Title
Analysis of genomic variants via gene networks

Permalink
https://escholarship.org/uc/item/01v0n0pf

Author
Hofree, Matan

Publication Date
2014

Supplemental Material
https://escholarship.org/uc/item/01v0n0pf#supplemental

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA, SAN DIEGO

Analysis of genomic variants via gene networks

A dissertation submitted in partial satisfaction of the requirements for
the degree Doctor of Philosophy

in

Computer Science

by

Matan Hofree

Committee in charge:
Professor Ideker Trey, Chair
Professor James Fowler
Professor Kelly Frazer
Professor Yoav Freund
Professor Lawerence Saul

2014
The Dissertation of Matan Hofree is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

________________________________________

________________________________________

________________________________________

________________________________________

________________________________________

________________________________________

Chair

University of California, San Diego

2014
DEDICATION

This doctoral dissertation is dedicated to the cherished memory of Doron Golan, a great friend and source of inspiration. In our time together, he taught me a life-changing lesson about dedication and the measure of perseverance. In my academic journey, I have asked myself for the guidance he would have given, the inspiration and creative insight he would offhandedly dispense. I forgave myself of follies, as he could not. I enjoy life, as he would not.
Imagination will often carry us to worlds that never were. But without it we go nowhere.
- Carl Sagan

We don't stop playing because we grow old; we grow old because we stop playing.
- George Bernard Shaw

Per Aspera Ad Astra.
- Obligatory latin saying
# TABLE OF CONTENTS

Dedication ................................................................................................................... iv

Epigraph ........................................................................................................................ v

Table of Contents........................................................................................................ vi

List of Abbreviations .................................................................................................... ix

List of Figures ............................................................................................................. xii

List of Tables .............................................................................................................. xv

List of Supplemental Files .......................................................................................... xvi

Acknowledgements ..................................................................................................... xvii

Vita ............................................................................................................................... xix

Abstract of the Dissertation ....................................................................................... xxi

Chapter 1 Introduction............................................................................................... 1
  1.1 Preface .................................................................................................................. 1
  1.2 A (brief) introduction to network biology ....................................................... 3
    1.2.1 Examples of using networks for biological inference ....................... 5

Chapter 2 Inferring sparse multivariate models to predict disease phenotype from genotype .................................................................................................................................. 8
  2.1 Background and Significance ........................................................................... 8
  2.2 Methods ............................................................................................................. 12
    2.2.1 Genotype-phenotype data ................................................................. 13
Chapter 4 Saturation and power analysis for detecting combinatorial gene association in cancer

4.1 Background and Significance ................................................................. 70
4.2 Methods .................................................................................................. 73
  4.2.1 Data processing ........................................................................... 73
  4.2.2 Cancer gene reference Lists ......................................................... 73
  4.2.3 Global power analysis: ................................................................. 74
  4.2.4 Estimating relative mutation frequency ......................................... 75
  4.2.5 Functional enrichment testing ..................................................... 76
  4.2.6 Conservation enrichment testing ................................................... 77
  4.2.7 Identifying individual enriched cancer genes ................................ 77
  4.2.8 Testing for pairwise enrichment of genes .................................... 78
  4.2.9 Switching permutation co-mutation background model .......... 79
  4.2.10 Pareto estimation of p-values .................................................... 80
4.3 Results .................................................................................................... 80
  4.3.1 Power analysis for detecting cancer genes ............................... 80
  4.3.2 The role of infrequent mutation in cancer ................................. 81
  4.3.3 Power and significance of co-mutation and exclusivity ............ 85
4.4 Discussion ............................................................................................. 87
4.5 Chapter Acknowledgements ................................................................. 90

Appendices .................................................................................................... 91

References ................................................................................................... 119
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADT</td>
<td>Alternating Decision Trees</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under the Curve</td>
</tr>
<tr>
<td>BD</td>
<td>Bipolar Disorder</td>
</tr>
<tr>
<td>BP</td>
<td>Gene ontology biological process</td>
</tr>
<tr>
<td>CAD</td>
<td>Coronary artery disease</td>
</tr>
<tr>
<td>CC</td>
<td>Gene ontology cellular component</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy number variations</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>FPR</td>
<td>False positive rate</td>
</tr>
<tr>
<td>GC</td>
<td>Type-1 Diabetes Genetics Consortium</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-Wide Association Studies</td>
</tr>
<tr>
<td>ICGC</td>
<td>International Cancer Genome Consortium</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>MAF</td>
<td>Minor allele frequency</td>
</tr>
<tr>
<td>MF</td>
<td>Gene ontology molecular function</td>
</tr>
<tr>
<td>MHC</td>
<td>Major histocompatibility complex</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger ribonucleic acid</td>
</tr>
<tr>
<td>NBS</td>
<td>Network-based stratification</td>
</tr>
<tr>
<td>NetNMF</td>
<td>Network regularized non-negative matrix factorization</td>
</tr>
<tr>
<td>NMF</td>
<td>Non-negative matrix factorization</td>
</tr>
<tr>
<td>NS</td>
<td>non-synonymous</td>
</tr>
<tr>
<td>OG</td>
<td>Oncogenes</td>
</tr>
<tr>
<td>PPI</td>
<td>Protein-protein interaction</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>SAM</td>
<td>Significance Analysis of Microarrays</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphisms</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>T1D</td>
<td>Type-1 Diabetes</td>
</tr>
<tr>
<td>TCGA</td>
<td>The Cancer Genome Atlas</td>
</tr>
<tr>
<td>TPR</td>
<td>True positive rate</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>TSG</td>
<td>Tumor-suppressor genes</td>
</tr>
<tr>
<td>VI</td>
<td>Variable importance</td>
</tr>
<tr>
<td>WT</td>
<td>Wellcome Trust Case-Control Consortium</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2.1. Accuracy of Adaboost on Type 1 diabetes. ...................................... 19
Figure 2.2. Example decision rules learned using Adaboost. ......................... 21
Figure 2.3. Proximity of predictive SNP sets.................................................. 23
Figure 2.4. 'Classical' vs 'Boosting' Manhattan plots.................................. 25
Figure 2.5. Gene and pathway overlap.......................................................... 27
Figure 2.6. A boosting SNP gene interaction map......................................... 31
Figure 3.1. Overview of network-based stratification.................................... 42
Figure 3.2. Exploring performance of stratification through simulation......... 52
Figure 3.3. Network-based stratification of somatic tumor mutations........... 54
Figure 3.4. Comparing data types................................................................. 57
Figure 3.5. A network view of genes with high network smoothed mutation scores in HumanNet subtype 1 relative to other subtypes...................... 58
Figure 3.6. From mutation-derived subtypes to expression signatures........... 60
Figure 3.7. Effects of different types of mutations on stratification................. 63
Figure 4.1. Recovering known cancer genes using mutation frequency relative to background................................................................. 82
Figure 4.2. Evidence for functional implications of mutations in cancer genes irrespective of mutation frequency......................................................... 84
Figure 4.3. Exploring interactions among mutated genes......................... 87
Figure S2.1. Example SNPs with poor genotype calls...........................................91
Figure S2.2. Accuracy of Adaboost on Type I diabetes........................................92
Figure S2.3. Proximity of SNP sets in results based on genotype calls.................93
Figure S2.4. Supporting evidence for interaction edges recovered from SNP disease association.................................................................94

Figure S3.1. An overview of the somatic mutation landscape of the TCGA ovarian cancer cohort.................................................................95
Figure S3.2. Simulating across different networks...........................................96
Figure S3.3. Ovarian cancer association with overall survival...........................97
Figure S3.4. Lung cancer association with overall survival.............................98
Figure S3.5. Uterine cancer association with histological type.........................99
Figure S3.6. Standard predictors of survival are independent of ovarian subtype.................................................................100
Figure S3.7. A network view of genes with high network smoothed mutation scores in ovarian subtype 2 relative to other subtypes, using HumanNet.......101
Figure S3.8. A network view of genes with high network smoothed mutation scores in ovarian subtype 3 relative to other subtypes, using HumanNet.......102
Figure S3.9. A network view of genes with high smoothed mutation scores in ovarian subtype 4 relative to other subtypes, using HumanNet..............103
Figure S3.10. A network view of genes with high smoothed mutation scores in uterine cancer subtype 1 relative to other subtypes, using STRING.............104
Figure S3.11. A network view of genes with high smoothed mutation scores in uterine cancer subtype 2 relative to other subtypes, using STRING.............105
Figure S3.12. A network view of genes with high smoothed mutation scores in uterine cancer subtype 3 relative to other subtypes, using STRING.................106

Figure S3.13. A network view of genes with high smoothed mutation scores in lung cancer subtype 1 relative to other subtypes, using HumanNet.................107

Figure S3.14. A network view of genes with high smoothed mutation scores in lung cancer subtype 2 relative to other subtypes, using HumanNet.................108

Figure S3.15. A network view of genes with high smoothed mutation scores in lung cancer subtype 3 relative to other subtypes, using HumanNet.................109

Figure S3.16. A network view of genes with high smoothed mutation scores in lung cancer subtype 5 relative to other subtypes, using HumanNet.................110

Figure S3.17. From mutation-derived subtypes to expression signatures.......111

Figure S3.18. Standard consensus clustering NMF used to recover subtypes in the Tothill et al. expression cohort of ovarian tumors.............................................112

Figure S3.19. Effect of permuting proportions of the lung cancer dataset.....113

Figure S4.1. Estimate of the number of samples needed to detect significantly mutated genes........................................................................................................115

Figure S4.2. Relative mutation frequency as surrogate predictor for mutSigCV genes.................................................................................................................................116

Figure S4.3. Functional score of cancer genes across the range of mutation frequency within cancer type.................................................................117

Figure S4.4. Conservation of cancer genes across the range of mutation frequency within cancer type.........................................................................................118
LIST OF TABLES

**Table 1.1.** A summary of several common varieties of biological network and some examples. ........................................................................................................... 4

**Table S3.1.** Summary of gene interaction networks.................................114
LIST OF SUPPLEMENTAL FILES

Supplemental File 1. Supplementary tables accompanying chapter 2.
ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Trey Ideker, for his encouragement, advice, and the great freedom to pursue my own questions and ideas. This work would not have been possible without his insight and support. I would like to thank Professor Yoav Feund for his guidance and advice; Professor Kelly Frazer for her insights and invaluable career advice. I would also like to thank Professor Hanah Margalit my undergraduate research mentor, for her patience, guidance, and encouragement. I also thank my friend and colleague, Dr. Janusz Dutkowski, for years of encouragement and advice and all members of the Ideker lab past and present. Finally, I thank the remaining members of my dissertation committee for their input and advice.

I would like to acknowledge individual co-authors for their contributions to this dissertation:

Chapter 1, contains material published as part of the following review, of which I am a co-author:


Chapter 2, contains material currently in submission as part of the following manuscript, of which I am a primary co-investigator and co-author:

Raj, A.*, Hofree, M.*, Ideker, T, Wiggins, C.,Freund Y. Inferring sparse multivariate models to predict disease phenotype from genotype
Chapter 3, is a reprint of material from the following published work, of which I am the primary investigator and author:


Chapter 34, contains material currently being prepared for submission for publication, of which I am the primary investigator and author:

Hofree, M., Carter H., and Ideker, T., Limitations of mutation frequency for detecting gene associations in cancer.
VITA

1998-2003 Officer (Lieutenant), Infantry Brigades, Israeli defense force

2007 Bachelor of Science, Hebrew University of Jerusalem, Israel

2007-2008 Software Engineer, Imperva, Ltd.

2011 Master of Science, University of California, San Diego

2014 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS


ABSTRACT OF THE DISSERTATION

Analysis of genomic variants via gene networks

by

Matan Hofree

Doctor of Philosophy in Computer Science

University of California, San Diego, 2014

Professor Trey Ideker, Chair

Genome-wide measurements of genomic state offer unprecedented opportunities for biological discovery, with potential to make dramatic impact on medicine and life. One fundamental challenge is associating complex phenotypes with genetic cause. Here, I will describe efforts to advance solutions to this challenge via analysis of gene networks.
Genome-wide association studies are designed link between a phenotype and genomic loci anywhere in the genome; however, applying standard statistics to such data has fallen far short of building accurate predictive models for disease. We use Adaboost, a large-margin classification algorithm, to predict disease status in two cohorts of diabetes and suggest a method for overcoming limitations arising from correlation between genetic variants. We uncover a novel set of 163 disease-associations, missed by ‘classic’ statistics.

Classification of cancer remains predominantly organ based and fails to account for considerable heterogeneity of outcomes. Tumor genomes provide a new source of data for uncovering subtypes, but are difficult to compare, as tumors share few mutations in common. We introduce network-based stratification (NBS), a method for integrating somatic genomes with networks encoding biological knowledge. This allows for identification of cancer subtypes by clustering tumors with mutations in similar network regions. We demonstrate NBS in multiple cancer cohorts, identifying subtypes predictive of clinical features and outcomes, and highlighting sub-networks characteristic of each.

Current approaches for identifying cancer genes rely on the idea that particular perturbations, occurring in a subset of genes unique to each cancer type, are selected for by conferring a survival advantage to tumor cells. Such genes are expected to be enriched for mutations when examined across a population. Here we show that 30-50% of well-known cancer genes
are not significantly elevated in mutation frequency. Despite this lack of enrichment, known cancer genes are enriched for mutations causing changes in amino-acid composition, protein structure properties and conservation. Furthermore, we observe 15-30% of cancer genes have altered mutation rates conditioned on other genes, each individually spanning the range of single-gene mutation frequencies, implicating a large genetic interaction network underlying human cancer. This suggests a substantial number of cancer genes will never be identified by frequency alone.
CHAPTER 1
INTRODUCTION

1.1 Preface

The biological sciences are undergoing an information explosion. A ‘Moore’s law’ of sequencing is beating its computational variant (1), the $1000 genome is here (2), new technologies are emerging which allow us to peer into the exact states of single cells (3), while technologies for quantification of a multitude of ‘omics have matured. This information downpour offers many computational and biological challenges on multiple fronts.

One important challenge is to find robust and efficient ways to combine measurements into accurate predictions. Biological data sets are often under-constrained problems; involving many, possibly redundant, often noisy, measurements of a biological property of interest. In contrast, the sample size for said measurements is in general smaller than the feature set size (D >> n). This issue has been extensively studied in the field of machine learning, however, tailoring existing methodologies to the specific characteristics of problems in the biological domain, remains an important topic.

A second challenge is the analysis of the interaction space of such measurements in a principled and efficient manner. Our current
understanding of biological processes suggests that much of the complexity of life is hidden in complex regulatory circuitry and high-order interactions. To illustrate the importance of this property, one may consider for example the nematode *Caenorhabditis elegans* with its fully charted 959 cell body has nearly the same number of genes as a human, which has nearly half as many distinctly different cell types (4). An elegant demonstration of this point is presented in a series of experiments examining two closely related strains of the yeast species *Saccharomyces cerevisiae*. The two strains, differing genetically to the same extent as two humans, exhibit differences in 57 genes whose deletion causes death in one strain but not the other. The experimenters go on to show that in most cases these conditionally essential genes occur as a result of more than three background-specific modifiers in each strain. Despite the clearly apparent importance of considering this interaction space, considerable biological study is still devoted to the study of simple relationships between a single gene and a single phenotype.

A third prominent challenge involves the systematic encoding and use of existing biological knowledge to inform on new experiments and improve prediction and modeling in the lab and clinical setting. One methodology for addressing this challenge encodes experimental results as graphs or hierarchical ontology structures describing different types of relationships between genes or proteins. Efficiently using these structures to better inform classification and drive new discovery remains at the forefront of networks and systems biology.
In the following chapters, I will describe some of my work aimed at making headway on addressing the above challenges. The rest of the chapter will briefly survey key concepts and examples of network biology. Chapter 2 will describe a method for uncovering associations between genetic loci and complex phenotypes that go beyond simple gene-phenotype association. Chapter 3 will describe our efforts to use prior biological knowledge in the form of gene interaction networks together with mutation profiles to discover novel subtypes of cancer. Chapter 4 will present an analysis of the role of rarely mutated genes and higher order interaction effects in cancer.

