EXPLORING THE GENETIC CHARACTERISTICS UNDERLYING A MULTIDIMENSIONAL LATENT CHEMOTHERAPY SYMPTOM BURDEN

by

Winston W. H. Eng

BSE, Case Western Reserve University, Cleveland, Ohio, 2016

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Winston W. H. Eng

It was defended on

April 13th 2018

and approved by

Daniel E. Weeks, PhD, Professor, Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh

Yan Lin, PhD, Research Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Stewart J. Anderson, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Thesis Advisor: Daniel E. Weeks, PhD, Professor, Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh

Daniel E. Weeks, PhD

EXPLORING THE GENETIC CHARACTERISTICS UNDERLYING A MULTIDIMENSIONAL LATENT CHEMOTHERAPY SYMPTOM BURDEN

Winston W. H. Eng, MS

University of Pittsburgh, 2018

ABSTRACT

The incidence rate of cancer is expected to increase within the coming decades. While related mortality is expected to decrease with improving treatments, oncology patients are still expected to experience harmful physiological and psychological symptoms. This "symptom burden" has been shown to permanently extend into the patients' lives, and researchers have hypothesized that a genetic component may dictate the severity of its presentation. From a public health perspective, sustaining an expanding concentration of those considered "symptom burdened" will create unsustainable stress on the current healthcare system. Hoping to describe the heterogeneity in this oncology experience, researchers have relied on statistical clustering to generate patient subgroups differing in quality-of-life. This study's objective is to assess if there exists any association between Single Nucleotide Polymorphisms (SNP) and oncology "symptom burden" following subcategorization; more specifically, it aims to compare analyses of the latent class phenotypes using the "default" method of Multinomial Logistic Regression with those of the "novel" method of Dirichlet Regression.

A four category latent class was generated while adjusting for site from symptom clusters measured on 2111 subjects from sites $UCSF_{total}$ and TOR_1. Genotyping occurred using two versions of the Illumina exome chip: `HumanCoreExome-24v1-0` (Group A) and `HumanExome-12v1-1_A` (Group B). Group A had 944 $UCSF_{total}$ individuals, while Group B had 669 $UCSF_{total}$ and 498 TOR_1 subjects. Following quality control, there were 1272 and 415 for the $UCSF_{total}$ and TOR_1 cohorts respectively. Covariates included within the re-

gression models were total number of comorbidities, Karnofsky Performance Status (KPS), sex, and the first four principal components for population substructure.

After applying both the Multinomial Logistic Regression and Dirichlet Regression approaches, neither method demonstrated statistically significant genetic association ($P < 5 \times 10^{-8}$). However, issues concerning SNPs with very low minor allele frequencies appeared to plague both approaches, indicating that methodological corrections may be necessary in future studies. Additionally, covariate selection, sample size, and imputation may be areas of future inquiry with regards to rectifying some of the issues presented. Future aims should involve simulation and power calculations to determine how appropriate these proposed methods are for assessing genetic association in relation to oncology "symptom burden."

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I wanted to graciously thank my advisor, Dr. Daniel E. Weeks for providing me the opportunity to learn about the wonders of genetics; working as his research associate afforded me the opportunities and skills necessary to tackle this project. Additionally, I would like to thank Dr. Yan Lin and Dr. Stewart J. Anderson for serving on my proposal and offering incredibly helpful advice during its process. Lastly, I wish to acknowledge Dr.Christine Miaskowski for providing the data and necessary background information to conduct my analysis and Dr. Bruce Cooper for providing assistance with the latent class analysis.

1.0 INTRODUCTION

1.1 SPECIFIC AIMS

From their disease and subsequent treatment plans, oncology patients often report experiencing a sweeping number of psychological and physiological symptoms. Previous studies have assessed this extensive subject-level variability and have reported various subgroups each representing different levels of a patient's quality-of-life. Past methods have focused on latent class analysis on the patient symptom burden. This method, while capable of creating patterns of association in the symptoms, has relied on a maximum likelihood estimation to calculate the probabilities associated with a particular set of latent classes. Each individual case is subsequently assigned to the latent class with the highest probability associated with it.

This project has provided experience working with this latent class analysis method in relation to a Genome-Wide Association Study (GWAS). In essence, the patients in our study have been stratified into different response level groups via latent class analysis based on their qualitative symptom traits during chemotherapy. My focus has been to see if there are any genetic differences amongst the different response groups as demonstrated by single nucleotide polymorphisms (SNPs).

As compared to my previous work, my proposal seeks to change the method in which these different response groups are assigned. Rather than condense each case to a single latent class based on the highest probability assignment, we propose using Dirichlet component regression to consider the probabilities assigned to the set of latent classes as a probability simplex. Subsequently, we aim to perform a GWAS to assess whether there are any SNPs

associated with differences in these multidimensional phenotypes derived from the set of qualitative symptom traits.

