10709*</td>
<td>ENSMUSG00000074516</td>
<td>ALVKPQAIKP</td>
<td>2.03</td>
<td>0.026</td>
<td>3</td>
<td>0.064</td>
</tr>
<tr>
<td>Pfln</td>
<td>ENSMUSG00000018293</td>
<td>DSLIQDGEFTMDLR</td>
<td>0.36</td>
<td>0.028</td>
<td>5</td>
<td>0.076</td>
</tr>
<tr>
<td>Hsp90ab1</td>
<td>ENSMUSG00000023944</td>
<td>YIDQVELNK</td>
<td>0.54</td>
<td>0.029</td>
<td>26</td>
<td>0.07</td>
</tr>
<tr>
<td>Krt13</td>
<td>ENSMUSG00000044841</td>
<td>ITMQNENDR</td>
<td>0.51</td>
<td>0.032</td>
<td>7</td>
<td>0.074</td>
</tr>
<tr>
<td>Ezr</td>
<td>ENSMUSG000000052397</td>
<td>FYPEDVAELIQDTQK</td>
<td>8.33</td>
<td>0.033</td>
<td>13</td>
<td>0.098</td>
</tr>
<tr>
<td>Gm9755*</td>
<td>ENSMUSG00000030735</td>
<td>YEEIDNAPIER</td>
<td>0.39</td>
<td>0.035</td>
<td>15</td>
<td>0.064</td>
</tr>
<tr>
<td>Hist3h2a</td>
<td>ENSMUSG00000078851</td>
<td>VAGAPVYLAAYLTAIEEELAGNAAR</td>
<td>0.31</td>
<td>0.037</td>
<td>3</td>
<td>0.076</td>
</tr>
<tr>
<td>Pdlim1</td>
<td>ENSMUSG00000055044</td>
<td>VTPPEGYDVTVFR</td>
<td>0.71</td>
<td>0.037</td>
<td>3</td>
<td>0.078</td>
</tr>
<tr>
<td>Ndufs6</td>
<td>ENSMUSG00000021606</td>
<td>EVNENFAIDLIAQQPVNEVEHR</td>
<td>0.47</td>
<td>0.037</td>
<td>3</td>
<td>0.096</td>
</tr>
<tr>
<td>Set</td>
<td>ENSMUSG000000054766</td>
<td>VEVTEFEDIK</td>
<td>0.21</td>
<td>0.037</td>
<td>4</td>
<td>0.126</td>
</tr>
<tr>
<td>Pdha1</td>
<td>ENSMUSG00000031299</td>
<td>TREEQEVIR</td>
<td>0.35</td>
<td>0.039</td>
<td>4</td>
<td>0.076</td>
</tr>
<tr>
<td>Gm10053</td>
<td>ENSMUSG000000058927</td>
<td>GTWYGDIYTYLMELENPK</td>
<td>0.22</td>
<td>0.039</td>
<td>5</td>
<td>0.094</td>
</tr>
<tr>
<td>Pamb4</td>
<td>ENSMUSG00000005779</td>
<td>QPVLSQTLEAR</td>
<td>1.73</td>
<td>0.039</td>
<td>3</td>
<td>0.128</td>
</tr>
<tr>
<td>Psma6</td>
<td>ENSMUSG00000021024</td>
<td>LYQVEYAFK</td>
<td>1.54</td>
<td>0.039</td>
<td>3</td>
<td>0.146</td>
</tr>
<tr>
<td>Hnmpa3</td>
<td>ENSMUSG000000059005</td>
<td>GFAFVFTDDHDTVDK</td>
<td>0.59</td>
<td>0.040</td>
<td>6</td>
<td>0.062</td>
</tr>
<tr>
<td>Dnpep</td>
<td>ENSMUSG000000026209</td>
<td>GTPEPGLAYDER</td>
<td>0.50</td>
<td>0.040</td>
<td>3</td>
<td>0.068</td>
</tr>
<tr>
<td>Ap1b1</td>
<td>ENSMUSG00000009090</td>
<td>LGAPISSGSLSDLTLTSGVTSYGAPK</td>
<td>8.39</td>
<td>0.040</td>
<td>3</td>
<td>0.094</td>
</tr>
<tr>
<td>Gene</td>
<td>ENSMUSG00000000000</td>
<td>Protein</td>
<td>Fold Change</td>
<td>p-Value</td>
<td>q-Value</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>--------------------</td>
<td>---------</td>
<td>-------------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>Usp5</td>
<td>ENSMUSG00000038429</td>
<td>IVILPDYLEIAR</td>
