Senior Thesis in Mathematics

Extracting hitherto unseen variant signals from the cancer genome using data de-sparsification strategies

Author:
Ethan Ashby

Advisor:
Dr. Johanna Hardin

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 14, 2021

Abstract

Mining rare variant signals from the cancer genome is relevant to a variety of clinical problems like classifying cancers of unknown primary site. Somatic variant mutation analysis has traditionally been restricted to cancer-associated genes with frequent occurrences, neglecting potential signals encoded in the vast "hidden genome" of rare and hitherto unseen mutations. Indeed, future tumor samples are bound to generate hitherto unseen mutations which may contain clinically-relevant signals. Smoothed Good-Turing frequency estimation is an intriguing statistical method that uses mutation richness to estimate probabilities of encountering hitherto unseen mutations. While previous research illustrated the potential of this method, analysis was restricted to the level of frequently mutated genes, omitting from consideration any signal in more sparsely mutated genes (which encompass the preponderance of the cancer exome). To include a wider cross-section of the cancer exome, this thesis explores the possibility of using de-sparsification strategies to aggregate mutation data in thoughtfully-constructed gene groups, for which Good-Turing probabilities can be calculated reliably. This thesis presents two de-sparsification strategies: 1) a higher-variance method that learns gene groupings directly from somatic mutation probability patterns and 2) a higher-bias method that aggregates mutation data within known biological pathways. Generating gene groups that contain cancer type specific hitherto unseen mutation probabilities may improve the ability to harness the "hidden genome" of unseen somatic mutation for important clinical tasks.

Contents

Chapter 1

Cancer: A Mutational Malady

Mutation is the unifying hallmark of cancer etiology, though mutation is a remarkably heterogeneous process that varies between cancer types and between patients. Moreover, the vast majority of somatic variant mutations (the most abundant mutations in human cancer) are extremely rare in databases of sequenced cancer genomes. This chapter provides an overview of cancer genetics, defines relevant terminology, and lays out the overarching, biomedically-oriented goal of this thesis: extracting clinically relevant signals from the abundance of rare mutation in the cancer genome.

1.1 A brief overview of cancer genetics

Human cancer is a tremendous burden on public health on a global scale. Even with the advent of modern medicine and improved health care, cancer has proven an especially difficult malady to treat. Indeed, cancer is the second leading cause of death worldwide (Hassanpour & Mohammadamin 2017). The difficulty of diagnosing and treating cancer lies in its many different causes and forms. Cancer is an umbrella term referring to more than 277 different kinds of disease (Hassanpour & Mohammadamin 2017) and is the result of a confluence of environmental and genetic factors (Lodish et al. 2003). However, the underlying cause of cancer is an accumulation of inherited or acquired alterations to an individual's DNA. For the purposes of this thesis, I will focus on the acquired *molecular alterations* that cause cancer, rather than the epidemiological or behavioral factors that increase disease risk.

Cancer is a loss of cellular regulation caused by an accumulation of mutations to DNA. Cancer development is based on two processes. First, the continuous, more-or-less random acquisition of mutation in individual cells throughout their lifetimes. Second, natural selection acting on the phenotype conferred by those mutations (Stratton et al. 2009). In most cases, random mutations are either neutral, meaning they have no impact on an individual's fitness, or deleterious, which disposes mutated cells to negative selection. However, in some cases, mutation can confer a growth advantage to cells, which allows them to proliferate uncontrollably, spread to other tissues, and cause disease.

There are two main classes of mutations in human cancer. Non-heritable mutations that are obtained through a cell's lifetime are termed **somatic mutations**. Somatic mutations comprise the majority of cancer causing genetic alterations. Germline, or heritable, mutations are passed down through family pedigrees (Stratton et al. 2009). There are many different kinds of somatic alterations, ranging from relatively simple single nucleotide DNA changes to exotic rearrangements of large DNA regions. The catalogue of somatic mutation is a rich resource to understand the etiology and mutational processes acting in human cancers (Stratton et al. 2009). For the purposes of this thesis, I will primarily focus on the most common mutation type: single nucleotide somatic alterations, or **somatic variant mutations** that occur in the protein-coding, or **exonic** regions of the genome. [1]

Each mutation in cancer may be classified based on its impact vis-à-vis oncogenesis (cancer development). Mutations can either be classified as *drivers* or *passengers*. Drivers, or causal mutations, confer a growth advantage that allows cancer cells to proliferate unchecked. Passenger mutations are biologically inert and do not contribute to tumor development (Stratton et al. 2009). It is important to note that a single mutation is rarely ever sufficient to induce oncogenesis alone. Most scientists agree on a "multi-hit" hypothesis, where carcinogenesis is caused by a sequence of mutations that create a rapidly proliferating cell type that evades normal checks on cell growth. This creates a positive feedback cycle, where unchecked cell growth decreases genomic stability and permits further mutation (Lodish et al. 2003). Indeed, as tumors develop, cancer cells acquire mutations that differentiate

[1] With the advent of whole genome sequencing, non-coding mutations are increasingly becoming interesting to researchers. However, I focus on somatic variants in cancer exomes, due to more available data and greater functional enrichment of mutations in exons.

them from their progenitors, leading to differences in treatment responses within cells of the same tumor (Croce 2008). Additionally, the"multi-hit" hypothesis explains why most cancers develop later in life, as a perfect storm of mutations steadily acquired through life are required to initiate cancer development (Lodish et al. 2003).

How do mutations to DNA give rise to a cancerous phenotype? Altering the DNA sequence of genes can lead to aberrant proteins which disregulate important pathways involving cell proliferation and survival. There are six main phenotypic hallmarks of cancer: 1) loss of contact inhibition, 2) exaggerated response to growth regulating signals, 3) failure to undergo programmed cell death in response to genetic damage, 4) immortalization 5) ability to evade immune defenses and 6) the production of factors promoting increased vascularization of the tumor (angiogenesis) (Hanahan & Weinberg 2000).

Genes involved in cancer induction can be classified into two major categories, *proto-oncogenes* and *tumor suppressor genes*. Both proto-oncogenes and tumor suppressor gene classes are involved in the maintenance of the cell cycle, or the endogenous system that controls cell division. Mutations to proto-oncogenes are typically dominant, gain of function mutations; that is, mutation to only one genomic copy of the gene is sufficient to induce cancer (Lodish et al. 2003). When a proto-oncogene is mutated, it becomes "activated" to an oncogene which can contribute to carcinogenesis. Tumor suppressor genes are a broad class of genes that encode cell cycle inhibitors, receptor proteins for inhibitory hormones, checkpoint proteins that halt the cell cycle upon DNA damage, and proteins that promote apoptosis (Lodish et al. 2003). Under normal function, tumor suppressor genes act to restrain cell division. Mutation to these genes permits inappropriate growth. Mutations in tumor suppressors are typically recessive, loss of function mutations, meaning that mutations (hits) to both gene copies are required to inactivate tumor suppressor genes (Lodish et al. 2003). An important subclass of tumor suppressor genes are *DNA repair genes*, which are responsible for ensuring genomic stability. In short, DNA repair genes fix DNA damaged by copying errors or mutagens, preventing the accumulation of potentially deleterious or cancerous mutations.

A useful analogy for carcinogenesis is the progressive breakdown of a car. In a normal human cell (much like a functioning car), there are multiple layers of control to ensure the cell operates predictably. Proto-oncogenes, like the gas pedal of the cell, provide the go signals for cell division. Tumor

suppressor genes, like the breaks, halt cell growth until appropriate conditions for division are met. DNA repair genes, like your mechanic, ensure that the cellular machinery are functioning properly. All of these components work in concert to tightly control the cell's passage through the cell cycle. But over time, an accumulation of inherited and acquired defects (like manufacturing errors and wear and tear on a car) can cause the internal checks on cell growth to fail. This can causes the cell to divide uncontrollably resulting in tumor growth.

1.2 Mutational heterogeneity in cancer

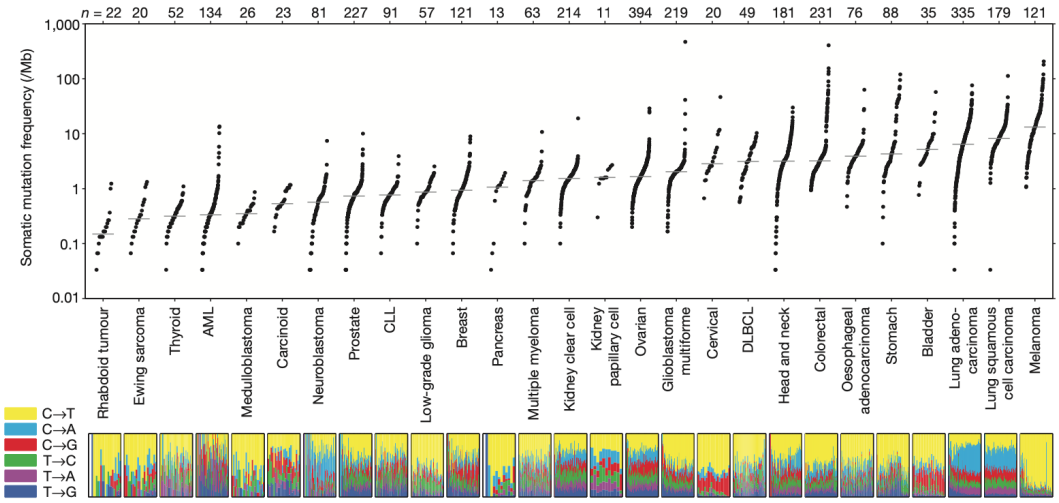


Figure 1.1: Figure from Lawrence et al. (2013). Main panel: frequency of mutation (# mutations/megabase) illustrates remarkable heterogeneity in mutation rates between and within cancer types. Minor panel: relative proportion of base substitution signatures within and between cancer types.

While mutation is the unifying molecular cause of cancer, it is not a monolithic process. The types and frequencies of mutation vary dramatically across the genome and across different cancer types[2]. Some mutations are

[2]Cancer types are defined according to tissue of origin and cell type (histology).

highly specific to particular cancer types, while others are present in many cancers. For example, the gene *TP53* is found mutated at high frequencies in many cancers, while the *BRCA1* gene is highly specific to breast and ovarian cancers (Schaefer & Serrano 2016).

In a study by Lawrence et al. (2013), the exomes (protein coding regions of the genome) of 3083 tumor samples spanning 27 different cancer types were sequenced. The authors assessed mutational heterogeneity on several different levels: mutation frequencies between tumors of the same cancer type, mutation frequencies between tumors of different cancer types, mutation signatures between tumors of different cancer types, and mutation frequencies between different locations of the genome.

Analysis of the 27 cancer types revealed that median mutation frequencies varied up to 1000-fold between cancer types as seen in the main panel of Figure 1.1. About half the variation in mutation frequency was explainable by cancer type. The study also found marked variation in mutation frequency between tumors of the same cancer types, also visible on the main panel of Figure 1.1. For example, the cancers with the highest mutation frequencies are lung cancers and melanomas, due to the potency of the carcinogens at work (cigarette smoke and UV radiation respectively) (Vogelstein et al. 2013). The results of Lawrence et al. (2013) demonstrate the remarkable variability in mutation frequency between different types of cancer and between different patients.

More complex mutational phenomena also show tissue specific patterns. The frequency of catastrophic "hypermutation" events are known to occur with varying frequencies in different cancer types. Complex mutation relationships such as co-mutation or mutual exclusivity are also known to be remarkably specific to certain cancer types (Schneider et al. 2018). Studies of somatic mutation have illustrated remarkable heterogeneity in mutational phenomena, which can be largely explained by differences in tissue type and background mutation rate.

Different cancer types also show marked variability in mutation spectra, or relative contributions of each single base substitution signature in the cancer genome. The minor panel of Figure 1.1 shows the relative proportions of the 6 possible single nucleotide base substitution signatures. A remarkable example is the last three panels, which show the mutation signature contributions in Lung adenocarcinoma, Lung squamous cell carcinoma, and Melanoma. In the two lung cancers, the majority of mutations are C$\rightarrow$A changes, as indicated by the large blue shaded area. These mutations are

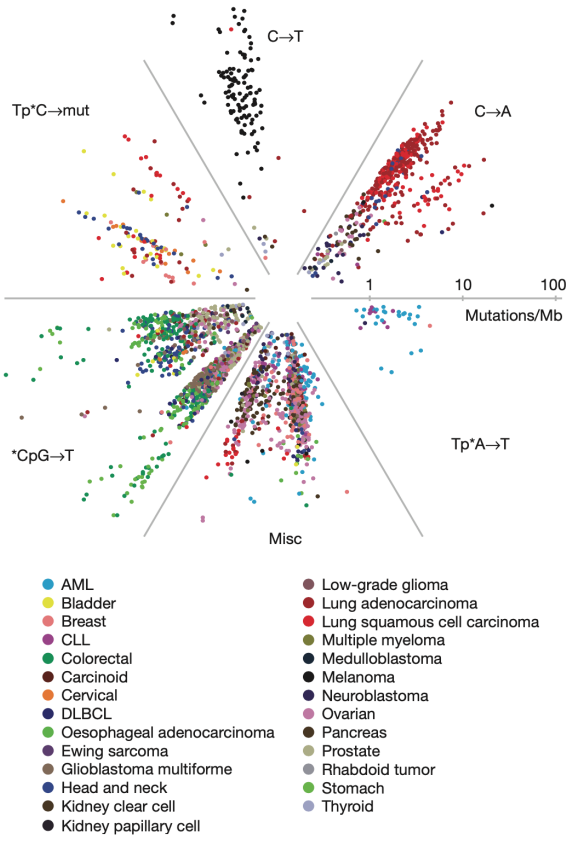


Figure 1.2: Mutation spectra in different cancer types reveals rich variation among different cancer types and natural grouping among cancers of the same types. Distance from the center represents the total mutation frequency (in mutations/Mb) while the angle indicates the relative contributions of each spectra (Lawrence et al. 2013).

consistent with exposure to the hydrocarbons found in cigarette smoke. In contrast, the melanomas were composed of an overwhelming number of C→T mutations, consistent with exposure to excess UV radiation.

The heterogeneity in mutation signatures between cancer types is further illustrated in the radial plot in Figure 1.2. A dimension reduction algorithm (non-negative matrix factorization) was used to summarize each tumor sample according to the relative contributions of each of the 96 single nucleotide mutations signatures[3]. Plotting tumors according to their mutation frequencies and the relative contributions of the mutation spectra identified by the algorithm illustrates a grouping of tumors according to their cancer type. Figure 1.2 illustrates a link between relative contributions of different mutation types and cancer type.

Lastly, the study by Lawrence et al. (2013) identified that mutation frequency varied regionally across the genome, with differences in mutation frequency varying between 5-fold and 10-fold at different positions along the genome. The authors contended that a failure to account for mutational heterogeneity (and therefore background mutation rates) was a major hindrance to identifying mutations that were statistically significant from the background rate (and therefore relevant to cancer). In short, mutation is not a monolithic process in human cancer. Mutation varies between cancer patients, cancer types, and within the genome itself.

1.3 Harnessing rare mutations: a clinically-relevant endeavor

In addition to improving understanding of cancer etiology, biomolecular study of cancer genomes can help inform diagnosis and patient treatment. The idea underlying the field of precision oncology is that molecular characterization of a patient's tumor can help predict an individual patient's response to specific treatments. These genotype-directed therapies have led to dramatic improvements in patient outcomes, providing a positive sign for the future of personalized cancer medicine (Scholl & Fro 2019).

Comprehensive sequencing efforts to uncover the mutational drivers of

[3]There are 96 single base substitution signatures, because there are 6 single nucleotide mutation types (as shown in the minor panel of 1.1), 4 possible bases (A,C,T,G) to flank the mutated locus on the 5' end, and 4 possible bases to flank the locus on the 3' end.

human cancer are complicated by a "long-tail" phenomenon, where sparsely mutated genes vastly outnumber commonly mutated ones (Vogelstein et al. 2013). Sequencing efforts have documented a small number of "mountains" and a large number of "hills", i.e., a small number genes that acquire a large number of mutations, and a large number of genes that are mutated infrequently (Vogelstein et al. 2013). Mutational "mountains", such as genes like *TP53* and *KRAS*, have been studied intensively, are extremely likely to drive carcinogenesis, and in some cases have associated treatments designed to target the particular gene or gene product. However, the challenge lies in the many "hills". Sparsely mutated genes dominate the landscape of mutation in cancer, and several likely play important roles in cancer development. A good example is the rare gene fusions in the *NTRK* family of receptor kinases, which allow cancers to evade normal checks on cell growth (Scholl & Fro 2019). This implies that many clinically-relevant yet rare mutations remain undiscovered. Thus, the statistical task of distinguishing sparsely mutated cancer genes from the background mutational noise is both extremely important and difficult.

Several different approaches have been used to identify and illustrate the clinical-relevance of rare mutations. One large scale analysis of nearly 25,000 cancers identified over 1000 mutational hotspots, of which 26% were novel (Chang et al. 2018). This brute force approach demonstrated the importance of large genome sequencing efforts sufficiently powered to identify rare driver variants. A clever statistical approach to identifying rare cancer drivers involves identifying mutational signals among proximal groups of "hills". For example, the HotNet2 model (Leiserson et al. 2015) assembles genes into small interacting networks and identifies rare drivers by identifying subnetworks with substantially higher-than-expected mutation scores. HotNet2 groups individual "mountains" into mountain ranges and their associated "foothills", enabling more powerful identification of rare driver genes.

1.4 Cancers of unknown primary site, clonal origin of metastasis, and liquid biopsy

Study of mutational heterogeneity of different cancer types (especially rare mutations) is relevant to clinical problems like diagnosing cancers of unknown primary site, identifying the clonal origin of a metastatic cancer, and in

emerging liquid biopsy technologies.

1.4.1 Cancers of unknown primary

Cancers of unknown primary site (CUPs) comprise 3-5% of cancer diagnoses globally and are typically associated with poor health outcomes (Pavlidis & Khaled 2015). Cancers of unknown primary site are often metastatic cancers for which the anatomical site of origin is unknown after detailed investigation. Once thought to be their own class of cancers, it is now recognized that CUPs are a heterogeneous class of tumors, each retaining a genomic signature of their tissue of origin. The advent of improved imaging techniques, immunohistochemical testing, and genomic and proteomic sequencing tools have sophisticated our approach to diagnosing and treating these cancers. Thus, cancers of unknown primary are a prime target for personalized medicine, as treatment can be informed by the molecular profile of individual patients (Varadhachary & Raber 2014).

1.4.2 Clonal origin of metastasis

A major challenge for pathologists is determining whether a tumor is a metastatic tumor or independent occurrence. This task is typically done by comparing the histological characteristics of tumor cells, but genetic markers have become relevant to this decision. The idea is that two tumors that metastatically evolved from the same progenitor will have some somatic mutations in common, while unrelated tumors will display different mutation profiles. However, the significance of shared mutation depends precisely on how common the observed mutations are. Thus, using genomic data, particularly rare variant data, to test for clonal relatedness is a important task (Ostrovnaya et al. 2015).

1.4.3 Liquid biopsy

Liquid biopsy is the analysis of cell-free, circulating tumor DNA obtained through a minimally invasive procedure such as a blood draw or urine sample (Wan et al. 2017). Liquid biopsies could be clinically beneficial in a variety of settings: early or population-wide cancer detection screens, prognosis assessment, treatment selection, and treatment monitoring. Liquid biopsy approaches are still an emerging technology, and identifying better ways to

match mutational information in circulating tumor DNA fragments to specific cancer diagnoses will render these tests more efficacious and more widely used.

1.5 Summary

The biomedical goal of this thesis is to improve extraction of cancer type specific signals from large-scale sequencing data (particularly from sparsely mutated genes) in the cancer genome. These signals could potentially be relevant to the clinical tasks outlined in the previous section. The major statistical challenge, of which the remainder of my thesis will attempt to address, involves harnessing the signal encoded in the preponderance of rare mutation in the cancer genome.

Chapter 2

Unseen variant probability estimation

How do you estimate the probability of a hitherto unseen event? This chapter details a statistical approach to this problem, initially developed in the field of computational linguistics. Then I detail the work of Chakraborty, Arora, Begg & Shen (2019*b*), who apply this method to mutations in the cancer genome, generating clinically-relevant results.

2.1 Smoothed Good-Turing Frequency Estimation

In linguistics, there are essentially infinitely many words and word combinations. To that end, a finite sample of language data will fail to capture the multitudinous linguistic possibilities, and future samples are bound to encounter new or previously-unseen units. Important tasks in computational linguistics such as spelling correction, sense disambiguation, and machine translation can be improved by precisely estimating the probability of hitherto unseen units by assigning them a nonzero probability (Gale 1995).