### 1.2 A (brief) introduction to network biology

Considerable efforts in the study of biology have been directed into systemizing, organizing and encoding, the results of several decades’ worth of experiments and domain knowledge into more accessible forms, which allow for efficient application of computational methods.

One approach toward overcoming this challenge, has taken the path of organizing biological knowledge in the form of graph structures, often referred to as (gene or protein) networks. Biological networks can be defined using the formal definitions of a graph – $G(V,E)$. Where $V$ is a set biological entities, most often proteins or genes (5-7), but could also possibly be for regulatory regions in the genome (8), enzymatic reactions, substrates and metabolites, etc.$E \subset V \times V$, is usually used to describe the set of observed relationships between the entities. These could capture relationship of genes
or proteins known to interact physically (as in the case protein interaction events), in a regulatory role such as in the case of transcription factor activating the transcription of some other gene, or even merely to be part of related cellular mechanisms which are often mentioned together in the exiting literature. Table 1.1 summarizes a brief survey of some common types of biological networks.

Table 1.1. A summary of several common varieties of biological network and some examples.

<table>
<thead>
<tr>
<th>Network type</th>
<th>Nodes</th>
<th>Edges</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein-protein interaction (PPI)</td>
<td>Proteins</td>
<td>Physical interactions</td>
<td>HPIN(9)</td>
</tr>
<tr>
<td>Structurally resolved PPI</td>
<td>Protein</td>
<td>Physical interactions</td>
<td>HSIN(9), SIN(10), Interactome3D(11), INstruct (12)</td>
</tr>
<tr>
<td>Protein-DNA interaction</td>
<td>Transcription factors</td>
<td>Transcription factor DNA binding</td>
<td>(13, 14)</td>
</tr>
<tr>
<td>Co-expression</td>
<td>Proteins</td>
<td>Common expression</td>
<td>(15)</td>
</tr>
<tr>
<td>Genetic interaction (GI)</td>
<td>Genes</td>
<td>Common function</td>
<td>(16, 17)</td>
</tr>
<tr>
<td>Difference</td>
<td>Genes</td>
<td>Differential function</td>
<td>(18)</td>
</tr>
<tr>
<td>Metabolic</td>
<td>Enzymes, Metabolites</td>
<td>Biochemical reactions</td>
<td>(19)</td>
</tr>
<tr>
<td>Non-coding RNAs</td>
<td>miRNA, lincRNA, asRNA, target genes</td>
<td>Physical interactions, common function</td>
<td>(20)</td>
</tr>
<tr>
<td>Integrated</td>
<td>Any</td>
<td>Any</td>
<td>HumanNet(21), BioGrid(22)</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>Any</td>
<td>Any</td>
<td>Nexo(23)</td>
</tr>
</tbody>
</table>

Integrated networks are constructed using a variety of different methods, using both low and high throughput experimental techniques. In the work discussed below, I use networks, which combine evidence from multiple
different experimental techniques, as well as evidence from systematic mining of existing literature. One difficulty of using such networks is that they often include certain types of bias (e.g. over representation of well-studied genes) and a variety of sources of error. To improve the signal-to-noise ratio of using these networks we often prefilter the networks to include only a fraction of the most high confidence edges. In internal benchmarks using the GO ontology as a reference gold-standard, I have observed an improved classification performance (AUC) after performing such pruning (data not shown).

1.2.1 Examples of using networks for biological inference

Networks provide a framework for deriving information from a set of relationships among biological entities. In models of sub-cellular biological processes, network nodes are typically genes, proteins, nucleic acids or metabolites, and edges represent physical interactions or a rich variety of functional associations (Table 1.1). Hybrid networks that are mixtures of different types of relationships are prevalent as well.

Biological network models can be constructed from systematic genome-wide unbiased screens or focused interrogation of distinct biological functions. For complex disorders that are poorly characterized, mapping candidate genes and mutations implicated by association studies onto holistic network models can implicate underlying biological processes (Table 2). In a recent GWAS of coronary artery disease (CAD), Deloukas et al. identified subnetworks enriched for genes implicated by variable expression with or physical proximity to SNPs in a larger protein-protein interaction (PPI)
network \cite{15}. Subsequent gene set analysis to determine functional enrichment of the subnetworks, and analysis of subnetwork overlap with canonical pathways implicated crosstalk between lipid metabolism and inflammatory pathways as underlying the pathogenesis of CAD.

If the disease is better understood, focused models may enable development of specific biological hypotheses about the mechanisms by which alterations cause disease. For example, Chu et al. constructed a network of protein interactions involved in angiogenesis, which they dub “the angiome”, in order to study diseases related to irregular blood vessel formation \cite{16}. In another example, a network of human-HIV protein complexes constructed by affinity tagging and purification mass spectrometry has provided a near-comprehensive view of how HIV evades host cell defenses \cite{17}. While focused approaches represent only a partial view of the cell, the resulting networks provide an intelligent framework for constraining hypothesis testing to proteins most relevant to a disease. On the other hand, focused screens may miss systems level trends, for example cross-talk between biological processes that can play a role in disease \cite{18}.

Network edges can also represent abstract relationships derived from biological knowledge. Gilman et al. built a network where all pairs of proteins are connected by a weighted edge representing the \textit{a priori} expectation that the proteins participate in the same phenotype. Edge weights were based on evidence sources such as tissue-specific expression, pathway membership, common functional annotations and similar domain composition \cite{19}. They
then searched over this network to identify the most functionally similar genes affected by de novo copy number variants (CNVs) in autism cases.
CHAPTER 2
INFERRING SPARSE MULTIVARIATE MODELS TO PREDICT DISEASE PHENOTYPE FROM GENOTYPE

2.1 Background and Significance

Genome-Wide Association Studies (GWAS) spurred by rapid, technological advances and decreasing costs in high-throughput sequencing and genotyping, have become a promising tool to answer fundamental questions in the study of the genetic basis of complex traits and disease. Built on the foundations of the ‘Common Disease - Common Variant’ hypothesis, GWAS are driven by recent advances in quantifying patterns of inheritance of Single Nucleotide Polymorphisms (SNP) – the most common form of genomic variation – among different populations (24). These large-scale studies aim at inferring the genotype of hundreds of thousands of common genetic polymorphisms for several hundreds of cases and controls for a phenotype of interest and elucidating those variants that have strong, significant association with the phenotype of interest. Armed with phenomenal amounts of genomic data, geneticists can then infer the genetic architecture of common traits. In particular the number of genetic loci that underlie variation in heritable traits. The distribution of their effect sizes (the contribution of each locus to variation in a phenotype). The complex mechanisms of action (additive/dominant), possible epistatic interactions, and the dependence on environmental
conditions that characterize disease-associated variants. The phenotypes of interest include common, large spectrum diseases such as Type-1 Diabetes (T1D), Bipolar Disorder (BD), and Autism, and quantitative traits such as height, BMI, and blood cholesterol level.

Unlike hypothesis-driven linkage and family-based studies, GWAS are hypothesis-generating studies discovering polymorphisms that have significant association with the phenotype of interest, also called risk variants. Given the strong linkage disequilibrium structure in the human genome (25) and the limitations on the number of SNPs that current chips can hold, most studies choose a suitable set of common tag SNPs, with minor allele frequency (MAF) greater than 5%, that well approximate the variation in the human genome. Since it is unknown whether the causal variants are included in the genotyped set of SNPs, the inferred risk variants can only suggest genomic regions that contribute to the phenotype of interest; fine mapping and expression studies focused on these regions can then identify rarer, putative causal variants and their relevant causal pathways (26, 27). Most GWAS to-date have employed traditional single-variate statistical tools based on multi-factorial analysis of variance or case-control statistics to resolve traits into associated loci on molecular marker maps. Over the last decade, several large scale GWAS have discovered a number of common SNPs to be strongly implicated in age-related macular degeneration (28), Type-1 diabetes (29), obesity (30, 31) and several other traits and diseases (32, 33).
The use of single-variate statistics for such high-dimensional problems, however, demands tight statistical constraints to correct for the inevitable multiple-hypothesis nightmare and to weed out spurious correlation signals between SNPs and phenotype (34). This often leads to reduced statistical power, thus requiring larger and larger sample sizes, careful meta-analysis (30) or multi-tiered studies (35) to detect putative associations that did not pass the necessary statistical constraints. Furthermore, the single SNP tests often assume that disease-predisposing variants function additively, thus inevitably neglecting information in their joint distribution. Despite several successes of GWAS, an oft-mentioned failure is the low fraction of sibling recurrent risk that is accounted for by the risk variants that have been detected to-date for several heritable traits (36). Finally, very few studies have explored the inference of models from genotypic data that are predictive of genetic risk of disease. Most of these studies have focused on building a predictive model purely from associated variants that passed stringent statistical controls (37) (38); a striking feature of their results is the extremely poor predictive power conferred by those risk variants detected by current study sizes (30, 37-39).

The restricted success of GWAS to strongly heritable diseases, despite large study sizes (37), suggests that student t-tests, case control studies with p-values, and other cornerstones of orthodox statistics simply are not the appropriate high-dimensional statistical approaches to build disease predictive models and reveal disease relevant genetic variants for complex diseases. The variety of sequence loci constitute an overwhelmingly large
number of features yet given a typical GWAS experimental study, the number of individuals and the diversity of phenotypic variation are not sufficient to reveal which of these hundreds of thousands of covariates constitute the predictive risk loci. Despite the fact that clinicians widely recognize the insufficiency of existing statistical approaches (34), genetics remains firmly entrenched in low-dimensional or one-dimensional statistical tools, which do little to help us escape the above-mentioned multiple hypothesis nightmare.

More recently, there has been some attempt to move away from the single-variate statistical tools that have been extremely popular in GWAS. Some studies have analyzed groups of SNPs for associations with disease (40-42), reducing the exponentially growing space of combinatorial features by searching only over groups of variants that are co-located with the same gene, groups of genes sharing the same ontology or in some common biological pathway. While these techniques allow for the detection of disease-relevant epistatic effects, results obtained from these studies will be strongly biased towards well-characterised parts of the human genome — uncharacterized intergenic regions that might play a role in disease via changes in gene regulation will be completely ignored. Large-margin classification algorithms, like Support Vector Machines (SVMs) and Random Forests, that have been popular in other applications in computational biology, have also been used to build black-box models (43) for predicting disease. While some models learned with these algorithms have demonstrably achieved high prediction accuracies, additional metrics need to be designed
to quantify the effect size of each SNP (or groups of SNPs) in the model, making biological interpretation of these models difficult (44, 45).

Developing the appropriate statistical framework — one that is both predictive (for clinical goals) and interpretable (for basic science goals) — presents a deep machine learning challenge. Adaboost (46) is a popular, iterative large-margin classification ‘meta-algorithm’ that has successfully been used to learn predictive, interpretable models in other applications of computational biology (47, 48). Alternating Decision Trees (ADT) are tree-structured linear models built from simple decision rules (49) that allow us to represent combinatorial interactions between SNPs. Using Adaboost to learn ADTs from the vast amounts of genotypic data, we can learn models highly predictive of disease risk whilst allowing the algorithm to automatically infer model complexity (i.e., size of the model and presence of epistatic interactions). We compare the predictive accuracy of models learned using this algorithm in Type-1 diabetes with that achieved by other statistical tools, and also identify predictive, functionally relevant genomic regions selected by Adaboost.

2.2 Methods

Our overall goal is to infer associations between genotype and disease phenotype, for any given disease. To this end, we develop a model that can predict the disease state of an individual given their genotype measurements across several SNPs. An ideal model is one that is simple and easy to interpret, whilst incorporating putative combinatorial interactions between SNPs. In
addition, the interpretability of the results is improved if we have a simple, intuitive learning algorithm that can be easily verified and allows us to infer the correct model complexity.

2.2.1 Genotype-phenotype data

The data used in this study were obtained from two large GWAS for Type-1 diabetes conducted by the Wellcome Trust Case-Control Consortium (WT) and Type-1 Diabetes Genetics Consortium (GC). The raw data from the WT consortium included 1963 cases and 2938 controls genotyped using Affymetrix GeneChip 500K Mapping Array Set while the GC consortium data included 3577 cases and 2570 controls from amongst a similar population and genotyped on the Illumina 550K platform. Commonly, association testing is performed using the allelic states of an individual at every SNP on the genotyping array. This state is estimated using a statistical model which combines a priori assumptions based on population genetic models with allele intensity measurements to estimate a likely genotype call (50) Giannoulatou (51), Korn (52) Marchini, Spencer, Teo, & Donnelly, 2007 (53). Despite pruning the SNP set with a MAF threshold of 0.05 and HWE test cut-off of 10^{-5}, we observed the SNPs selected by our algorithm as highly predictive to have erroneous patterns in their genotype calls (Figure S2.1).

To avoid such pathological SNPs and other biases specific to the different genotype calling algorithms Hong (54-56) we designed a straightforward representation of the data using the raw allele intensity measurements. Specifically, for each case/control, given the log normalized
signal intensity for the two alleles, $I_A$ and $I_a$, of a given SNP, we represent that SNP by the angle computed as $x = \tan^{-1} \frac{I_a}{I_A}$. Thus, each case/control is now represented by a vector of angles, instead of a vector of genotypes. While this helps us avoid the aforementioned biases, we are unable to use a classifier trained on one cohort to directly test on another cohort; instead we rely on within dataset cross validation to control for over-fitting.

### 2.2.2 Adaboost

Adaboost learns a linear, discriminative model built from simple binary-valued functions by iteratively selecting the function that maximally decreases the exponential loss (57). Adaboost optimizes the exponential loss by searching over the space of Alternating Decision Trees (ADT) (49) built from one-sided binary-valued functions. These functions are called ‘one-sided’ because they contribute to a prediction only if the rule associated with them is satisfied. Although each binary-valued function could potentially be a combination of several single-SNP decision rules, in this work we only consider single-SNP functions. Thus, the space of functions has a size complexity of $O(ND)$ at each iteration, where $D$ is the number of SNPs and $N$ is the sample size. (see supplementary text for more details on the model and algorithm).

The accuracy of the ADT model, at each round of boosting, is evaluated using the Area Under the Curve (AUC). Here the ‘curve’ is the Receiver Operating Characteristic (ROC) that traces the true positive rate (TPR) vs false positive rate (FPR) of the ADT for each value of a real-valued discrimination threshold; the AUC score is defined as the area under this ROC.
curve. This non-parametric test measures the separability of two classes (cases and controls) based on a real-valued metric (in our case the ADT prediction score) assigned to each element of the two classes.

### 2.2.3 Calculating variable importance

We use a variable importance (VI) measure based on unscaled permutation importance, exchanging the accuracy criteria used in Random Forest models (58) AUC. Specifically, we define the VI measure for a variable $v$ as the mean change in ADT accuracy, as measured by AUC, produced by permuting the variable. Let $A_T$ denote the AUC on a held out set of a given ADT $T$ and $A_T(v^*)$ be the AUC of the ADT after permuting the variable $v$. Also, let $N(v)$ denote the number of ADTs which contain the variable $v$: $N(v) = |\{T: v \in T\}|$. Formally, we define a VI measure $\Delta AUC$ as

$$\Delta AUC(v) = \frac{1}{N(v)} \sum_{T: v \in T} A_T - A_T(v^*)$$

Note that each ADT in the above sum is associated with a different test set and thus a different permutation of the variable $v$. We compute the mean of this score over 1000 random permutations and report that as the importance measure for the variable.

### 2.2.4 SNP to gene mapping

All SNPs used in the study were mapped to the human genome based on locations reported in dbSNP132. All validated genes with Entrez gene IDs from GRCh37.p5 assembly of the human genome were used in the study. SNPs
were mapped to genes if they occur within the gene or within 20Kb upstream or downstream of the gene, an arguably conservative mapping.

2.2.5 Gene interaction mapping

For every pair of SNPs selected by boosting, we tested for the presence of a significant interaction term in a logistic regression model predicting case from control. A p-value for each SNP pair was estimated using a likelihood ratio test of a model with an interaction term compared to the model with the linear terms only. If a pair of SNPs has a significant interaction term in both cohorts (tested independently), then the genes linked to this pair of SNPs will have an edge in the gene interaction map. Similarly, we construct an interaction map using top SNPs selected by single-SNP tests, restricting the number of SNPs to be the same as that selected by boosting.