<td>1.65</td>
<td>0.040</td>
<td>4</td>
<td>0.102</td>
</tr>
<tr>
<td>Aldh2</td>
<td>ENSMUSG000000239455</td>
<td>VVGNPPDSR</td>
<td>0.63</td>
<td>0.041</td>
<td>16</td>
<td>0.074</td>
</tr>
<tr>
<td>Fcgbp</td>
<td>ENSMUSG00000047730</td>
<td>VLIDGVLQYLPFQAAGGK</td>
<td>3.06</td>
<td>0.041</td>
<td>23</td>
<td>0.08</td>
</tr>
<tr>
<td>Lasp1</td>
<td>ENSMUSG00000038366</td>
<td>QQSELQSQVR</td>
<td>2.17</td>
<td>0.041</td>
<td>4</td>
<td>0.124</td>
</tr>
<tr>
<td>Arf6</td>
<td>ENSMUSG00000044147</td>
<td>FNYWDVGGQDK</td>
<td>1.46</td>
<td>0.041</td>
<td>4</td>
<td>0.138</td>
</tr>
<tr>
<td>Ap2b1</td>
<td>ENSMUSG000000335152</td>
<td>LAPPLVTLLSGEPIVQYVALR</td>
<td>3.26</td>
<td>0.042</td>
<td>5</td>
<td>0.114</td>
</tr>
<tr>
<td>Actr3</td>
<td>ENSMUSG00000026341</td>
<td>ISEELSGGR</td>
<td>0.54</td>
<td>0.043</td>
<td>6</td>
<td>0.112</td>
</tr>
<tr>
<td>A1R1</td>
<td>ENSMUSG00000060904</td>
<td>ILILGLDGAGK</td>
<td>1.46</td>
<td>0.044</td>
<td>3</td>
<td>0.11</td>
</tr>
<tr>
<td>Aldh9a1</td>
<td>ENSMUSG00000026687</td>
<td>VEPVDASGTIEK</td>
<td>0.42</td>
<td>0.045</td>
<td>4</td>
<td>0.1</td>
</tr>
<tr>
<td>Acox1</td>
<td>ENSMUSG00000020777</td>
<td>DVTGLGVLGR</td>
<td>1.59</td>
<td>0.045</td>
<td>7</td>
<td>0.112</td>
</tr>
<tr>
<td>Anxa1</td>
<td>ENSMUSG00000024659</td>
<td>GGPGSAYSPYPSFNVSSDVAALHK</td>
<td>0.64</td>
<td>0.046</td>
<td>5</td>
<td>0.078</td>
</tr>
<tr>
<td>Rpl13</td>
<td>ENSMUSG00000040074</td>
<td>KGDSAAELK</td>
<td>0.46</td>
<td>0.046</td>
<td>5</td>
<td>0.126</td>
</tr>
<tr>
<td>Gm8226</td>
<td>ENSMUSG00000064652</td>
<td>VAPAPAVVK</td>
<td>1.93</td>
<td>0.046</td>
<td>5</td>
<td>0.154</td>
</tr>
<tr>
<td>Actg1</td>
<td>ENSMUSG00000062825</td>
<td>QEYDESGPSIVHR</td>
<td>0.54</td>
<td>0.047</td>
<td>5</td>
<td>0.116</td>
</tr>
<tr>
<td>Hnmpa2b1</td>
<td>ENSMUSG0000004980</td>
<td>GFGFVVTFDDHDHPVDK</td>
<td>0.63</td>
<td>0.047</td>
<td>8</td>
<td>0.126</td>
</tr>
<tr>
<td>Gm10145</td>
<td>ENSMUSG00000042088</td>
<td>IYHPNIDIEK</td>
<td>1.40</td>
<td>0.048</td>
<td>3</td>
<td>0.112</td>
</tr>
<tr>
<td>Eef2</td>
<td>ENSMUSG00000034994</td>
<td>VFDAIMNFR</td>
<td>1.37</td>
<td>0.048</td>
<td>20</td>
<td>0.156</td>
</tr>
<tr>
<td>Uqerc2</td>
<td>ENSMUSG00000038884</td>
<td>GGLGLAGAK</td>
<td>0.50</td>
<td>0.049</td>
<td>13</td>
<td>0.046</td>
</tr>
<tr>
<td>Clca6</td>
<td>ENSMUSG00000068547</td>
<td>AEYHFTPFDPVLGR</td>
<td>7.75</td>
<td>0.049</td>
<td>4</td>
<td>0.156</td>
</tr>
<tr>
<td>Suclg1</td>
<td>ENSMUSG00000052738</td>
<td>HILGLPVFNTVK</td>
<td>0.64</td>
<td>0.050</td>
<td>3</td>
<td>0.098</td>
</tr>
</tbody>
</table>
Appendix II