1.2 CANCER PATIENT SYMPTOM BURDEN

Cancer has historically been linked to high mortality rates worldwide. Within the United States, it is the second leading cause of death with a projected 600,000+ loss of life in 2018 [Siegel et al., 2018]. Increased awareness of risk factors, novel and more efficacious treatments, and improved diagnostic testing have led to an overall 26% decrease in cancer mortality over the past two decades as well as an significant increase in the 5-year relative survival rate [Siegel et al., 2018]. However from their disease and subsequent treatment plans, oncology patients often report experiencing a sweeping number of physical, psychological, and temporal burdens to their everyday lives. Reports of fatigue, pain, and nausea/vomiting as well as increased time needed to address these and other side effects demonstrate that the "burden" cancer brings upon the patient extends beyond the disease itself [Henry et al., 2008, Harrington et al., 2010]. An international study of over 29 epidemiological cancer-burden-related studies found a reported pain prevalence of at least 14%, with the majority of studies reporting a major depressive disorder rate of between 10-25% (25% of studies had lower rates, 17% had higher). Additionally, these symptoms do not occur exclusively in isolation; cancer-related fatigue, for instance, was seen to be correlated with psychological symptoms such as depression [Carr et al., 2002]. Most alarmingly, patients have reported the assumption that pain is an "inevitable part of dealing with cancer", which has prevented them from reporting or seeking more intensive treatment. Moreover, systemic barriers such as lack of coordination especially "during the transition from cure to hospice mode" have hindered patients during the entire symptom management cycle [Patrick et al., 2003].

From a patient's perspective, the term "symptom burden" can defined as "a loss of functional abilities along with psychological suffering, both of which are affected by the impact of severe symptoms" [Gill et al., 2012]. Previously, advanced cancer individuals suffering from an intense "symptom burden" have been reported to be associated with increased hospital

stay length and subsequent unplanned hospital readmission within 90 day [Nipp et al., 2017]. Moreover, it is often seen that this "symptom burden" extends post-treatment and significantly impacts survivors for the rest of their lives. Reports of increased quality of life (QOL) without "symptom burden" improvement has implied that patients often must compromise their lifestyles to work around the limitations set by their illness [Yang et al., 2012].

1.3 PUBLIC HEALTH IMPACT

Knowing the reality of how detrimental cancer and its "symptom burden" is the first step in acknowledging the greater public health implications. From 2010 to 2030, the overall US cancer incidence rate is projected to rise by approximately 45% with the aging US population [Smith et al., 2009]. By 2022, 18 million patients are expected to be cancer survivors. [De Moor et al., 2013] As technology develops and life expectancy increases, sustaining an increased survivorship will become an increasingly pressing public health concern. If issues surrounding physical, mental, and social challenges are not directly addressed, the current healthcare system will experience unsustainable stress. Through studying and identifying genetic components via a Genome-Wide Association study, it may be possible to discover novel biomarker targets related to this concerning cancer-derived "symptom burden." Moreover, another possibility would be to create a predictive model capable of accepting a subject's information and determining which specific "treatment group" would be most appropriate. As opposed to receiving the default treatment plan available to her, an assigned patient would have the opportunity for more specialized care designed to be more efficacious for her condition. Lastly, contributing to this research could allow us to further strengthen the forthcoming "personalized medicine" approach leading to not only safer treatments but also increased efficiency and productivity with regards to developing these therapeutics [Ginsburg and McCarthy, 2001].

1.4 ASSESSING CANCER SUBGROUPS

It is imperative to highlight that "symptom burden" is not experienced equally by all oncology patients. Previous assessments have concluded that tailored treatments either to individual or cancer type may be more accurate in managing symptom burden [Deshields et al., 2014]. Past methods have relied on analysis of latent variables as a way of subcategorizing patients based on their symptom burdens. From techniques such as latent class analysis, studies have been able to identify clusters of patients who are at a "higher risk for multiple co-occurring symptoms and diminished Quality-of-Life [QoL]" as well as highlight demographical differences amongst gender, ethnicity, and age [Astrup et al., 2017, Miaskowski et al., 2014, Miaskowski et al., 2015].

Though Latent Class Analysis (LCA) is capable of creating patterns of association from qualitative symptom traits, its methods are fallible under certain conditions. Via maximum likelihood estimation, LCA calculates the probabilities associated with a particular set of latent classes and assigns each individual case to the latent class with the highest associated probability. As a result, there is no assumed difference between individuals who have a more nuanced differences amongst the probability simplex and those subjects who are more clear cut. For instance, consider the example:

Table 1: **Latent Class Example with Posterior Probabilities**

	Subject I	**Subject II**
	Posterior Probabilities	Posterior Probabilities
Latent Class I	0.1	0.3
Latent Class II	0.1	0.3
Latent Class III	0.8	0.4

In this case, both subjects (I & II) will be assigned to "Latent Class III". As a result, it may be possible to consider that there exists some heterogeneity within and amongst the latent class groups instead of clear-cut separation. This may especially be a major concern in

situations where LCA is used to cluster individuals before running a genome-wide association study.