Gene interactions networks from STRING (59), HumanNet (6), and FunCoup (7) are used as benchmarks with which the interaction maps constructed above can be validated. We calculate the fraction of inferred interactions that are reported in each of these networks and the hypergeometric p-value of observing the given number of interactions from all possible interaction pairs in the network.

Functional GO similarity was calculated using wang method with the implementation given in the GoSemSim R package (60). This metric is a semantic similarity measure which scores the distance between two annotation terms based on the location of these terms in an ontology directed acyclic graph and their relation to ancestor terms. Wang et al.
extend this similarity function to genes by taking the average of all maximum similarities between pairs in the set of annotations for two genes.

2.3 Results

In this study, we use Adaboost, a large-margin classification algorithm to infer models predictive of disease phenotype using raw SNP signal intensity data (see Methods for details) for Type–1 diabetes (T1D) cohorts genotyped by the Wellcome Trust Case-Control Consortium (WT) (53) and Type–1 Diabetes Genetics Consortium (GC) (29). We demonstrate the success of Adaboost in selecting a sufficient set of highly predictive SNPs – crucial for the clinical goal of predicting disease risk – and in learning models with high prediction accuracies using these SNP sets, and compare it to the predictive performance of traditional single-variate statistical tools that have been extremely popular in the GWAS community. We evaluate how well boosting replicates results between two GWAS conducted on similar populations, using different measurement platforms, and show that the predictive genomic regions inferred by boosting from the two studies are statistically significant. While the algorithm says nothing about the biological relevance of each individual SNP selected into the model, we illustrate, using gene interaction maps, how predictive genomic regions identified by these SNPs contain disease relevant genes and pathways missed by traditional association studies, explaining different aspects of the etiology of the disease. Finally, we elucidate the biological importance of the set of all selected SNPs using a number of different functional annotations of the human genome. For the
datasets used in this study, while highly predictive models could be learned using genotype data (Figure S2.2), we observed inaccuracies in the genotype calls for SNPs selected by the algorithm (Figure S2.1). These observations motivated our use of the raw SNP measurement data, represented as angles (see 2.2.1 for details) in the learning algorithm.

2.3.1 Adaboost infers sparse models with high predictive accuracies

In Figure 2.1, we compare the predictive accuracy achieved by Adaboost on held-out data (10-fold cross validation) using stumps (ADTs with depth 1) on the two different datasets (shown in red), with the accuracy of predictive models reported in the literature using the WT dataset. The prediction accuracy indicated by the dashed line was achieved using SVMs with radial basis function kernels on the genotype data of a subset of SNPs (43), while the prediction accuracy indicated by the dot-dashed line was achieved using LASSO (L1-regularized logistic regression)(61). In both studies, only those SNPs whose p-value of association with disease crossed a pre-specified threshold and SNPs identified as disease-relevant in earlier studies were used in learning. Several studies (e.g., Jakobsdottir (37)) have strongly argued against using associated SNPs to build disease predictive models and it is unclear if the reported accuracies were inflated by such preselection of SNPs based on p-values using the entire data set.
Figure 2.1. Accuracy of Adaboost on Type 1 diabetes. Shown here is the predictive accuracy of Adaboost, measured by area under the ROC curve (AUC), on held out samples (10-fold cross validation), using ADTs [a, WT cohort and b, GC cohort]. Colors correspond to different boost-remove iterations (e.g. the red curve corresponds to the accuracy when no SNPs have been removed from the data, while for the dark blue curve, 250 SNPs previously selected by Adaboost have been masked out). The accuracy of boosting is compared with the predictive accuracy of LASSO (black dashed line) and SVM (black dot-dashed line) reported in the literature for Type 1 diabetes, one the WT dataset.

Given the strong LD structure in human genomes (25), it would be useful to see how much predictive signal is retained in the data if these highly predictive SNPs selected by Adaboost are masked. By iteratively masking sets of predictive SNPs selected by Adaboost, we can identify those genomic regions containing putative causal variants whose effects are correlated with that of the variants selected in the first round. In Figure 2.1, we plot the predictive accuracy of boosting as successive sets of SNPs selected by the algorithm are removed. We performed 25 boost-remove iterations, where for each boost-remove iteration, we run Adaboost until 50 SNPs are selected into
the model, and then mask these SNPs before boosting on the remaining dataset; each cross-validation fold is performed independently in this fashion. This results in 2281 and 2203 unique SNPs recovered from the GC and WT cohorts, respectively. The statistical and functional analysis described in the rest of this paper are based on these selected SNPs. Quite surprisingly, we notice that a significant amount of signal remains in the data even after 20 boost-remove iterations. Despite correcting for large-scale genetic confounders by including the top principal components of the relatedness matrix during learning (62), it is possible that small-scale population structure contributes to the relatively high predictive signal achieved by Adaboost at much higher boost-remove iterations.
Figure 2.2. Example decision rules learned using Adaboost. This plot visualizes the scatter of held-out data along the two SNPs selected by Adaboost in consecutive boosting rounds. Panels (a) and (c) correspond to cases and (b) and (d) correspond to controls. Darker points correspond to mis-classified points (lighter points were correctly classified) before Adaboost selected the SNPs plotted along the two axes. Panel (a) and (b) show example SNPs selected in rounds 1 and 2, panels (c) and (d) show the decision rules on SNPs selected in rounds 3 and 4. For each selected SNP, the threshold angle learned as part of the decision rule is shown as vertical and horizontal lines.

As an illustration of how Adaboost selects predictive rules, we plot in Figure 2.2 the raw angles of the first four SNPs selected by Adaboost, for cases and controls separately. In the top row, we plot SNPs selected in boosting rounds 1 and 2, while, in the bottom row, we plots SNPs selected in rounds 3 and 4. The angle thresholds learned by Adaboost for each SNP are indicated by solid lines. Darker points correspond to samples that were mis-classified.
before the SNPs (plotted on the axes) were selected by the algorithm. Since the datasets have more controls than cases, the optimal value of the intercept term in the model predicts all controls correctly, and cases incorrectly. All samples to the right of the first angle threshold satisfy the first decision rule; this was optimally chosen by the algorithm to correctly classify the cases while reducing the number of controls that would now be misclassified. We find it noteworthy that despite not having genotype calls, the angle thresholds in the decision rules quite neatly selects out samples with specific genotypes for the predictive SNP. The weights for each such decision rules can be interpreted as effects of the corresponding SNP that deviate from an additive model. Since the same SNP can be selected into the final model with different thresholds, these models, in general, also allow for additive effects.

Conditional on having selected this first rule, the second rule is chosen to further increase the accuracy on cases (samples below the threshold now satisfy the decision rule) while mis-classifying as few controls as possible. In the bottom row, before selecting a new decision rule, we observe that there are more mis-classified controls than cases; thus, the new decision rule is selected to improve accuracy on these mis-classified points. Each new decision rule improves accuracy by providing information `orthogonal` to that provided by the previous decision rule. This property enables Adaboost to learn highly predictive combinations of SNPs that traditional single variate association studies would find difficult to detect. This property also enables Adaboost to
learn sparse models; from Figure 2.1, we see that the accuracy achieved by Adaboost reaches close to the maximum with \( \approx 30 \) SNPs. In contrast, the best models learned using LASSO assigned non-zero weights to over 100 SNPs.

### 2.3.2 Significant overlap between predictive regions inferred from different studies.

![Proximity of predictive SNP sets](image)

**Figure 2.3.** Proximity of predictive SNP sets. The figure compares the distribution of physical distance (normalized to chromosome length) between sets of SNPs selected by Adaboost on the two cohorts (shown in red), with the distribution of relative physical distance between random sets of SNPs drawn uniformly (shown in blue). The left subfigure corresponds to SNPs selected in the first boost-remove iteration, while the right subfigure corresponds to the tenth iteration.

Encouraged by the high accuracies on held-out data, we tested the ability of Adaboost to replicate results across different studies. Specifically, we tested if the SNP sets learned on data from the two T1D GWAS were closer than expected by chance. In figure **Figure 2.3**, we plot the histogram of the distances of SNPs selected by boosting on the GC dataset, to the set of SNPs selected from the WT dataset (shown in red). For this analysis, we defined the
distance between two sets of SNPs to be the minimum distance (using hg18 coordinates) over all pairs of SNPs, one chosen from each set for each pair of SNPs. We compare this with the histogram of distances between pairs of SNP sets, of the same size as that selected by boosting, drawn randomly from the typed variants in the two studies (shown in blue). The plots show a significant shift of the medians between pairwise distances of the true and randomly sampled SNPs (Wilcoxon Ranksum test, \( P = 10^{-30} \)); i.e., the variants selected by Adaboost from the two studies are significantly closer than we would expect by chance, indicating that Adaboost selects genomic regions consistently across different studies. We performed the same analysis on SNP sets selected by Adaboost using the processed genotype calls for these data (Figure S2.3) and observe the selected SNPs to still be closer than expected, but to a lesser extent.

2.3.3 A complementary predictive landscape revealed by boosting

Next, we compared the genomic markers recovered using boosting with those recovered using standard statistical tools. Figure 2.4a shows a Manhattan plot of the single-SNP p-values from both datasets, as reported by the two studies. Contrast this ‘classical statistics’ view with a ‘machine learning’ view shown in Figure 2.4b, with p-values replaced by \( \Delta \text{AUC} \), a variable importance score similar to that used in Random Forest models (58). While the two plots show substantial concordance (e.g., MHC region on chromosome 6), highly predictive regions selected using boosting abound across the genome in both datasets and show limited correlation with p-value
based results. Furthermore, several SNPs selected by Adaboost, that also have high importance scores, lie in genomic regions reported in T1DBase (version 4.10), a database aggregating known genes implicated with T1D (63). This suggests that single SNP statistical tests with strict significance thresholds are inadequate tools to accurately identify many causal loci. Another important observation is that highly predictive genomic regions are chosen as predictors in both early and late boost-remove iteration.

Figure 2.4. ‘Classical’ vs ‘Boosting’ Manhattan plots. Panel (a) shows the classic Manhattan plot of single SNP p-values as a function of chromosomal location. Results for SNPs in both the WT and GC cohorts are overlaid on top of each other. The red dotted line shows the Bonferroni corrected p-value threshold used in both studies. Panel (b) shows a Manhattan plot of $\Delta AUC$ as a function of chromosomal location, for every SNP selected by boosting in the two datasets. Shaded in blue are WT SNPs and in red GC SNPs. Color intensity corresponds to the median boost-remove iteration the SNP was selected across cross-validation folds. Marker size corresponds to the number of unique cross-validation folds in which this marker is chosen as a predictor.
2.3.4 Gene overlap between studies

Next, we examine the gene overlap of SNPs selected by Adaboost from the two cohorts. We map the typed SNPs to genes and rank the genes mapped to boosting selected SNPs based on the maximum ΔAUC over all SNPs mapped to a gene. Similarly, we rank genes mapped to SNPs selected by p-value based methods, using the minimum p-value over SNPs mapped to the gene. For a given rank threshold, we calculate the Jaccard index (the ratio of the intersection over the union of two sets) of the top genes in the two gene sets, and plot this score against the rank threshold in Figure 2.5a. A background overlap distribution is calculated by choosing SNPs uniformly and mapping them onto genes; we report the mean and twice the standard deviation from 1000 random draws. We observe that boosting recovered genes between cohorts have a strongly significant overlap, while only the most important genes recovered by boosting and most significant genes recovered by simple statistical tests have a significant overlap both within cohorts and between cohorts. We also compare the gene sets recovered through boosting with known T1D gene associations, as reported in T1DBase. As shown in Figure 2.5b, in all, 302 of 521 genes reported in T1DBase are found by boosting on either data set and are also reported to be associated in T1DBase, of which 157 are found by boosting on both cohorts (hypergeometric p-values of $P = 2.225 \times 10^{-308}$ and $P = 3.331 \times 10^{-16}$, for GC and WT respectively).
Figure 2.5. Gene and pathway overlap. Panel (A) shows the Jaccard index of genes selected by Adaboost ranked by ∆AUC and p-value genes ranked by p-values between and within the GC and WT cohorts. Panel (B) shows the overlap of genes tagged using SNPs selected by boosting with known T1D associations as reported in T1DBase. Panel (C) visualizes the overlap of genes in the set of pathways found enriched using a hypergeometric test (FDR < 0.05). The number of p-value SNPs used was chosen to match the number of boosting genes. Blue denotes genes detected by boosting but not by p-value methods, red denotes genes detected using p-value based methods but not by boosting, and green denotes genes detected by both methods. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways are denoted with a preceding (K) and Reactome pathways are denoted with a preceding (R). 'B' denotes a pathway found to be significant using boosting in either GC or WT,'P' denotes the same for p-value based methods.

2.3.5 Significantly enriched T1D pathways

In order to show that the genes selected by Adaboost are part of diabetes-relevant biological processes, we test the recovered genes from each cohort for enrichment in pathways from the MSigDB v.3 (64). For this analysis we use the SNPs selected by boosting and an identical number of SNPs chosen according to their p-value ranks, and map each SNP to genes resulting in four gene sets (Boosting GC, Boosting WT, P-value GC, P-value WT). In Figure 2.5c, we illustrate those REACTOME/BIOCARTA/KEGG pathways that...
are enriched with genes from at least one of the four disease-relevant gene sets along with the number of genes in each pathway that overlaps with any of the four sets (Hypergeometric enrichment test with a FDR < 0.05). The subset of genes denoted as 'boosting only' represents genes which are recovered in either boosting method and neither p-value method, 'p-value only' if vice versa, and 'boosting and p-value' if found in one or both cohorts using either method. Over half of all pathways include an equal or greater number of 'boosting only' genes. It seems of note that several enriched MSigDB pathways are directly related to T1D and our understanding of the biological mechanisms underlying the disease, such as KEGG Type I Diabetes Mellitus and REACTOME TCR signaling. Finally, we observe that seven pathways are significantly recovered in boosting but are not recovered when using p-value based methods including extremely pertinent pathways such as CLASSIC PATHWAY (65), while a further set of seven pathways are only found significant using p-value methods.

2.3.6 Cross cohort replicated genes not recovered using p-value based methods

Next, we compile a list of genes tagged by SNPs recovered by boosting in both cohorts and exclude any gene tagged by a SNP with a p-value below the reported significance threshold of $5 \times 10^{-7}$ (Supplementary File 1-Table S2.1). We find 163 such genes, of which 15 genes are implicated with a T1D association in T1DBase (Supplementary File 1-Table S2.2), and 48 genes are found associated with 3 or more T1D related key words on a Pubmed
literature search, performed using PubMatrix (66) [Supplementary File 1-Table S2.3]. The key words used for this search were ‘diabetes mellitus type 1’, ‘T1D’, ‘diabetes’, ‘insulin’, ‘immune response’, ‘glucose’, ‘glycemia’, ‘hyperglycemia’, ‘hypoglycemia’, and ‘autoimmune’. Amongst these hits are several genes with unremarkable p-value scores in both GC and WT cohorts, which have been mentioned extensively in the diabetes literature. For example, RYR2, co-cited with the terms ‘diabetes’ and ‘insulin’ 55 and 66 times respectively in Pubmed, is a gene coding a calcium channel protein in the cellular endoplasmic reticulum. Blockage of RYR2 pumps has been shown to induce apoptosis of pancreatic islets beta-cells Johnson (67, 68). RYR2 has an unremarkable p-value of 0.5 and 0.2 in WT and GC; yet, it is implicated by boosting in both cohorts, in each it is associated with two recovered SNPs, albeit with a small effect size. As a second example, PARP1, detected with above threshold p-values (0.019 and 0.28), is a gene known to be involved in the recovery from DNA damage. It has been show in human populations to be involved in pathophysiology related directly with T1D (69, 70) and has been implicated as an important causal factor of T1D in mouse models (71) and as an aggravating factor in Type–2 diabetes in humans (72).