Apc-Cdkn1a signaling:

Associated Data
Table I. List of proteins found differentially expressed in the villi and crypts of Apc<sup>1638N+/−</sup> mice (compared to the villi and crypts, respectively, of wild-type mice).

<table>
<thead>
<tr>
<th>Spot #</th>
<th>Location</th>
<th>Gene Symbol</th>
<th>Swiss Prot Acc</th>
<th>Coverage (%)</th>
<th>Mass (kDa)</th>
<th>pI</th>
<th>Score</th>
<th>t-test p-value</th>
<th>LIMMA p-value</th>
<th>Fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>442</td>
<td>Villi</td>
<td>ALDH1L1</td>
<td>Q8R0Y6</td>
<td>58</td>
<td>98647</td>
<td>5.64</td>
<td>3906</td>
<td>0.184</td>
<td>0.039</td>
<td>1.64</td>
</tr>
<tr>
<td>442</td>
<td>Villi</td>
<td>VIL1</td>
<td>Q62486</td>
<td>47</td>
<td>92743</td>
<td>5.72</td>
<td>4724</td>
<td>0.184</td>
<td>0.039</td>
<td>1.64</td>
</tr>
<tr>
<td>463</td>
<td>Villi</td>
<td>EEF2</td>
<td>P58252</td>
<td>62</td>
<td>92743</td>
<td>5.72</td>
<td>8484</td>
<td>0.012</td>
<td>0.002</td>
<td>1.871</td>
</tr>
<tr>
<td>479</td>
<td>Villi</td>
<td>GAPD2</td>
<td>Q64521</td>
<td>51</td>
<td>95253</td>
<td>6.41</td>
<td>5624</td>
<td>0.001</td>
<td>0.004</td>
<td>1.893</td>
</tr>
<tr>
<td>665</td>
<td>Villi</td>
<td>HSAP5</td>
<td>P20029</td>
<td>64</td>
<td>72377</td>
<td>5.07</td>
<td>11781</td>
<td>0.09</td>
<td>0.051</td>
<td>0.627</td>
</tr>
<tr>
<td>692</td>
<td>Villi</td>
<td>LCP1</td>
<td>Q3V0K9</td>
<td>60</td>
<td>70363</td>
<td>7.21</td>
<td>6060</td>
<td>0.02</td>
<td>0.003</td>
<td>2.357</td>
</tr>
<tr>
<td>782</td>
<td>Villi</td>
<td>EPHX2</td>
<td>P34914</td>
<td>17</td>
<td>62475</td>
<td>5.85</td>
<td>1316</td>
<td>0.034</td>
<td>0.031</td>
<td>1.507</td>
</tr>
<tr>
<td>902</td>
<td>Villi</td>
<td>HSPD1</td>
<td>P63038</td>
<td>42</td>
<td>60917</td>
<td>5.91</td>
<td>2937</td>
<td>0.016</td>
<td>0.009</td>
<td>1.57</td>
</tr>
<tr>
<td>1026</td>
<td>Villi</td>
<td>KRT8</td>
<td>Q61463</td>
<td>80</td>
<td>54531</td>
<td>5.42</td>
<td>14772</td>
<td>0.055</td>
<td>0.009</td>
<td>0.414</td>
</tr>
<tr>
<td>1110</td>
<td>Villi</td>
<td>MDH2</td>
<td>P08249</td>
<td>50</td>
<td>35589</td>
<td>8.93</td>
<td>1268</td>
<td>0.053</td>
<td>0.017</td>
<td>1.867</td>
</tr>
<tr>
<td>1110</td>
<td>Villi</td>
<td>GAPDH</td>
<td>P04797</td>
<td>48</td>
<td>35760</td>
<td>8.44</td>
<td>1387</td>
<td>0.053</td>
<td>0.017</td>
<td>1.867</td>
</tr>
<tr>
<td>1115</td>
<td>Villi</td>
<td>GAPDH</td>
<td>P16858</td>
<td>64</td>
<td>35787</td>
<td>8.44</td>
<td>4484</td>
<td>0.055</td>
<td>0.008</td>
<td>1.913</td>
</tr>
<tr>
<td>1176</td>
<td>Villi</td>
<td>OAT</td>
<td>P29758</td>
<td>71</td>
<td>48324</td>
<td>6.19</td>
<td>9578</td>
<td>0.039</td>
<td>0.003</td>
<td>1.515</td>
</tr>
<tr>
<td>1226</td>
<td>Villi</td>
<td>OAT</td>
<td>P29758</td>
<td>52</td>
<td>48324</td>
<td>6.19</td>
<td>3253</td>
<td>0.028</td>
<td>0.002</td>
<td>1.32</td>
</tr>
<tr>
<td>1375</td>
<td>Villi</td>
<td>KRT8</td>
<td>P11679</td>
<td>55</td>
<td>53210</td>
<td>5.42</td>
<td>4713</td>
<td>0.076</td>
<td>0.038</td>
<td>0.571</td>
</tr>
<tr>
<td>1402</td>
<td>Villi</td>
<td>HNRNPF</td>
<td>P19001</td>
<td>51</td>
<td>45701</td>
<td>5.31</td>
<td>3244</td>
<td>0.076</td>
<td>0.038</td>
<td>0.571</td>