Table 2.1 shows a common pattern in linguistic data. The table contains summary statistics from the abstract of this thesis. r, the frequency, denotes how frequently a word appears in my abstract. For example, a word that occurs with $r = 1$ appears once in my abstract, a word that occurs with $r = 2$ appears twice, and so on. N_r, the frequency of frequencies, counts how many words occur at a particular frequency, r. For example, N_1 denotes the

frequency	frequency of frequencies
r	N_r
1	100
2	25
3	9
4	4
5	2
6	0
7	1
8	1
9	0
10	2

Table 2.1: A table illustrating the word frequencies of words in the abstract of this thesis. The right skew in distribution of N_r illustrates that linguistic data is dominated by rare terms (occur with frequency $r = 1$).

number of words that occur only once in my abstract (there are 100 of these), N_2 denotes the number of objects that occur twice (25 of these), and so on. Words included in the N_1 category represent infrequent or rare words like "restricted", while words in N_{10} might include common words like "the". The distribution of N_r values shows a substantial right skew, where the majority of words occur at low frequencies ($r = 1, 2$), and as r increases, N_r decays and becomes more unstable/noisy. In short, linguistic data is dominated by the rare terms, and grows noisier as r gets larger.

What is absent from Table 2.1 is the row corresponding to $r = 0$, N_0. This row corresponds to the number of *unseen* objects. Maximum likelihood estimation (which produces probability estimates that maximize the likelihood of the observed data) would predict the the probability of a word that occurs with frequency r to be $\frac{r}{N}$ where $N = \sum rN_r$. Since $r = 0$ in the case of unseen objects, maximum likelihood would assign every unseen object a 0 probability.

However, assigning unseen objects 0 probability is often undesirable. For instance, consider the abstract of this thesis as a sample of linguistic data.

The main body of this thesis has produced new words that were not present in my abstract; in short, the occurrence of new words does not happen with 0 probability. In applied scenarios that require realistic modeling of human language, such as sense disambiguation. machine translation, and text prediction, assigning 0 probability to unseen words can lead to poor performance. In short, maximum likelihood approaches fail to account for the large number of unseen objects in linguistic-type data, motivating the need for probability estimation methods that account for the units unseen in a finite sample.

Good-Turing estimation is a popular strategy for dealing with the unseen unit estimation problem. Good-Turing estimation relies on the idea that the number of objects that occur with rate $(r + 1)$ is typically smaller than the number of objects that occur with rate r. We saw this behavior in Table 2.1: as r increases, the N_r value decays. Qualitatively, the Good-Turing method takes a portion of the probability occupied by objects that occur with rate $r + 1$ and divides the probability evenly among the objects that occur with rate r. The probability occupied by all objects $i \in \{1, \ldots, N_r\}$ that occur with rate r is estimated by:

$$\sum_{i=1}^{N_r} P_{GT}(r) = \frac{(r+1)N_{r+1}}{N} \tag{2.1}$$

To obtain the probability of an individual object that occurs r times, we need to spread the above probability evenly among the N_r objects:

$$P_{GT}(r) = \frac{(r+1)N_{r+1}}{N \cdot N_r} \tag{2.2}$$

This is the Good-Turing estimator for the probability of encountering an object that appears r times!

2.1.1 Where does smoothing come in?

The Good-Turing estimator works well when r is small, but fails when operating in the noisy tail of the frequency distribution (see larger rows of r in Table 2.1). For example, suppose using the data from the thesis abstract, we want to estimate the probability of encountering a word occurring 5 times $(r = 5)$ in Chapter 2. The Good-Turing formula would assign the probability like so:

$$P_{GT}(r=5) = \frac{(r+1)N_{r+1}}{N \cdot N_r} = \frac{(5+1)}{N} \cdot \frac{N_6}{N_5} = \frac{(6)}{N} \cdot \frac{0}{2} = 0 \tag{2.3}$$

The problem is that N_r values become unreliable for large values of r. In a finite (often limited) sample of linguistic data, there will exist gaps in our distribution of N_r, which need to be filled in to obtain an accurate representation of the true word frequency distribution.

Smoothing can help alleviate the effect of noise on N_r. Gale (1995) provided a simple smoothing of the N_r values that also performed best in a series of Monte Carlo simulations. In fact, the ease of their smoothing method has increased the popularity of applying Good-Turing methods in practice. Intuitively, Gale's method smooths the N_r distribution asymmetrically, preserving probabilities for the better-estimated regions of the N_r distribution (for example, when r is small) and adjusting the probabilities for sparser regions of the N_r distribution (for example, when r is large).

Specifically, Gale uses an averaging transform to generate quantities Z_r which represent averages of each nonzero N_r with the zeros around it:

$$Z_r = \frac{N_r}{0.5(t-q)} \tag{2.4}$$

Where q, r, t are successive indices of non-zero N_r values. If N_r is in a well-estimated (non-noisy) part of the frequency distribution, $(t-q) = 2$ and $Z_r = \frac{N_r}{(0.5)(2)} = N_r$. If N_r is noisy, i.e., lots of gaps in the N_r distribution, Z_r can differ from N_r by several orders of magnitude.

Once these Z_r values are calculated, a simple linear regression of $\log Z_r$ on $\log r$ suggests a simple linear smooth of these average transformed frequency values (Figure 2.1). By averaging nonzero and zero N_r, we allow Z_r to take on continuous values, allowing the linear trend to continue past $N_r = 1$. Using this linear regression, we can impute Z_r values for any desired value of r.

Thus, the **smoothed Good-Turing frequency estimator** includes this simple and effective smoothing step. Here, Z_r is represented by $S(N_r)$, denoting the result of the smoothing of the N_r. The probability of encountering an a single object that occurs r times is as follows:

$$P_{GT}(r) = \frac{(r+1)}{N} \cdot \frac{S(N_{r+1})}{S(N_r)} \tag{2.5}$$

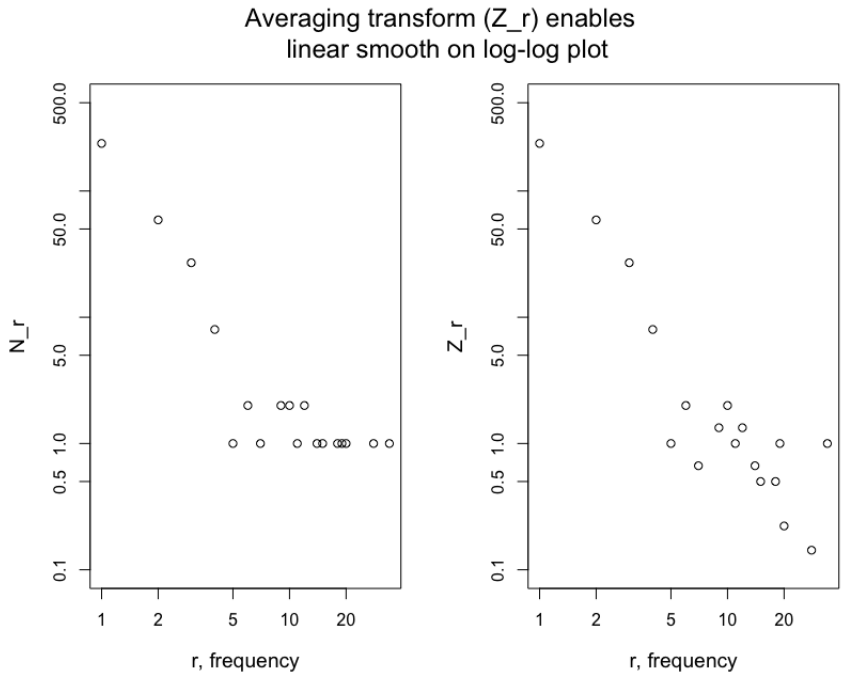


Figure 2.1: Average transform (Z_r) permits simple linear smooth of frequencies of frequencies on log-log scale. Data represent word frequencies from the first two pages of this thesis.

2.2 Application to the cancer genome

At this point, smoothed Good-Turing frequency estimation may seem like a useful tool in estimating probabilities of words or word combinations. But how is this method applicable to mutations in the cancer genome? Perhaps more importantly, how could it assist in the task of classifying cancers of unknown primary sites?

2.2.1 Motivating Good-Turing and cancer classification with a novelistic analogy

Suppose you are interested in classifying novels into two categories: Gothic and science fiction. You hunt around in your library and obtain a training set of novels. Two Gothic novels — *Jane Eyre* by Charlotte Brontë and

Wuthering Heights by Emily Brontë — and two science fiction novels — *Fahrenheit 451* by Ray Bradbury and *1984* by George Orwell.

You find a loose page from an unknown novel, and you would like to assign it to either your Gothic or science fiction bookshelf. To do this, you might construct a table of word frequencies from your training set of books like so:

Category	Gothic		SciFi	
Word	JE	WH	F451	1984
Romance	47	96	2	11
Machine	0	2	41	62
Love	101	167	23	52
...	...	...	...	...

Table 2.2: Toy example illustrating made-up word frequencies for different documents and document categories (genres).

If your loose page contained an abundance of words like "Romance" or "Love", that would suggest that the word composition is most similar to *Jane Eyre* and *Wuthering Heights* and is most likely a Gothic novel. On the other hand, if your loose page contained words like "Machine", that would suggest it is more likely to be a Science Fiction book.

However, this reasoning is conditional on your loose page containing a word like "Love" or "Machine". There is no guarantee that these highly discriminative words appear on your loose page. In fact, many words in your loose page will not have appeared in your training set of novels at all. Rather than ignoring these many new words, one could use Good-Turing frequency estimation to estimate the probabilities of encountering these previously unseen words in your various documents and document categories. In this case, Good-Turing probabilities are a proxy for *vocabulary richness*. In this novelistic setting, Good-Turing probabilities can help which genre is most likely to generate a previously unseen word, or which genre has the richer vocabulary. By estimating a category's propensity for generating new words, you can use hitherto unseen words to aid the assignment of your loose piece of paper to a category.

This document classification example is directly analogous to an example in cancer classification. Although instead of considering words and word

frequencies, we consider genes and their mutation frequencies. And instead of novels, we consider tumor samples that have been "read" using whole exome sequencing. And instead of novel categories, we consider cancer type categories.

In the case of The Cancer Genome Atlas (TCGA), there are 32 different cancer types (categories), 10268 tumor samples, and 19441 genes mutated at various frequencies. The table of genes, mutations, and cancer sites is shown in Table 2.3 below:

Genes	ACC (92)	BLCA (411)	...	UVM (80)
A1BG	0	0	...	0
A1CF	2	1	...	0
...	...	...	...	...
ZZZ3	1	0	...	1

Table 2.3: Table denoting mutations in 19020 cancer genes (rows) in 32 different cancer types (columns) with number of tumors of each particular cancer type in parentheses. Note that the matrix is very sparse (a number of 0 entries).

Smoothed Good-Turing frequency estimation is applicable to somatic variant mutation data because mutation data emulates text data in a critical ways. Like the linguistic data described above, somatic mutations are count data. Mutations (like words and word combinations) are essentially infinite in varieties and numbers. Also, like text data, mutation data are dominated by rare objects. Figure 2.2 demonstrates the right skewed frequency distribution characteristic of gene mutation and linguistic data, wherein the vast majority of units are rare (occur with frequency 1 in the dataset) and N_r decays with increasing r. Validation on an external dataset illustrated that $\geq$ 66% of mutations im major cancer genes were new (hitherto unseen in the TCGA cohort) (Chakraborty, Arora, Begg & Shen 2019*b*), illustrating that future sequenced tumor samples will generate many new mutations, highlighting the potential utility of the Good-Turing method.

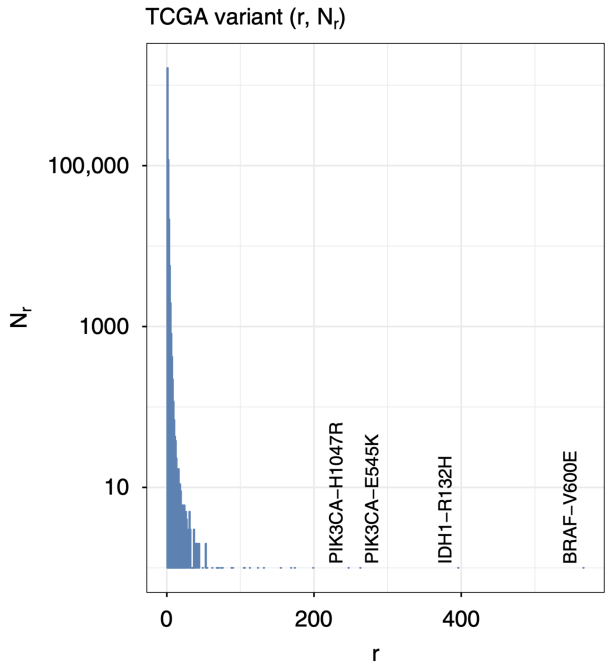


Figure 2.2: Figure from Chakraborty, Arora, Begg & Shen (2019*b*). In large sequencing cohorts (The Cancer Genome Atlas shown), rare mutations are comprise the vast majority of mutations. The x-axis is the frequency, r, at which a mutation is observed across more than 10,000 tumors, and the y-axis is how many mutations occur at that frequency, N_r (on log scale). The substantial right skew indicates that the number of rare mutations vastly outnumber the common mutations.

2.2.2 Applying smoothed Good-Turing frequency estimation to mutations in the cancer genome

Chakraborty, Arora, Begg & Shen (2019*b*) define the Good-Turing formula as follows[1]:

$$P_{GT}(r) = \frac{r+1}{m+1}\frac{S(N_{r+1})}{S(N_r)} \tag{2.6}$$

Where $P_{GT}(r)$ measures the probability of occurrence in a randomly se-

[1]For the derivation of this formula, please see the supplement section 7.3

lected new tumor of a variant that occurs r times in m previous tumor samples. Note that this formula estimates the probability of a particular somatic variant mutation (e.g., C→T mutation at position 20,000 on Chromosome 6).

Note that estimating the probability of a hitherto unseen mutation ($r = 0$) requires knowledge of N_0, of the total number of unseen mutations in the population of cancers. This is a very difficult parameter to estimate; how do we know the total number of unobserved mutations possible in all human cancers? Chakraborty, Arora, Begg & Shen (2019*b*) circumvent this problem by considering the probability of observing *at least one hitherto unseen variant* (π_0):

$$\pi_0 = 1 - \prod_{\forall \text{variants}; r=0} (1 - P_{GT}(r = 0)) \tag{2.7}$$

Per the Good-Turing formula in equation 2.6, $P_{GT}(r = 0)$ can be written as:

$$P_{GT}(r = 0) = \frac{1}{m+1}\frac{N_1}{N_0} \tag{2.8}$$

so

$$\pi_0 = 1 - \left(1 - \frac{N_1/(m+1)}{N_0}\right)^{N_0} \tag{2.9}$$

Using the limit definition of the exponential function:

$$e^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n \tag{2.10}$$

and supposing N_0 is large, we can rewrite (2.9) as:

$$\hat{\pi}_0 \approx 1 - \exp\left[-\frac{N_1}{m+1}\right] \tag{2.11}$$

This formula allows calculation of the probability of encountering **at least one previously unseen variant** in a set of sequenced tumor samples of size m that contain N_1 singleton mutations (without requiring knowledge of N_0).

2.2.3 A brief note on assumptions

In Chakraborty, Arora, Begg & Shen (2019*b*), a product binomial model is used to derive the Good-Turing estimator, requiring that variants occur independently of one another. This is an overly simplistic assumption, since certain genes are known to demonstrate co-mutating or mutually exclusive mutation patterns. However, Chakraborty, Arora, Begg & Shen (2019*b*) contend that violating independence doesn't have a major influence on results, since the patterns were validated on an external dataset.

2.2.4 Results of applying smoothed Good-Turing frequency estimation to mutations in the cancer genome

In Chakraborty, Arora, Begg & Shen (2019*b*), equation 2.11 was applied to genes mutated in more than 3% of tumor samples, by considering N_r values per gene and per cancer type. Interestingly, the Good-Turing probability estimates varied between genes and between cancer types. In Figure 2.3, the x-axis denotes the 32 different cancer types in the TCGA data, and the y-axis denotes a select group of 13 genes. The bubbles are sized based on the probability of encountering at least one previously unseen mutation in that particular gene in a future sequenced tumor of a particular cancer type.

For instance, the gene *VHL* almost exclusively produces previously unseen mutations in the KIRC cancer type, a particular form of renal cancer. Thus, if a tumor of unknown primary site contained a previously uncatalogued mutation in the *VHL* gene, that would lend some evidence that the tumor may belong to the KIRC type. Chakraborty, Arora, Begg & Shen (2019*b*) found that unseen variant signals were statistically significant[2] across hundreds of genes and could be validated on an external mutation dataset. They also found that the task of cancer type classification could be improved by incorporating unseen variant probabilities into a machine learning classifier (Chakraborty et al. 2020). This illustrates the potential utility of the "hidden iceberg" of unseen mutations in important clinical tasks.

[2]Statistical significance was determined by permutation test, comparing Normalized Mutual Information (NMI) of true unseen variant probabilities to a null distribution of NMIs generated by permuting variant tissue labels.

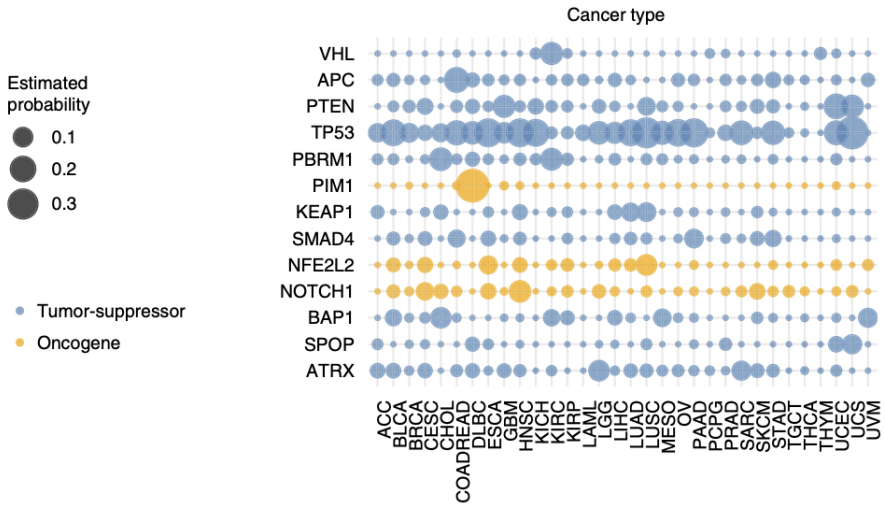


Figure 2.3: Patterns of probabilities of encountering at least one hitherto unseen variant (calculated using equation 2.11) in different genes and different cancer types.

2.2.5 The tip of the mutational iceberg: limitations and proposed problem

Chakraborty, Arora, Begg & Shen (2019*b*) highlight that while they focused their analysis to single nucleotide, non-synonymous somatic variant mutations in non-hypermutated tumors, their analysis is extendable to more complex mutation types and all but the most severely hypermutated tumor types.

Most importantly, the authors mention that they restricted their analysis to only relatively frequently mutated genes (mutated in at least 3% of all tumor samples). This amounted to fewer than 500 genes, or approximately 2.5% of the cancer exome, which mostly represented primary cancer genes. Only considering cancer genes does not effectively represent the genome as the whole. Indeed, it is believed that oncogenic mutations display a "long-tail" phenomenon, where the vast majority of genes driving cancer are sparsely mutated (Leiserson et al. 2015).

In order to go beyond the tip of the "hidden iceberg" of unseen somatic mutations, we must be able to identify unseen variant signals in the 98.5% of the cancer exome that is sparsely mutated. This is the problem that is the central focus of my thesis, and that I intend

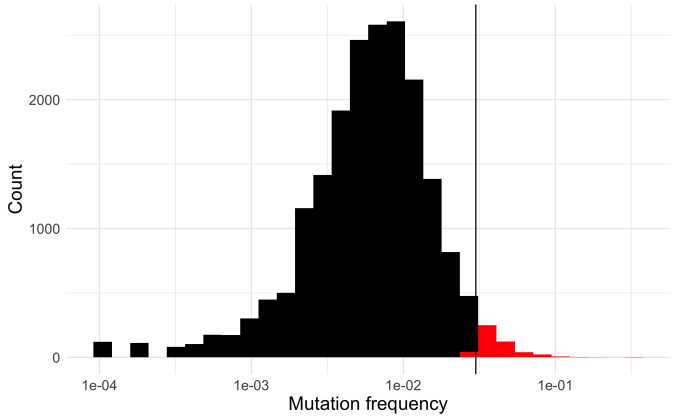


Figure 2.4: The tip of the mutational iceberg (shown in red), are the genes analyzed by Chakraborty, Arora, Begg & Shen (2019*b*). The dark part of the iceberg – variants in sparsely mutated genes – remains unexplored.

to address by aggregating mutational signals in gene groups.