1.5 OBJECTIVE

Rather than condense each case to a single latent class based on the highest probability assignment, the primary objective of this study revolves around using a Dirichlet Component Regression method to consider the entirety of the probability simplex when detecting genetic association. Formally, this study aims to answer the following question:

How do the dirichlet-derived phenotypes perform compared to those of the most-likely latent class phenotypes when evaluating association?

2.0 METHODS

2.1 STUDY DESIGN

2.1.1 DATA DESCRIPTION

All subjects were 18 years or older and had a diagnosis of breast, gastrointestinal, gynecological, or lung cancer. Each subject had already received treatment in the form of Cyclophosphamide (CTX), a standard chemotherapy, at least four weeks prior to taking part in the study [Miaskowski et al., 2017]. Each patient filled out a demographical questionnaire alongside a self-administered comorbidity questionnaire. Topics including but not limited to age, ethnicity, gender, marital status, living arrangements, education, employment status, income status, and comorbid status (occurrence, treatment, functional impact). To further determine functional status, patients filled out a Karnofsky performance status (KPS) scale evaluation [Mor et al., 1984, Yates et al., 1980]. Additionally, subjects self-reported on the Memorial Symptom Assessment Scale (MSAS) questionnaire, demonstrating certain symptom occurrence within the past week. [Portenoy et al., 1994] Finally, to evaluate quality of life (QOL), subjects underwent the Medical Outcomes Study-Short Form-12 and Quality of Life Scale-Patient Version [QOL-PV] for general and disease-specific measures respectively [Ware Jr et al., 1996, Ferrell et al., 1989].

In total, there were two separate cohorts from here on referenced as UCSF and TOR1. The UCSF group involved 1343 patients gathered from an ongoing longitudinal symptoms experience study in the San Francisco Bay Area, while the TOR1 group included 534 individuals gathered from Norway. Both groups have been previously assessed in past symptom

cluster studies [Astrup et al., 2017, Miaskowski et al., 2014, Miaskowski et al., 2015, Miaskowski et al., 2017].

Before recruiting and obtaining written informed consent from subjects, Miaskowski, *et al* received approval from the related Human Subject Committees. More intensive details regarding the recruitment process may be found in a previous study [Illi et al., 2012].

2.1.2 GENOTYPE INFORMATION

The Johns Hopkins University Genetic Resources Core Facility (GRCF) SNP Center performed the genotype calling for the Illumina Exome Chip Data which was split into two different groups from here on referenced as Group A and Group B. Group A had 944 subjects and their SNPs associated with the `HumanCoreExome-24v1-0` array and hg19 Genome Build; Group B had 2,330 subjects and their SNPs associated with the `HumanExome-12v1-1_A` array and hg19 Genome Build. The UCSF cohort was split between Group A and Group B, while the TOR1 cohort was only in Group B (Table 2). For ease of nomenclature, we will call the UCSF subgroups as "UCSF$_A$" and "UCSF$_B$" for Group A and Group B UCSF subjects respectively and leave "TOR1" as self-referential since it was only present in a Group B. For analyses requiring the entire UCSF cohort ($UCSF_A + UCSF_B$), it shall be referred as $UCSF_{total}$.

Additionally, outside of the UCSF and TOR1 cohorts were individuals who had been previously collected from the same sample population as well as assessed in other studies

Table 2: **Distribution of UCSF and TOR1 Samples as Separated by Group**

	$UCSF_{total}$		TOR1	
Group Assignment	Group A	Group B	Group A	Group B
Short-hand Reference	UCSF$_A$	UCSF$_B$	NA	TOR1
Pre-Filtering Count	944	669	0	498
Post-Filtering Count	730	542	0	415

[Miaskowski et al., 2014, Miaskowski et al., 2015]. These individuals were included to increase the sample size and subsequently make the genotype calling more accurate, a technique often used in population genetic studies [Fumagalli, 2013]. GRCF's calling algorithm relied on GenomeStudio version 2011.1, Genotyping Module version 1.9.4, and GenTrain Version 1.0.

2.1.3 DATA STORAGE

All data is securely stored in the 'Gattaca' cluster currently under Dr. Weeks's supervision.

2.1.4 DATA CLEANING

The "GWASTools" package was utilized to clean the data before any formal analysis was attempted [Gogarten et al., 2012]. Each dataset, A and B, underwent a filtering pipeline. Initially, subjects who had missing call rates over all SNPs $\geq 3\%$ (missing.e2) were excluded. To determine discrepancies between genetic and annotated sex, mean allelic intensities of SNPs on both the X and Y chromosomes were assessed. Females with low X intensities and males with low Y intensities were assumed to be the result of potential sample misidentification and were not retained. Additionally, relatedness was assessed. The given pedigree structure insinuated that subjects in this study were not related; however, from the genotype data, individuals who were determined to be duplicates or familial were excluded. As both Groups A and B included subjects from multiple studies, it was the case that there were, in fact, individuals who were in multiple studies. In this section of the data cleaning stage, those with the lowest missingness were retained, while their duplicate counterparts were excluded. B Allele Frequency (BAF) and Log R Ratio were assessed to determine if there were any chromosomal aberrations such as duplications or (partial and full) deletions. Additionally, subjects who had an annotated missing gender were excluded. Finally, subjects with BAF values greater than 5mb (indicating a missingness greater than 5,000,000 bases) were dropped as well.