</tr>
<tr>
<td>1431</td>
<td>Villi</td>
<td>ACTR3</td>
<td>Q9J4Y9</td>
<td>68</td>
<td>47327</td>
<td>5.61</td>
<td>4497</td>
<td>0.006</td>
<td>0.012</td>
<td>1.703</td>
</tr>
<tr>
<td>1533</td>
<td>Villi</td>
<td>OAT</td>
<td>P29758</td>
<td>71</td>
<td>48324</td>
<td>6.19</td>
<td>9578</td>
<td>0.039</td>
<td>0.003</td>
<td>1.515</td>
</tr>
<tr>
<td>1552</td>
<td>Villi</td>
<td>MDH1</td>
<td>P14152</td>
<td>42</td>
<td>36454</td>
<td>6.16</td>
<td>1597</td>
<td>0.009</td>
<td>1.74</td>
<td></td>
</tr>
<tr>
<td>1552</td>
<td>Villi</td>
<td>MDH1</td>
<td>P14152</td>
<td>42</td>
<td>36454</td>
<td>6.16</td>
<td>1597</td>
<td>0.009</td>
<td>1.74</td>
<td></td>
</tr>
<tr>
<td>1689</td>
<td>Villi</td>
<td>SULT1B1</td>
<td>Q9QWG7</td>
<td>64</td>
<td>34905</td>
<td>6.33</td>
<td>3318</td>
<td>0.144</td>
<td>0.049</td>
<td>1.879</td>
</tr>
<tr>
<td>1793</td>
<td>Villi</td>
<td>KRT18</td>
<td>P05784</td>
<td>43</td>
<td>47509</td>
<td>5.22</td>
<td>3682</td>
<td>0.006</td>
<td>0.007</td>
<td>0.481</td>
</tr>
<tr>
<td>1885</td>
<td>Villi</td>
<td>NDUFS8</td>
<td>Q5M9P5</td>
<td>45</td>
<td>24023</td>
<td>5.89</td>
<td>1689</td>
<td>0.037</td>
<td>0.079</td>
<td></td>
</tr>
<tr>
<td>1902</td>
<td>Villi</td>
<td>KRT19</td>
<td>P19001</td>
<td>57</td>
<td>44515</td>
<td>5.28</td>
<td>17414</td>
<td>0.013</td>
<td>0.012</td>
<td>0.49</td>
</tr>
<tr>
<td>1943</td>
<td>Villi</td>
<td>ATP5H</td>
<td>Q9Y1K9</td>
<td>65</td>
<td>18738</td>
<td>5.52</td>
<td>2127</td>
<td>0.001</td>
<td>0.018</td>
<td>0.597</td>
</tr>
<tr>
<td>1943</td>
<td>Villi</td>
<td>KRT20</td>
<td>Q9D312</td>
<td>44</td>
<td>49004</td>
<td>5.31</td>
<td>3791</td>
<td>0.001</td>
<td>0.018</td>
<td>0.597</td>
</tr>
<tr>
<td>2011</td>
<td>Villi</td>
<td>KRT19</td>
<td>P19001</td>
<td>60</td>
<td>44515</td>
<td>5.28</td>
<td>3948</td>
<td>0.005</td>
<td>0.042</td>
<td>0.426</td>
</tr>
<tr>
<td>442</td>
<td>Crypt</td>
<td>ALDH1L1</td>
<td>Q8R0Y6</td>
<td>58</td>
<td>98647</td>
<td>5.64</td>
<td>3906</td>
<td>0.029</td>
<td>0.069</td>
<td></td>
</tr>
<tr>
<td>442</td>
<td>Crypt</td>
<td>VIL1</td>
<td>Q62486</td>
<td>47</td>
<td>92743</td>
<td>5.72</td>
<td>4724</td>
<td>0.029</td>
<td>0.069</td>
<td></td>
</tr>
<tr>
<td>628</td>
<td>Crypt</td>
<td>PDIA3</td>
<td>P27773</td>
<td>63</td>
<td>56586</td>
<td>5.98</td>
<td>4423</td>
<td>0.043</td>
<td>0.019</td>
<td>1.534</td>
</tr>
<tr>
<td>906</td>
<td>Crypt</td>
<td>KRT8</td>
<td>P11679</td>
<td>16</td>
<td>54531</td>
<td>8</td>
<td>1117</td>
<td>0.009</td>
<td>0.015</td>
<td>1.625</td>
</tr>
<tr>
<td>930</td>
<td>Crypt</td>
<td>MEF2C</td>
<td>P11679</td>
<td>16</td>
<td>54531</td>
<td>8</td>
<td>1117</td>
<td>0.009</td>
<td>0.015</td>
<td>1.625</td>
</tr>
<tr>
<td>1661</td>
<td>Crypt</td>
<td>CAPZB</td>
<td>A2AMV7</td>
<td>71</td>
<td>33746</td>
<td>6.02</td>
<td>3020</td>
<td>0.165</td>
<td>0.044</td>
<td>1.804</td>
</tr>
</tbody>
</table>
Table II. List of proteins found differentially expressed in the villi and crypts of \textit{Cdkn1a}^{-/-} mice (compared to villi and crypts, respectively, of wild-type mice). P-values less than 0.05 are highlighted in yellow. Upregulation is highlighted in green; downregulation is highlighted in red.