Chapter 3

Cancer classification and data de-sparsification

In order to identify cancer type-specific unseen variant signals in the 98.5% of the cancer exome that is sparsely mutated, we must aggregate somatic mutation data in a statistically or biologically principled manner, or else lose out on potential discriminatory signals. This chapter provides an overview on machine learning-based cancer classification and the role of data de-sparsification in cancer classification efforts.

It's worth noting that most attempts to identify cancer biomarkers and molecular patterns for the purposes of cancer classification rely on transcriptomic (gene expression) data. However, DNA sequencing (as opposed to RNA-Seq or microarray technologies) is most commonly used in clinical practice. Somatic mutation data is also considerably sparser than gene expression data. Some research has been conducted on data-desparsification in the realm of cancer subtype classification using somatic mutation data which is detailed in the next section.

3.1 Cancer classification using somatic mutations

To my knowledge, the first attempt at using a machine learning classifier to predict cancer type using somatic mutation data was achieved by Chen et al. (2015). Using a large dataset of 6751 samples and approximately 21,000 genes and a support vector machine (SVM) approach, the authors achieved

moderately high accuracy of 62% in predicting cancers spanning 17 subtypes (i.e., tissues and histologies). However, the features used in their models were rather crude; each mutation was labeled according to the gene it occurred in, its mutation type (single nucleotide, indel, etc.), and the chromosome it appeared on.

Instead of considering individual mutations as features, Soh et al. (2017) considered somatic variants at the gene level. In other words, the genes in which the mutations occurred are used as predictors for building machine learning models. Using 100 genes as predictors and a SVM approach, the authors achieved an accuracy of nearly 50% in predicting 28 cancer types. In the approach of Soh et al. (2017), the number of genes is much smaller than the number of mutations, and the number of genes is a known quantity. Thus, Soh's approach dramatically decreased the dimension of the learning problem and could handle the problem of previously unseen mutations, simply by mapping a mutation to the gene it occurred in.

Both Chen and Soh's approaches are reasonable ways of approaching the cancer classification problem. Chen's approach considered mutations as features, where each mutation was encoded using to a few summary features. Soh's approach aggregated mutations within genes and considered genes as the features for the learning task. However, both of these approaches ignored variant-level information that could be critical to classifying cancer types. For example, the *KRAS* G12C variant is primarily associated with lung adenocarcinoma while *KRAS* G12R is almost exclusively associated with pancreatic cancer (Chakraborty et al. 2020). The frequencies of different *kras* mutations between different cancer types are shown in Figure 3.1, reflecting how different mutations in the same gene can encode different cancer type specific signals. This insight motivates the need to incorporate variant-level information in a prediction algorithm.

Recognizing the limitations of a gene-centric approach, recent efforts to classify cancers by examining broader patterns of mutation have achieved greater success. Salvadores et al. (2019) segmented the genome into several 1 megabase segments to capture the regional mutation density of passenger mutations. Using regional mutation density and trinucleotide context as features, their approach classified cancer primary site with 92% accuracy as compared to 36% for the models built on driver genes only. Another recent study by Jiao et al. (2020) also segmented the genome into 1 megabase segments and found that regional passenger mutation density was the most predictive feature for predicting cancer primary site. A deep learning classifier built on

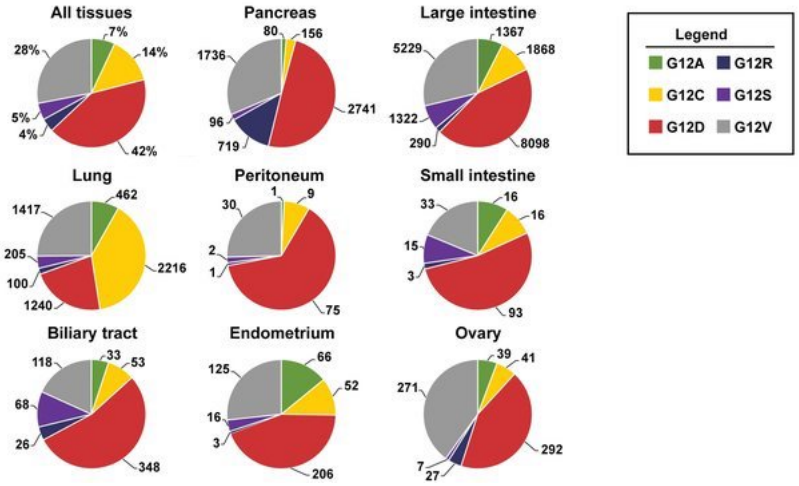


Figure 3.1: Figure from Pantsar et al. (2018). Proportions of various mutations to the *KRAS* gene across 9 cancer types reveals that different mutations show preferences for different tissues.

passenger mutation distribution and mutation type achieved classification accuracies on the order of 80-90%, twice that of trained histopathologists. It is critical to note that these studies by Salvadores and Jiao rely on *Whole Genome Sequencing*, and the efficacy of their approaches were weakened when considering only exome data. Nevertheless, improved classification using 1 Mb genomic regions indicates that breaking from a gene-centric perspective with respect to classifying cancers according to their somatic mutation profiles may improve classification importance.

A recent study by Young et al. (2020) took the approaches of Salvadores and Jiao a step further. Recognizing that the 1Mb segments were rather ad hoc, Young developed a dynamic programming approach to optimally segment the genome into regions to maximize differences in relative mutation rates across different cancer types. This "genome gerrymandering" approach achieved a 20% reduction in the number of mutations needed to discriminate between cancer types.

3.2 De-sparsifying somatic mutation data

Classifying cancers using somatic mutations is extremely challenging because the data are sparse: tumor exomes only possess a handful of mutations, and many represent passenger mutations that aren't directly involved in the progression of carcinogenesis. Typically, machine learning classifiers don't perform well on sparse data. Thus, methods to "de-sparsify" somatic mutation data in a principled manner to facilitate the learning task are useful to classification efforts in cancer.

3.2.1 Data-driven de-sparsification methods

This subsection details several data-driven de-sparsification methods for somatic mutation data that I've come across in the literature. I emphasize "Data-driven", as these methods rely on minimal biological assumptions, and instead use data-encoding tricks, feature selection strategies, or neural networks to de-sparsify their somatic mutation data.

Yuan et al. (2016) developed an algorithm for de-sparsifying somatic mutation data and enhancing cancer type prediction using a clever data-encoding trick. The DeepGene algorithm de-sparsifies somatic mutation data using a hashing function that returns the index of the nonzero entries for each feature, which compresses a large, sparse vector of mutation counts to a small, dense vector of indices. For example, a sparse $N \times 1$ vector containing mutation counts of a particular gene over N samples can be reduced to a $b \times 1$ vector, where $b << N$ represents the number of nonzero entries in the initial vector. Once these indices are generated, they are supplied (along with a few other features) to train a Deep Neural Network. Their algorithm achieved a 24% boost in accuracy as compared to standard methods like SVM, k-nearest neighbors, and Naive Bayes, illustrating the utility of this method. In short, DeepGene de-sparsifies somatic mutation data by using a clever data-encoding trick, which acts to improve classification performance.

Dikaios (2020) used the same data-encoding trick as DeepGene, but modified DeepGene's approach for *a priori* feature selection. Dikaios used a L_1 penalized neural network approach to encourage a parsimonious model that retained groups of relevant features. This approach represented a more principled approach for selecting relevant features. Indeed, his sparse input neural network approach outperformed gradient boosting, deep neural networks, and support vector machines by achieving 73% accuracy on an

independent testing dataset in classifying cancers of 32 different types. In summary, Dikaois reprised the same hashing trick used in DeepGene, but used a more principled method for selecting the sets of genes to be included in the neural network model.

Another approach by Hasan & Lonardi (2018) eschewed the encoding trick and improved upon the performance of DeepGene using a logistic regression model with an information theoretic feature selection criterion. For the purposes of feature selection, the authors first clustered genes according to the cosine similarities of their mutation vectors, then selected the most relevant genes (top N genes sorted by normalized mutual information between the feature and class label) from each cluster to train a logistic regression model. Hasan and Lonardi's approach improved upon DeepGenes' accuracy in classifying 12 cancer types by approximately 2%. In summary, Hasan and Lonardi developed a more interpretable approach that used clustering and information theory to select the best subset of features for training a logistic regression model for cancer prediction.

An entirely different approach was employed Palazzo et al. (2019), who used an autoencoder (a neural network designed to find reduced and simple representations/embeddings of complex data) to find embeddings for tumor somatic mutation profiles. Embeddings are simply low dimensional, continuous valued representations of high-dimensional, discrete data, learned using a neural network[1]. Palazzo's autoencoder approach was able to represent somatic mutation profiles in a latent space of 50 dimensions that retained cancer type specific signals. The embeddings were used to build a SVM which improved performance in classifying 35 out of 40 cancer types compared to models agnostic to these embeddings. In short, Palazzo reduced data-sparsity by training a neural network to learn efficient, low-dimensional representations of somatic mutation profiles of tumor samples.

However, the classification strategies outlined above struggle with interpretability. DeepGene's de-sparsification strategy effectively compresses the sparse mutation data, but ignores variant-level information by consider mutations at the gene-level. Dikaios' method falls into the same trap. Hasan and Londardi's method selects a set of highly relevant features (genes), likely omitting potentially genes that are sparsely mutated but relevant to carcino-

[1]Embeddings are often used in the case of linguistic data, where high dimensional matrices of word and word frequencies can be reduced to smaller, dense representations that retain semantic meaning.

genesis. Palazzo's method maps tumor mutation profiles to a latent space, which while effective for de-sparsification, lacks interpretability.

3.2.2 Biologically-motivated de-sparsification methods

An alternative to these data-driven approaches involves methods that use biological datasets as priors to inform the data de-sparsification strategy. Indeed, it is widely accepted that cancer is not a disease characterized by individual mutations or genes, but of multiple functionally-connected genes working in concert to promote excess cell proliferation (Hanahan & Weinberg 2000, Hofree et al. 2013). The following subsection details several of these "biologically-motivated" methods encountered in the literature.

Pathways and networks are similar yet related concepts. They both represent systems of interacting genes/biomolecules. Pathways are composed of small numbers of genes/biomolecules that play roles in well-studied processes. Pathways represent consensus systems supported by decades of research and are typically visualized in linear diagrams (Creixell et al. 2016). A classic example of a pathway is the electron transport chain in mitochondria: a sequence of proteins that collaborate to generate a chemical gradient that drives the synthesis of the universal energy currency of cells. Networks comprise genome/proteome-wide interactions detected through high throughput screens or integration of multiple datasets. Networks are abstractions of cellular logic, noisy, and challenging to visualize. However, networks contain information that is not covered in well-defined pathways (Creixell et al. 2016).

Hofree et al. (2013) developed a method to reduce somatic mutation data sparseness by smoothing mutations over a gene-gene interaction network. In their approach, a sample of genes and patients are projected onto a gene interaction network, smoothed, and clustered using network NMF. The procedure is repeated several times, with the consensus matrix representing the final clustering of the genes. This approach relies on the idea that a mutation in one gene implicates all other genes it interacts with, i.e., a "guilt by association". The network propagation approach de-sparsifies the mutation data by smoothing or "filling in" the mutational gaps in a gene-gene network. Their approach effectively identifies cancer subtypes predictive of survival, treatment response, and histology, and identifies network regions characteristic of each subtype.

Kuijjer et al. (2018) developed a method called SAMBAR (Subtyping Agglomerated Mutations By Annotation Relations) which summarizes muta-

tions using pathway annotation scores. The authors focused on 2219 genes either directly involved in carcinogenesis or with functional connections to such genes, and summarized the somatic mutations in 1135 canonical pathway gene signatures. They computed gene specific scores corresponding to the proportion of the total mutation rate (number of mutations per nucleotide) consumed by each gene. Then pathway mutation scores were obtained by summing the individual gene scores, correcting for the number of pathways each gene belongs to and the number of genes present in a pathway.

De-sparsification using networks and pathways each present their own benefits and challenges. Our knowledge of gene and protein interaction networks are more flexible and extensive than our knowledge of pathway maps. Also, gene-gene networks have empirically proven useful in identifying biomarkers and important subnetworks involved in carcinogenesis. However, these networks rely on a set of prior interactions that may not be relevant to the tumor or cancer type at hand. Network-based methods may miss genes that interact indirectly via a larger pathway, and exhibit ascertainment bias by overemphasizing highly connected genes or genes connected to highly mutated genes. Alternatively, pathway-based methods consider suites of genes with known relevance to cancer biology. Pathways are better characterized, provide the best snapshot of functional relationships between genes, and are unlikely to produce irrelevant interactions. However, pathways represent a much more restrictive biological prior, as the vast majority of genes (which may be relevant to carcinogenesis) do not map to important cancer pathways.

Thus, network and pathway-based de-sparsification strategies represent competing sides of the bias-variance tradeoff. Network-based analyses will exhibit *higher variance*: rich interaction networks may over-represent the number of true interactions among genes in the tumor samples in our dataset. Pathway-based approaches will exhibit *higher bias*: potentially underfitting the gene interactions in our tumor samples.

3.2.3 Why de-sparsify?

From a practical point of view, de-sparsifying somatic mutation data could enable more precise calculation of Good-Turing probabilities, as the accuracy of Good-Turing probabilities is predicated on the precision of the N_1 estimate (or the estimated number of singleton variants). Additionally, a diagnostically relevant gene group is more likely to be mutated than a diagnostically relevant gene; de-sparsifying in groups of genes can increase the likelihood of

observing a mutation in a relevant predictor. Lastly, machine learning classifiers in general do not perform well with sparse data. Grouping genes can also reduce the dimensionality of the data, potentially reducing the complexity of a machine learning model and increasing generalization performance.

De-sparsifying somatic mutation data in a biologically principled manner also presents a number of other benefits. In high-dimensional inference scenarios, aggregating mutations within groups can increase the likelihood of passing a statistical detection threshold while mitigating multiple testing. Second, grouping genes in a biologically principled manner can aid interpretation, by connecting genomic alterations to concepts like the cell cycle or apoptosis. Third, gene groupings can facilitate comparisons across other datasets by providing a common feature space. Lastly, principled groupings can facilitate integrating other data types relevant to cancer biology (Creixell et al. 2016).

3.3 De-sparsification in this thesis

This thesis will explore two different somatic mutation de-sparsification strategies with differing degrees of bias and variance.

The first method relies on few biological assumptions. The dataset will first be filtered for genes with known association to cancer to eliminate passenger genes. Good-Turing probabilities for encountering hitherto unseen mutations per cancer type will be computed on a per gene basis, and pairwise correlations between genes will be computed. Hierarchical clustering with a hybrid branch cut will be performed on the matrix of angular distances[2] between genes to identify and extract modules of genes with similar mutation probability patterns. Good-Turing probabilities will then be computed for gene clusters which will be subjected to downstream analyses.

The second method relies more on biological prior assumptions. The dataset will first be filtered for genes with known association to cancer to eliminate passenger events. Genes will be grouped according to pathway memberships, where pathways represent "gold-standard" annotations of biomolecular interactions in cells. Good-Turing probabilities will be computed on a per-pathway basis and will be subjected to downstream analyses.

These two methods will be described in detail in the following chapter.

[2]The angular distance is the dissimilarity metric, defined as the arcosine of the Pearson correlation.

Chapter 4

Methods

This chapter presents a full overview of the experimental pipeline, including description of the datasets used, de-sparsification methods used, computation of Good-Turing probabilities, and downstream analyses and visualizations.

4.1 Somatic mutation dataset

The mutation dataset used in this thesis is a publically-available dataset of non-synonymous single nucleotide somatic variant mutations in 10,295 tumor samples spanning 32 different primary site types. The dataset was made available through the R package, *variantprobs* (Chakraborty, Begg & Shen 2019). I restricted analysis to tumors with known cancer type labels and which belonged to mutation signatures with low to moderate mutation rates, which prevented confounding effects of hypermutated tumor signatures. I ultimately included 6689 Non-hypermutated tumors, 810 APOBEC (2, 13) signature tumors, 1120 Smoking (4) signature tumors, and 1050 MMR (6, 15, 20, 26) signature tumors in my dataset for analysis[1]. This produced a filtered dataset of 9669 tumor samples. A snapshot of the dataset is included below (Figure 4.1) along with a summary table illustrating the number of tumor samples per cancer type (Table 4.1).

[1]The numbers within parentheses indicate the SBS numbers belonging to each dominant signature group

Cancer Code	Disease	Number of tumors
ACC	Adrenocortical carcinoma	92
BLCA	Bladder urothelial Carcinoma	403
BRCA	Breast invasive carcinoma	1013
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	289
CHOL	Cholangiocarcinoma	34
COADREAD	Colorectal adenocarcinoma	543
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	37
ESCA	Esophageal carcinoma	184
GBM	Glioblastoma multiforme	393
HNSC	Head and Neck squamous cell carcinoma	499
KICH	Kidney Chromophobe	64
KIRC	Kidney renal clear cell carcinoma	368
KIRP	Kidney renal papillary cell carcinoma	278
LAML	Acute Myeloid Leukemia	136
LGG	Brain Lower Grade Glioma	522
LIHC	Liver hepatocellular carcinoma	363
LUAD	Lung adenocarcinoma	564
LUSC	Lung squamous cell carcinoma	480
MESO	Mesothelioma	80
OV	Ovarian serous cystadenocarcinoma	408
PAAD	Pancreatic adenocarcinoma	174
PCPG	Pheochromocytoma and Paraganglioma	177
PRAD	Prostate adenocarcinoma	491
SARC	Sarcoma	228
SKCM	Skin Cutaneous Melanoma	67
STAD	Stomach adenocarcinoma	429
TGCT	Testicular Germ Cell Tumors	149
THCA	Thyroid carcinoma	488
THYM	Thymoma	122
UCEC	Uterine Corpus Endometrial Carcinoma	466
UCS	Uterine Carcinosarcoma	49
UVM	Uveal Melanoma	79

Table 4.1: Cancer categories and cohort sizes in filtered TCGA dataset.

	patient_id	Hugo_Symbol	Variant	Cancer_Code	MS
1:	TCGA-02-0003-01A	TACC2	TACC2__10__123810032__C__T	GBM	Non-hypermutated
2:	TCGA-02-0003-01A	PANX3	PANX3__11__124489539__G__A	GBM	Non-hypermutated
3:	TCGA-02-0003-01A	SPI1	SPI1__11__47380512__G__T	GBM	Non-hypermutated
4:	TCGA-02-0003-01A	NAALAD2	NAALAD2__11__89868837__C__T	GBM	Non-hypermutated
5:	TCGA-02-0003-01A	MTERFD3	MTERFD3__12__107371855__A__G	GBM	Non-hypermutated
---					
1991484:	TCGA-ZX-AA5X-01A	SMC5	SMC5__9__72914973__C__T	CESC	Non-hypermutated
1991485:	TCGA-ZX-AA5X-01A	SMC5	SMC5__9__72914979__G__T	CESC	Non-hypermutated
1991486:	TCGA-ZX-AA5X-01A	TRPM3	TRPM3__9__73477937__G__A	CESC	Non-hypermutated
1991487:	TCGA-ZX-AA5X-01A	THOC2	THOC2__X__122829875__T__G	CESC	Non-hypermutated
1991488:	TCGA-ZX-AA5X-01A	SRPX2	SRPX2__X__99925883__G__A	CESC	Non-hypermutated

Figure 4.1: Snapshot of TCGA dataset. "patient_id" denotes the unique tumor sample identifier, "Hugo_Symbol" denotes the gene name, "Variant" denotes the unique variant identifier (gene, chromosome, index, and nucleotide change), "Cancer_Code" denotes the cancer type, and "MS" denotes the mutation signature label.