2.1.5 GENETIC IMPUTATION

Genetic Imputation was performed using the Minimac3 imputation engine available via the Michigan Imputation Server [Das et al., 2016]. For chromosomes 1-22, the Haplotype Research Consortium (HRC) Version r1.1 2016 was the reference panel and Eagle v2.3 was used for Phasing. The X chromosome utilized the same HRC reference panel; however, it required the ShapeIT v2.r790 (unphased) option instead of Eagle v2.3. All data was secured via Advanced Encryption Standard (AES) 256 encryption.

2.1.6 LATENT CLASSES DERIVATION

For $UCSF_{total}$ and TOR_1 groups, latent class analysis (LCA) was performed on their common 22 qualitative traits as defined in Table 3. All variables were chosen from the the 32 total variables found within the MSAS questionnaire as each had a rate of occurrence $\geq$ 40% within both the UCSF and TOR1 subjects. Additionally, each was dichotomous; patients responded with either a "Yes" or "No" to indicate the presence of absence of the trait as a part of their symptom burdens.

Initially, Dr. Bruce Cooper, a collaborator from University of California, San Francisco, performed the LCA using the same methodology and MPlus software as previously conducted in past publications regarding symptom cluster burden and latent classes [Miaskowski et al., 2014, Miaskowski et al., 2015, Muthén and Muthén, 2012]. In this analysis, the latent classes were estimated using the combined UCSF and TOR1 data while adjusting for site, thereby allowing latent class phenotypes to be equivalent in definition. Additionally, we provided quality assurance by verifying the phenotypes ourselves; this was conducted via the 'poLCA' R package [Linzer and Lewis, 2011].

poLCA initially maximizes the log-likelihood function of each specified latent class model via the expectation-maximization (EM) algorithm with a Newton-Ralphson step. After deleting cases where there are missing observations for the predictor variables, the function then esimates the latent class regression. Initially, the LCA regression was run with up to a max of 10 possible latent classes; the poLCA algorithm additionally calculated each model's associated Bayesian Information Criteria (BIC) and Akaike Information Criteria (AIC).

Table 3: **The 22 Qualitative Traits Utilized in Latent Class Analysis**

	Variable	Definition
1	dfcno	Difficulty Concentrating
2	paino	Pain
3	enrgo	Lack of Energy
4	cougho	Cough
5	nrvso	Feeling Nervous
6	drymo	Dry Mouth
7	nauso	Nausea
8	drwsyo	Feeling Drowsy
9	numbo	Numbness or Tingling in Hands or Feet
10	dfslpo	Difficulty Sleeping
11	bloto	Feeling Bloated
12	sado	Feeling Sad
13	swtso	Sweats
14	wryo	Worrying
15	sxlo	Problems with Sexual Interest or Activity
16	aptito	Lack of Appetite
17	dizzyo	Dizziness
18	irrito	Feeling Irritable
19	hrlso	Hair Loss
20	cnstpo	Constipation
21	tasto	Change in the Way Food Tastes
22	myslfo	I Do Not Look Like Myself

Subjects responded with either a "Yes" or "No" indicating the presence of the trait as a part of their symptom burdens.

However, one thing to note is that despite LCA models with different number of classes being technically considered nested models, it is not appropriate to utilize a likelihood ratio test to distinguish statistically significant differences due to the failure of meeting the regularity conditions [Nylund et al., 2007]. Computing the difference in likelihood between a model with K classes and K-1 classes does not, in fact, follow a χ^2 distribution [McLachlan and Peel, 2000]. However, [Lo et al., 2001] have built upon the work of [Vuong, 1989]to provide an alternative method which approximates the likelihood ratio test distribution and can be used for nested latent class regression models. Unlike the LRT, the VLMR test is capable of discerning whether adding an additional class within the model will lead to a statistically significant improvement [Nylund et al., 2007]. Therefore, in his analysis, Dr. Cooper utilized the the VLMR test and determined that the 4 Latent Class Regression Model was most appropriate for our dataset.

2.2 LATENT CLASS ANALYSIS

2.2.1 LATENT VARIABLE

A latent variable is defined as a variable that is not directly observed but rather inferred from existing observations. Latent Class models utilize categorical variables to determine mutually exclusive subgroups (or latent classes) that theoretically comprise a population. For "g" latent classes, an individual has a set of "g" assigned posterior probabilities. After the probability calculations, the subject is assigned to the latent class with the highest posterior probability.