<table>
<thead>
<tr>
<th>Spot #</th>
<th>Location</th>
<th>Gene Symbol</th>
<th>SwissProt ID</th>
<th>Coverage (%)</th>
<th>Mass (kDa)</th>
<th>pI</th>
<th>Score</th>
<th>t-test p-value</th>
<th>LIMMA p-value</th>
<th>Fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1552</td>
<td>Villi</td>
<td>ANXA2</td>
<td>P07356</td>
<td>59</td>
<td>39839</td>
<td>5.96</td>
<td>2160</td>
<td>0.04</td>
<td>1.32</td>
<td></td>
</tr>
<tr>
<td>1039</td>
<td>Villi</td>
<td>ATP5A1</td>
<td>Q03265</td>
<td>49</td>
<td>59716</td>
<td>9.22</td>
<td>4277</td>
<td>0.09</td>
<td>0.05</td>
<td>1.50</td>
</tr>
<tr>
<td>479</td>
<td>Villi</td>
<td>EEF2</td>
<td>P58252</td>
<td>51</td>
<td>95253</td>
<td>6.41</td>
<td>5624</td>
<td>0.04</td>
<td>0.05</td>
<td>1.53</td>
</tr>
<tr>
<td>1039</td>
<td>Villi</td>
<td>GLUD1</td>
<td>Q8C273</td>
<td>61</td>
<td>61298</td>
<td>8.05</td>
<td>5415</td>
<td>0.09</td>
<td>0.05</td>
<td>1.50</td>
</tr>
<tr>
<td>1552</td>
<td>Villi</td>
<td>GPD1</td>
<td>P13707</td>
<td>50</td>
<td>37560</td>
<td>6.75</td>
<td>2198</td>
<td>0.04</td>
<td>1.32</td>
<td></td>
</tr>
<tr>
<td>1793</td>
<td>Villi</td>
<td>KRT18</td>
<td>P05784</td>
<td>43</td>
<td>47509</td>
<td>5.22</td>
<td>3682</td>
<td>0.04</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Villi</td>
<td>KRT19</td>
<td>P19001</td>
<td>60</td>
<td>44515</td>
<td>5.28</td>
<td>3948</td>
<td>0.01</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>1032</td>
<td>Villi</td>
<td>KRT8</td>
<td>Q3TJE1</td>
<td>83</td>
<td>54514</td>
<td>5.42</td>
<td>13024</td>
<td>0.04</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>1623</td>
<td>Villi</td>
<td>MDH1</td>
<td>P14152</td>
<td>42</td>
<td>36454</td>
<td>6.16</td>
<td>1597</td>
<td>0.04</td>
<td>1.30</td>
<td></td>
</tr>
<tr>
<td>1552</td>
<td>Villi</td>
<td>OTC</td>
<td>P11725</td>
<td>50</td>
<td>38652</td>
<td>7.55</td>
<td>3748</td>
<td>0.04</td>
<td>1.32</td>
<td></td>
</tr>
<tr>
<td>463</td>
<td>Villi</td>
<td>VIL1</td>
<td>Q62468</td>
<td>62</td>
<td>92743</td>
<td>5.72</td>
<td>8484</td>
<td>0.01</td>
<td>0.02</td>
<td>1.50</td>
</tr>
<tr>
<td>691</td>
<td>Crypts</td>
<td>HSPA9</td>
<td>P38647</td>
<td>56</td>
<td>73483</td>
<td>5.91</td>
<td>7584</td>
<td>0.16</td>
<td>0.03</td>
<td>1.86</td>
</tr>
</tbody>
</table>
Appendix III