4.2 Implementing Good-Turing probability estimation

Good-Turing probabilities were calculated per feature (either gene or gene group) and per cancer type using the "goodturing_probs" function available through the R package, *variantprobs* (Chakraborty, Begg & Shen 2019). The function takes as input a vector of v_f, denoting how frequently variants are observed in the training set for a particular feature (gene or gene group) and cancer type. "goodturing_probs" also takes as input m, denoting the number of tumor samples in that particular cancer type. For example, a $v_f = (1, 4, 7)$ indicates that the first variant appeared once in the training set, the second variant appeared four times, and the third variant appeared seven times across the training samples.

The function outputs probabilities of encountering variants at various values or r, including an estimate of the probability of encountering *at least one previously unseen variant* in a future tumor sample of that particular cancer type (as calculated via Equation 2.11). The following work considers the Good-Turing probabilities formula as:

$$\hat{\pi}_0 \approx 1 - \exp\left[-\frac{N_1}{m+1}\right] \tag{2.11}$$

Note that the probability estimate ($\hat{\pi}_0$) described in equation 2.11 corresponds to the probability of observing **at least one** hitherto unseen variant

in a future tumor sample. In other words, equation 2.11 refers to the *total probability* occupied by the unseen variants in the sample space. For the remainder of this thesis, probabilities of encountering at least one hitherto unseen variant in a randomly selected future tumor sample will be generally referred to Good-Turing probabilities.

4.2.1 Validating Good-Turing probabilities

In an effort to validate the Good-Turing probabilities generated using each de-sparsification method, I conducted a 50-50 train-test dataset split of the 16 most common cancer types: BLCA, BRCA, COADREAD, GBM, HNSC, KIRC, KIRP, LGG, LIHC, LUAD, LUSC, OV, PRAD, STAD, THCA, UCEC. I focused on the most common cancer types so that there would be sufficient sample sizes in both training and testing datasets. The 7708 tumors were split into 2 datasets containing 3854 tumors each.

For the training dataset, I ran the specific de-sparsification procedure and calculated Good-Turing probability estimates per module[2]/pathway and per cancer type. For the test dataset, I measured the observed proportions of tumors of each cancer type that generated a hitherto unseen variant per module/pathway. For example, if out of 100 tumors of cancer type A in the test set, 50 contained mutations in Module 1 that were not observed in the cancer type A samples in the training dataset, then the observed proportion of tumors that produced a hitherto unseen variant in Module 1 would be $\frac{1}{2}$.

I measured the concordance of Good-Turing estimates and observed proportions of previously unseen variants using Lin's concordance correlation coefficient (CCC) (Lin 1989). Lin's CCC measures the agreement between an estimate (in this case, Good-Turing probabilities estimated from the training set) and a gold-standard measurement (in this case, the observed probabilities of encountering previously unseen mutations in the test set). A Lin's CCC close to 1 would indicate that the de-sparsification and estimation procedure is generating reliable estimates of hitherto unseen variant probabilities that agree with observed proportions of previously unseen mutations.

[2] "Module" refers to a cluster of genes defined using the hierarchical clustering de-sparsification procedure.

4.3 De-sparsification methods

This section details the two somatic mutation data desparsification strategies explored in this thesis.

4.3.1 Data-driven strategy: hierarchical clustering of mutation probability correlations (HCOMPC)

This de-sparsification strategy represents a lower bias/higher variance approach. It relies on no biological prior knowledge, operating by clustering genes according to the correlations between their mutation probabilities over cancer types.

A priori filter for putative cancer genes

To eliminate passenger genes which encode noisy (i.e., non-cancer-type-specific) mutation events, only 2352 cancer-associated genes from the Catalogue of Somatic Mutations in Cancer (COSMIC) (Tate et al. 2019) and Supplemental Table 3 from Östlund et al. (2009) were subject to analysis. This gene list was available through the *SAMBAR* R package (Kuijjer 2021).

Computation of matrix of pairwise distances between genes

HCOMPC works by first calculating Good-Turing probabilities for each gene of interest across each cancer type. This creates a $n \times 32$ matrix containing the Good-Turing probabilities for n genes across 32 cancer types. For example, matrix entry $x_{i,j}$ represents the probability of observing a hitherto unseen mutation in Gene i in cancer type j. Then the Pearson correlation is calculated between probability vectors for each pair of genes; this produces a $n \times n$ matrix M_S of *probability correlations*. Since hierarchical clustering algorithms in R require a dissimilarity matrix, I convert the matrix of correlations (a similarity measure) to a matrix of angular distances using equation 4.1:

$$d_{\text{Cor}}(x_{i,\cdot}, x_{j,\cdot}) = \arccos(\text{Cor}(x_{i,\cdot}, x_{j,\cdot})) \tag{4.1}$$

Angular distance has the desired property of forcing highly positively correlated genes to have a d_{Cor} close to 0, while making uncorrelated and

negatively correlated genes have a higher d_{Cor}. For uncorrelated genes $d_{\text{Cor}} = \frac{\pi}{2}$.

Applying equation 4.1 to M_S, the matrix of correlations between Good-Turing probabilities for genes, produces a matrix of angular distances, M_D. M_D can now be supplied to the hierarchical clustering algorithm *hclust* in R, which creates a dendrogram delineating the distances between the Good-Turing probability profiles of different genes.

Identification of gene clusters using dynamic hybrid cut

In hierarchical clustering, modules/clusters are represented as branches of a dendrogram/tree. The most common approach for defining clusters in a dendrogram is choosing a *fixed cut height* and continuous branches of objects produced by the cut are considered distinct clusters. However, the fixed height cut approach is heuristic and doesn't perform well in complicated dendrogram structures (especially with nested clusters).

To obtain more accurate representation of clusters in a dendrogram, there exist tree cut algorithms that instead consider *branch shape*. One such method that I use in this thesis is an agglomerative (bottom up) approach called "Dynamic hybrid" (Langfelder et al. 2008) to generate clusters.

The first step in Dynamic hybrid is cluster detection. Preliminary clusters are defined by the following criteria:

1. Clusters must exceed a minimum threshold size.
2. The tip of each branch (the core) should be tightly connected.
3. Each cluster must de distinct from its surroundings.
4. Objects that are too far from the clusters are excluded.

In order for a branch to pass the first cluster criterion, the branch must contain more than N objects, where N is a user-specified parameter indicating the minimum cluster size. In order to pass the second cluster criterion, we introduce the *core scatter* statistic, $\bar{d}$, which is the average of the pairwise dissimilarities between objects in a core. In order to be a cluster, the core scatter $\bar{d} < d_{max}$, a user-specified parameter indicating the maximum permissible core scatter. In order to pass the third cluster criterion, we introduce the *cluster gap* statistic, g, which represents the difference between $\bar{d}$ and the joining height where the proposed cluster attaches to the rest of the

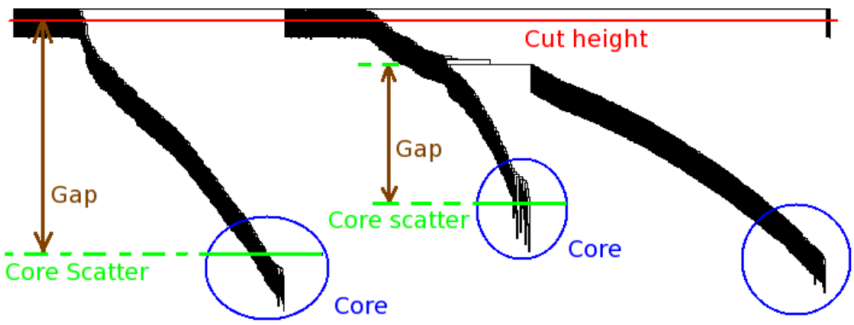


Figure 4.2: Figure from Langfelder et al. (2009). Figure illustrating cluster cores (blue), the core scatter (average dissimilarity amongst the core, green), the gap (difference between scatter and join height to rest of dendrogram, brown), and the max cut height (red).

dendrogram. In order to be designated a cluster, $g < g_{max}$, where g_{max} is a user-specified parameter indicating the maximum permissible gap. In order to pass the fourth criterion, all joining heights must be at most h_{max}, which is a user-specified parameter denoting the maximal join height (Langfelder et al. 2009). For a visual interpretation of cores on an example dendrogram, gap, core scatters, and max cut height, see Figure 4.2.

There exists no clear method for choosing "optimal" parameter values, although the authors include an argument, *deepSplit*, which allows users to choose from sets of parameters that encourage different sensitivities of cluster splitting. I heuristically chose a minimum cluster size of 3 and varied the value of *deepSplit* between 0 and 3, providing a range of sensitivity to cluster splitting. Varying *deepSplit* generated sets of clusters that varied widely in terms of size.

The algorithm steps through the dendrogram from the bottom up, merging either two objects, merging an object to a branch, or merging two branches. Merging two objects generates a new branch. When merging two branches, if at most one branch passes the cluster criteria, the branches are merged. If both branches pass the criteria, the algorithm passes to the next step (Langfelder et al. 2009).

The second optional step in the algorithm involves the assignment of unassigned objects to existing clusters. The average dissimilarity of an object (single object or tiny branch) to existing clusters is computed, and

then the object is merged to its nearest cluster. Since Partitioning Around Medoids (PAM) clustering involves assigning objects to their nearest medoid, the merging of unassigned objects represents a hybrid of hierarchical clustering and PAM (Langfelder et al. 2009). The authors note that the decision to include the PAM step should be determined by whether sensitivity or specificity is favored. I prioritize specificity, i.e., high confidence that genes in the same cluster belong together, so I excluded the PAM assignment step in my runs of Dynamic Hybrid. This resulted in several genes being unassigned to clusters.

Aggregating mutation data among clusters and Good-Turing probability estimation

Once clusters of genes are identified, I will **aggregate or sum** the mutation data among clusters. For instance, if Genes i and j are in the same cluster, the aggregated mutation data for their cluster would be represented by $x_{i,\cdot} + x_{j,\cdot}$. After aggregating mutations for each cluster, I will recalculate the Good-Turing probabilities as outlined in the previous section. Instead of treating genes as the genomic units of interest, HCOMPC considers gene clusters as the genomic units of interest.

4.3.2 Biologically-driven strategy: pathway membership based grouping (PMBG)

The PMBG de-sparsification strategy represents a higher bias/lower variance approach. It considers mutations aggregated/summed within known biological pathways.

A priori filter for putative cancer genes

To eliminate passenger genes which encode noisy (i.e., non-cancer-type-specific) mutation events, only 2352 cancer-associated genes from the Catalogue of Somatic Mutations in Cancer (COSMIC) (Tate et al. 2019) and Supplemental Table 3 from Östlund et al. (2009) were subject to analysis. This gene list was available through the *SAMBAR* R package (Kuijjer 2021).

Pathways dataset

PMBG works by aggregating (summing) somatic mutation data within pathways. Pathways represent groups of genes that share direct, functional interactions that determine a biologically relevant process. Pathways are often defined through years of rigorous experimentation on the part of domain experts and represent the gold standard of biomolecular interactions. While quite complex, knowledge of pathways are incomplete, and current annotated pathways probably underestimate the true complexity of biomolecular interactions within cells.

Gene set files containing pathways were acquired from MSigDb (Subramanian et al. 2005) in .gmt format[3]. The gene set file used for PMBG was the complete Canonical Pathways gene set ("c2.cp.v7.3.symbols.gmt"), containing 2887 canonical pathways derived from several pathway databases including BioCarta, KEGG, Matrisome Project, Pathway Interaction Database, Reactome, SigmaAldrich, Signaling Gateway, SuperArray SABiosciences, and Wikipathways. The gene set file were converted to binary adjacency matrices A where $A_{i,j} = 1$ indicates the presence of Gene j in Pathway i using the *SAMBAR* R package (Kuijjer 2021).

Aggregating mutation data among pathways and Good-Turing probability estimation

Pathways that contained fewer than 3 genes were eliminated. Mutation counts were **aggregated/summed** identically to the HCOMPC procedure, except instead of summing mutations per cluster, mutations are summed within known pathways. Then Good-Turing probabilities were calculated per pathway and per cancer type according to the procedure outlined in Section 4.2. Instead of treating genes as the genomic units of interest (per Chakraborty, Arora, Begg & Shen (2019*b*)), or clusters of genes as the genomic units of interest (per HCOMPC), PMBG considers biological pathways defined by domain experts as the units of interest.

[3].gmt format refers to gene matrix transposed format, often used to describe gene groups.

4.4 Normalized Mutual Information

Normalized mutual information (NMI) is the method used to rigorously measure the association between Good-Turing probabilities and cancer types in this thesis. In other words, NMI is a measure of the **cancer type specificity** of the hitherto unseen variant probabilities.

4.4.1 Motivation

The following section is adapted from Chakraborty, Arora, Begg & Shen (2019*a*)

In general, Normalized mutual information (NMI) measures the nonlinear dependence between two random variables. In the context of this thesis, NMI quantifies the association between the probability of occurrence of a particular variant and cancer type. NMI is calculated according to equation 4.2 below, where y_j denotes the presence (1) or absence (0) of the j-th variant and let C denote the cancer type associated with the tumor $C = 1, \ldots, 32$.

$$NMI(y_j, C) = \frac{MI(y_j, C)}{\sqrt{H(y_j)H(C)}} \tag{4.2}$$

and $MI(y_j, C)$ is the mutual information between y_j and C:

$$\begin{aligned} MI(y_j, C) &= \sum_{y=0}^{1}\sum_{k=1}^{K} P(y_j = y, C = k) \log \frac{P(y_j = y, C = k)}{P(y_j = y)P(C = k)} \\ &= \sum_{y=0}^{1}\sum_{k=1}^{K} P(y_j = y|C = k)P(C = k) \log \frac{P(y_j = y|C = k)}{\sum_{i=1}^{K} P(y_j = y|C = i)} \end{aligned} \tag{4.3}$$

and $H(y_j)$ and $H(C)$ are the Shannon entropies of y_j and C:

$$\begin{aligned} H(y_j) &= \sum_{y=0}^{1} \log P(y_j = y)P(y_j = y) \\ &= \sum_{y=0}^{1}\sum_{k=1}^{K} \log \left(\sum_{k=1}^{K} P(y_j = y|c = k)P(c = k) \right) P(y_j = y|C = k)P(C = k) \end{aligned} \tag{4.4}$$

and

$$H(c) = \sum_{k=1}^{K} \log P(c = k) P(c = k) \tag{4.5}$$

In the above formulas, the cancer type specific variant probabilities $P(y_j = y|c = k)$ are calculated using the Good-Turing estimation approach, and the cancer type probabilities $P(C = k)$ are estimated by the proportion $\frac{m_k}{\sum_{i=1}^{K} m_i}$ where m_k denotes the number of tumors in the dataset of type k. For more information on how to explicitly calculate these values, please see Chakraborty, Arora, Begg & Shen (2019*a*).

It can be useful to consider normalized mutual information as the nonlinear kin to the Pearson correlation. Indeed, the Pearson correlation is a normalized version of the covariance, scaled by the variance to encompass values in the range $[-1, 1]$. However, Pearson correlation can only measure *linear* association between two random variables. Normalized mutual information is capable of measuring *nonlinear* association between two random variables. As shown in equation 4.2, the NMI is equal to MI scaled by square root of the product of the marginal Shannon entropies (a measure of uncertainty in a random variable, akin to the variance). Normalized mutual information takes on values in the range $[0, 1]$, with 0 denoting independence and 1 indicating determinance.

NMI values were calculated using the "calc_minfo" function available in the R package *variantprobs* (Chakraborty, Begg & Shen 2019).

4.4.2 Generating null distribution of NMIs

It is of great interest to not only quantify the cancer type specificity of unseen variant probabilities, but also determine if the observed cancer type specificities exceed thresholds that could be explained by random chance. To do this, I employed a permutation testing approach which generates a set of NMI reference values under the null hypothesis that there is no association between cancer type and variant occurrence. I impose this null hypothesis by randomly permuting the cancer label for all tumor samples in the dataset before calculating Good-Turing probabilities and NMI values. Note that this approach ensures that the number of cancers of each type remain constant.

In this thesis, I generate a reference distribution of NMIs under 1000 cancer label permutations, and compare the 99th quantile of the null dis-

tribituion of NMIs to the observed values. Any observations that exceed this value are considered significantly cancer type specific with 99% confidence.

4.5 Visualization

All visualizations were constructed in R using the *ggplot* R plotting frameworks (Wickham 2016) the "heatmap" function through the R package *stats* (R Core Team 2020).

Chapter 5

Results

5.1 De-sparsification by hierarchical clustering of mutation probability correlations (HCOMPC)

The goal of the following section is to evaluate the performance of the HCOMPC de-sparsification algorithm in generating clusters with significantly cancer type specific and reproducible Good-Turing probabilities.

5.1.1 Hierarchical clustering, dynamic hybrid pruning, and module detection

In order to de-sparsify the somatic mutation data, genes need to be grouped into clusters in a principled manner. This subsection details the results of hierarchically clustering genes' hitherto unseen mutation probabilities according to a correlation distance.

Hierarchical clustering of the pairwise angular distances (as calculated via equation 4.1) between Good-Turing probabilities of the 2352 cancer-associated genes was performed. The resulting dendrogram was pruned using the dynamic hybrid cut method (Langfelder et al. 2008) with a minimum cluster size set to 3 genes and varying values of the *deepSplit* parameter which modulates the sensitivity to cluster splitting. Summary statistics including the total number of clusters, the maximum module size, the median module size, and the number of genes unassigned to a cluster are shown in Table 5.1.

deepSplit	# of modules	Med module size	Max module size	# unassigned
0	196	9	71	40
1	286	6	44	40
2	401	5	21	78
3	485	4	21	80

Table 5.1: Summary statistics for different runs of the Dynamic Hybrid pruning approach while varying values of the *deepSplit* parameter. Summary statistics illustrate that as *deepSplit* increases, sensitivity to branch splitting increases.

As anticipated, increasing the *deepSplit* parameter increased the number of modules and decreased module sizes, demonstrating a greater sensitivity to cluster splitting.

Visualization of the different pruning runs on the dendrogram is shown in Figure 5.1. The different color bars below the dendrogram denote the different pruning runs while varying the *deepSplit* (ds) parameter value. The colors in the color bars denote module membership. For the remainder of this section, clusters of genes will be referred to as **modules**.

5.1.2 Good-Turing probabilities per module

Once modules were detected using hierarchical clustering with hybrid branch pruning, probabilities of encountering hitherto unseen variants in a randomly selected future tumor sample were estimated for each module. Modules that were uncorrelated with background mutation rates were identified. This subsection details the output of Good-Turing probability calculations of modules identified using the HCOMPC approach and details the identification of modules uncorrelated with background mutation rate patterns.

After module detection, Good-Turing probabilities were calculated on a per module and per cancer type basis, by summing the mutation frequencies per cancer type for all genes in each module. Good-Turing probabilities were visualized on the heatmap shown in the right half of Figure 5.2. Examining the overarching probability patterns shows red bands (indicative of higher probabilities) across some cancer types: BLCA, CESC, COAD-READ, LUAD, LUSC, STAD, and UCEC. The lack of diversity in probability patterns among modules suggests that Good-Turing probabilities may be tracking patterns in background mutational rate, or total mutational bur-

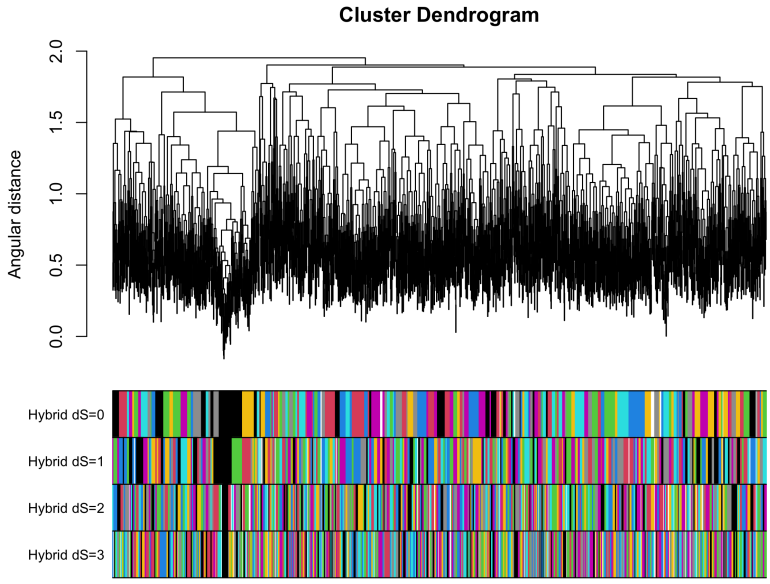


Figure 5.1: Dendrogram illustrating the angular distances between Good-Turing probabilities of different genes. Modules were determined from the dendrogram using dynamic hybrid branch pruning approach and varying the value of the deepSplit ('ds') parameter from 0 (least sensitive cuts) to 3 (most sensitive cuts). The color bars at the bottom denotes the module membership of each gene under the different runs of the pruning approach.

den (TMB), among different cancers. In other words, in many modules, the Good-Turing probabilities may be simply measuring differences in background mutation rates between different cancer types.