2.2.2 LATENT CLASS MODEL

Given that in the actual analysis, it was inferred that there were four latent classes from 22 separate categorical variables, we will now detail the Latent Class Model asumming there are four classes. First, assume that there are 22 categorical variables $(A_1, A_2, ..., A_{21}, A_{22})$ consisting of $(C_1, C_2, ..., C_{21}, C_{22})$ classes respectively. [Goodman, 1974] Let $\pi_{(c_1 ... c_{22})}$ denote

the probability that an individual will be at level $(c_1, ..., c_{22})$ with respect to the joint variable $(A_1, A_2, ..., A_{21}, A_{22})$ where $(c_1 = 1, ..., C_1; c_2 = 1, ..., C_2; ...; c_{21} = 1, ..., C_{21}; c_{22} = 1, ..., C_{22})$. Suppose there exists a latent polygamous variable X, consisting of 4 classes that can explain the relationships among manifest variables $(A_1, A_2, ..., A_{21}, A_{22})$. As demonstrated in [Goodman, 1974], we can define the following:

$$\pi_{(c_1...c_{22})} = \sum_{t=1}^{4} \pi_{(c_1...c_{22})t}^{(A_1...A_{22})X} \tag{2.1}$$

where

$$\pi_{(c_1...c_{22})t}^{(A_1...A_{22})X} = \pi_t^X \prod_{w=1}^{22} \left(\pi_{c_w t}^{A_w|X} \right) \tag{2.2}$$

In this case, we can consider $\pi_{(c_1...c_{22})t}^{(A_1...A_{22})X}$ to be the probability that the individual will be at level $(c_1, ..., c_{22}, t)$ with respect to the joint variable $(A_1, ..., A_{22}, X)$. In the definition of $\pi_{(c_1...c_{22})t}^{(A_1...A_{22})X}$ in Equation (2.2), π_t^X denotes the probability that an individual will be at level t with respect to variable X. Additionally, $\pi_{c_1 t}^{A_1|X}$ represents the conditional probability that an individual will be at level c_1 with respect to variable A_1, given that he is at level t with respect to variable X. $\pi_{c_2 t}^{A_2|X}$,..., $\pi_{c_{22} t}^{A_{22}|X}$ represent similar conditional probabilities. Equation (2.1) demonstrates that all individuals can be classified into 4 mutually exclusive and exhaustive latent classes, while Equation (2.2) explains that within the t-th latent class, the manifest variables of $A_1...A_{22}$ are mutually independent.

Additionally, consider the following formulae:

$$\sum_{t=1}^{T} \pi_t^X = 1 \tag{2.3}$$

$$\forall i \in \{1, ..., 22\}, \quad \sum_{c_i=1}^{T} \pi_{c_i t}^{A_i|X} = 1 \tag{2.4}$$

$$\pi_t^X = \sum_{c_1,...,c_{22}} \pi_{(c_1...c_{22})t}^{(A_1...A_{22})X} \tag{2.5}$$

$$\pi_t^X \pi_{c_1 t}^{A_1|X} = \sum_{c_2,...,c_{22}} \pi_{(c_1...c_{22})t}^{(A_1...A_{22})X} \tag{2.6}$$

From (2.3), (2.4), and (2.5), (2.6) can be generated for A as (2.6) has been summed over c_1. Similar equations can be generated for $\pi_t^X \pi_{c_2 t}^{A_2|X}$,..., $\pi_t^X \pi_{c_{22} t}^{A_{22}|X}$.

$$\pi_{(c_1 \ldots c_{22})t}^{(A_1 \ldots A_{22})|X} = \frac{\pi_{(c_1 \ldots c_{22})t}^{(A_1 \ldots A_{22})X}}{\pi_{c_1 \ldots c_{22}}} \tag{2.7}$$

Let $\pi_{(c_1 \ldots c_{22})t}^{(A_1 \ldots A_{22})|X}$ describe the conditional probability that an individual is in latent class t, given that he was at level $(c_1, ..., c_{22})$ with respect to the joint variable $(A_1, ..., A_{22})$. Therefore, we can rewrite (2.5) and (2.6) as:

$$\pi_t^X = \sum_{c_1, \ldots, c_{22}} \pi_{c_1 \ldots c_{22}} \pi_{(c_1 \ldots c_{22})t}^{(A_1 \ldots A_{22})|X} \tag{2.8}$$

and

$$\pi_{c_i t}^{A_1|X} = \frac{\sum_{(c_2, \ldots, c_{22})} \pi_{c_1 \ldots c_{22}} \pi_{(c_1 \ldots c_{22})t}^{(A_1 \ldots A_{22})|X}}{\pi_t^X} \tag{2.9}$$

respectively. Equations similar to (2.9) can be obtained for $\pi_{c_2 t}^{A_2|X}$,...,$\pi_{c_{22} t}^{A_{22}|X}$.