2.3.7 An interaction map of T1D genes

One explanation for the difference in results observed between genes selected by boosting and genes recovered via standard statistical tests is that the ADT model can capture higher order interactions amongst SNPs or genes using complex combinations of decision rules. To verify this hypothesis, among
all SNPs selected by boosting, we tested for the existence of an interaction term between every pair of SNPs in a logistic regression model predicting disease status in each cohort separately. The map in figure **Figure 2.6** summarizes the set of interaction terms replicated in both cohorts with an FDR < 0.01. We observe the presence of several genes linked to SNPs with non-significant p-value but which exhibit replicable interactions with other genes or genomic loci (e.g. TMPRSS3↔1q4, ZNRD1↔SNORD117). We further tested the genes recovered in the map for significantly enriched genes sets, and find several immune response complexes some of which have been implicated in T1D and other autoimmune diseases — MHC II complex, TAP complex, Classic complement pathway, and a set of NF-kappaB binding proteins with FDR adjusted P-values of $3.91 \times 10^{-12}$, $1.22 \times 10^{-8}$, $3.94 \times 10^{-3}$, and $1.13 \times 10^{-10}$ respectively.

We build a similar interaction map using top SNPs selected by single-SNP tests and assess the biological significance of these recovered interaction maps by comparing them to three curated gene interaction networks from STRING (5), HumanNet (6), and FunCoup (7) as shown in **Figure S2.1**. These networks aggregate various types of evidence from co-citation to conserved interactions in model organisms to build a genome wide network of gene interactions in humans. We observed that the map based on SNPs selected by boosting recover proportionally more known edges from all three networks with highly significant p-values estimated using
**Figure 2.6.** A boosting SNP gene interaction map. This figure shows a map of gene pairs linked with SNPs selected by boosting for which SNP-SNP interactions are significantly associated with disease status. Each node in the map corresponds to a gene linked to a SNP selected by Adaboost. SNPs recovered in boosting in were tested for a significant interaction term in a logistic regression model in both the WT and GC cohorts (FDR<0.01). A pair of nodes has an edge if they are linked to a pair of SNPs that have a significant interaction term. Square nodes denote genes which have also been implicated as T1D associated in T1DBase. Upper-left part of a node is colored green if the minimum p-value among all SNPs linked to that gene is less than 5x10^{-7} in the GC cohort, bottom-right is colored blue to denote the same in the WT cohort. In cases where no proximal genes were found to a SNP, the SNP was mapped to the nearest locus on the chromosome.
hypergeometric enrichment than the maps constructed using SNPs with low p-values.

Finally, as further validation of the interaction hypothesis, we use semantic distance of GO annotation ontologies as a measure of functional similarity between genes (73) (Error! Reference source not found.) to validate the two interaction maps. For each of the three major GO ontologies, we compare the histogram of functional distances of edges in the two constructed interaction maps. When compared against a background distribution of functional similarities, we observe both interaction maps to have a marked enrichment in functional similarities for the Biological Process (BP, P = 1.15 × 10^{-53}), Molecular Function (MF, P = 3.14 × 10^{-29}) and Cellular Component (CC, P = 4.11 × 10^{-28}) GO onotologies, using the one-sided KS test. When comparing sets of edges in the two interaction maps, we find that edges in the map from SNPs selected by Adaboost exhibit significantly greater similarities across all three categories (p-values of 3.17 × 10^{-4}, 4.40 × 10^{-4}, and 4.09 × 10^{-8} for BP, MF, CC respectively).

2.4 Discussion and Conclusion

In this study, we address the clinical goal of predicting disease risk by demonstrating the use of boosting to learn classifiers which accurately capture disease risk, using a sparse human interpretable decision tree model, in two independent cohorts for Type 1 diabetes. The algorithm selects a replicable, sufficient set of novel SNPs predictive of Type 1 diabetes,
unattainable through standard single-SNP statistical tools commonly used to analyze data in such large studies. While we cannot claim that each SNP selected into the model is biologically important, we illustrate how the selected set of SNPs are linked to gene sets that fit into coherent functional annotations and biological pathways enriched in known T1D and immune response genes. Our work has direct implications for every association study conducted to date and future GWAS research, as it demonstrates the importance of classification based model learning as a complementary method to existing statistical methods.

The alternating decision tree framework fits remarkably well to disease association studies as it allows our classifiers to capture a rich variety of association models using a combination of simple decision rules. Boosting allows predictors to be included in the model multiple times with different decision rules, allowing the full decision tree to infer the gamut of single site causative models, including additive, dominant only or recessive only effects. The full ADT model can capture complex interactions between different predictors; however, some appropriate form of regularization is necessary to avoid over-fitting and degradation of predictive performance.

Boosting, along with many other classification algorithms (e.g. SVMs), have a unique advantage over classical single-SNP methods used for association mapping. Most statistical tests used for GWAS involve implicit assumptions about the underlying distributions from which the random variables tested are drawn; rejecting the null hypothesis tells us that the
difference between the two distributions is substantial. Furthermore, case-control studies are designed to be powered enough to associate variants exhibiting penetrance or effect-size above a certain threshold within the entire tested population (74, 75). In classical GWAS, such tests are applied repeatedly, in every instance making that same assumption, examining observed vs. expected differences in the complete sample space. In contrast, boosting methods have the benefit of reweighting the relative importance of the sample in each step of the classification process. In this way, the boosting algorithm focuses iteratively on subspaces of the sample space in which it is currently classifying poorly, and identifies a predictor which best classifies within this subspace. While this process if unchecked can result in over-fitting, it has substantial advantage when properly controlled through correct application of supervised learning methodologies (e.g., the use of cross-validation). It is well appreciated that the biological processes underlying complex phenotypes are the result of the complex interplay of a large number of factors (76) each of these factors has low penetrance; i.e., it may be a substantial factor in only a fraction of the overall population Park et al. (2010). To complicate things further, we are already aware of vast redundancies hard-wired into the biological network, resulting in a situation where multiple genetic differences can result in very similar overall phenotypes. In such a regime, the benefit of the reweighting scheme of boosting quickly becomes apparent, allowing us to discover new biologically relevant factors important in only subsets of the tested population.
2.5 Chapter Acknowledgments

This chapter contains material currently in submission as part of the following manuscript, of which I am a primary co-investigator and co-author:

Raj, A.*, Hofree, M.*, Ideker, T. Wiggins, C.,Freund Y. Inferring sparse multivariate models to predict disease phenotype from genotype
3.1 Background and Significance

Cancer is a disease that is not only complex, i.e. driven by a combination of genes, but also extremely heterogeneous, in that gene combinations can vary greatly between patients. To gain a better understanding of these complexities, major projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) are systematically profiling thousands of tumors at multiple layers of genome-scale information, including mRNA and microRNA expression, DNA copy number and methylation, and DNA sequence(77-79). There is now a strong need for informatic methods that can integrate and interpret genome-scale molecular information to provide insight into the molecular processes driving tumor progression. Such methods are also of pressing need in the clinic, where the impact of genome-scale tumor profiling has been limited by the inability of current analysis techniques to derive clinically-relevant conclusions from the data(80, 81).

One of the fundamental goals of cancer informatics is tumor stratification, whereby a heterogeneous population of tumors is divided into clinically-meaningful subtypes based on similarity of molecular profiles. Most prior attempts to stratify tumors with molecular profiles have used mRNA
expression data(78, 82-85), resulting in the discovery of informative subtypes in diseases such as glioblastoma and breast cancer. On the other hand, in TCGA cohorts including Colorectal Adenocarcinoma and Small-Cell Lung Cancer, subtypes derived from expression profiles do not correlate with any clinical phenotype including patient survival and response to chemotherapy(78, 86). These results might be due to limitations of expression-based analysis that have been noted(87) such as issues with RNA sample quality, lack of reproducibility between biological replicates, and ample opportunities for overfitting of data.

A promising new source of data for stratification is the somatic mutation profile, in which next-generation sequencing is used to compare the genome or exome of a patient’s tumor to that of the germline to identify mutations that have become enriched in the tumor cell population (88). As this set of mutations is presumed to contain the causal drivers of tumor progression (89), similarities and differences in mutations across patients could provide invaluable information for stratification. While individual mutations in well-established cancer genes have long been used to stratify patients in a straightforward manner (90-93), stratification of the entire mutation profile of a patient has been more challenging. Somatic mutations are fundamentally unlike other data types such as expression or methylation, in which nearly all genes or markers are assigned a quantitative value in every patient. Instead, somatic mutation profiles are extremely sparse, with typically fewer than 100 mutated bases in an entire exome (Figure S3.1). They are also remarkably
heterogeneous, such that it is very common for clinically-identical patients to share no more than a single mutation (78, 94, 95). For these reasons, it is not surprising that standard approaches for clustering fail to produce meaningful stratification results.

Here, we report the discovery that these problems can be largely overcome by integrating somatic mutation profiles with knowledge of the molecular network architecture of human cells. It is widely appreciated that cancer is a disease not of individual mutations, nor of genes, but of combinations of genes acting in molecular networks corresponding to hallmark processes such as cell proliferation and apoptosis (96, 97). We postulated that, although two tumors may not share any mutations in common, they may share remarkable similarity in the networks impacted by these mutations (as per Waddington’s original theory of ‘genetic canalization’ (98)). Although current cancer pathway maps are incomplete, much relevant information is available in the current public databases of human protein-protein, functional, and pathway interactions. An increasing number of approaches have had success in integrating these network databases with tumor molecular profiles to map the molecular pathways of cancer (99-103). Here, we focus on the orthogonal problem of using network knowledge to stratify a cohort into meaningful subsets. We now show that, using this knowledge, somatic mutation profiles can be clustered into robust tumor subtypes with strong association to clinical outcomes such as patient survival time and emergence of drug resistance. As proof of principle, we apply this
method to stratify the somatic mutation profiles of three major cancers catalogued in TCGA: ovarian, uterine and lung.

3.2 Methods

3.2.1 Overview of network-based stratification

The technique of Network-based Stratification (NBS) combines genome-scale somatic mutation profiles with a gene interaction network to produce a robust subdivision of patients into subtypes (Figure 3.1a). Briefly, somatic mutations for each patient are represented as a profile of binary (1,0) states on genes, in which a ‘1’ indicates a gene for which mutation has occurred in the tumor relative to germline (i.e. a single nucleotide base change or the insertion or deletion of bases). For each patient independently we project the mutation profiles onto a human gene interaction network obtained from public databases (6, 104, 105). Next, the technique of network propagation (106) is applied to spread the influence of each subsampled mutation profile over its network neighborhood (Figure 3.1b). The result is a ‘network-smoothed’ profile in which the state of each gene is no longer binary but reflects its network proximity to the mutated genes in that patient, along a continuous range [0,1]. Following this ‘network smoothing’, patient profiles are clustered into a predefined number of subtypes \( k = 2 \ldots 12 \) using the unsupervised technique of non-negative matrix factorization(107) (NMF, Figure 3.1c). For NBS we use a variant of NMF which encourages the selection of gene sets supporting each subtype based on high network connectivity (NetNMF) (108). Finally, to promote robust cluster assignments we use the
technique of consensus clustering(109), in which the above procedure is repeated for 1000 different subsamples in which subsets of 80% of patients and genes are drawn randomly without replacement from the entire data set. The results of all 1000 runs are aggregated into a (patient x patient) co-occurrence matrix, which summarizes the frequency of times each pair of patients has co-segregated into the same cluster. This co-occurrence matrix is then clustered to recover a final stratification of the patients into clusters / subtypes (Figure 3.1d,e).
Figure 3.1. Overview of network-based stratification. (a) Flowchart of the approach. (b) An example illustrating smoothing of patient somatic mutation profiles over a molecular interaction network. Mutated genes are shown in yellow (patient 1) and blue (patient 2), in the context of a gene interaction network. Following smoothing, genes with high scores in both patients appear in green (dashed circle). (c) Clustering mutation profiles using Non-negative Matrix Factorization (NMF) regularized by a network. The input data matrix (F) is decomposed into the product of two matrices, one of subtype prototypes (W) and an assignments matrix of each mutation profile to the prototypes (H). The decomposition attempts to minimize the objective function shown, which includes a network regularization term on the subtype prototypes. (d–e) The final tumor subtypes are obtained from the consensus (majority) assignments of each tumor after 1000 applications of this procedure to samples of the original data set. A darker blue color coincides with higher co-clustering for pairs of patients. The overall outcome of network-based stratification (d) is to capture informative clusters within somatic mutation data, in contrast to standard consensus clustering (e) which generally fails to produce such clusters.

3.2.2 Processing of patient mutation profiles

High-grade serous ovarian cancer, uterine endometrial carcinoma, and lung adenocarcinoma somatic mutation data were downloaded from the TCGA data portal on August 8th, 2012, January 1st, 2013 and January 1st,
2013, respectively. Only mutation data generated using the Illumina GAIIx platform were retained for subsequent analysis, and patients with fewer than 10 mutations were discarded. This left 356 patients with mutations in 9,850 genes for the TCGA ovarian cohort, 248 patients with mutations in 17,968 genes for the TCGA uterine endometrial cohort, and 381 patients with mutations in 15,967 genes in the TCGA lung adenocarcinoma cohort. Patient mutation profiles were constructed as binary vectors such that a bit is set if the gene corresponding to that position in the vector harbors a mutation in that patient. Additional details on processing and organization of the data are available in a previous TCGA publication (78).

3.2.3 Sources of molecular network data

Patient mutation profiles were mapped onto gene interaction networks from three sources: STRING v.9 (104), HumanNet v.1 (6), and PathwayCommons (105). All network sources comprise a combination of interaction types, including direct protein-protein interactions between a pair of gene products and indirect genetic interactions representing regulatory relationships between pairs of genes (e.g. co-expression or TF activation). The PathwayCommons network was filtered to remove any non-human genes and interactions and all remaining interactions were used for subsequent analysis. Only the most confident 10% of interactions for both the STRING and HumanNet networks were used for this work, ordered according to the quantitative interaction score provided as part of both networks. This threshold was chosen using an independent ROC analysis with respect to a set of Gene
Ontology derived gold standards (data not shown). After filtering of edges all networks were used as unweighted, undirected networks.

### 3.2.4 Network smoothing

After mapping a patient mutation profile onto a molecular network, network propagation\(^{(106)}\) is applied to ‘smooth’ the mutation signal across the network. Network propagation uses a process that simulates a random walk on a network (with restarts) according to the function:

\[
F_{t+1} = \alpha F_t A + (1 - \alpha) F_0
\]

\(F_0\) is a patient-by-gene matrix, \(A\) is a degree-normalized adjacency matrix of the gene interaction network, created by multiplying the adjacency matrix by a diagonal matrix with the inverse of its row (or column) sums on the diagonal. \(\alpha\) is a tuning parameter governing the distance that a mutation signal is allowed to diffuse through the network during propagation. The optimal value of \(\alpha\) is network-dependent (0.7, 0.5 and 0.7, for HumanNet, PathwayCommons and STRING respectively), but the specific value seems to have only a minor effect on the results of NBS over a sizable range (e.g. 0.5 - 0.8). The propagation function is run iteratively with \(t = [0, 1, 2, \ldots]\) until \(F_{t+1}\) converges (the matrix norm of \(F_{t+1} - F_t < 1 \times 10^{-6}\)). Following propagation, the rows of the resultant matrix \(F_t\) are quantile normalized to ensure that the smoothed mutation profile for each patient follows the same distribution.
3.2.5 Network-regularized NMF

Network-regularized NMF is an extension of non-negative matrix factorization (NMF) that constrains NMF to respect the structure of an underlying gene interaction network. This is accomplished by minimizing the following objective function using an iterative method (107, 108, 110):

$$
\min_{W,H>0} \|F - WH\|^2 + \text{trace}(W^tKW)
$$

$W$ and $H$ form a decomposition of the patient x gene matrix $F$ (resulting from network smoothing as described above) such that $W$ is a collection of basis vectors, or ‘metagenes’, and $H$ is the basis vector loadings. The $\text{trace}(W^tKW)$ function constrains the basis vectors in $W$ to respect local network neighborhoods. The term $K$ is the Graph Laplacian of a nearest-neighbors influence distance matrix(99) derived from the original network. The degree to which local network topology versus global network topology constrains $W$ is determined by the number of nearest neighbors. We experimented with neighbor counts ranging from 5 to 50 to include in the nearest network and observed only small changes in outcome (data not shown). For the work presented in this manuscript, the 11 most influential neighbors of each gene in the network as determined by network influence distance were used.