To determine if modules were tracking patterns in background mutation rate, I measured the Pearson correlation between each module's Good-Turing probabilities and a vector of average mutational burden (i.e., average number of mutations per tumor) per cancer type and visualized them in a histogram shown in the left half of Figure 5.2. The left-skewed distribution indicated that the majority of modules had Good-Turing probabilities that were highly correlated with patterns in background mutation rates (the peak of the histogram is at approximately 0.75). However, there were some modules that are moderately correlated, weakly correlated, and uncorrelated with patterns in background mutation rate.

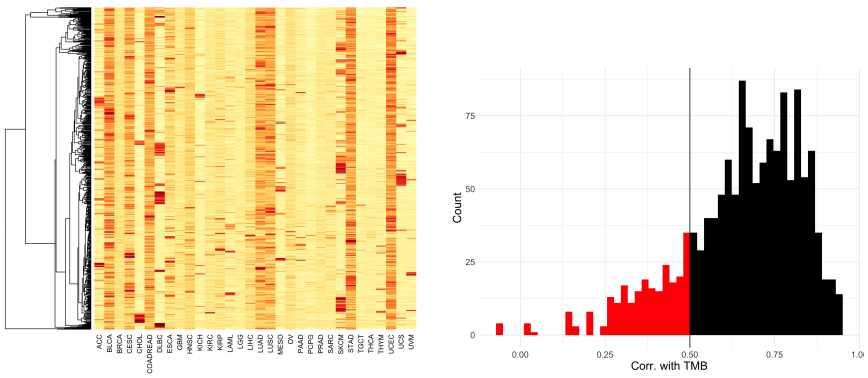


Figure 5.2: Left: Heatmap of Good-Turing probabilities for modules defined using hierarchical clustering of the gene-specific probabilities. Red tiles indicate high probabilities while yellow tiles indicate low probabilities. Red banding illustrates that many Good-Turing probabilities per modules may be tracking patterns in background mutation rates. Right: Histogram of correlations between Good-Turing probabilities for each module with the vector of average background mutation rate per tumor. The left-skewed distribution shows the majority of modules are tracking patterns in background mutation rate. The bars to the left of the vertical line, colored red, have correlation < 0.5 with average total mutational burden patterns.

To filter out modules that were highly correlated with average total mutational burden patterns, I used an strict *ad hoc* correlation threshold of 0.5. All modules that showed a correlation with average total mutational burden < 0.5 were designated as sufficiently uncorrelated with background mutation rates. 245 modules had correlations under the threshold. Next, the 245 modules below the correlation threshold were subjected to simulations to assess the cancer type specificity of their unseen variant probabilities.

5.1.3 Assessing cancer type specificity of modules using simulation

A module is relevant to clinical tasks like primary site classification if its mutation rates vary significantly between cancer types. In this subsection, the cancer type specificities of the hitherto unseen variant probabilities of modules are measured using Normalized Mutual Information (NMI) and tested for significance using a simulation approach.

Figure 5.3: Left: distribution of null (blue) and 245 observed (red) NMI values for modules generated using the HCOMPC procedure (uncorrelated with background mutation rates). The dotted black line indicates the 99th quantile value on the null distribution. The relative right shift in the red distribution illustrates that the observed modules are more cancer type specific than those values generated by random chance. Right: Bubbleplot illustrating Good-Turing probabilities over different cancer types for modules uncorrelated with average Total Mutational Burden (TMB). Bubbles are sized according to probability of encountering a hitherto unseen mutation in a future randomly selected tumor sample of that cancer type.

To test if Good-Turing probabilities calculated per module and cancer type were significantly cancer type specific, I generated a null distribution of Normalized Mutual Information (NMI) values under the assumption of no association between variants and cancer type according to the procedure described in Methods Subsection 4.4.2. Comparing the observed NMIs for the 245 modules sufficiently uncorrelated with background mutation rates to the null distribution of NMI values, I identified that all modules exceeded the 99th quantile of the null NMIs (Right panel of Figure 5.3). Thus, I conclude with 99% confidence that the observed cancer type specificities in Good-Turing probabilities were inconsistent with those expected under random chance.

5.1.4 Exploring patterns of hitherto unseen variant probabilities

The previous subsection illustrated that the Good-Turing probabilities of modules constructed using the HCOMPC procedure were significantly cancer type specific. This subsection explores the patterns of unseen mutation probabilities between modules identified by HCOMPC.

Focusing specifically on 245 modules sufficiently uncorrelated with background mutation rates, I plotted Good-Turing probabilities per module and per cancer type. I examined the Good-Turing probability values, and selected 15 modules with high NMI values that displayed different cancer type specific patterns and displayed them in the bubbleplot shown in the right panel of Figure 5.3. Some clear patterns emerge from this graphic. For instance, "Module 197" (fourth from top) shows a high propensity for encountering hitherto unseen mutations in the DLBC cancer type (Large B cell Lymphoma). "Module 491" (second from top) shows a high propensity for producing hitherto unseen variants in the UCS cancer type (Uterine Carcinosarcoma). "Module 77" (fourth from bottom) shows a high propensity for generating hitherto unseen variants in the MESO cancer type (Mesothelioma). And "Module 21" (bottom) shows a high probability of generating hitherto unseen mutations in SKCM cancer type (Skin Cutaneous Melanoma). This statistically significant variability in hitherto unseen variant probabilities between cancer types suggests that these modules may encode clinically relevant signals.

5.1.5 Assessing reproducibility of Good-Turing probability estimates

Probability estimates of encountering hitherto unseen variants must be reliable to warrant consideration for important clinical tasks. This subsection assesses the reproducibility of Good-Turing probability estimates for HCOMPC-defined modules using a train-test set split.

In an attempt to validate Good-Turing probabilities generated through the HCOMPC algorithm, I conducted a train-test split procedure as described in Subsection 4.2.1. I focused on the top 16 most common cancer types (BLCA, BRCA, COADREAD, GBM, HNSC, KIRC, KIRP, LGG, LIHC, LUAD, LUSC, OV, PRAD, STAD, THCA, UCEC), which accounted for 7708 tumor samples. I split this dataset into a training dataset of 3854 tumors and a testing dataset of equal size. For the training dataset, I conducted

the HCOMPC procedure, generating Good-Turing probability estimates for 743 unique modules identified through the HCOMPC procedure (with *deepSplit* values ranging between 0 and 3). For the test dataset, I calculated the observed relative frequency of tumors samples that presented at least one hitherto unseen mutation (not seen in the training dataset). The concordance between the estimates and observed frequencies of hitherto unseen variants were illustrated in Figure 5.4. The x-axis denotes the Good-Turing probability estimates per module calculated per cancer type, while the y-axis denotes the observed proportions of each cancer type in the test dataset that presented at least one hitherto unseen variant per module. The 45° line indicates perfect prediction, and Lin's concordance correlation measures how strongly the data fit the 45° line. The Lin's coefficient of 0.905 indicated that the observed and estimated probabilities were highly concordant, illustrating that the Good-Turing estimates for these HCOMPC-defined modules are reproducible for cancer types with reasonably large sample sizes.

For large probabilities, there appears to be a slight overestimate of unseen variant probabilities by the Good-Turing procedure (as evidenced by more points under the 45° line). This bias is also evident in varying degrees according to cancer type (particularly for UCEC), which may be due to limited sample sizes or heterogeneous singleton mutation counts within cancer types. For a faceted version of the plot with Lin's correlations calculated per cancer type, I refer the reader to the Supplementary Figures section, Figure 7.1.

In summary, when the HCOMPC procedure was applied to known cancer genes to define modules with similar mutation patterns, these resulting modules displayed cancer type specific patterns in probabilities of hitherto unseen variants. Many modules showed Good-Turing probabilities patterns that merely tracked patterns in background mutation rates, but others only showed moderate or weak correlations to these prevailing mutation patterns. Good-Turing probabilities generated using the HCOMPC procedure were significantly cancer type specific as assessed by comparison to a null distribution of Normalized Mutual Information (NMI) values under the assumption of no variant-cancer type relationship. Additionally, Good-Turing probabilities of modules defined using HCOMPC were highly reproducible between training and test datasets for cancer types with reasonably large sample sizes. In other words, modules of cancer genes defined by HCOMPC contained cancer type specific signals that may be relevant to cancer classification tasks.

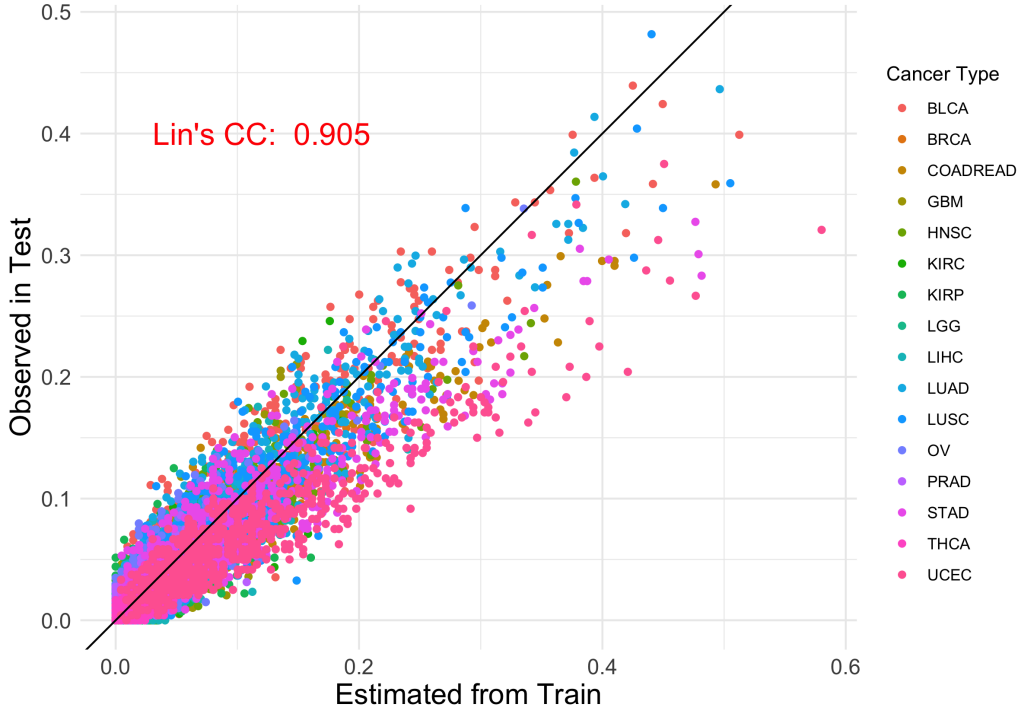


Figure 5.4: Assessing the concordance between Good-Turing probability estimates and observed proportions of unseen mutations for modules defined using the HCOMPC procedure. The x-axis denotes the estimated Good-Turing probabilities per module per cancer type in a training dataset of 3854 tumors. The y-axis denotes the observed probabilities of encountering a hitherto unseen variant per module per cancer type in a held-out testing dataset of 3854 tumors. The 45° line denotes perfect estimation. The Lin's Concordance Correlation (Lin's CC) measures the agreement between the estimated probabilities and observed proportions. Points are colored based on cancer type.

5.2 De-sparsification by pathway membership based grouping (PMBG)

The goal of the following section is to evaluate the performance of the PMBG de-sparsification algorithm in producing pathways with significantly cancer type specific and reproducible unseen variant probability estimates.

5.2.1 Grouping genes into biological pathways

This subsection details the construction of the biological pathways and the aggregation of mutation data within pathways. Summary statistics detailing pathway sizes are also provided.

2352 known cancer associated genes were assigned to 2922 known biological pathways. Mutation data were aggregated by summing the mutation counts among genes in the same pathway.

Summary statistics for pathways are shown in Table 5.2. Pathways were eliminated that contained 3 genes or less. PMBG yielded 1915 pathways that contained more than 3 genes. The median pathway size was 10 genes, and the maximum size was 332 genes. 523 genes out of the 2352 considered were not assigned to any pathway and were therefore excluded from analysis.

# pathways	Med. path. size	Max path. size	# unassigned
1915	10	332	523

Table 5.2: Summary statistics for pathways derived from the PMBG procedure. Pathways were were summarized according to the number of pathways, the median pathway size, the maximum pathway size, and the number of genes that were not assigned to any pathway.

In short, grouping genes according to pathway membership produced a large number of pathways that range in size from small (4 genes) to large (up to 332 genes). The majority of known cancer genes were assigned to biological pathways.

5.2.2 Good-Turing probabilities per pathway

This subsection illustrates the results of computed Good-Turing probabilities per pathway. High level patterns in hitherto unseen variant probabilities are examined and pathways uncorrelated with background mutation rates are identified.

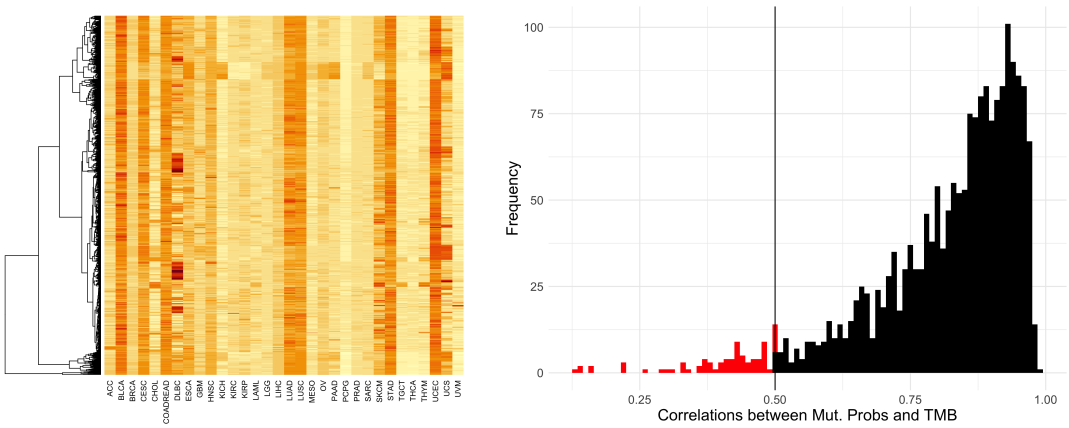


Figure 5.5: Left: heatmap of Good-Turing probabilities of 1915 pathways (rows) over 32 cancer types (columns). The darker/redder a tile is, the higher the Good-Turing probability in that particular pathway and cancer type. The dendrogram along the rows is determined by hierarchical clustering of the mutation profiles according to Euclidean distance. Right: histogram of correlations between pathway-derived Good-Turing probabilities and total mutational burden patterns. The left-skew indicates that the majority of pathways are highly correlated with the total mutational burden pattern. The vertical line denotes the 0.5 correlation threshold.

Probabilities of observing at least one hitherto unseen variant were calculated per pathway according to equation 2.11, and visualized in the heatmap shown in the left panel of Figure 5.5 to examine broader patterns in unseen mutational probabilities. The heatmap illustrated highly consistent patterns in hitherto unseen mutation probabilities among the different pathways. Indeed, the cancer types BLCA, CESC, COADREAD, LUAD, LUSC, STAD, and UCEC consistently show higher unseen variant probabilities than the other cancer types. These cancer types also typically have the highest frequencies of mutation. The consistent behavior of Good-Turing probabilities

in pathways suggests that the pathways may be tracking patterns in background mutation rate.

To confirm whether pathways were tracking patterns in background mutation rates, I measured the Pearson correlations between each pathway's Good-Turing probabilities and a vector of the average total mutational burden per tumor across the 32 tumor types. A histogram of the correlations is shown in the right panel of Figure 5.5. The left-skewed distribution confirmed that the vast majority of pathways show Good-Turing probabilities highly correlated to patterns of background mutation rate. However, not all pathways were highly correlated with background mutation rate. Using an *ad hoc* correlation cutoff of 0.5, I identified 80 pathways that were sufficiently uncorrelated with background mutation rates. In the following subsection, the cancer-type specificity of the 80 pathways were tested for statistical significance.

5.2.3 Assessing cancer type specificity of pathways obtained using PMBG procedure

This subsection details how the cancer-type-specificity of Good-Turing probabilities were tested for statistical significance using simulation.

To measure the statistical significance of the cancer type specificities observed in the Good-Turing probabilities per pathway, I generated a null distribution of Normalized Mutual Information (NMI) values under the assumption of no association between variants and cancer type according to the procedure described in Methods Subsection 4.4.2. I summarized the distribution of null NMIs along with the observed NMIs calculated from the un-permuted data in a relative frequency histogram shown in the right panel of Figure 5.6. The dotted black line represents the 99th quantile of the null distribution of NMIs. All 80 pathways sufficiently uncorrelated with background mutation rate patterns exceeded this 99th quantile cutoff. Thus, I conclude with 99% confidence, that the observed cancer-type-specificities in Good-Turing probabilities were statistically significant for the 80 pathways.

In short, all 80 pathways (sufficiently uncorrelated with background mutation rates) had cancer type specific unseen variant signals that were incompatible with a null model of no association between variants and cancer type. Thus, the differences in unseen variant probabilities between cancer types is statistically significant for the 80 pathways surveyed.

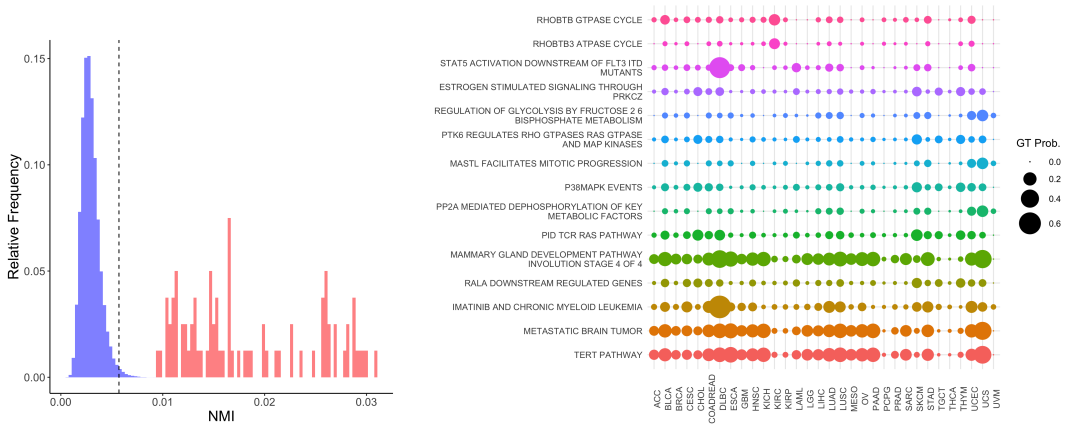


Figure 5.6: Left: distribution of null (blue) and 80 observed (red) NMI values for pathways produced by the PMBG procedure (uncorrelated with background mutation rates). The dotted black line indicates the 99th quantile value on the null distribution. The right shift in the red distribution illustrates that the observed pathways are more cancer type specific than those values generated by random chance. Right: Bubbleplot illustrating Good-Turing probabilities over different cancer types for 15 select pathways uncorrelated with average Total Mutational Burden (TMB). Bubbles are sized according to probability of encountering a hitherto unseen mutation in a future randomly selected tumor sample of that cancer type.

5.2.4 Exploring patterns of hitherto unseen variant probabilities in pathways

The previous subsection illustrated that the Good-Turing probabilities of pathways identified using the PMBG procedure were significantly cancer-type-specific. This subsection explores the patterns of unseen mutation probabilities between pathways identified by PMBG.

Focusing specifically on 80 pathways sufficiently uncorrelated with background mutation rates, I plotted Good-Turing probabilities per pathway and per cancer type. I examined the Good-Turing probability values, and selected 15 pathways that displayed different cancer type specific patterns and displayed them in the bubbleplot shown in the right panel of Figure 5.6.