Additionally, by letting $\pi_t^X > 0$ and $\pi_{c_1 \ldots c_{22}} > 0$, we can determine the maximum likelihood estimates of the latent-class model specified above. From (2.1) and (2.2), let $\hat{\pi}_{c_1 \ldots c_{22}}$ be defined as:

$$\hat{\pi}_{c_1 \ldots c_{22}} = \sum_{t=1}^{T} \hat{\pi}_{(c_1 \ldots c_{22})t}^{(A_1 \ldots A_{22})X} \tag{2.10}$$

where

$$\hat{\pi}_{(c_1 \ldots c_{22})t}^{(A_1 \ldots A_{22})X} = \hat{\pi}_t^X \prod_{i=1}^{22} \hat{\pi}_{c_i t}^{A_i|X}, \tag{2.11}$$

and from (2.7)

$$\hat{\pi}_{(c_1 \ldots c_{22})t}^{(A_1 \ldots A_{22})|X} = \frac{\hat{\pi}_{(c_1 \ldots c_{22})t}^{(A_1 \ldots A_{22})X}}{\hat{\pi}_{c_1 \ldots c_{22}}} \tag{2.12}$$

If $\mathbf{p_{c_1 \ldots c_{22}}}$ is defined as the observed proportion of individuals at level $(c_1, ..., c_{22})$ with respect to the joint variable $(A_1, ..., A_{22})$, then the maximum likelihood estimates must satisfy the following system of equations:

$$\hat{\pi}_t^X = \sum_{c_1, \ldots, c_{22}} p_{c_1 \ldots c_{22}} \hat{\pi}_{(c_1 \ldots c_{22})t}^{(A_1 \ldots A_{22})|X}, \tag{2.13}$$

$$\hat{\pi}_{c_1 t}^{A_1|X} = \frac{\sum_{c_2,\ldots,c_{22}} p_{c_1\ldots c_{22}} \hat{\pi}_{c_1\ldots c_{22} t}^{(A_1\ldots A_{22})|X}}{\hat{\pi}_t^X}, \tag{2.14}$$

$$\hat{\pi}_{c_2 t}^{A_2|X} = \frac{\sum_{c_1,c_3,\ldots,c_{22}} p_{c_1\ldots c_{22}} \hat{\pi}_{c_1\ldots c_{22} t}^{(A_1\ldots A_{22})|X}}{\hat{\pi}_t^X}, \tag{2.15}$$

$$\vdots$$

$$\hat{\pi}_{c_{21} t}^{A_{21}|X} = \frac{\sum_{c_1,\ldots,c_{20},c_{22}} p_{c_1\ldots c_{22}} \hat{\pi}_{c_1\ldots c_{22} t}^{(A_1\ldots A_{22})|X}}{\hat{\pi}_t^X}, \tag{2.16}$$

$$\hat{\pi}_{c_{22} t}^{A_{22}|X} = \frac{\sum_{c_1,\ldots,c_{21}} p_{c_1\ldots c_{22}} \hat{\pi}_{c_1\ldots c_{22} t}^{(A_1\ldots A_{22})|X}}{\hat{\pi}_t^X}, \tag{2.17}$$

In order to determine the maximum likelihood estimate $\hat{\pi}$, let π act as the vector comprised of the parameters of the latent class model $(\hat{\pi}_t^X, \hat{\pi}_{c_1 t}^{A_1|X}, \ldots, \hat{\pi}_{c_{22} t}^{A_{22}|X})$. Initially, starting with $\pi(0)$, which in turn is equivalent to $\{\hat{\pi}_{c_1 t}^{A_1|X}(0), \ldots, \hat{\pi}_{c_{22} t}^{A_{22}|X}(0)\}$, it is possible to determine the value for (2.11), (2.12), and subsequently (2.13) through (2.17). This process can be performed iteratively with different trial values, and when an estimate is equal to zero, the corresponding latent class is dropped. By assessing different initial values for $\hat{\pi}$, we can determine which solution minimizes the chi-squared statistic based upon the likelihood ratio

$$\chi^2 = 2 \sum_{c_1,\ldots,c_{22}} f_{A_1\ldots A_{22}} \log\left(\frac{f_{c_1\ldots c_{22}}}{\hat{F}_{c_1\ldots c_{22}}}\right) \tag{2.18}$$

where

$$f_{c_1\ldots c_{22}} = n p_{c_1\ldots c_{22}}, \quad \hat{F}_{c_1\ldots c_{22}} = n \hat{\pi}_{c_1\ldots c_{22}}, \tag{2.19}$$

with $\hat{\pi}_{c_1\ldots c_{22}}$ coming from (2.10) and n being the observed number of cases. Whichever trial value leads to the lowest value in (2.18) will have yielded the maximum likelihood estimate $\hat{\pi}$.

2.2.3 CLASS COMPARISON METHOD

It is necessary to select the fewest number of latent classes capable of most accurately explaining the relationships between the observed variables. While the Akaike Information

Criterion and Bayesian Information Criterion were calculated and evaluated for each of the latent class models, the VLMR test, as specified in Subsection 2.1.6, was actually utilized to determine which latent class regression model would be most appropriate.