Apc-Cdkn1a Signaling:

Additional Methods
Methods for Forming a Filtered Protein-Protein Interaction Network

Eliminating False Positives via a logistic regression model: Analysis based on the agreement of PPI and their expression data shows less than half of these interactions are biologically relevant. In our logistic regression model, we incorporate (1) the number of times an interaction between two proteins was observed, (2) the Pearson correlation of expression measurements for the corresponding genes, (3) the proteins’ small world clustering coefficient, and (4) the protein subcellular localization data of interacting partners. As expected, the protein subcellular localization data would eliminate interactions among proteins that are unlikely to come into contact with each other. Given the four input variables, $X = (X_1, X_2, X_3, X_4)$, the probability of a true interaction between two proteins $i$ and $j$, $Pr(I_{ij})$ under the logistic distribution is

$$Pr(I_{ij} | X) = \frac{1}{1 + \exp(-\beta_0 - \sum_{i=1}^{4} \beta_i X_i)}.$$  

Given positive and negative training data sets, one can optimize the parameters to maximize the likelihood of a true interaction. We acquired randomly selected 1000 PPIs from the MIPS database of interactions, an accepted gold standard as our positive data set. The negative training set was composed of 1000 randomly selected PPIs not contained in the MIPS database, since false positives are in high abundance in the training data sets (similarly employed in). These experiments were repeated 1000 times and a cut off point for the probability of true interactions was determined.