3.2.6 Consensus clustering

Clustering was performed with a standard consensus clustering framework, discussed in detail by Monti et al. (109) and used in previous TCGA publications (78, 94, 111). Briefly, we used network-regularized NMF (see
above) to derive a stratification of the input cohort. In order to ensure robust clustering, network-regularized NMF was performed 1000 times on subsamples of the dataset. In each subsample, we sampled 80% of the patients and 80% of the mutated genes at random without replacement. The set of clustering outcomes for the 1000 samples was then transformed into a co-clustering matrix. This matrix records the frequency with which each patient pair was observed to have membership in the same subtype over all clustering iterations in which both patients of the pair were sampled. The result is a similarity matrix of patients, which we then used to stratify the patients by applying either average linkage hierarchical clustering or a second symmetric NMF step.

### 3.2.7 Simulation of somatic mutation cohorts

We used simulations to determine the ability of NBS to recover subtypes from somatic mutation profiles. In order to quantify the performance of NBS we needed a cohort with specified subtypes as a “ground truth” reference, while allowing us to control over the properties of the signal to be detected. We simulated a somatic mutation cohort as follows. Patient mutation profiles were sampled with replacement from the TCGA ovarian dataset. For each patient, the mutation profile was permuted while keeping the per-patient mutation frequency invariant, resulting in a background mutation matrix with no subtype signal. To simulate an underlying network structure for NBS to detect, a network-based signal was added to the patient-by-mutation matrix as follows. First, we established a set of network communities (i.e. connected
components enriched for edges shared within community members) in the input network (STRING, HumanNet, or PathwayCommons) using the network community detection algorithm Qcut(112). Next, we divided the patient cohort randomly into four equal-sized subtypes (four was selected as reasonable due to the four expression-based subtypes that have been identified for glioblastoma, ovarian and breast cancers(78, 94, 111, 113)). Each subtype was assigned a small number (e.g. 1-6) of network modules that together had a combined size $s$ ranging from 10 to 250 genes. These network modules represent ‘driver’ sub-networks characterizing the subtype. For each patient, we reassigned a fraction of the patient’s mutations $f$ to genes covered by the driver modules for that patient’s subtype. This procedure resulted in a patient x gene mutation matrix with underlying network structure, while maintaining the per-patient mutation frequency.

A plausible range for the number of driver mutation in a tumor was recently proposed to be between 2 to 8 driver mutations(114). We note that in our simulation framework a 4% mutation rate corresponds to between 1 and 9 mutations with a median of 3, on par with the aforementioned estimates. In order to estimate the appropriate size of cancer pathways ($s$) we examined the known cancer pathways in the NCI-Nature cancer interaction database (115). We observe that pathways in the database are of varying sizes, 2 – 139 genes, with a median size of 34, and over 23% of pathways include over 50 genes.
3.2.8 Identifying differentially mutated subnetworks

After applying NBS, we identified genes that were enriched for mutation in each of the subtypes relative to the whole cohort. To do this we applied the Significance Analysis of Microarrays (SAM) method(116), on the network smoothed mutation profiles. This is a non-parametric method developed for discovering differentially expressed genes in microarray experiments. We used a rank based Wilcoxon statistic to compare each subtype against the remaining cohort. We assess significance using the SAM permutation scheme with 1000 permutations. The resulting set of genes for each subtype was overlaid on the network used for network smoothing.

3.2.9 Survival analysis

Survival analysis was performed using the R ‘survival’ package. We fit a Cox-proportional hazards model(117) to determine the relationship between the NBS-assigned subtypes and patient survival. A likelihood ratio test and associated p-value is calculated by comparing the full model, which includes subtypes and clinical covariates, against a baseline model that includes covariates only. Clinical covariates available in TCGA and included in the model were age, grade, stage, residual surgical resection, and mutation rate, as well as cigarette smoking status for the lung cancer cohort.

3.2.10 Shrunken centroid prediction on expression profiles

We used shrunken centroids to derive an expression signature equivalent to the somatic mutation-based NBS subtypes. Expression data
were provided by Győrffy et al. (118) who aggregated several expression datasets as part of a meta-analysis of ovarian cancer. In this analysis, all data were regularized using quantile and MAS5 normalization. We performed this analysis on the Tothill et al. (ovarian serous samples only), Bonome et al., and TCGA datasets, as well as across the full meta-analysis cohort. We use the ‘pamr’ R package, with default parameters to train a shrunken centroid model (119) on mRNA expression levels for all genes in the TCGA ovarian dataset with subtype assignment as the class label. The trained model was next used to predict subtype labels on the held-out Tothill et al. and Bonome et al. data or the full meta-analysis expression cohorts.

3.2.11 Missense mutation Scoring

Missense mutations were scored using three methods: CHASM (89), VEST (120) and MutationAssessor (121). CHASM and VEST use supervised machine learning to score mutations. The CHASM training set is composed of a positive class of driver mutations from the COSMIC database and a negative class of synthetic passenger mutations simulated according to the mutation spectrum observed in the tumor type under study. The VEST training set comprises a positive class of disease mutations from the Human Gene Mutation Database (122) and a negative class of variants detected in the ESP6500 (http://evs.gs.washington.edu/EVS/) cohort with an allele frequency > 1%. MutationAssessor uses patterns of conservation from protein alignments of large numbers of homologous sequences to assess the functional impact of missense mutations. CHASM and VEST scores were obtained from the CRAVAT
webserver (120) (www.cravat.us). MutationAssessor precomputed mutation scores were downloaded from mutationassessor.org.

### 3.2.12 Replication timing

RepliSeq (123) data for GM12878 were downloaded from the ENCODE project website (March 2013, (124)). Summed normalized tag densities were used as a proxy for replication time (higher counts indicating that a transcript was replicated earlier in the cell cycle). Normalized tag densities for RefSeq protein coding regions were retrieved using bigWigAverageOverBed (125) with RefSeq gene sequence features in gff3 format downloaded from (126). Tag densities were averaged for each transcript and the longest transcript was selected to represent each gene.

### 3.3 Results

#### 3.3.1 Input networks

To evaluate the impact of different sources of network data, we used three interaction databases for this analysis: STRING (104), HumanNet (6) or PathwayCommons (105). STRING integrates protein-protein interactions from literature curation, computationally-predicted interactions, and interactions transferred from model organisms based on orthology. HumanNet uses a naïve bayes approach to weight different types of evidence together into a single interaction score focusing on data collected in humans, yeast, worm and fly. PathwayCommons aggregates interactions from several pathway and interaction databases, focused primarily on physical protein-protein interactions.
interactions (PPIs) and functional relationships between genes in canonical regulatory, signaling, and metabolic pathways (including hallmark pathways of cancer). Supplementary Table 1 summarizes the number of genes and interactions used in our analysis from each of these three networks.

### 3.3.2 Initial benchmarking and performance analysis

As an initial exploration of NBS, we simulated a somatic mutation dataset based on the structure of the TCGA ovarian tumor mutation data and the STRING gene interaction network (Figure 3.2a). Mutation profiles were permuted and patients were divided randomly and uniformly into a predefined number of subtypes \((k = 4)\). Next, a fraction of mutations in each patient was reassigned to fall within genes of a single ‘network module’ characteristic of that patient’s subtype (the ‘driver’ mutation frequency \(f\), varied from 1 to 15%), and the remaining mutations were left to occur randomly. The network modules were selected randomly from the set of all network modules in STRING, defined as sets of densely-interacting genes with size range \(s = 10\) to 250 (see Online Methods for details and justification for the ranges of \(k, f\) and \(s\)). Although it is unknown whether these assumptions completely mirror the biology of cancer, they provide a reasonable model of a pathway-based genetic disease which is[1] driven by genetic circuits embedded in a molecular network whose activity can be altered by mutations at multiple genes and is[2] characterized by many additional mutations that are non-causal ‘passengers’. 
Figure 3.2. Exploring performance of stratification through simulation. (a) TCGA somatic mutations for ovarian cancer (top left) are combined with the STRING human protein interaction network (bottom left) to generate simulated mutation datasets embedded with known network structure (center right). (b) Accuracy with which NBS clusters recover simulated subtype assignments, evaluated with and without network smoothing and using NMF versus hierarchical clustering. Accuracy is calculated as the Adjusted Rand Index of overlap between the clusters and correct subtype assignments, for which a score of zero represents random overlap. Simulation was performed with a driver mutation frequency $f = 7.5\%$ with a single network module assigned to each subclass. (c) The accuracy landscape of NBS across varying driver mutation frequency and module size. (d) As in c, for a standard non-network-based clustering approach (i.e., no network smoothing and substituting NMF for NetNMF). (e) As in c, using a permuted network. Using this simulation framework, we measured
the ability of NBS to recover the correct subtype assignments in comparison to a standard consensus clustering approach not based on network knowledge (i.e., the same NBS pipeline in Figure 3.1a without network smoothing and substituting NMF for NetNMF). NBS showed a striking improvement in performance, especially for large network modules as these can be associated with any of numerous different mutations across the patient population (Figure 3.2b). As module size decreases, the chance of observing the same mutated gene in patients of the same subtype increases and the standard clustering algorithm performs increasingly well. We found that the high performance of NBS depends not only on network smoothing but also NMF clustering approach, as substitution with an alternative method such as hierarchical clustering performs relatively poorly (Figure 3.2b).

Next, we investigated how NBS performance was impacted as a function of mutation frequency and network module size (Figure 3.2c). Standard consensus clustering was sufficient for high mutation frequencies and small modules, for which there is substantial overlap in mutations among patients of the same subtype (Figure 3.2d); however NBS was able to accurately recover the correct subtypes for a much larger range of both variables. Applying NBS on a permuted network resulted in poor performance (Figure 3.2e), on par with that observed with standard consensus clustering. These results were qualitatively similar when using multiple network modules per patient (2–6) and/or a different network (Figure S3.2).
3.3.3 Network-based stratification of tumor mutations

Motivated by the performance of NBS in simulation, we next sought to apply NBS to stratify patients profiled by TCGA full exome sequencing, separately for three different cancers – uterine(79), ovarian(78), and lung(77). In all three cancers, we observed that NBS resulted in robust subtype structure, whereas standard consensus clustering was unable to stratify the patient cohort (Figure 3.3a for uterine cancer, Figure S3.7a and Figure S3.8a for ovarian and lung). Similar results were obtained when using any of the three human networks considered in this study (STRING, HumanNet, PathwayCommons).

To determine the biological importance of the identified subtypes, we investigated whether they were predictive of observed clinical data such as histological appearance and patient survival time. In uterine cancer, NBS subtypes were closely associated with the recorded subtype based on histology (Figure 3.2b–c and Figure S3.9). Survival analysis was not possible due to low mortality rates for this cohort. In ovarian cancer, the identified subtypes were significant predictors of patient survival time (Figure 3.2d–e and Figure S3.7b,c). The most aggressive ovarian tumor subtype had a mean survival of approximately 32 months, while the least aggressive subtype had a mean
Figure 3.3. Network-based stratification of somatic tumor mutations. (a) Co-clustering matrices for uterine cancer patients, comparing NBS [STRING] to standard consensus clustering. (b) Uterine cancer: Association of NBS subtypes with histology. (c) Uterine cancer: composition of NBS subtypes (k = 3) in terms of histological type and tumor grade. (d) Ovarian cancer: Association of NBS subtypes (HumanNet) with patient survival time (Cox proportional hazard model). (e) Ovarian cancer: Kaplan-Meier survival plot for NBS subtypes (k = 4). (f) Lung cancer: Association of NBS subtypes (HumanNet) with patient survival time. (g) Lung cancer: Kaplan-Meier survival plot for NBS subtypes (k = 6). Hazard R. = Hazard Ratio.

Moreover, these subtypes were predictive of survival independently of clinical survival of more than 80 months, a 2.5-fold difference (Figure S3.7d,e).
covariates including tumor stage, age, mutation rate and residual tumor after surgery (Figure S3.10, Likelihood ratio test, \( P = 3.75 \times 10^{-5} \)). Furthermore, subtypes were predictive of time until the onset of platinum resistance (Figure S3.7f), as measured using a Kaplan-Meier analysis of platinum free survival(127). Finally, in lung cancer the identified subtypes were also found to be significant predictors of patient survival time (Figure 3.3f–g, median survival of 12 months versus approximately 50 months for the best surviving subtype, Figure S3.8). As for ovarian cancer, the lung cancer subtypes had predictive value beyond known clinical covariates such as tumor stage, grade, mutation frequency, age at diagnosis and smoking status (Likelihood ratio test, \( P = 3.3 \times 10^{-4} \)). Finally, stratification using a network in which the mapping between mutated genes and the network was permuted, disrupting the relationship between mutations and network structure, resulted in degraded predictive performance (Figure 3.3b,d,f).

Next, we compared these results to subtypes derived from other data types in TCGA, including CNV, methylation, mRNA expression, microRNA expression, and protein profiles. For ovarian cancer, all other data types had inferior ability to predict survival beyond what can be predicted from clinical covariates (Figure 3.4a) and led to very different subtype assignments than NBS (Figure 3.4b). In lung cancer, NBS subtypes and those based on mRNA-seq both had good predictive power (Figure 3.4c) and had some overlap in terms of patient assignments (Figure 3.4d), whereas other data types were not predictive of survival. In uterine cancer, subtypes derived from all data types
were highly predictive of histology (Figure 3.4e, CNVs had highest predictive power overall) and also had very high overlap with NBS subtype assignments (Figure 3.4f).

Figure 3.4. Comparing data types. (a,c) A comparison of the predictive value for patient survival as estimated using a Cox proportional-hazards model, and association with histological type (e), across different data types and methods. Subtypes resulting from clustering of data from CNVs, mRNA, microRNA (miRNA), methylation and reverse phase protein arrays (RPPA) were obtained from the Broad Firehose web portal\textsuperscript{59}. These subtype definitions were compared to the subtypes identified by network-based stratification of somatic mutations using HumanNet with four subtypes for ovarian (b), HumanNet with six for lung (d) and STRING with three for uterine (f). In each case (b,d,f), the p-value of significance is reported from a $\chi^2$ test of association between the assignment of patients to subtypes for each data type with NBS subtypes of a fixed number of subtypes.
3.3.4 Distinct network modules associate with each tumor subtype

Figure 3.5. A network view of genes with high network smoothed mutation scores in HumanNet subtype 1 relative to other subtypes. Subtype 1 has the lowest survival and highest platinum resistance rates amongst the four recovered subtypes. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes included in the COSMIC cancer gene census.

We next sought to identify the regions of the network that are most responsible for discriminating the somatic mutation profiles of tumors of different subtypes. Focusing on ovarian cancer as a proof-of-principle, for each subtype we identified genes for which the network-smoothed mutation state differs significantly for patients of that subtype versus the others (FDR < 0.05). This set of genes was projected onto the HumanNet network and visualized using Cytoscape(91). The network for subtype 1 (Figure 3.5), which had the worst overall survival and shortest platinum-free interval, contained over 20 genes in
the fibroblast growth pathway, which has previously been implicated as a
driver of tumor progression and associated with resistance to platinum and
anti-VEGF therapy(128). The network for subtype 2 was enriched in DNA
damage response genes including ATM, ATR, BRCA1/2, RAD51 and CHEK2
(Figure S3.11). Collectively these are characteristic of a functional deficit in
response to DNA damage, which has been referred to as ‘BRCAness’(83, 129).
Consistent with this finding, this subtype also included the vast majority of
patients with BRCA1 and BRCA2 germline mutations (15/20 and 5/6 patients in
the cohort, respectively). The network for subtype 3 was enriched for genes in
the NF-κB pathway (Figure S3.12), while subtype 4 was enriched for genes involved in cholesterol transport and fat
and glycogen metabolism (Figure S3.13). A similar analysis in uterine and lung
cancers produced other subnetworks with unique characteristics, including
enrichments for DNA damage response, WNT signaling and histone
modification to name a few (Figure S3.14-16).