2.3 DIRICHLET REGRESSION

2.3.1 DIRICHLET DISTRIBUTION

Let a $(k-1)$ dimensional probability simplex be a surface in R^k space with k non-negative components which sum to 1. As each value in the probability simplex is bounded between 0 and 1, they can be equivalently thought of as probability mass functions (pmfs). The Dirichlet distribution is defined as the probability distribution of these pmfs, effectively making it a distribution of distributions.

Moreover, the Dirichlet distribution is a generalization of the beta distribution beyond two dimensions.

For $\{W_1, ..., W_k\}$ where $0 < W_i < 1$ and $\sum W_i = 1$ for all the {i = 1,...,k}, the density function is defined as

$$D(w|\alpha) = f(w_1, ..., w_k|\alpha_1, ..., \alpha_k) = \frac{\Gamma\left(\sum(\alpha_i)\right)}{\prod \Gamma(\alpha_i)} \prod_i w_i^{a_i - 1} \tag{2.20}$$

with parameters $\{a_1, ..., a_k\} > 0$ for all $\{i = 1, ..., k\}$.

It can be thought of a distribution comprised of multinomials as $\sum W_i = 1$ for all $\{i = 1, ..., k\}$, and it is the conjugate prior of the multinomial distribution.

2.3.2 DIRICHLET REGRESSION

From (2.20), we can determine the full log-likelihood of the model:

$$l_i(w|\alpha) = \log \Gamma\left(\sum_{i=1}^{k} (\alpha_i)\right) - \sum_{i=1}^{k} \log \Gamma\left(\alpha_i\right) + \sum_{i=1}^{k} (\alpha_i - 1) \log\left(w_i\right) \tag{2.21}$$

However, given that w_i may take any value [0,1], a correction proposed by Smithson and Verkuilen (2.22) must be applied in order to account for data taking a value of 0 or 1 e.g. given $w_i = 0$, $log(w_i) = -\infty$, while $w_i = 1$, $log(w_i) = 1$.

$$w^* = \frac{w(N-1) + \frac{1}{k}}{N} \tag{2.22}$$

N, in this case, represents the number of observations within the dataset. Additionally, this correction constricts the data symmetrically around 0.5. As $N \to \infty$, the restriction increasingly relaxes indicating that larger datasets may be less affected by this correction.

The resulting link function $g(w_i)$ can be defined as the combination of the predictor matrix $X^{[i]}$ and the matrix comprising of the regression coefficients in each dimension $\beta^{[i]}$.

$$g(w_i) = X^{[i]}\beta^{[i]} \tag{2.23}$$

2.3.3 DIRICHLET REGRESSION INTERPRETATION

Espin-Garcia (2014) [Espin-Garcia et al., 2014] initially suggested a likelihood model to relate the genetic information and response variable from a Dirichlet Regression:

$$L = \prod_{i=1}^{n} \left[\gamma(\Lambda(\mathbf{s}_i) \prod_{k=1}^{4} \frac{y_{ij}^{\lambda_i(s_i)-1}}{\lambda_j(\mathbf{s}_i)} \right] \tag{2.24}$$

where $\lambda_j(\mathbf{s}_i) = \lambda_{ij} > 0, \quad \Lambda(\mathbf{s}_i) = \Lambda_i = \sum_{j=1}^{4} \lambda_j(\mathbf{s}_i)$ and $\Gamma(\cdot)$ represents the gamma function.

Additionally, let $\lambda_j(\mathbf{s}_i)$ be defined using a logarithm link,

$$\log(\lambda_j(\mathbf{s}_i)) = \log(\lambda_{ij}) = \sum_{m=1}^{M} \beta_{jm} s_{im} = \mathbf{s}_i \beta_j \tag{2.25}$$

where $j = 1, 2, 3, 4$ or the number of components within the probability simplex, M is the number of covariates within the model, and B_j is the vector or regression coefficients that account for the effects of the covariates on the j^{th} component in the log scale.

Therefore, two models (additive and adjusted) can be constructed.

Model 1: $\log(\lambda_{ij}) = \alpha_j^{M1} + \beta_j^{M1} g_i^k$ (additive)

Model 2: $\log(\lambda_{ij}) = \alpha_j^{M2} + \beta_j^{M2} g_i^k + \text{FAM}_i \delta_j^{M2}$ (adjusted)

For the additive model, g_i^k is the number of minor allele copies of the k^{th} SNP for the i^{th} individual. FAM_i describes the i^{th} row of the contrast matrix for the pedigree number, and $\theta_j^h = (\alpha_j^h, \beta_j^h, \delta_j^{h^t})^t$ is the vector of regression coefficients on the j^{th} component.

From these models, we could construct a series of Wald tests to evaluate the two-sided null hypothesis of no association between individuals SNPs and the response or more formally:

$H_0 : \beta = 0; \quad H_A : \beta \neq 0$ where $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$

2.3.4 MODEL COMPARISON METHOD

To test the differences between two consecutive models, a likelihood ratio test can be utilized. For example, a full model D and its nested model W will have likelihood values of L_D and L_W respectively.