The bubbleplot illustrates some interesting patterns. For instance, "TERT PATHWAY" (bottom) refers to the pathway containing the hTERT gene

which encodes the enzyme telomerase. Telomerase is found to be activated in 80-90% of cancers and can contribute to cancer immortality by repairing frayed ends of chromosomes. The TERT pathway shows high probabilities of generating hitherto unseen mutations in DLBC (B-cell Lymphoma), ESCA (Esophageal Carcinoma), KICH (Kidney Chromophobe), LUSC (Lung Squamous Cell Carcinoma), and especially UCS (Uterine Carcinosarcoma). The "IMATINIB AND CHRONIC MYELOID LEUKEMIA" pathway (third from bottom) is associated with disregulation and oncogenesis of the myeloid (B-cell) lineage (and potential evasion of the immune checkpoint inhibitor imatinib). As one might expect, the Imatinib and Chronic Myeloid Leukemia pathway shows a high and specific probability of generating hitherto unseen mutations in DLBC (B-cell Lymphoma) cancers. In this case, the biology of the pathway matches the cancer type specificity observed in the hitherto unseen variant probabilities, as a pathway with known disregulation in myeloid cancers shows mutation probabilities specific to a myeloid cancer. The "RHOBTB3 GTPASE CYCLE" (top) and "RHOBTB3 ATPASE CYCLE" (second from top) involves a Ras-like protein and all the downstream targets that it phosphorylates. RHOBTB3 has been hypothesized as a tumor suppressor, as its expression is downregulated in many cancers, including renal carcinomas in humans. The RHOBTB3 GTPase/ATPase Cycle pathways show cancer-type-specificity in unseen variant probabilities for KIRC (Kidney Renal Cell Carcinoma).

5.2.5 Assessing reproducibility of Good-Turing probabilities for pathways

Probability estimates of encountering hitherto unseen variants must be reliable to warrant consideration for important clinical tasks. This subsection assesses the reproducibility of Good-Turing probability estimates generated by PMBG-derived pathways using a train-test set split.

To validate the signals produced by this approach, I conducted a 50-50 train-test split of tumors from the 16 most common cancer types according to the procedure described in Methods Subsection 4.2.1. For the training dataset, I conducted the PMBG procedure, de-sparsified the somatic mutation data according to pathway membership, and calculated Good-Turing probabilities per pathway and cancer type according to the procedure described in Methods subsubsection 4.3.2. Then for the test dataset, I calcu-

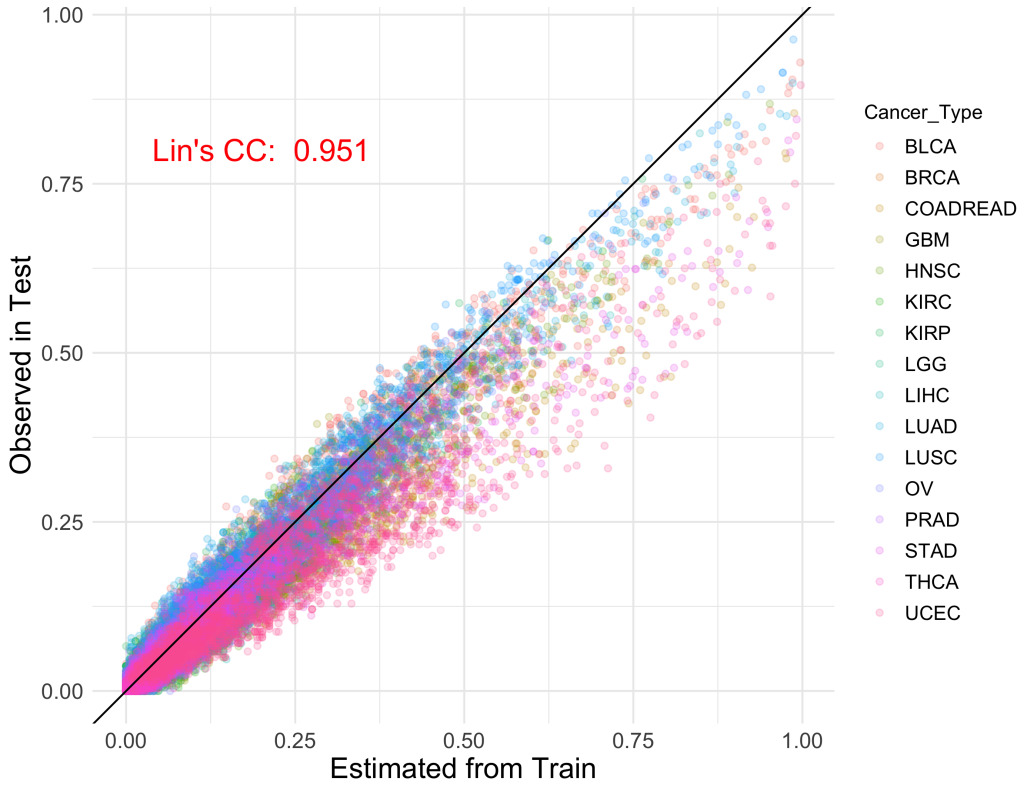


Figure 5.7: Assessing the concordance between Good-Turing probability estimates and observed proportions of unseen mutations for pathways defined using PMBG. The x-axis denotes the estimated Good-Turing probabilities per pathway per cancer type in a training dataset of 3854 tumors. The y-axis denotes the observed proportions of tumors that generated a hitherto unseen mutation in each pathway in a held-out testing dataset of 3854 tumors. The 45° line denotes perfect concordance between estimated probabilities and observed proportions. The Lin's Concordance Correlation (Lin's CC) measures the agreement between the estimated and observed probabilities. Points are colored based on cancer type.

lated the observed proportions of tumor samples of each cancer type that produced at least one hitherto unseen variant. I visualized the concordance between the estimated and observed probabilities of hitherto unseen variants in Figure 5.7, and used the Lin's Concordance Correlation to quantify how well the observed probabilities were reproduced by the Good-Turing estimates. The high Lin's Concordance Correlation (0.951) suggests that the estimates reproduce the true probabilities quite well. However, there is a consistent bias towards overestimating large probabilities (with some cancer types showing more bias than others). For a faceted version of the plot with Lin's correlations calculated per cancer type, I refer the reader to Supplementary Figure 7.2.

In summary, probabilities of encountering *at least one* hitherto unseen mutations in known pathways defined using the PMBG procedure mostly track patterns in background mutation rate between tumors. However, there are some pathways with mutational patterns sufficiently uncorrelated with background mutation. These pathways display mutation patterns with evident and statistically significant cancer type-specific signals. A train-test set split experiment showed that these unseen variant probabilities are highly reproducible among cancer types with reasonably large sample sizes. These results indicate that Good-Turing probabilities for known biological pathways could be relevant to certain clinical tasks.

Chapter 6

Conclusions and Future Work

In summary, HCOMPC (a higher variance method that clustering genes according to their mutation probability patterns and estimates Good-Turing probabilities for the resulting modules) and PMBG (a higher bias method that groups genes and estimates hitherto unseen variant probabilities within known biological pathways) both mostly produce gene groups that track patterns in background mutation rates. However, some gene groups defined using HCOMPC and PMBG are uncorrelated with background mutation rate patterns and are significantly cancer type specific. Good-Turing probabilities produced from these two approaches are highly reproducible, per a train-test set split experiment.

This section includes the main conclusions of this thesis, limitations, and possible future directions with this project.

6.1 Conclusions

Somatic variant mutation data is dominated by rare variants. In the TCGA dataset, more than 90% of mutations were observed only once across the more than 10,000 tumor samples in the dataset. This finding illustrates that the preponderance of mutations in human cancer are rare, and future sequenced tumor samples are bound to contain mutations that haven't yet been observed in the existing cohort of sequenced tumor genomes. Rather than discarding novel variants for which we have no prior information, using statistical methods to estimate the mutational richness of different cancer types can help extract clinically-relevant signals from these previously ne-

glected mutational data.

Previous analysis of cancer genomes using mutation richness estimators like smoothed Good-Turing frequency estimation were restricted to genes commonly mutated in human cancer, which represented only a small cross-section (approximately 2.5%) of the cancer exome (Chakraborty, Arora, Begg & Shen 2019*b*). The exclusive focus on commonly mutated genes was due to the sparsity of somatic mutation data; the vast majority of genes were too sparsely mutated to reliably estimate hitherto unseen variant probabilities.

This thesis sought to expand the scope of cancer-specific analyses of mutation richness by exploring methods for de-sparsifying the somatic mutation data in a statistically and biologically principled manner. Two de-sparsification methods were proposed:

1. HCOMPC: a "lower bias-higher variance" method that learned modules of genes with similar mutation probability patterns using hierarchical clustering and a dynamic branch pruning method.

2. PBMG: a "higher bias-lower variance" method that estimated hitherto unseen variant probabilities for a set of predefined pathways, including pathways with known roles in canonical and cancer related processes.

To control for mutational noise (i.e., passenger mutations that don't show reliable cancer type specific patterns), I restricted my analysis to a set of 2352 known cancer-related genes (i.e., genes with known roles to cancer or genes with functional connections to such genes) (Kuijjer et al. 2018). I applied both de-sparsification approaches to the somatic mutation data for these 2352 genes. Both approaches produced many gene groups with Good-Turing probabilities that merely tracked patterns in background mutation rate. While background mutation rate is a cancer type specific feature relevant to cancer-classification, it represents a high level signal that doesn't require a mutation richness analysis to detect. These gene groups were not of great interest.

HCOMPC and PMBG (to a somewhat lesser extent) identified several gene groups with Good-Turing probabilities that were sufficiently uncorrelated with patterns in background mutation rates. Using Normalized Mutual Information and a null simulation approach, I showed that the cancer-type-specificities of hitherto unseen variant probabilities were statistically significant for both gene groups generated by HCOMPC and PMBG.

A 50-50 train-test split of tumor samples for the 16 most common cancer types was performed, and the reproducibility of the Good-Turing probabili-

ties were assessed by comparing the probabilities estimated from the training set to the probabilities observed in the test set. In general, high Lin's Concordance Correlations illustrated that for cancer types with reasonably large sample sizes, the Good-Turing probabilities were highly reproducible between the train and test data.

In summary, significant and reproducible cancer type-specific signals for some gene groups were obtained using both the HCOMPC and PMBG approaches. These approaches represent an effort to expand the application of statistical estimators of "mutation richness" to a broader cross-section of the cancer genome in statistically and biologically principled manners. HCOMPC and PMBG also defined broader genomic features with clinically-relevant patterns in hitherto unseen mutation frequencies. HCOMPC and PMBG may offer benefits such as improving the precision of hitherto unseen variant estimation from sparse data, and reducing dimension and data sparsity to facilitate the development of machine learning classifiers of human cancers.

Ultimately, this thesis offers novel methods to potentially expand the application of statistical estimators of "mutation richness" to a broader cross-section of the cancer genome in statistically and biologically principled ways. Continued research in this area will further unlock the "hidden genome" of rare and hitherto unseen variation, which could have important clinical implications.

6.2 Limitations

One major limitation of this research is the unbalanced nature of the dataset, as many cancer types had a small number of samples. Precisely estimating unseen variant probabilities is directly dependent on how well one can estimate the number of singleton variants (N_1) per tumor. Larger sample sizes would indeed help estimate these quantities more accurately. The generation, standardization, and dissemination of more larger cancer genomics datasets will help improve the quality of the estimates and aid efforts at mining mutation data for clinically-relevant signals.

This thesis only considered tumors of certain mutation signatures (Non-hypermutated, APOBEC (2,13), Smoking (4), and MMR (6,15,20,25) signatures) and certain cancer types (32 were considered). These cancers represented relatively broad cancer types with low to moderate mutational bur-

dens, as I wanted to avoid cancer signatures like POLE (10), which are known for catastrophic hypermutation events that make Good-Turing estimation unreliable (Chakraborty, Arora, Begg & Shen 2019*a*). Furthermore, my validation analyses were restricted to the 16 most common cancer types. Thus, the results reported in this thesis are not readily generalizable to many cancer types or different mutation signatures not included in the TCGA dataset.

Another important limitation of this research is that tumor heterogeneity is not fully explained by cancer type alone. While cancers may share tissues of origin or broad histological features, there are latent characteristics of tumors which have direct influence on cancer biology, disease outcome, and response to different treatments. Characterizing cancers in greater detail using other omics data types (e.g., single-cell methods), imaging data, and biomarkers, and integrating the data in principled ways will yield more comprehensive and accurate portraits of individual cancers. There is an entire field of research focused on cancer subtype discovery aimed at addressing these questions. The complexity of cancer cannot be emphasized enough; no two cancers are alike, and even individual cancers are composed of heterogeneous populations of cells. It is important to note that this thesis examined cancers on the basis of cancer type, a useful but reductionist perspective of cancer biology.

One other limitation of this work concerns the biological prior information used in the de-sparsification approaches. Conventional scientific knowledge posits that the majority of mutated cancer genes are "passenger events", that are akin to mutational noise. In other words, the majority of mutations don't play an active role in driving the cancerous phenotype. Feature screening/selection is critical in distinguishing genes with cancer-type specific mutational signals from the background mutational noise. This thesis addressed this problem by focusing on a list of over 2000 known cancer-associated genes compiled by COSMIC (Tate et al. 2019) and Östlund et al. (2009). However, focusing only on genes with known links to cancer probably did not capture all the genes truly relevant to carcinogenesis in the TCGA dataset. Moreover, the PMBG approach relied on grouping genes according to known biological pathways and other biologically defined pathways. Even though many different pathways were considered, they likely are an underfit model of the true molecular circuits underlying cancer initiation and progression. Increased sample sizes of cancer genomics datasets will enable scientists to distinguish rare cancer driver genes from background noise, and improved knowledge of cancer genetics improve characterization of cancer-relevant pathways.

6.3 Future Work

There are a variety of potential future research opportunities related to this project. One important question involves how to integrate hitherto unseen variants into machine learning classifiers of cancer type. Work like that of Chakraborty et al. (2020) use a "meta-regression" step, which uses the information in previously observed variants to estimate the logistic regression coefficients associated with hitherto unseen variants. Essentially, the approach imputes the predictive effect of hitherto unseen variants using observed variants. While the meta-regression step incorporates information from hitherto unseen variants, it does not directly incorporate Good-Turing probabilities.

To directly incorporate Good-Turing probabilities into the machine learning scheme, I propose a relatively straightforward Naive-Bayes classification algorithm. Naive Bayes are simple probabilistic classifiers that assign objects to classes by maximizing the posterior likelihood associated with each class. In the context of this thesis, Naive Bayes can convert mutation probabilities (previously observed or hitherto unseen) across different cancer types into probabilities of cancer classes *given* the occurrence of mutational events. Future efforts that harness of the "hidden genome" of previously unseen variation could facilitate important clinical tasks like cancer primary site prediction.

Another future direction is finding ways to adjust Good-Turing probabilities for background mutation rates. The majority of gene groups produced using HCOMPC and PMBG methods merely tracked patterns in background mutation rates, which limits the clinical utilities of these approaches. I was unable to identify adjustments to the Good-Turing formula that account for background event rates, so this line of research may require the development of novel statistical methods[1].

Another possible extension of this thesis is considering other possible criteria for grouping genes. One possible method for grouping genes could be gene-specific variables termed "metafeatures" that were shown by Lawrence et al. (2013) to influence somatic mutation rate. Metafeatures include the gene's average expression level, GC content, replication time, non-coding mutation rate, and presence in an open/closed chromosomal compartment. Clustering genes according to relevant features may help identify gene groups

[1]I considered an alternative Good-Turing formula, which while controlling for background mutation rate, generated signals that were likely artifacts. See Section 7.4 for more.

with clinically-relevant rare variant signals. Another potential grouping criterion to guide variant aggregation are networks. Networks are higher variance criteria than pathways, and they illustrate relationships among genes or proteins as edges drawn between nodes on a network diagram. Network methods have proven extremely valuable to cancer research in a variety of manners: identified gene modules implicated in cancer (Leiserson et al. 2015), as a somatic variant mutation de-sparsifying strategy (Hofree et al. 2013), and in classifying breast cancer metastases (Chuang et al. 2007). The potential utility of networks in defining groups of genes that share rare variant signals is a worthy task with potentially useful clinical implications.

6.4 Acknowledgements

I would like to thank Professor Johanna Hardin (Pomona College) for her insights and mentorship throughout this senior thesis project and throughout my years at Pomona. I owe her a debt of gratitude for helping me discover my passion for Biostatistics research. I would also like to thank Professor Saptarshi Chakraborty (SUNY Buffalo) and Dr. Ronglai Shen (Memorial Sloan Kettering Cancer Center) for their technical guidance throughout this project. I would like to thank Professor Vin De Silva (Pomona College) for his strong leadership of this senior thesis cohort. I would like to thank Kathy Sheldon (formerly of Pomona College) for her central role in the Pomona Math department over my four years at Pomona. I would like to thank all my fellow Pomona math seniors for supporting me through this unusual and unprecedented senior year. Lastly, I would like to thank my family — my mom, Mary Jo, my dad, Mark, and my brother, EJ — for everything they have done for me. I am all the better for their wisdom, the balance they instill in my life, and their unconditional love and support.

Chapter 7

Supplementary materials

This chapter contains supplementary figures, derivations of the Good-Turing formula, and additional analyses using alternative grouping criteria and Good-Turing formula.

7.1 Supplementary Figures

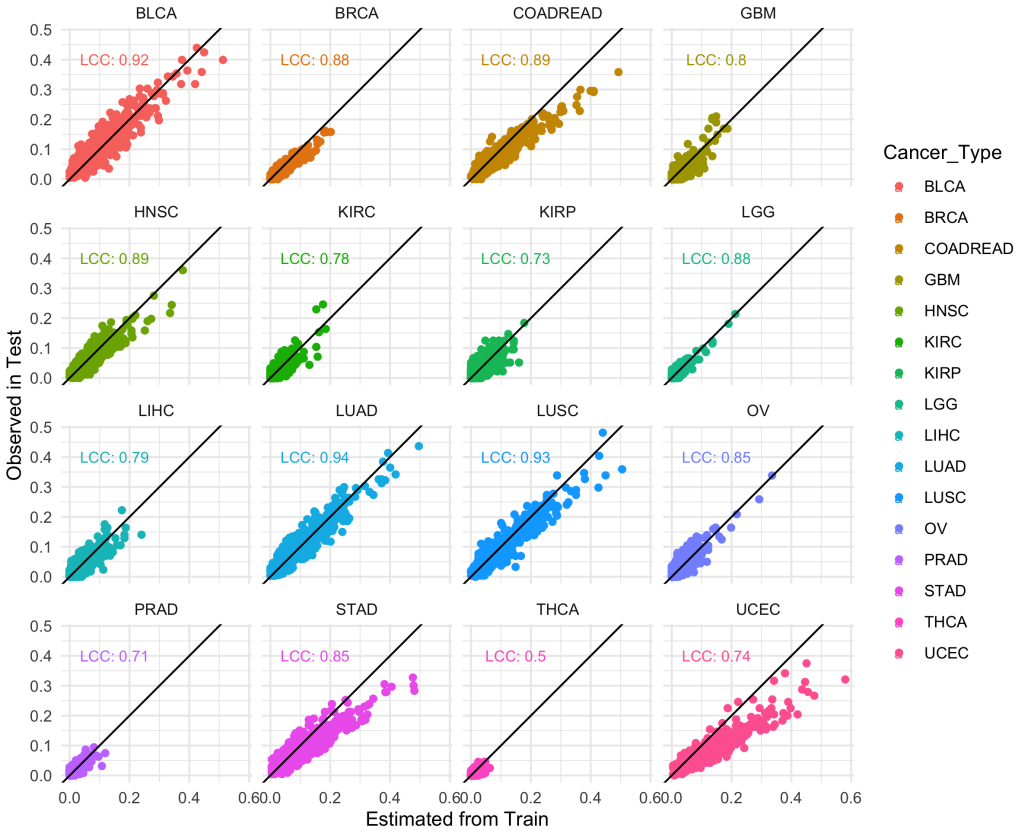


Figure 7.1: Concordance between Good-Turing probabilities for modules defined using HCOMPC procedure on training dataset compared to the observed proportions of hitherto unseen variants in a test dataset, faceted by cancer type. The 45° line denotes perfect prediction of probabilities. Lin's Concordance Correlations, which measures the strength of correlation along the 45° line, per cancer type are shown in each panel.