Thus, the NBS approach is not only capable of stratifying patients into
clinically informative subtypes but may also assist in identifying the molecular
network regions commonly mutated in each subtype.

### 3.3.5 Translation to predictive signatures

For network-based stratification to be generally applicable to new patients
not in the TCGA, it is necessary to complement it with a procedure for
assigning a single patient to one of the existing NBS subtypes. For this purpose,
we explored the nearest shrunken centroid approach(119), a standard
method for sample classification which summarizes each subtype with a class ‘centroid’ and assigns new samples to the subtype with closest centroid. We found that this method was able to classify the network-smoothed mutation profile of an individual patient with over 95% accuracy (Figure 3.6a, 10-fold cross-validation).

![Figure 3.6](image)

**Figure 3.6.** From mutation-derived subtypes to expression signatures. (a) Classification accuracy (1 − classification error) when using a supervised learning method to learn a signature based on either somatic mutation profiles or gene expression, showing training error and cross-validation error. Dashed line shows the accuracy for a random predictor. (b) Kaplan-Meier survival plots for the TCGA ovarian cancer cohort patients when predicted using a classifier trained on subtype labels derived from network-based stratification of mutation data in TCGA. (c) Applying the same classifier to serous ovarian cancer samples from Tothill et al.

However, mRNA expression data are presently much more widely available than full genome or exome sequences, such that there are
numerous existing cohorts of cancer patients that have been profiled in mRNA expression but not in somatic mutations (83, 118, 130, 131). We therefore sought to test whether, having used NBS to define subtypes within TCGA somatic mutation data, it was possible to assign a new patient to these subtypes using an expression signature. To explore this idea, we used the mRNA expression profiles available for the TCGA ovarian tumor cohort to learn an expression signature for each subtype defined earlier by NBS. For this purpose, we again used the nearest shrunken centroid approach (119). We found that expression performed as an adequate surrogate for mutation profile, albeit at a reduced accuracy (Figure 3.6a, >95% for mutations, ~60% for expression, ~30% at random). This expression signature was nonetheless able to recover stratification predictive of survival (Figure 3.6b).

We also examined the predictive value of this gene expression signature in two independent studies of serous ovarian tumors by Tothill et al. (131) and Bonome et al. (132), as well as a meta-analysis including over 1000 patients, which subsumes Tothill, Bonome and TCGA samples which included expression profiles but lacked somatic mutation profiles (118) (Figure 3.6c, Figure S3.17). It is further noteworthy that this meta-analysis includes an unknown number of non-serous ovarian cancer samples. Using the expression signature, all patients were assigned to one of the four NBS subtypes. In the Tothill dataset, the subtype assignments were found to be significantly predictive of patient survival and platinum drug resistance (Logrank $P = 6.1 \times 10^{-3}$ and $1.65 \times 10^{-6}$ respectively, Figure 3.6c and Figure S3.17), following the
same trends observed in the original TCGA cohort. In the Bonome and the meta-analysis datasets, the gene expression signature again recovered four subtypes that were significantly associated with patient survival (Logrank $P=1.40 \times 10^{-3}$ and $1.22 \times 10^{-4}$, respectively, Figure S3.18). As a final control, we performed clustering of the Tothill expression profiles independent of NBS subtypes, resulting in a different set of subtypes that associated with survival to a more limited extent ($P=0.01$, Figure S3.18). These results show that tumor subtypes identified by NBS are identifiable in independent data sets using gene expression as a surrogate biomarker.

### 3.3.6 Effects of different classes of mutation on stratification

Finally, we studied the impacts of different classes of somatic mutation on the NBS approach. Although TCGA somatic mutations are profiled only in coding regions (i.e. exomes), a fraction of these mutations do not alter the protein sequence and thus are classified as ‘synonymous’ (e.g. 23% in ovarian cancer). Therefore, we tested the effect on NBS of disrupting synonymous mutations by reassigning them to new randomly-chosen gene locations. For uterine and lung cancer (Figure 3.7a and Figure S3.19), disruption of synonymous mutations had little effect on NBS performance, in sharp contrast to disruption of non-synonymous mutations or all mutations which greatly affected performance. Interestingly, in the ovarian cancer cohort (Figure 3.7b), disruption of both synonymous and non-synonymous mutations was detrimental to stratification performance.
Figure 3.7. Effects of different types of mutations on stratification. (a–b) The effects of permuting a progressively larger fraction of mutation per patient for different types of somatic mutation, for the uterine (a) and ovarian (b) tumor cohorts. Lines show the median performance and colored regions represent the median absolute deviation (MAD). (c–e) Different types of filters were applied as a preprocessing step prior to running NBS on the uterine (c), ovarian (d) and lung (e) cohorts. In blue the full dataset, in red we filter all synonymous mutations, in orange and yellow we filter the top 2% late to replicate and long genes respectively (long*: top 2% long genes omitting any COSMIC genes). In green are three types of filters based on predictors of the functional effect of mutation, in cyan the performance when permuting all mutation within each patient separately as a control. (a–e) We report the median $\chi^2$ statistic for uterine, and for ovarian and lung we report the median likelihood difference of a full model to a base model including just clinical covariates (age, grade, stage, mutation rate, residual tumor after surgery).

We also studied the effect of removing mutations judged to be non-functional in cancer by methods such as MutationAssessor(121), CHASM(89) and VEST(120), which use features such as sequence conservation and protein structural information to assess the likely impact of mutations. After using each method to score all mutations across all patients, we picked a permissive threshold for retaining mutations to use for NBS (retaining the top
75% of mutations as scored by CHASM and VEST, and using MutationAssessor with the ‘low threshold’ setting). Nonetheless, filtering mutations with these tools resulted in decreased association of NBS subtypes with patient survival in all three cancers (Figure 3.7c–e, with possible exception of VEST for ovarian tumors, Figure 3.7d). Finally, we studied the effect of removing long genes or genes with late cell-cycle replication times, both of which have been postulated to accrue high numbers of mutations that may be unrelated to tumor progression(133). We found that removal of long genes substantially degraded the ability to identify ovarian and lung subtypes predictive of survival (Figure 3.7d–e). However, removal of late-replicating genes had little effect, and in the case of the lung tumor cohort actually increased predictive power (Figure 3.7e).

3.4 Discussion and Conclusions

Here we have reported the discovery that, using prior knowledge captured in molecular networks, a set of tumor mutation profiles can be stratified into subtypes that are both biologically and clinically informative. These subtypes are distinct from those recovered through stratification of other types of data and are independent of other clinical markers known to associate with survival. We are able to highlight network modules distinctly characteristic of each subtype, providing new insight into the biological mechanisms driving tumor progression. To our knowledge, this is the first time that full exome somatic mutation profiles have been used to stratify patients and uncover subtypes in an unsupervised fashion.
A compelling proof-of-principle for Network-Based Stratification is provided by its ability to recapitulate the major histological subtypes of uterine cancer (Figure 3.3b,c) whereas clustering in the absence of network information fails to identify meaningful substructure in the data (Figure 3.7a,b). In ovarian cancer, subtypes defined by somatic mutations are predictive of survival, whereas subtypes derived by other measurements such as CNVs, mRNA and microRNA expression are not. Similarly for lung cancer, NBS subtypes are predictive of survival well beyond other data layers, with the exception of mRNA-seq which also performs well.

One might consider at least three potential reasons for the good performance of NBS. First, somatic mutations represent a digital signal in that a given gene can be considered either mutated or not, whereas most other data layers are analog signals representing measurements of continuous values. In general digital systems have improved accuracy and reproducibility, and are more robust to noise, in comparison to analog systems(134). Second is the issue that somatic mutation profiles are differential measurements between tumor and normal tissue, whereas expression and other ‘omics profiles are absolute measurements in each patient. The differential analysis filters out any mutations or variants present in the patient’s germline, leaving only tumor specific changes. In contrast, it has been difficult to identify a true ‘baseline’ gene expression state for a tissue as these measurements are dynamic and highly context specific. Finally, the somatic mutation profile captures the causal genetic events underlying tumor
progression, whereas 'omics profiles such as mRNA and protein expression are a functional readout of the current cell state and are influenced by external factors which may be unrelated to the underlying tumor biology.

The network modules identified as characteristic for each tumor subtype provide new insights into the biology of cancer and raise many new questions. One particularly promising finding was the prominence of the FGF pathway in ovarian tumor subtype 1 (Figure 3.5). This pathway has been implicated as playing an important role in tumor proliferation and angiogenesis, and currently many inhibitors of the FGF pathway are in clinical development(135). Specifically it has been shown that increased expression of FGF1 is associated with poor survival in ovarian cancer(136), and inhibition of FGF1 and FGF2 increases sensitivity to cisplatin in ovarian cancer cell lines(128). An intriguing question for future work is whether subtype 1 patients are particularly responsive to chemotherapy directed at network-identified targets such as FGFR inhibitors.

Another interesting observation is that several network modules are enriched for long genes. For example, for ovarian tumor subtype 2, a total of 12 of 176 genes in the module are in the top 2% by length ($P = 2.3 \times 10^{-4}$). One prominent example is TTN, the longest known coding gene. While prominent 'gold-standard' catalogs of cancer genes, such as COSMIC(137) and the list of Vogelstein et al.(114), are also enriched for long genes (e.g. 17 of 125 in the Vogelstein list, $P = 5.11 \times 10^{-10}$), there remains some controversy about the roles these genes may play in cancer. On the one hand, it is possible that long
genes are highly mutated not because they are drivers of cancer, but simply due to chance since they are a bigger ‘target’ to hit. On the other hand, there is no definitive evidence that mutations in long genes are not functional or do not contribute to tumor progression.

Our present analysis provides some evidence that these long genes should not be ignored. In the molecular network, long mutated genes are highly interconnected to other functionally-related genes of all lengths, which are also found to be mutated in patients of that subtype. For example, the network region for ovarian tumor subtype 1 (Figure 3.5) shows TTN interconnected to genes such as NEB, ANK1, and MYOM2, all of which are also mutated in patients of this subtype. These genes encode components of the cytoskeleton thought to have both structural and signaling roles(138). Although TTN is a long gene and thus might accrue mutations by chance, it is striking that other members of the same protein interaction neighborhood are also found to be mutated in tumors of the same subtype. Using permutation analysis we estimate that the chance of TTN having a network neighborhood with this same number of mutations is roughly P < 0.0001. Thus, one possibility is that the TTN and other cytoskeletal components are required for platinum-induced, p53 independent apoptosis, and that mutation in either structural or signaling proteins in this pathway lead to platinum resistance. In support of this theory is prior work demonstrating that cell shape is associated with chemotherapy response in ovarian cancer(139).
Another interesting observation is that synonymous mutations, while dispensable for stratification of uterine and lung tumors, appear to have some predictive power in stratification of ovarian tumors. In support of this finding, a number of high profile studies have suggested that synonymous mutations may indeed play a causal role in cancer progression(140-143). Further study is needed to understand whether ovarian cancer is indeed the outlier in this respect, and whether and how synonymous mutations truly function in this disease.

Finally, we see many opportunities to improve upon the basic concept of network-based stratification in future work. First, integrating multiple layers of information beyond somatic mutations (e.g. copy number variations, epigenome, transcriptome etc.) into a composite stratification method might further expand our ability to identify subtypes with clinically relevant differences. Second, although we have shown the utility of three sources of gene-gene interactions, there are several other types of networks worth exploring, such as those involved in signaling, metabolism, or transcription. Finally, although this study focused on uterine, ovarian, and lung cancer as proofs of concept, the NBS method is broadly applicable to any cohort of cancer patients for which somatic mutations are known. Finally, analyzing NBS subtypes across all cancers simultaneously (i.e. a pan-cancer analysis) will offer the intriguing opportunity to explore whether the genes and networks underlying the progression of a tumor are more informative than its tissue of origin.
3.5 Chapter Acknowledgements

Chapter 3, is a reprint of material from the following published work, of which I am the primary investigator and author:

CHAPTER 4
SATURATION AND POWER ANALYSIS FOR DETECTING COMBINATORIAL GENE ASSOCIATION IN CANCER

4.1 Background and Significance

The term Cancer is widely agreed to describe a diverse collection of conditions marked by the aberrant and uncontrolled proliferation of cells within a living organism (144). It is the result of a series of perturbations to the genome of a single progenitor lineage of cells accumulating over the span of years (114, 145). This multigenerational process of perturbation and selection on a population cells, draws many parallels with the process of evolution by natural selection (145). The primary mode of perturbations is thought to be single nucleotide changes to the genome (145), although other types of perturbation certainly play an important role. Some important examples include: local insertions and deletions, copy number changes of chromosomal regions, fusion events of different genomic regions, transposable elements, virus integration, and changes to epigenetics (146). The cause of these changes is equally variable among patients and cancers and includes inherited gene variation, germline mutations, changes due to error in gene replication, external environmental factors and reactive molecules produced with in a living cell.
Large-scale sequencing of cancer patients has allowed us to peer to a greater resolution than ever before at the mutational landscape of cancer (147). The emerging landscape of mutation frequencies is characterized by towering peaks, a smattering of foothills and vast plains (148) – one survey showed over 70% of genes mutated in less than 1% of patients and a mere 6 genes occurring in over 20% (149). Another prominent feature is great heterogeneity, which holds true on multiple levels. From a global whole genome perspective, mutations vary greatly among different cancer types (149) (e.g. melanoma is greatly enriched for cytidine to thymidine (C \rightarrow T) characteristic of UV radiation damage (150)). Furthermore, within any cancer type mutational frequency may vary by orders of magnitude among patients (151), and substantial heterogeneity with functional implication has been observed among individual cells from the same patient (152). Finally, across different regions of a genome the probability of mutation is thought to vary substantially with significant correlation to a variety of cellular factors (e.g. replication time, expression level) (151).

Despite this complexity, systematic identification of genes mutated in cancer remains a central goal of large scale sequencing efforts. Frequently mutated genes, often referred to as cancer drivers, are distinguished from the rest of the genome in which mutation are expected to act as ‘passengers’ and confer no meaningful proliferative advantage. Current opinion holds that tumorigensis is the result of multiple such successive steps of mutation and selection each of which confers a small survival advantage to the tumor (145).
Multiple methods have been developed to identify such `cancer drivers`, necessitating parameter-rich models trying to precisely estimate a gene and cancer-specific background mutation distribution (150, 151, 153).

Beyond single cancer genes, mounting evidence from the study of model organisms (154) and genetic disease (155) demonstrate the importance of higher order, multi-gene effects. Cancer specifically is often regarded as a disease of dysregulated pathways (97, 114) and interacting networks of genes (156), in which epistasis and synthetic lethality have been shown to play a prominent role (157, 158). Multiple methodologies have been specifically developed for identifying dysregulated subnetworks (99), modules exhibiting exclusivity patterns among genes (101, 159) and higher order predictive logic among interacting genes (103). These multi-gene effects may be expected to exert selective pressure on tumorigensis and cancer progression resulting in complex co-occurrence and exclusivity patterns among perturbed genes within individual patients.

In this work we explore limitations of analysis of mutational frequency to catalog the mutational landscape of cancer. First, we show that infrequent mutations of important oncogenes (OG) and tumor-suppressor genes (TSGs) abound. We show how using single gene focused approaches these would remain difficult to identify, even considering monumental sequencing efforts. Next, we show evidence that these rare genes play an important role in individual cancers. Finally, we explore the relationship between such cancer genes and higher-order interaction effects.
4.2 Methods

4.2.1 Data processing

We use curated somatic mutation data provided as supplementary to work by Lawerence et al. (149). The data was downloaded from http://tumorportal.org (160) on January 22nd 2014.

The data set contains tumor somatic variants and short insertion and deletions for a set of 4742 patients from 21 cancer types, processed using a standardized annotation pipeline (149).