Recovering false negative interactions via homology: A protein family is a group of evolutionarily related proteins based on sequence similarity. Additionally, it has been also observed that sequence-wise similar proteins share similar interaction patterns in the same organism, suggesting that proteins within the same family are likely to have similar
interaction patterns. The Protein Family database was downloaded from Pfam\textsuperscript{146}, and we inferred an edge between two proteins if (1) they are not already known to interact in the PPIN, and (2) there exists at least one interaction between the families of these two proteins.

**GO Annotations and Association Rule Mining:** Biological annotations, e.g., Gene Ontology\textsuperscript{195} Biological Process (GO-BP) annotations provide a basis to find functionally similar proteins. We map proteins in both known signaling pathways and protein-protein interaction networks to their annotations. In this study, the training data of known signaling pathways is collected from various publicly available pathways databases\textsuperscript{144, 162, 196, 197}, and pathway connections are converted to tuples of interacting proteins. Next, GO-BP annotations of pathway proteins are collected and kept as functionality sets. For each protein, associations between gene product and GO terms are queried from the GO Database; the Biological Process ontology terms acquired are leaf nodes on the directed acyclic GO term graph. Each annotation of a protein is linked with its interacting neighbor’s annotations and a network of annotation links is formed. All possible combinations are examined since they represent all possible functional associations. Association rule mining is then used to discover rules of GO annotation pairs, collecting the underlying patterns of signaling pathways to form a library of templates. The data used in this study generated results with 28572 observed association rules when the support value is 0.000003 and the confidence is 0.001 (Refer to\textsuperscript{194} on how to pick a threshold). These rules and parameters are used to evaluate candidate pathway segments for possible occurrences of these rules.

**Interactions with weight assignments:** A weighted PPIN is formed by calculating Pearson correlation coefficient of the interacting pairs’ gene expression levels. In this study, the absolute value of $corr(e)$ is used to capture correlation ($r=|corr(e)|>0.7$). Usually the
correlation of the expression genes provides some evidence as to whether the produced proteins are biologically related.

**Searching for pathway segments:** Our hypothesis is that, given association rules that capture the characteristics of some known pathways, and a weighted PPIN, a pathway segment should belong to a pathway if (1) it contains at least a certain number of these rules and (2) the average weight of interactions is above a given threshold. For filtering by GO annotation association rules, each pair of interacting proteins’ functional annotations is checked for a match with a tuple from the association rules set (the “rules” being defined from *a priori* known signaling pathways, e.g. KEGG). For each selected path, an average absolute expression correlation coefficient is also calculated, which is then compared to the threshold ($|r|>0.7$). This extra filtering improves the outcome since true interactions often exhibit stronger correlation with expression levels. These values are more apparent when examining signaling paths, i.e. chains of proteins. The candidate paths with $p$-values higher than the threshold are returned as query results.
References


36. Heard, E. et al. Ten years of genetics and genomics: what have we achieved and where are we heading? Nat Rev Genet 11, 723-733.


109. Holla, V.R., Backlund, M.G., Yang, P., Newman, R.A. & Dubois, R.N. Regulation of 
prostaglandin transporters in colorectal neoplasia. Cancer Prev Res (Phila) 1, 93-99 
(2008).

signaling in rheumatoid synovial fibroblasts: involvement of Epac1 and PKA. 
Prostaglandins Other Lipid Mediat 89, 26-33 (2009).

111. Dube, N. et al. The RapGEF PDZ-GEF2 is required for maturation of cell-cell 

112. Bailey, C.L., Kelly, P. & Casey, P.J. Activation of Rap1 promotes prostate cancer 

113. Nuoffer, C., Davidson, H.W., Matteson, J., Meinkoth, J. & Balch, W.E. A GDP-
bound of rab1 inhibits protein export from the endoplasmic reticulum and transport 

114. Grigoriev, I. et al. Rab6 regulates transport and targeting of exocytic carriers. Dev 


116. Junutula, J.R. et al. Rab14 is involved in membrane trafficking between the Golgi 

117. Garrison, W.D. et al. Hepatocyte nuclear factor 4alpha is essential for embryonic 

118. Yuan, X. et al. Identification of an endogenous ligand bound to a native orphan 


191. Deng, M., Sun, F. & Chen, T. in *Pac Symp Biocomput* 140--151Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, 1042 West 36th Place, Los Angeles, CA 90089-1113, USA.; 2003).