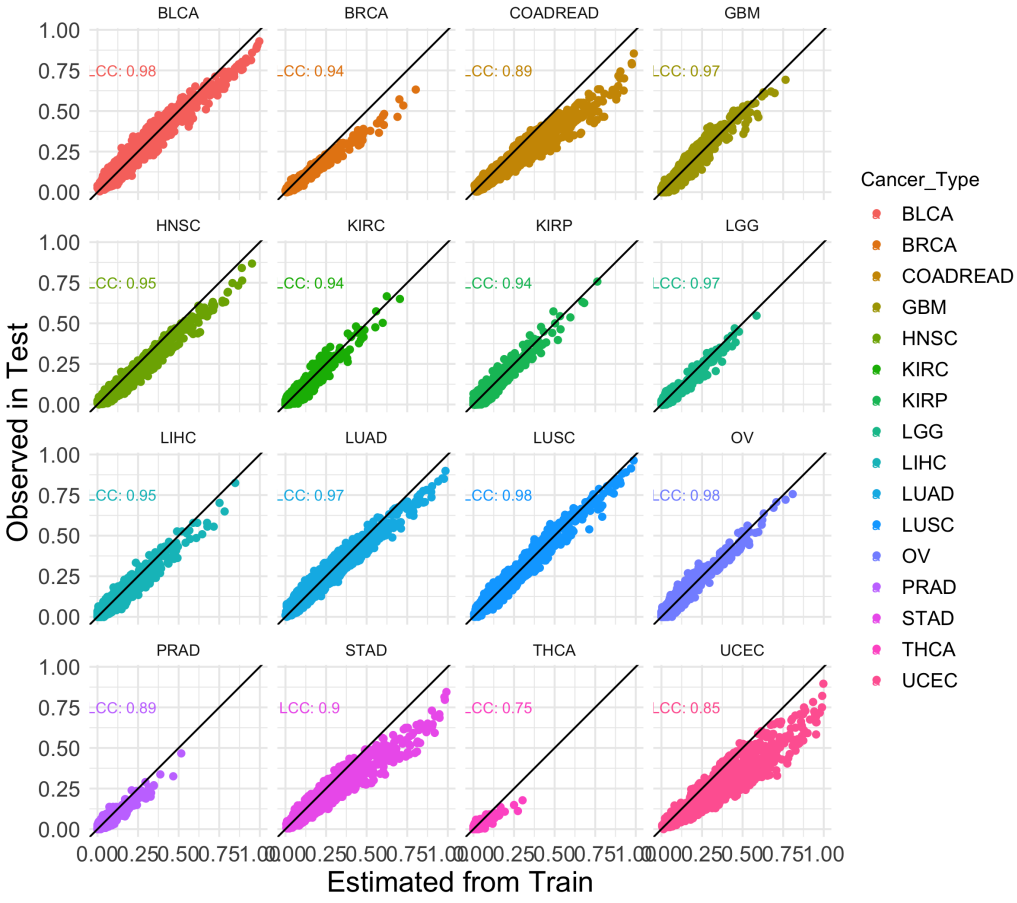


Figure 7.2: Concordance between Good-Turing probabilities for modules defined using HCOMPC procedure on training dataset compared to the observed proportions of hitherto unseen variants in a test dataset, faceted by cancer type. The 45° line denotes perfect prediction of probabilities. Lin's Concordance Correlations, which measures the strength of correlation along the 45° line, per cancer type are shown in each panel.

7.2 Applying PMBG with non-pathway gene sets

The PMBG de-sparsification strategy considered mutations aggregated/summed within known biological pathways. Biological pathways are considered the "gold standards" of biomolecular interactions. However, our knowledge of biological pathways likely underfits the true extent of gene-gene and protein-protein interactions occurring within cells. In an effort to capture interactions not included in canonical pathway annotations, I explore the possibility of running the PMBG de-sparsification algorithm on alternative gene sets.

7.2.1 Method

Identical to the procedure outlined in the Methods chapter, 2352 cancer-associated genes from the Catalogue of Somatic Mutations in Cancer (COSMIC) (Tate et al. 2019) and Supplemental Table 3 from Östlund et al. (2009) were considered for grouping.

Gene set files containing pathways were acquired from MSigDb (Subramanian et al. 2005) in .gmt format [1]. The following 5 gene set files were used as criteria for grouping genes, listed roughly in order of decreasing bias/increasing variance:

1. Positional criteria ("c1.all.v7.3.symbols.gmt"): contains 299 positional gene sets, where genes are grouped according to chromosome and cytogenic band (regions of the chromosome with actively expressed DNA; often appear different under karyotypic staining).

2. Computational Gene Set criteria ("c4.all.v7.3.symbols.gmt"): contains 858 gene sets obtained by previous computational mining of cancer microarray data. These gene sets include gene neighborhoods surrounding 380 cancer-related genes and cancer modules significantly implicated in a variety of cancer outcomes.

3. Ontology Gene Set criteria ("c5.all.v7.3.symbols.gmt"): contains 14,996 gene sets where genes are annotated according to their gene ontology

[1] .gmt format refers to gene matrix transposed format, often used to describe gene groups.

(GO) terms, i.e., to which biological processes, cellular components, and molecular functions to genes belong to.

4. Oncogenic Signature Gene Set criteria ("c6.all.v7.3.symbols.gmt"): contains 189 oncogenic signature gene sets, representing cellular pathways which are often dis-regulated in cancer. The majority were generated directly from microarray data or from MSigDb's unpublished profiling experiments involving perturbation of known cancer genes.

5. Cell Type Signature Gene Set criteria ("c8.all.v7.3.symbols.gmt"): contains 673 gene sets that contain curated cluster markers for cell types identified in single-cell sequencing studies of human tissue. The gene sets in this collection contain cell types from Heart, GI Tract, Pancreas, Kidney, Liver, the Immune system, Retina, Olfactory tissue, and the Brain.

The gene set file were converted to binary adjacency matrices A where $A_{i,j} = 1$ indicates the presence of Gene j in Pathway i using the *SAMBAR* R package (Kuijjer 2021).

Gene sets were eliminated that contained 3 or fewer genes. Mutation counts were aggregated/summed within gene sets per cancer type. Then Good-Turing probabilities were calculated per gene set and per cancer type according to the procedure outlined in Section 4.2. The procedure outlined above treats *gene sets* as the genomic units of interest.

7.2.2 Grouping genes into gene sets

Summary statistics for gene sets produced using the different grouping criteria are shown in Table 7.1. The Positional dataset, which groups genes according to chromosome and cytogenic band, produced 183 gene sets with more than 3 genes. The gene set sizes ranged from 4 to 71 genes. The Computational dataset, which groups genes according to groups identified by computational mining of cancer microarray data, produced an intermediate number of pathways (715) that varied in size between 4 and 151 genes. The Gene Ontology dataset, which groups genes according to their GO terms (involvement in biological processes, cellular components, or molecular functions), produced 7965 gene sets that varied in size between 4 and 495 genes. The Oncogenic Signature dataset, where groupings are defined based on microarray data and perturbation experiments of known cancer genes, produced

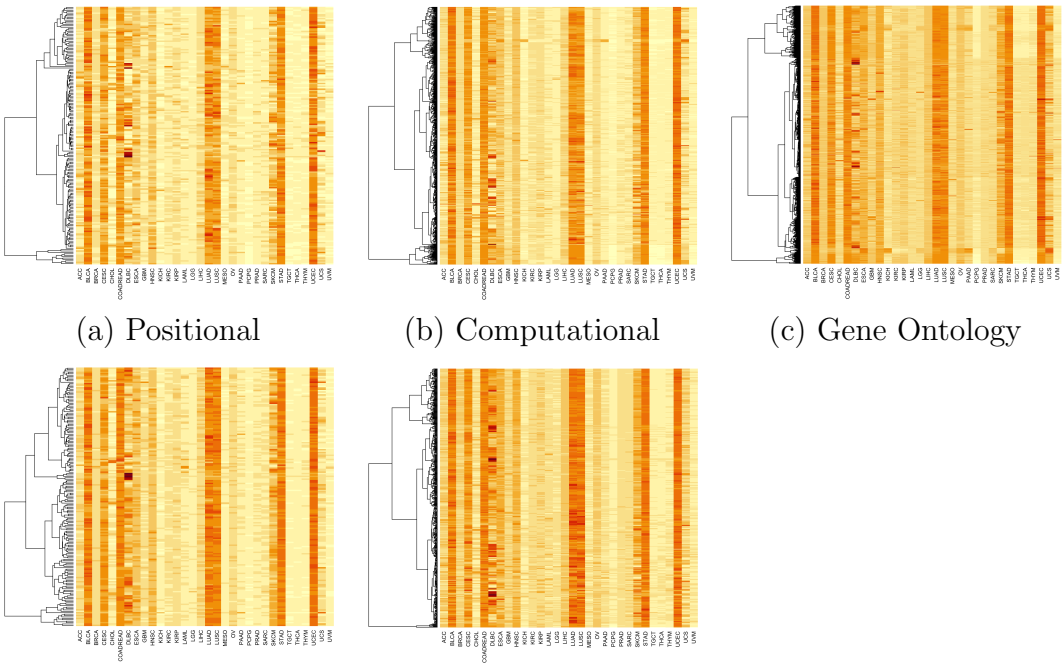
189 gene sets, ranging from 4 genes to 57 genes in size. The Cell Type Specific dataset, where groupings are markers for various cell types derived from scRNA-seq produced 543 gene sets that varied in size from 4 to 335 genes.

Group. criteria	# groups	Min group size	Med group size	Max group size
Positional	183	4	8	71
Computational	715	4	17	151
Gene Ontology	7965	4	12	495
Onco. Sig.	189	4	22	57
Cell Type Sig.	543	4	20	335

Table 7.1: Summary statistics for gene sets derived from the PMBG procedure. Pathways were broken into groups according to their parent datasets: Positional, Computational, Gene Ontology, Oncogenic Signature, and Cell Type Signature.

7.2.3 Good-Turing probabilities per gene set

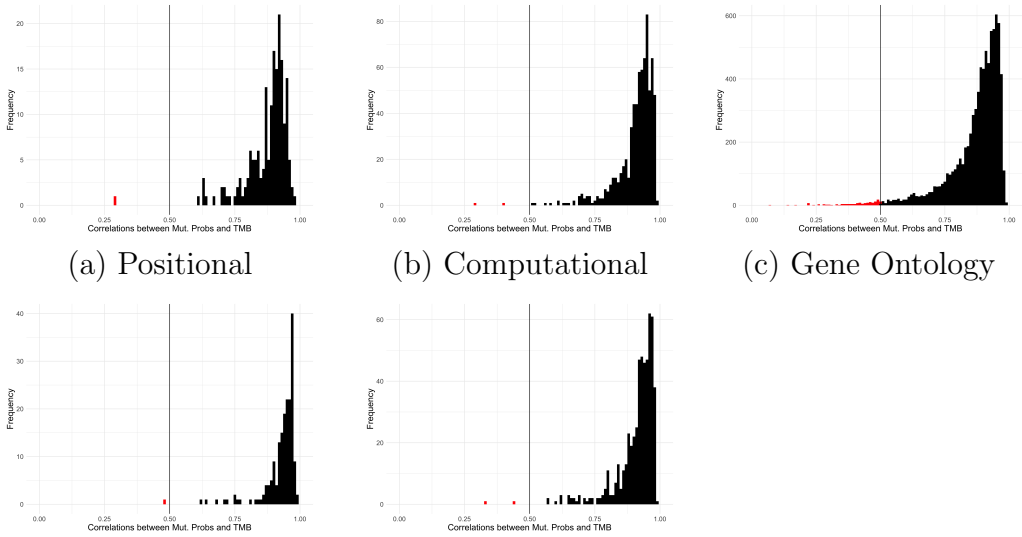
Probabilities of observing at least one hitherto unseen variant in a future randomly selected tumor sample were calculated per gene set and per cancer type according to equation 2.11 and visualized in the heatmaps shown in Figure 7.3 to examine broader patterns in unseen mutational probabilities. The heatmap illustrates highly consistent patterns in hitherto unseen mutation probabilities among the different gene sets. The cancer types BLCA, CESC, COADREAD, LUAD, LUSC, STAD, and UCEC consistently show higher probabilities than the other cancer types and typically have the highest frequencies of mutation. The consistent patterns in Good-Turing probabilities suggests that the gene sets may be tracking patterns in background mutation rate or total mutational burden (TMB).



(a) Positional (b) Computational (c) Gene Ontology

(d) Oncogenic Signature (e) Cell Type Signature

Figure 7.3: Heatmaps of Good-Turing probabilities per gene set per cancer type for each gene grouping criteria. Columns represent cancer types and rows represent gene sets. Darker tiles indicate higher Good-Turing probabilities.



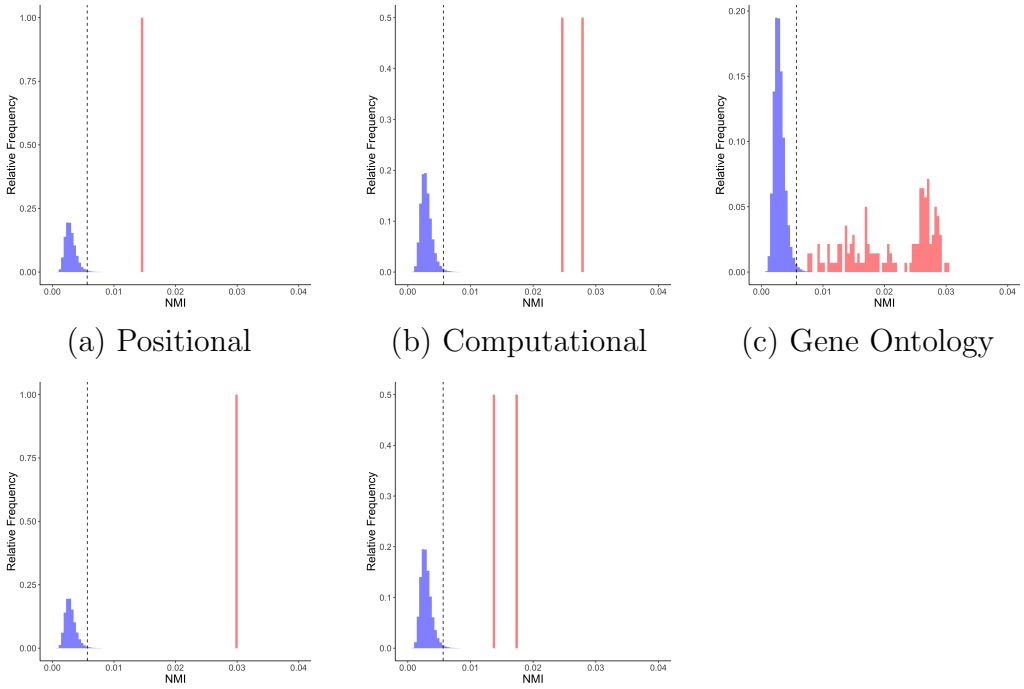
(a) Positional (b) Computational (c) Gene Ontology

(d) Oncogenic Signature (e) Cell Type Signature

Figure 7.4: Histograms illustrating the distribution of correlations among gene sets defined via (a) positional, (b) computational, (c) gene ontology, (d) oncogenic signature, and (e) cell type signature criteria. The vertical black line denotes the Pearson correlation cutoff of 0.5. Bars to the left of that cutoff are colored red.

To confirm if gene sets were tracking patterns in background mutation rate, I measured the Pearson correlations between each gene set's Good-Turing probabilities and a vector of the average total mutational burden per tumor across the 32 tumor types. Histograms of the correlations values per grouping criterion is shown in Figure 7.4. The left-skewed distributions confirmed that the vast majority of gene sets show Good-Turing probabilities highly correlated to patterns of background mutation rate. A few gene sets achieved correlations under the *ad hoc* cutoff of 0.5: 1 gene set for positional, 2 for computational, 140 for gene ontology, 1 for oncogenic signature, 2 for cell type signature. In all, 146 gene sets cleared this threshold.

7.2.4 Assessing cancer type specificity of hitherto unseen variant probabilities of gene sets



(a) Positional (b) Computational (c) Gene Ontology

(d) Oncogenic Signature (e) Cell Type Signature

Figure 7.5: Distributions of null (blue) and observed (red) NMI values for gene sets (uncorrelated with background mutation rates). The dotted black line indicates the 99th quantile value on the null distribution. The right shift in the red distribution illustrates that the observed gene sets are more cancer type specific than those values generated by random chance.

To assess the statistical significance of the cancer type specificities observed in the Good-Turing probabilities per gene set, I generated a null distribution of Normalized Mutual Information (NMI) values under the assumption of no association between variants and cancer type according to the procedure described in Methods Subsection 4.4.2. I summarized the distribution of null NMIs along with the observed NMIs calculated from the un-permuted data in a relative frequency histograms shown in Figure 7.5. The dotted black line represents the 99th quantile of the null distribution

of NMIs. All 146 pathways sufficiently uncorrelated with background mutation rate patterns exceeded this 99th quantile cutoff. Thus, I conclude with 99% confidence, that the observed cancer-type-specificities in Good-Turing probabilities were statistically significant for the 146 pathways.

7.2.5 Exploring patterns of hitherto unseen variant probabilities

Examining the patterns of hitherto unseen variant probabilities in the gene sets sufficiently uncorrelated with background mutation rates yielded some intriguing findings. Shown in Figure 7.6 are a select group of gene sets produced by the positional, computational, gene ontology, oncogenic signature, and cell type signature grouping criteria.

The the single positional gene set uncorrelated with background mutation rates ("chr6p25", bottom), the two computational gene sets ("MORF MYST2" and "MORF PDPK1", second and third from bottom), several gene ontology gene sets ("Regulation of brown fat cell differentiation" and "Ribosomal small subunit binding" shown), the single oncogenic signature gene set ("IL.2.UP.V1.UP", third from top), and the cell type signature gene sets ("Durante adult olfactory neuroepithelium olfactory horizontal basal cells" and "Travaglini lung Natural Killer T cell", top two) all produced hitherto unseen variant probability patterns with high signal in DLBC (B cell Lymphoma). Two gene ontology gene sets ("Negative regulation of transcription from RNA Polymerase II Promoter in response to stress" and "Ubiquitin Ligase Substrate Adaptor Activity") showed increased hitherto unseen variant probabilities in the KIRC cancer type (Kidney Renal Clear Cell Carcinoma). The GO Ontology gene set "Abnormal pupillary function" shows higher Good-Turing probabilities in BLCA (Bladder Urothelial Carcinoma), CHOL (Cholangiocarcinoma), MESO (Mesothelioma), and UVM (Uveal Melanoma).

7.3 Deriving Good-Turing estimator using binomial likelihoods

The following derivation of the Good-Turing estimator is borrowed from Chakraborty, Arora, Begg & Shen (2019*a*).

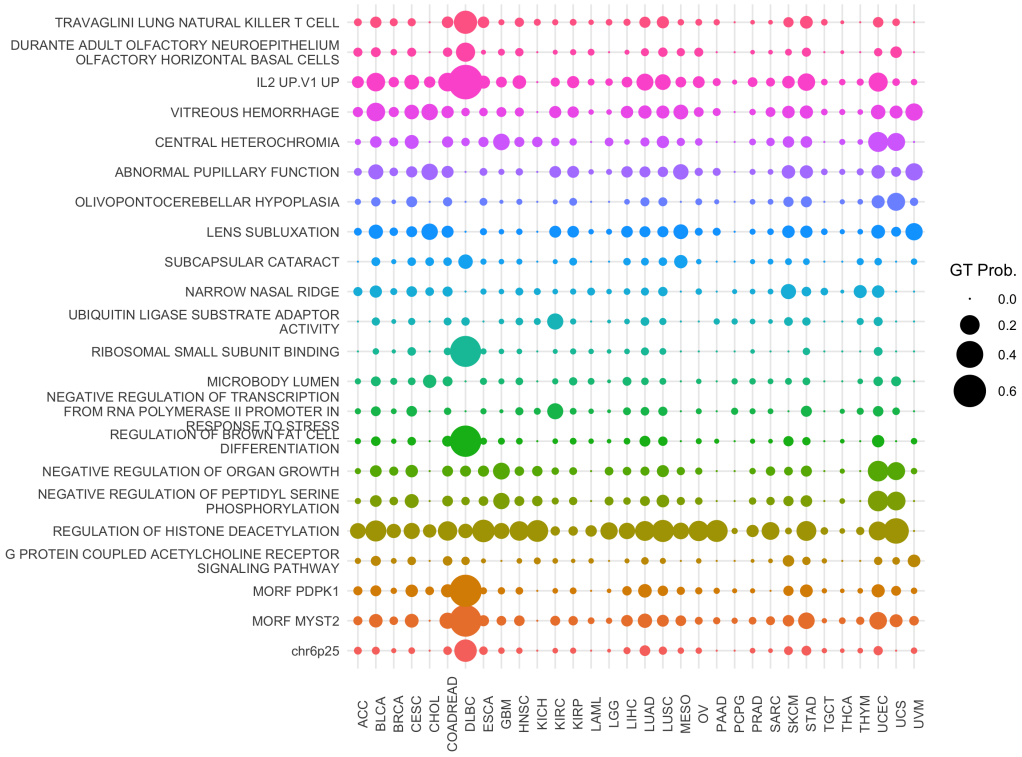


Figure 7.6: Probabilities of observing at least one hitherto unseen mutation in a select group of gene sets per cancer type. Bubbles are sized according to Good-Turing probabilities.