4.2.2 Cancer gene reference Lists

We use the following sources as reference lists of cancer genes:

1. Cosmic genes. 480 Genes. The catalog of somatic mutations in cancer (161) v68.
2. Vogelstein genes. 138 Genes. Cancer genome landscapes (114)
3. Suspected cancer related. A list of 1055 genes implicated as cancer related from the above two sources plus the following:
   a. Lawrence et al., Discovery and saturation analysis of cancer genes across 21 tumor types (149).
   b. Kandoth et al., Mutational landscape and significance across 12 major cancer types (162).
   c. Tamborero et al., Comprehensive identification of mutational cancer driver genes across 12 tumor types (163).
   e. Firehose analyses 2014_01_15, list of Gistic2 narrow peak significant amplification and deletions across cancer types (165).
4.2.3 Global power analysis:

Analysis of the statistical power to identify cancer genes follows the analysis described in Lawerence et al. (149). Briefly, we use a test of power for difference in proportions (166) to estimate the sample size needed to detect with power of 90% a gene at a genome wide significance alpha (P<5x10^{-6}), for different background mutation frequencies $\mu$.

The proportions are estimated based on difference between two frequencies calculated using the following equations. In the additive model (Figure S4.1) the background $P_0$ proportion is estimated using:

$$P_0 = 1 - \left(1 - \mu f_g\right)^{3L}$$

The proportion for the signal we wish to detect:

$$P_1 = P_0 + r(1 - m)$$

Where:

$\mu$ A variable representing the background mutation frequency.

$f_g=3.9$ A gene-specific mutation rate factor estimated using MutSigCV. The chosen value corresponds to the 90\textsuperscript{th} percentile of genes.

$L=1,500$ The length of the gene in coding bases. Representing the 90\textsuperscript{th} percentile of genes.

$\frac{3}{4}$ Corresponds to the observed typical proportion of silent to non-silent mutations (of 1:3). The model only considers non-silent mutations to be of interest.
\( m = 0.1 \quad \text{A fixed misdetection rate parameter.} \\
\)

\( r \quad \text{A variable representing the signal.} \\
\)

The multiplicative model presented in Figure 4.1 uses the same background equation as above and the following equation for the signal we are attempting to detect:

\[
P_1 = P_0 + rP_0
\]

### 4.2.4 Estimating relative mutation frequency

Relative mutation frequency is calculated using the approach used to model background as part of the MutSigCV method (151).

For every gene and patient the non-synonymous (NS) mutation frequency is calculated as the ratio of observed NS mutations \( (n_{ns}) \) for a gene \( (g) \) and patient \( (p) \) to the number of bases in the gene which may harbor a NS mutation \( (N_{ns}) \):

\[
M_{p,g}^{ns} = \frac{n_{p,g}^{ns}}{N_{p,g}^{ns}}
\]

A background mutation frequency is estimated using the MutSigCV local regression bagel method as described in (151). Briefly, a background mutation frequency is estimated for every gene, using a local neighborhood of genes (‘bagel’) and the set of silent (synonymous) mutations in a gene. The neighborhood is found by using a covariance normalized (Mahalanobis) distance on a set of covariates shown to affect mutation frequency (gene expression, replication timing, and HiC compartment). Neighboring genes are sequentially added to the bagel ordered by distance, after each additional
gene is added the new estimate is tested for the ‘fit’ of the estimate to the observation for the gene and stopping when this fit becomes poor according to a beta-binomial probability function. The gene for which this estimate was made for is added to the bagel and the estimate is scaled by a patient specific mutation frequency factor.

We use the implementation and covariates provided as part of MutSigCV v1.4 (downloaded June 22\textsuperscript{nd} 2014).

4.2.5 Functional enrichment testing

VEST scores for every mutation in the analysis are extracted using the Cravat 3.1 web service (http://www.cravat.us/) (120). 559,742 (an additional 195,473 are silent variants and do not have VEST scores) of 845,988 are successfully scored using the service. Mutations which are not score by VEST (~10\% of all variants) are excluded from the functional analysis. Genes are assigned to evenly spaced bins in the observed range of [-4,9] based on the estimated relative mutation frequency for each gene. Two statistical tests are used to determine the significance of enrichment among a reference list (COSMIC, 480 genes) and a list of non-cancer genes (measured genome minus list of ‘suspected cancer related’ genes, 17,854 genes):

(1) A one-tailed Wilcoxon ranksum test for a difference in medians is used to compare the reference set of genes with the set of non-cancer ones. Functional scores corresponding to every mutation in the cohort for each gene are included as independent events.
(2) As the populations are highly skewed in terms of relative size we perform a second permutation based test which controls for the number of genes being compared among the two lists within each bin. We sample from the non-cancer list a set of genes of matching number to the reference set (in the bin). We calculate a median value for all mutations occurring in the sampled set of genes. Repeating this procedure 10,000 times, we calculate the empirical p-value for observing a median value greater than the median value for the reference set.

P-values are corrected using a family-wise error rate correction (Bonferroni) and declared significant if it is found lower than 0.05.

### 4.2.6 Conservation enrichment testing

We use conservation from the UCSC phylo46way track, calculated based on alignment of 45 vertebrate genomes to the human genome (167). We use the conservation score provided for each mutation by Lawerence et al. together with the mutation data. The score is used as above (4.2.5) in place of the functional score (VEST).

### 4.2.7 Identifying individual enriched cancer genes

Enriched cancer genes are tested using a one sided Wilcoxon ranksum test as described above, where every gene in a COSMIC reference list which has over five unique patients with a mutations is compared to the background distribution of genes (within the matching mutation frequency...
bin). We report genes significant with an FDR<0.1 calculated using a Benjamini&Hochberg FDR procedure.

4.2.8 Testing for pairwise enrichment of genes

Pairwise enrichment for gene co-occurrence or mutual exclusivity is tested using a two-step procedure based on a background derived by a switching permutation procedure that preserves both patient and gene marginal distributions (described below).

We resort to using a two-step procedure as described below due to the prohibitive computational cost (~14,000 choose 2 unique tests). of exhaustively testing every pair using a Fisher’s exact test.

1. Global pairwise scan using a G-statistic

   a) For every gene pair \( i,j \) we calculate a G-statistic comparing the observed and expected times the genes are co-mutated

   \[
   G_{i,j}^{1,1} = O_{i,j}^{1,1} \log \left( \frac{O_{i,j}^{1,1}}{E_{i,j}^{1,1}} \right)
   \]

   \[
   O_{i,j}^{1,1} = \sum_p I\{m_{pi} = 1\} \cdot I\{m_{pj} = 1\} \text{, the number of patients which are mutated in both gene } i \text{ and gene } j.
   \]

   \[
   E_{i,j}^{1,1} = N \cdot P(i = 1) \cdot P(j = 1) = \frac{\{p:m_{pi}=1\} \cdot \{p:m_{pj}=1\}}{N} \text{, the product of the marginal probability of mutation of genes } i \text{ and } j \text{ and the population size.}
   \]

   b) The G statistic is calculated on the true data and 1000 random switching permutations.
c) A pareto approximation is used to derive a p-value estimate for genes pairs with extreme values (as described below) if the true G statistic is within the set of 20 highest or lowest.

d) Gene pairs with and FDR < 0.1 are retained for step (2) testing, estimated using the Benjamini & Yekotile FDR procedure for correcting non-independent hypothesis (168).

e) Due to a relatively high false negative rate (primarily due to failures to successfully perform the pareto fit), steps (a-d) are repeated 10 times.

2. Testing of all pairs recovered in step (1) using a one-tailed fisher exact test.

   a) A p-value is calculated using a one sided fisher exact test for every pair recovered in step (1).

   b) A corresponding background of p-values is calculated using the same fisher test for 10,000 random switching permutations.

   c) We report the empirical p-value resulting from this procedure and declare an interaction significant, if it is below the FWER (Bonferroni) correction with an alpha=0.05, for the tests performed in the second stage procedure.

4.2.9 Switching permutation co-mutation background model

We use the switching permutation procedure previously described in the MEMo method (101). Briefly, the mutations matrix of patients by genes is treated as a bipartite graph. Edges in this graph are switched while preserving degrees. After a sufficient number of such switches, the result is a permuted mutation matrix which maintains the row and column distributions of the original matrix.

Using fast sparse matrix routines in Matlab and a number of heuristic improvements to the method we can perform this procedure quickly while
obtaining the same level of randomization. Randomization is estimated by comparing the resulting matrix jaccard as suggested in Gobbi et al. (169).

4.2.10 Pareto estimation of p-values

Calculating empirical p-value meeting significance thresholds after multiple hypothesis correction would make this problem prohibitively expensive to compute. Instead we use a procedure for estimating p-values by fitting extreme values in the sample with the Generalized-Pareto distributions, and extrapolating a p-value estimate based on the fit as outlined by Knijnenburg et al. (170).

4.3 Results

4.3.1 Power analysis for detecting cancer genes

In a recent work, Lawrence et al. estimate the saturation and power for detecting mutation occurring at a frequency above a background mutation rate (Figure S4.1). The model presented is based on a reasonable parameterization of mutation rates in cancer; however it assumes a cancer mutation which confers a proliferation advantage results in an additive deviation from the baseline mutation frequency. An alternative model may be that a significant deviation should have a frequency with a multiplicative factor of the background mutation rate (Figure 4.1a), modelling the property that cancers with higher mutation frequencies would have greater opportunity to collect mutations in any cancer genes (which once acquired would be selected to dominate the tumor). Due to limitation of current
sample sizes it remains difficult to ascertain which model better describes the biology of cancer. Regardless of model choice it is apparent that in order to be powered enough to discover mutations of modest enrichment of x1.5 (or 0.1 difference) over background; the majority of new cancer patients over the span of multiple years would have to be carefully sequenced.

4.3.2 The role of infrequent mutation in cancer

Next, we estimate the frequency relative to the background for every gene using the ratio of observed non-synonymous mutation with an estimated background for the gene. Testing the ability of this simple statistic to recover significant association identified with the full MutSigCV model we observe a median area under the receiver operator curve (AUC) of 91%, when testing on 20 different cancer cohorts (Figure S4.2). When used on the full PanCancer cohort we observe a similar remarkable AUC of 91%.
Figure 4.1. Recovering known cancer genes using mutation frequency relative to background. (a) Estimate of the number of samples needed to detect significantly mutated genes. As a function of a tumor type’s median background mutation frequency and a cancer gene’s mutation rate above background. The number of samples needed to achieve 90% power for 90% of genes (y axis). Grey vertical lines indicate tumor type median background mutation frequencies (x axis). Blue dots indicate the 10 year incidence for that cancer (in the US). (b-e) Compare the relative mutation frequency with the number of known cancer genes (cosmic, blue bars, left vertical axis) and the frequency of non-cancer genes (green line, right vertical axis). Dotted black vertical line indicates the median mutation frequency for the set of discovered using mutSigCV. Greyed region highlights the mutation frequency below background. Genes in the most extreme bins from each histogram are indicated. (b) Pancancer analysis over 20 cancer types. Same for (b) Melanoma (d) Breast (e) Lung adeno. (f) Calculating the maximum relative frequency across the 20 cancer cohorts.
Using relative mutation frequency, we compare the distribution of known cancer genes (COSMIC) to all non-cancer genes (Figure 4.1b). We observe that half (52%) of all known cancer genes have mutation frequencies lower than their background. Furthermore, we observe that ‘canonical’ cancer mutations, often considered ‘gatekeepers’ of tumorigenesis (e.g. TP53, PTEN, KRAS, CDKN2a and PI3CA), deviate remarkably from the background using this metric. We observe similar results when examining individual cancer cohorts (Figure 4.1c-e) or when using an alternative reference list of cancer genes. When comparing the maximum relative frequency across 20 different cancers we observe that most genes (cancer or otherwise) deviate from the background in some cohort (Figure 1f).

Next, we examine the functional implication of mutations occurring in different relative frequency bins. To avoid possible cancer biases we use VEST, a tool for scoring variants in heritable diseases (171). Comparing the vest scores of genes in the different relative frequency bins we observe that known cancer genes are significantly enriched for mutation with higher functional scores (Figure 4.2a). A comparison within individual cancer types (Figure 4.2b-d and Figure S4.3), or by using conservation of mutated bases in place of the functional score (Figure S4.4), shows the same result.
Figure 4.2. Evidence for functional implications of mutations in cancer genes irrespective of mutation frequency. (a) Boxplot compares the functional score (VEST) of known cancer genes (blue, COSMIC), with scores for non-cancer genes (green outline), within bins of relative mutation frequency (x-axis). The top and bottom of each box indicates the respective quartile, center line of box indicates the median value. Stars above box pairs indicate a statistically significant difference (Bonf. corrected $P<0.05$). Grey regions denote bins with mutation frequency below background. (b-d) Compares functional score in different mutation bins, for (b) melanoma, (c) breast and (d) lung adenocarcinoma. Lines indicate median values, solid denote known cancer, dashed non cancer, solid box markers indicate significance using both ranksum and permutation test based tests, diamonds indicate significance in one but not both tests (Bonf. corrected $P<0.05$). Error bars indicate standard error. (e-g) Show a selection of genes found significant (FDR<0.05), occurring in bins below or around one for each respective cancer type as in (b-d). Colored violins indicate the background value for non-cancer genes occurring in the bin, blue circles indicate score for individual mutation occurring in the gene. Reported are the uncorrected p-values for the difference in medians between circles and the distribution in violin (one-tailed Wilcoxon-ranksum p-values).
We can further examine the set of known cancer genes and identify specific examples of known cancer genes with relative mutation frequency below or around background that deviate significantly from other non-cancer genes in their relative frequency neighborhoods (Figure 2e-g and supplementary figure). For example mutations in TRAPP a PIKK family kinase specifically implicated in melanoma (172) exhibits relatively low relative mutation frequency, but a highly significant functional score compared to genes of similar mutation frequency (Wilcoxon Ranksum \( P=2.31 \times 10^{-6} \)). A second example in melanoma, the NF1 a tumor suppressor from the RAS pathway has been shown to play an important functional role specifically in patients with mutations in the oncogene BRAF(173). A third example ZNF521 a cancer gene implicated primarily in acute lymphoblastic leukemia, is significantly enriched in VEST score (Wilcoxon Ranksum \( P=2.93 \times 10^{-6} \)) while having a low relative mutation frequency. Mutations in this gene have not been implicated in playing a role in lung adenocarcinomas, however it has been shown to be differentially expressed in mesenchymal stem cells derived from non-small cell lung cancers (174).

### 4.3.3 Power and significance of co-mutation and exclusivity

Next, we examine all pairwise interactions among mutations occurring in at least 20 patients (~14K genes). We find that co-mutation patterns are highly skewed from the expected frequencies as estimated by their individual frequencies (i.e. marginal occurrence probabilities, Figure 4.3a). To account for this bias we compare the co-occurrence using a permutation scheme.
which preserves both patient and gene frequencies (101), and find that much of the skew towards co-occurrence is explained away using this alternative null model (Figure 4.3b). Focusing on exclusivity patterns we examine all pairs significantly mutually exclusive (FDR<0.1, see method for details). We use sub sampling to estimate the saturation for putative interacting pairs given different sample sizes and observe no saturation for genes in any joint expected mutation frequency (Figure 4.3c).

We examine the set of 7,000 putative pairwise exclusive pairs. Comparing the relative mutation frequencies for each gene of the interacting pairs, we observe a marked enrichment for pairs involving a gene of high relative frequency and another of neutral or lower frequency (Figure 4.3d). When we compare how the pairs overlap with known cancer status, we find that a similar trend where known cancer genes (particularly well studied oncogenes and tumor-suppressors) tend to interact preferentially with genes which have not previously been implicated in cancer (Figure 4.3e). One possible explanation for this is that this reflects the enrichment shown in relative frequency (Figure 4.3d). Finally, we examine the overlap of genes that participate in a mutually exclusive pairwise interaction with known cancer genes (Figure 4.3f). We find remarkably significant enrichment with all cancer gene sets examined. We find it noteworthy that this enrichment is most significant when we compared against a reference list of over 1,000 suspected cancer genes.
Figure 4.3. Exploring interactions among mutated genes. (a-b) Compares (log_{10}) observed and expected frequencies, hexagons are colored by the frequency of pairs falling in a respective bin. In panel (a) expected is calculated naively as the product of the product of probabilities of observing each gene to be mutated (product of the marginal probabilities, assuming independence of the mutation events). In panel (b) the expected is calculated as the mean observed co-occurance across a sample of 1,000 switching permutation. (c) Significantly mutually exclusive pairs, in black overall and in color for different expected co-mutation frequencies. (d) Indicated how the mutually exclusive pairs (black line in C), distribute across the relative mutation frequency space (additive model). (e) Shows how the interacting pairs distribute in terms of cancer status (Non-C.- non-cancer, Can. - suspected cancer gene, OG-known oncogene, TSG-known tumor suppressor gene). (f) Show a Venn diagram comparing the set unique genes recovered in C with genes in different reference cancer lists.

4.4 Discussion

The results presented above call to question a number of assumptions used in the study and treatment of cancer. The first of these is the notion that mutation frequency is an adequate proxy for the functional importance of a cancer gene. The evidence presented here for genes, known to be involved
in oncogenic processes, exhibiting negative or neutral selection while enriched for functional mutations, show that for many cancer genes frequency is a poor measure for importance.

We can offer multiple biological explanations for why frequency makes for a poor proxy to the importance of a gene. First, as we have shown some proportion of these genes are involved in higher order interactions effects. This suggests that such neutral/negative genes may confer little proliferative advantage or may even be detrimental to a tumor on their own; however in concert they may confer a benefit (156). An additional factor could be interactions with germline perturbation (175), copy-number changes (176) or with specific factors unique to a tumor micro-environment (177) (e.g. lack of specific signaling factors, nutrients and oxygen). Such differences may cause a perturbation to a specific gene to remain neutral in some patients, while highly selected for in others.

Furthermore, the presented results challenge the notion of dichotomizing genes as cancer ‘drivers’ and ‘passengers’ based on arguments about a population of cancer tumors. Due to the combinatorics of the space, every tumor is in fact unique in terms of its perturbation set. Mutations to individual cells in a tumor involve a random walk in a fitness landscape; nothing dictates that a step in this landscape should be either locally or globally optimal. Prevalence in this scenario may have little to do with selection. This seems particularly plausible in light of estimates that the majority of functional perturbations confer only a weak selective advantage
(178). As a result, a patient may be ‘fortunate’ to harbor a common well studied ‘driver’ with a specific targeted therapy, or unfortunate to have a tumor ‘driven’ by multiple genes acting in concert that have never been observed together before (and will never be observed again).

The disconnect between prevalence and selection adds to the mounting evidence that targeted therapy in cancer is not necessarily cancer type specific (179, 180, 181), and should rely instead on genetic testing (182). However, testing for a panel of commonly mutated ‘driver’ genes (164) occurring within a cancer type, may be nothing more than a cost saving surrogate to genome-wide tests (183). Given current sequencing costs it seems doubtful that testing limited gene panel will remain the cost effective solution for long, particularly when considering the potential return of using genome-wide results in retrospective studies.

A sobering corollary of these results is that ever larger sequencing efforts, while of paramount importance to the study and treatment of cancer, will not of themselves furnish us with a comprehensive catalog or parts-list for these diseases. The complex and varied nature of this disease, of itself suggests that assuming anything else will lead us to a similar folly as expecting complex heritable disease would be solved by genome wide association studies (184, 185). The size of the search space of even simple higher order interactions with cancer, dictates we are likely to remain under-powered to detect such effects using currently used methods. This will remain the case
even under the rosiest scenario of full patient sequencing as the standard of care and full unfettered access to this data by the research community.

4.5 Chapter Acknowledgements

I would like to acknowledge the support of the Ideker lab cancer team and in particular Dr. Hannah Carter, Dr. John P. Shen and Andy Gross, who provided great insight and advice, which made this work possible.
Figure S2.1. Example SNPs with poor genotype calls.

Called genotype: 0 1 2
Case: ○ ○ ○
Ctrl: × × ×

Figure S2.1. Example SNPs with poor genotype calls. Shown are scatter plots of the raw intensity measurements for a subset of QC-filtered, disease-predictive SNPs selected by Adaboost when using genotype calls (computed by the WTCCC, using CHIAMO) as input features. In each of the subplots, a point corresponds to a sample and the color corresponds to the genotype call for that sample at that SNP. Subplot A illustrates a SNP where the calling was accurate while the other subplots show anomalies in the genotype calls.
A.2 Figure S2.2. Accuracy of Adaboost on Type I diabetes.

Figure S2.2. Accuracy of Adaboost on Type I diabetes. This plot illustrates the predictive accuracy of Adaboost using processed genotype calls, in predicting Type I diabetes on held out data from two different cohorts. Colors correspond to different boost-remove iterations. Each iteration performs 50 rounds of boosting and then masks out the selected set of SNPs, in all subsequent iterations. The red curve corresponds to the accuracy when no SNPs have been removed from the data, while for the dark blue curve, 250 SNPs previously selected by Adaboost have been masked. In the WT set (a) accuracy of boosting is compared with the predictive accuracy of LASSO (black dashed line) and SVM (black dot-dashed line) as previously reported.
A.3 Figure S2.3. Proximity of SNP sets in results based on genotype calls.

Figure S2.3. Proximity of SNP sets in results based on genotype calls. Each subfigure compares the distribution of physical distance relative to chromosome length (in log space) between sets of SNPs selected by Adaboost using processed genotype calls on the two cohorts (shown in red), with the distribution of relative physical distance between random sets of SNPs drawn uniformly (shown in blue). We observe weak evidence (Wilcoxon ranksum, P=0.03) that the histograms shown in red and blue were sampled from the different distribution.
A.4 Figure S2.4. Supporting evidence for interaction edges recovered from SNP disease association. Figure S2.4. Supporting evidence for interaction edges recovered from SNP disease association.

Figure S2.4. Supporting evidence for interaction edges recovered from SNP disease association. The figure shows violin plots comparing the distribution of GO annotation similarities (measured using Wang et al. similarity) and the distribution of protein tissue expression correlation similarity as reported in the Protein Atlas database. Black violin denotes the background distribution for each metric, red violins denote the distance distribution for edges recovered from a network based on p-value recovered SNPs, blue violins denote the distance distribution for edges recovered from a network based on boosting SNPs. In both cases, a gene pair in the network contains an edge if each gene is linked to SNP that have a significant disease associated interaction term (FDR<0.01). Green squares denote the median value of similarity measure for the distribution.
APPENDIX B – Supplementary items for Chapter 3

B.1 Figure S3.1. An overview of the somatic mutation landscape of the TCGA ovarian cancer cohort.

**Figure S3.1.** An overview of the somatic mutation landscape of the TCGA ovarian cancer cohort. (a) Somatic mutations along the length of chromosome 17. (b) A histogram summing the frequency of mutations per gene for the entire exome. (c) A histogram summing the frequency of genes mutated per patient in the cohort.
B.2 Figure S3.2. Simulating across different networks.

Figure S3.2. Simulating across different networks. In this simulation network modules from the NCI-Nature cancer pathways network were used for the simulation and were recovered by NBS using the HumanNet network. Each subtype included between 2-6 driver modules totaling the specified size of genes and the driver gene frequency. Driver frequencies of 10%, 7.5%, 5% and driver modules comprising 100-120, 60-80, 20-40 were used in panels (a), (b) and (c) respectively. Furthermore, a subset (0-4) of the modules was assigned to overlap across multiple subtypes.
B.3 Figure S3.3. Ovarian cancer association with overall survival.

(a) Co-clustering matrices for ovarian cancer patients, comparing NBS (HumanNet) to standard consensus clustering. (b-c) Cox proportional hazards model logrank statistic for STRING and PathwayCommons. (d) Hazard ratio of each of the HumanNet subtypes compared to subtype 2 with confidence intervals (0.75, 0.8, 0.6 denoted in blue, yellow and orange respectively). (e) Mean and S.E survival in months for each of the subtypes. (f) A Kaplan-Meier plot of the probability of developing platinum drug resistance for HumanNet with four clusters, Logrank P=0.046. (Subtype 4 is dropped due to missing annotations for PFI for the majority of patients).
Figure S3.4. Lung cancer association with overall survival.

(a) Co-clustering matrices for lung cancer patients, comparing NBS (HumanNet) to standard consensus clustering.
(b) Lung cancer patient survival cox proportional hazard model logrank statistic for PathwayCommons. (c) A Kaplan-Meier survival plot with six subtypes.
B.5 Figure S3.5. Uterine cancer association with histological type.

Figure S3.5. Uterine cancer association with histological type. (a-c) Association with histological subtype vs. the number of clusters (K). (d-f) Association with tumor grade vs. the number of clusters (K). (g) Summary of histological types for each subtype. (h) Summary of tumor grade vs. each subtype.
Figure S3.6. Standard predictors of survival are independent of ovarian subtype. (a) Percentage of patients receiving an optimal surgical resection (defined as less than 10mm of residual tumor) does not vary significantly between subtypes ($\chi^2$ P-value = 0.77). (b) Federation of Gynaecological Oncologists (FIGO) tumor stage does not show evidence for dependence on tumor subtype ($\chi^2$ P-value = 0.48). (c) Age at diagnosis does not show dependence on tumor subtype (One-way ANOVA P-value = 0.89).
B.7 Figure S3.7. A network view of genes with high network smoothed mutation scores in ovarian subtype 2 relative to other subtypes, using HumanNet.

**Figure S3.7.** A network view of genes with high network smoothed mutation scores in ovarian subtype 2 relative to other subtypes, using HumanNet. Node size corresponds to smoothed mutation score. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes.
Figure S3.8. A network view of genes with high network smoothed mutation scores in ovarian subtype 3 relative to other subtypes, using HumanNet. Node size corresponds to smoothed mutation score. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes.
Figure S3.9. A network view of genes with high smoothed mutation scores in ovarian subtype 4 relative to other subtypes, using HumanNet. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes.
B.10  Figure S3.10. A network view of genes with high smoothed mutation scores in uterine cancer subtype 1 relative to other subtypes, using STRING.

Figure S3.10. A network view of genes with high smoothed mutation scores in uterine cancer subtype 1 relative to other subtypes, using STRING. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.
Figure S3.11. A network view of genes with high smoothed mutation scores in uterine cancer subtype 2 relative to other subtypes, using STRING.

Figure S3.11. A network view of genes with high smoothed mutation scores in uterine cancer subtype 2 relative to other subtypes, using STRING. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.
B.12 Figure S3.12. A network view of genes with high smoothed mutation scores in uterine cancer subtype 3 relative to other subtypes, using STRING.

Figure S3.12. A network view of genes with high smoothed mutation scores in uterine cancer subtype 3 relative to other subtypes, using STRING. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.
B.13  Figure S3.14. A network view of genes with high smoothed mutation scores in lung cancer subtype 2 relative to other subtypes, using HumanNet.

Figure S3.13. A. network view of genes with high smoothed mutation scores in lung cancer subtype 1 relative to other subtypes, using HumanNet. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.
B.14  Figure S3.14. A network view of genes with high smoothed mutation scores in lung cancer subtype 2 relative to other subtypes, using HumanNet.

Figure S3.14. A network view of genes with high smoothed mutation scores in lung cancer subtype 2 relative to other subtypes, using HumanNet. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.
Figure S3.15. A network view of genes with high smoothed mutation scores in lung cancer subtype 3 relative to other subtypes, using HumanNet. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.
Figure S3.16. A network view of genes with high smoothed mutation scores in lung cancer subtype 5 relative to other subtypes, using HumanNet.

Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.
Figure S3.17. From mutation-derived subtypes to expression signatures.

(a) A Kaplan-Meier analysis of the proportion of patients who acquire platinum resistance in the Tothill et al. expression cohort for subtypes defined in the TCGA dataset using somatic mutations and NBS. (b) Kaplan-Meier survival plots for the Bonome et al. ovarian cancer patients. (c) Kaplan-Meier survival plots for a metastudy of ovarian cancer patients by Győrffy et al. These subtypes were recovered using a shrunken centroid model trained on the TCGA expression data with somatic mutation NBS subtypes as labels.
Figure S3.18. Standard consensus clustering NMF used to recover subtypes in the Tothill et al. expression cohort of ovarian tumors. (a) Standard consensus clustering NMF was performed for 1000 rounds with random restarts on the top 4000 most variable genes in the cohort. Average linkage hierarchical clustering was performed on the co-occurrence matrix to recover subtypes. Kaplan-Meier plots are shown for three (b), four (c), and five subtypes (d).
Figure S3.19. Effect of permuting proportions of the lung cancer dataset. Permuting a progressively larger number of mutations uniformly from the entire lung cohort. We report the median likelihood difference of a full model to a base model including just clinical covariates (age, grade, stage, mutation rate, residual tumor after surgery, as well as smoking). The colored regions represent the median absolute deviation (MAD).
### B.20 Table S3. 1. Summary of gene interaction networks.

**Table S3. 1. Summary of gene interaction networks.** The table shows the networks used as part of our analysis. The HumanNet and STRING networks were filtered to include the top 10% of interactions according to the interaction weights. After filtering all edges were treated as unweighted.

<table>
<thead>
<tr>
<th>Network</th>
<th>Noes</th>
<th>Edges</th>
<th>Links and description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanNet v.1(6)</td>
<td>16,243 (7,949)</td>
<td>476,399 (47,641)</td>
<td><a href="www.functionalnet.org/humannet">www.functionalnet.org/humannet</a> A database of gene interactions, derived using a naïve Bayes approach by combining multiple lines of experimental evidence. Comprised of both protein-protein interactions (PPIs) and genetic interactions</td>
</tr>
<tr>
<td>STRING v.9(104)</td>
<td>16,560 (12,233)</td>
<td>1,638,830 (164,034)</td>
<td><a href="www.string-db.org/">www.string-db.org/</a> A database integrating a variety of evidence types including, experimental expression and literature mining approaches to derive a globally weighted network of gene interactions. Comprised of multiple types of gene interactions, including: PPIs, genetic and co-citation.</td>
</tr>
<tr>
<td>PathwayCommons(105)</td>
<td>14,355</td>
<td>507,757</td>
<td><a href="www.pathwaycommons.org/pc/">www.pathwaycommons.org/pc/</a> An aggregated repository of gene interactions from several sources including BioGrid, HPRD, IntAct and the NCI set of cancer specific pathways. Comprised of mostly physical PPIs.</td>
</tr>
</tbody>
</table>
Figure S4.1. Estimate of the number of samples needed to detect significantly mutated genes. As a function of a tumour type’s median background mutation frequency and a cancer gene’s mutation rate above background. The number of samples needed to achieve 90% power for 90% of genes (y axis). Grey vertical lines indicate tumour type median background mutation frequencies (x axis). Blue dots indicate the 10 year incidence for that cancer (in the US). For most tumour types, the incidence would be inadequate to reliably detect genes mutated at 0.1% or less above.
C.2 Figure S4.2. Relative mutation frequency as surrogate predictor for mutSigCV genes.

Figure S4.2. Relative mutation frequency as surrogate predictor for mutSigCV genes. The figure compares the relative mutation frequency with the number of known cancer genes (cosmic, blue bars, left vertical axis) and the frequency of non-cancer genes (green line, right vertical axis). Dotted black vertical line indicates the median mutation frequency for the set of discovered using mutSigCV. Greyed region highlights the mutation frequency below.
C.3 Figure S4.3. Functional score of cancer genes across the range of mutation frequency within cancer type.

**Figure S4.3.** Functional score of cancer genes across the range of mutation frequency within cancer type. This figure compares functional score (VEST) in different mutation bins, for 20 cancer types (ordered by median background mutation frequency). Lines indicate median values, solid denote known cancer, dashed non cancer, solid box markers indicate significance using both ranksum and permutation test based tests, diamonds indicate significance in one but not both tests (Bonf. corrected $P<0.05$). Error bars indicate standard error.
C.4  Figure S4.4. Conservation of cancer genes across the range of mutation frequency within cancer type.

Figure S4.4. Conservation of cancer genes across the range of mutation frequency within cancer type. This figure compares conservation score (phylo46way) in different mutation frequency bins, for 20 cancer types (ordered by median background mutation frequency). Lines indicate median values, solid denote known cancer, dashed non cancer, solid box markers indicate significance using both ranksum and permutation test based tests, diamonds indicate significance in one but not both tests (Bonf. corrected P<0.05). Error bars indicate standard error.
REFERENCES

1. Wetterstrand K (2012) DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. Available online.


126. Lab Y (bigWigAverageOverBed).