Let q_j denote the probability of encountering the j-th variant in a randomly selected tumor and r_j be the number of times that variant appears in the existing sample of m tumors. Let N be the total number of variants. Then $r \sim \text{Binomial}(m, q_j)$, and assuming independence between the observations, we obtain a *product binomial likelihood model*:

$$L(q_1, q_2, \ldots) = \prod_{j=1}^{N} \binom{m}{r_j} q_j^{r_j} (1 - q_j)^{m - r_j} \tag{7.1}$$

Let N_r denote the number of variants appearing exactly r times in the sample of m tumors.

Let the q_j's be a priori independent with a common prior distribution F on [0,1]. Under a non-parametric prior F, the posterior mean of q_j conditional on $r_j = r$ is given by:

$$E(q_j, r_j = r) = \frac{\int_0^1 q\binom{m}{r} q^r (1-q)^{m-r} dF(q)}{\int_0^1 \binom{m}{r} q^r (1-q)^{m-r} dF(q)} \tag{7.2}$$

Using the identity

$$q\binom{m}{r} q^r (1-q)^{m-r} = \frac{r+1}{m+1} \binom{m+1}{r+1} q^{r+1} (1-q)^{(m+1)-(r+1)} \tag{7.3}$$

in the posterior mean formula above, we obtain:

$$\begin{aligned} E(q_j, r_j = r) &= \frac{r+1}{m+1} \frac{\int_0^1 \binom{m+1}{r+1} q^{r+1} (1-q)^{(m+1)-(r+1)} dF(q)}{\int_0^1 \binom{m}{r} q^r (1-q)^{m-r} dF(q)} \\ &= \frac{r+1}{m+1} \frac{p_{m+1}(r+1)}{p_m(r)} \end{aligned} \tag{7.4}$$

Where $p_m(r)$ denotes the marginal probability of a variant frequency equal to r.

In the original Good-Turing estimation, $\frac{p_{m+1}(r+1)}{p_m(r)}$ is estimated by the ratio of the empirical frequencies $\frac{N_{r+1}}{N_r}$. However, the estimates are often unstable as N_r's for different values can be 0, making the estimation of $E(q_j, r_j = r)$ problematic. To overcome this, smoothing of raw N_r values is necessary.

Once the N_r values are smoothed ($S(N_r)$), we replace $\frac{p_{m+1}(r+1)}{p_m(r)}$ with $\frac{S(N_{r+1})}{S(N_r)}$, yielding the Good-Turing estimate:

$$P_{GT}(r) = \hat{q}^{GT} = \frac{r+1}{m+1}\frac{S(N_{r+1})}{S(N_r)} \tag{7.5}$$

We have derived the Good-Turing formula, that estimates the probability of encountering at least one hitherto unseen mutation in a future randomly selected tumor sample!

7.4 Can estimating probabilities of encountering a single hitherto unseen variant decouple Good-Turing probabilities and background mutation rate patterns?

Note that all Good-Turing probabilities presented in this section correspond to the probability of encountering a **particular** hitherto unseen mutation, according to the following equation:

$$P_{GT} = \frac{1}{m+1}\frac{\hat{N}_0}{N_1} \tag{7.6}$$

Where m is the number of tumor samples, N_1 is the number of singleton mutations, and $\hat{N}_0$ is the Chao estimate of the number of unseen species in a population Chao (1987).

Probabilities of observing *one particular* hitherto unseen variant were calculated per pathway according to equation 7.6, and visualized in the heatmap shown in the right panel of Figure 7.7 to examine overarching patterns in probabilities of encountering *a particular* hitherto unseen mutation. Unlike in Figures 5.5 and 5.2, this heatmap doesn't display consistent probabilities within particular cancer types. While many pathways have high probabilities in CHOL (Cholangiocarcinoma), there is appreciably more variation among pathways and among cancer types than previously observed. In other words, it appears that the majority of pathways show unseen variant probability estimates uncorrelated with patterns in background mutation rates. This is confirmed by a histogram of the correlations between Good-Turing probabilities and Total Mutational Burden vector (shown in the right panel of

Figure 7.7). Unlike in Figures 5.5 and 5.2, where the vast majority of pathways displayed Good-Turing probabilities that were highly correlated with background mutation rate. All probabilities in Figure 7.7 are uncorrelated to weakly negatively correlated with background mutation rate.

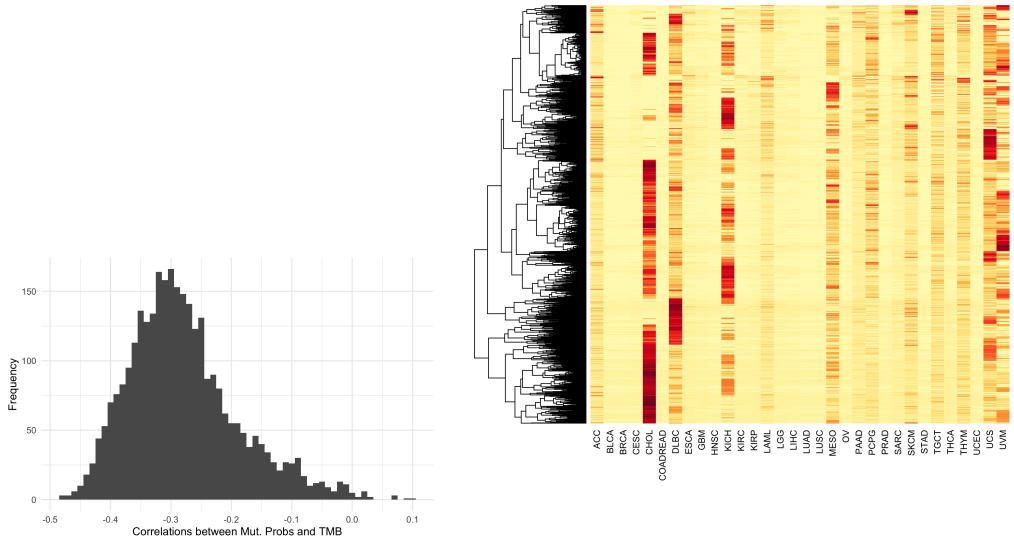


Figure 7.7: Left: correlations between unseen variant probabilities and average Total Mutational Burden (TMB) vector show weak negative correlation among preponderance of pathways. Right: heatmap of probabilities of encountering one particular hitherto unseen variant across 32 different cancer types across various pathways.

To evaluate the significance of the cancer type specific patterns in hitherto unseen variant probabilities for each pathway, I conducted a permutation test on the normalized mutual information (NMI) as outlined in the methods section. I repeated this procedure 1000 times and summarized the null NMIs along with the observed NMIs in a relative frequency histogram shown in the left panel of Figure 7.8. The dotted black line represents the 99th quantile of the null distribution of NMIs; thus with more than 99% confidence, NMIs greater than the black line show more cancer-type-specific patterns of hitherto unseen variant probabilities than would be expected by random chance.

There is a small right shift in the distribution of observed NMI values

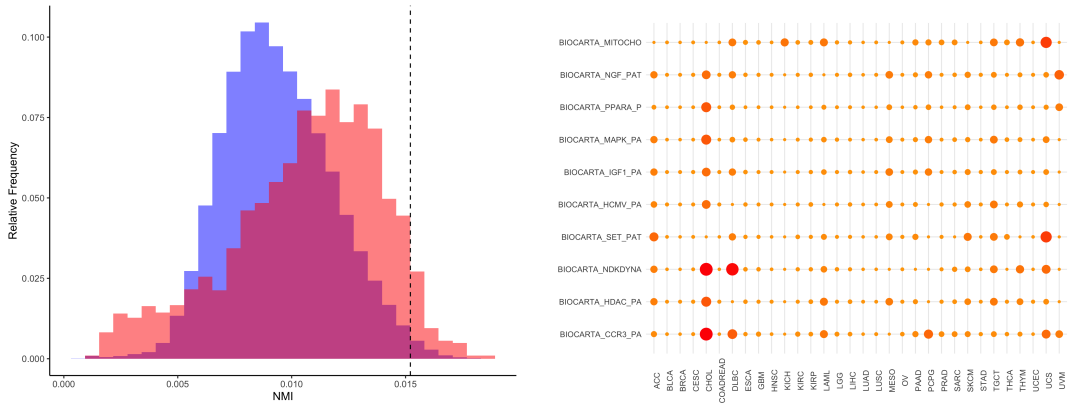


Figure 7.8: Left: Null reference distribution of NMIs (blue) compared to the observed NMIs (red) for probabilities of observing a particular previously unseen variant in known pathways. The dotted black line indicates the 99th quantile cutoff of the null distribution of NMI values. Right: bubbleplot illustrating probabilities of encountering a particular previously unseen mutation across different pathways and cancer types. This plot illustrates the 10 pathways with the highest NMIs. Bubbles are sized and colored according to their Good-Turing probabilities.

(red) relative to the null reference distribution (blue), suggesting that the observed pathways are more cancer type specific than would be expected by random chance. Using the 99th quantile as a significance cutoff, 150 pathways were deemed significantly cancer type specific. These pathways were subjected to further exploratory analysis.

To illustrate the patterns of mutation probabilities, I visualized probabilities of encountering a particular previously unseen mutant in the bubbleplot shown in the right panel of Figure 7.8. Plotted are the 10 pathways with the highest NMI values observed. There is some evident variability in the unseen variant probabilities between pathways. For instance, the "BIOCARTA_CCR3_PATHWAY" (bottom) shows a high affinity for the CHOL (Cholangiocarcinoma) cancer type in its Good-Turing probabilities. The "BIOCARTA_NDKDYNAMIN_PATHWAY" (third from bottom) shows high affinity for CHOL and DLBC (B-cell lymphoma). The "BIOCARTA_SET_PATHWAY" (fourth from bottom) shows high affinity for the UCS (Uterine Carcinosarcoma) and ACC (Adrenocortical carinoma).

It gives me pause that the highest probabilities of encountering a partic-

ular previously unseen mutation are found in cancer types with few samples. As shown in Table 4.1, CHOL($m = 34$), DLBC ($m = 37$), UCS ($m = 49$), and ACC ($m = 92$) (cancer types with high Good-Turing probabilities Figure 7.8) all represent cancer types with fewer than 100 samples. Perhaps these cancer type-specific signals are artifacts of the unbalanced nature of the dataset.

In summary, considering probabilities of encountering one particular previously unseen mutation resolves the problem of Good-Turing probabilities tracking background mutation rates, as evidenced by the de-coupling of pathway mutation probabilities with average mutational burden seen in Figure 7.7. However, the cancer type-specificity of these unseen variant probability patterns are weaker than in the "least one probability" case and are likely artifacts of the unbalanced nature of the dataset. What's more, considering a *particular* hitherto unseen mutation suffers from lack of interpretability, as identifying a particular mutation you haven't encountered before appears counterintuitive. Improved methods to adjust for background mutation rates warrant investigation.

Bibliography

Chakraborty, S., Arora, A., Begg, C. B. & Shen, R. (2019*a*), 'Supplementary Information for " Using Somatic Variant Richness to Mine Signals from Rare Variants in the Cancer Genome "', *Nature Communications* **10**, 1–20.

Chakraborty, S., Arora, A., Begg, C. B. & Shen, R. (2019*b*), 'Using somatic variant richness to mine signals from rare variants in the cancer genome', *Nature Communications* **10**, 1–9.
URL: *http://dx.doi.org/10.1038/s41467-019-13402-z*

Chakraborty, S., Begg, C. B. & Shen, R. (2019), *variantprobs: Computing Probabilities of Gene Variants*. R package version 0.1.0.
URL: *https://github.com/c7rishi/variantprobs*

Chakraborty, S., Begg, C. B. & Shen, R. (2020), 'Using the " Hidden " Genome to Improve Classification of Cancer Types', *Unpublished manuscript* pp. 1–24.

Chang, M. T., Bhattarai, T. S., Schram, A. M., Bielski, C. M., Donoghue, M. T. A., Jonsson, P., Chakravarty, D., Phillips, S., Kandoth, C., Penson, A., Gorelick, A., Shamu, T., Patel, S., Harris, C., Gao, J., Sumer, S. O., Kundra, R., Razavi, P., Li, B. T., Reales, D. N., Socci, N. D., Jayakumaran, G., Zehir, A., Benayed, R., Arcila, M. E., Chandarlapaty, S., Ladanyi, M., Schultz, N., Baselga, J., Berger, M. F., Rosen, N., Solit, D. B., Hyman, D. M. & Taylor, B. S. (2018), 'Accelerating Discovery of Functional Mutant Alleles in Cancer', *Cancer Discovery* **8**(February).

Chao, A. (1987), 'Estimating the Population Size for Capture-Recapture Data with Unequal Catchability', *Biometrics* **43**(4), 783–791.

Chen, Y., Sun, J., Huang, L.-C., Xu, H. & Zhao, Z. (2015), 'Classification

of Cancer Primary Sites Using Machine Learning and Somatic Mutations', *BioMed Research International* **2015**.

Chuang, H.-y., Lee, E., Liu, Y.-t., Lee, D. & Ideker, T. (2007), 'Network-based classification of breast cancer metastasis', *Molecular Systems Biology* **3**(140), 1–10.

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., Raphael, B. J., Marks, D. S., Ouellette, B. F. F. & Valencia, A. (2016), 'Pathway and Network Analysis of Cancer Genomes', *Nature Methods* **12**(7), 615–621.

Croce, C. M. (2008), 'Oncogenes and Cancer', *The New England Journal of Medicine* **358**(5), 502–511.

Dikaios, N. (2020), 'Sparse input neural networks to differentiate 32 primary cancer types based on somatic point mutations'.

Gale, W. A. (1995), 'Good-Turing Smoothing Without Tears', *Journal of Quantitative Linguistics* **2**, 1–24.

Hanahan, D. & Weinberg, R. A. (2000), 'The Hallmarks of Cancer', *Cell* **100**, 57–70.

Hasan, M. A. & Lonardi, S. (2018), mClass : Cancer Type Classification with Somatic Point Mutation Data, PhD thesis, University of California, Riverside.

Hassanpour, S. H. & Mohammadamin, D. (2017), 'Review of cancer from perspective of molecular', *Journal of Cancer Research and Practice* **4**(4), 127–129.
URL: *https://doi.org/10.1016/j.jcrpr.2017.07.001*

Hofree, M., Shen, J., Carter, H., Gross, A. & Ideker, T. (2013), 'Network-based stratification of tumor mutations', *Nature Methods* **10**(11), 1108–1114.

Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Danyi, A., Ridder, J. D., Herpen, C. V., Lolkema, M. P., Steeghs, N., Getz, G., Morris, Q. & Stein, L. D. (2020), 'A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns', *Nature Communications* **11**(728), 1–12.

Kuijjer, M. L. (2021), *SAMBAR: Subtyping Agglomerated Mutations By Annotation Relations*. R package version 0.3.

Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W. & Quackenbush, J. (2018), 'Cancer subtype identification using somatic mutation data', *British Journal of Cancer* (January).
URL: *http://dx.doi.org/10.1038/s41416-018-0109-7*

Langfelder, P., Zhang, B. & Horvath, S. (2008), 'Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R', *Bioinformatics* **24**(5), 719–720.

Langfelder, P., Zhang, B. & Horvath, S. (2009), Dynamic Tree Cut : in-depth description , tests and applications, Technical report, UCLA.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., Mckenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Auclair, D., Saksena, G., Voet, D., Noble, M., Dicara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-a., Wu, C. J., Melendez-zajgla, J., Hidalgomiranda, A., Koren, A., Mccarroll, S. A., Mora, J., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J. & Garraway, L. A. (2013), 'Mutational heterogeneity in cancer and the search for new cancer-associated genes', *Nature* **499**, 214–218.

Leiserson, M. D. M., Vandin, F., Wu, H.-t., Dobson, J. R., Eldridge, J. V., Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., Mclellan, M., Lawrence, M. S., Gonzalez-perez, A., Tamborero, D., Cheng, Y., Ryslik, G. A., Lopez-bigas, N., Getz, G., Ding, L. & Raphael, B. J. (2015), 'Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes', *Nature Genetics* **47**(2).

Lin, L. I.-K. (1989), 'A Concordance Correlation Coefficient to Evaluate Reproducibility', *Biometrics* **45**, 255–268.

Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, L. & Darnell, J. (2003), *Molecular Cell Biology*, fifth edit edn, Freeman.

Östlund, G., Lindskog, M. & Sonnhammer, E. L. (2009), 'Network-based Identification of Novel Cancer Genes', *Molecular and Cellular Proteomics* **9**(4), 648–655.

Ostrovnaya, I., Seshan, V. E. & Begg, C. B. (2015), 'Using somatic mutation data to test tumors for clonal relatedness', *Annals of Applied Statistics* **9**(3), 1533–1548.

Palazzo, M., Beauseroy, P. & Yankilevich, P. (2019), 'A pan-cancer somatic mutation embedding using autoencoders', *BMC Bioinformatics* **20**.

Pantsar, T., Rissanen, S., Dauch, D., Laitinen, T., Vattulainen, I. & Poso, A. (2018), 'Assessment of mutation probabilities of kras g12 missense mutants and their long-timescale dynamics by atomistic molecular simulations and markov state modeling', *PLoS Computational Biology* **9**(14).

Pavlidis, N. & Khaled, H. (2015), 'A mini review on cancer of unknown primary site : A clinical puzzle for the oncologists', *Journal of Advanced Research* **6**(3), 375–382.
URL: *http://dx.doi.org/10.1016/j.jare.2014.11.007*

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Salvadores, M., Masponte, D. & Supek, F. (2019), 'Passenger mutations accurately classify human tumors', *PLOS Computational Biology* **15**(4).

Schaefer, M. H. & Serrano, L. (2016), 'Cell type-specific properties and environment shape tissue specificity of cancer genes', *Scientific Reports* (February), 1–14.
URL: *http://dx.doi.org/10.1038/srep20707*

Schneider, G., Schmidt-Supprian, M., Rad, R. & Saur, D. (2018), 'Europe PMC Funders Group Tissue-specific tumorigenesis – Context matters', *Nat Rev Cance* **17**(4), 239–253.

Scholl, C. & Fro, S. (2019), 'Exploiting rare driver mutations for precision cancer medicine', *Current Opinion in Genetics and Development* **54**, 1–6.

Soh, K. P., Szczurek, E., Sakoparnig, T. & Beerenwinkel, N. (2017), 'Predicting cancer type from tumour DNA signatures', *Genome Medicine* **9**(104), 1–11.

Stratton, M. R., Campbell, P. J. & Futreal, P. A. (2009), 'The Cancer Genome', *Nature* **458**(7239), 719–724.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. (2005), 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences* **102**(43), 15545–15550.
URL: *https://www.pnas.org/content/102/43/15545*

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., Stefancsik, R., Thompson, S. L., Wang, S., Ward, S., Campbell, P. J. & Forbes, S. A. (2019), 'COSMIC: the Catalogue of Somatic Mutations in Cancer', *Nucleic Acids Research* **47**(D1), 941–947.

Varadhachary, G. R. & Raber, M. N. (2014), 'Cancer of Unknown Primary Site', *New England Journal of Medicine* **371**(8), 757–765.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Jr, L. A. D. & Kinzler, K. W. (2013), 'Cancer Genome Landscapes', *Cancer Genomics* **339**(March), 1546–1558.

Wan, J. C. M., Massie, C., Garcia-Corbacho, J. & Mouliere, F. (2017), 'Liquid biopsies come of age : towards implementation of circulating tumour DNA', *Nature Publishing Group* **17**, 223–238.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
URL: *https://ggplot2.tidyverse.org*

Young, A., Chmura, J., Park, Y., Morris, Q., Atwal, G., Classification, T. & Programming, D. (2020), Genome Gerrymandering: optimal division of the genome into regions with cancer type specific differences in mutation rates, *in* 'Pacific Symposium on Biocomputing', pp. 274–285.

Yuan, Y., Shi, Y., Li, C., Kim, J., Cai, W., Han, Z. & Feng, D. D. (2016), 'DeepGene : an advanced cancer type classifier based on deep learning and somatic point mutations', *BMC Bioinformatics* **17**, 243–303.
URL: *http://dx.doi.org/10.1186/s12859-016-1334-9*