A standard likelihood-ratio test (LRT) can then be constructed to test model W against D:

$$LRT_{DW} = -2log\left(\frac{L_W}{L_D}\right) = -2\left(l_W - l_D\right), \qquad \text{where} \quad LRT_{DW} \sim \chi^2_{n_D - n_W} \tag{2.26}$$

where n_D and n_W represent the different model parameters within models D and W respectively. In this case, the test statistic follows the asymptotic chi-squared distribution with $n_D - n_W$ degrees of freedom under the null hypothesis.

2.4 SNP ASSOCIATION

2.4.1 COVARIATE SELECTION

As part of the SNP association modeling process includes adjusting for covariates. From the data, there were a total of 11 variables including: (1) age, (2) body mass index (BMI), (3) Karnofsky Performance Status (KPS), (4) the number of comorbidities, (5) the time from cancer diagnosis, (6) Number of Prior Cancer Treatments, (7) Gender, (8) Education Level, (9) Marital Status, (10) Ethnicity, and (11) Cancer Diagnosis. For reproducibility we want

to adjust for some set of covariates in both analyses. From the *glmnet* package, we used a LASSO multinomial regression method with latent class as the outcome [Simon et al., 2011]. In both the $UCSF_{total}$ and TOR_1 groups, the KPS and the number of comorbidities variables were retained. In total, $UCSF_{total}$ contained additional variables after running the cross-validation; these included: sex, the marital status of the individual, the age, and the educational level. It is important to note that the $UCSF_{total}$ cluster was much larger than its TOR_1 counterpart, and while most of the variables asked for and received identical responses, not all variables were as comparable in the dataset we used at the time of analysis. For instance, the definition of "partnered" was a dichotomous option for the $UCSF_{total}$ group (Yes; No) instead of multinomial as it was for the TOR_1 group (Unmarried; Married/Living Together; Divorced; Widow; Separated). Additionally, race, which self-reported, was not included within the analysis. However, it is important to note that there currently does exist a dataset assembled by Dr. Miaskowski which sees the previously discrepant variables as harmonized; any future analysis beyond this work should favor using that complete dataset over what was utilized in this analysis.

2.4.2 PRINCIPAL COMPONENTS

Finally, to account for population substructure, principal components (PCs) were generated. For a study focused on the genetics of the individual, PCs provide a more accurate and appropriate measure of subjects' genetic ancestry due to differences in minor allele frequencies from genetically distant ancestries. Via the GWASTools R package, linkage disequilibrium (LD) pruning was applied in order to determine which SNPs have low levels of LD, missing.n1 < 0.03, and MAF < 0.05 [Gogarten et al., 2012]. Subsequenty, principal components analysis (PCA) was applied on the genotypes and the first 32 eigenvalues were calculated from this subset of SNPs. By including PCA-identified continental ancestry, it is possible to stratify samples by population group via treating the initial four eigenvectors (which account for the majority of the variation seen) as covariates within association analyses [Gogarten et al., 2012]. This additionally provides a solution to allele frequency differences between cases and controls which may arise from systematic ancestry differences [Price et al., 2006].

2.4.3 TRINCULO

Trinculo is an open source, C-based program capable of calculating an omnibus P-value of association for each variant against categorical phenotypes using a likelihood ratio test [Jostins and McVean, 2016]. It requires the "phenotype" (Latent Class Designation), "covariates"(as selected by LASSO), and "bed" (genotype calls at biallelic variants)/"bim" (additional variant information) /"fam" (sample information) files to calculate the multinomial logistic regression.

A multinomial regression approach within genetic association studies considers each individual i as either a control ($w_i = 0$) or one of D different phenotypes ($w_i = d|d \in (1, ..., D)$). The choice of "control" is arbitrary; however, by default, TRINCULO is programmed to chose to label the "control" as the first category it encounters within the data. In order to keep consistency, our analyses referred to the subset of individuals with the lowest "symptom burden" (e.g. the blue class in Figure 2) as the control reference. Additionally, each individual can be considered to have a specific genotype (g_i) at each locus which can fall under the designation of $g_i \in (0, 1, 2)$ depending on the absence, singular presence, or complete presence of the minor allele. By using the multinomial logit function, it is possible to assume that the probability of an individual having a "$w_i = d$" phenotype is related to their genotype at some locus with some β_{0d} and β_{1d} representing the intercept and effect size for each phenotype "$w_i = d$" where d $\in (1, ..., 4)$ respectively [Jostins and McVean, 2016].

$$Pr\left(w_i = d|g_i, \beta_0, \beta_1\right) = \frac{e^{\beta_{0d}+\beta_{1d}\times g_i}}{1 + \sum_{v=1}^{D} e^{\beta_{0v}+\beta_{1v}\times g_i}} \tag{2.27}$$

Subsequently, it is possible to utilize Equation (2.27) generate a formula which can take into account the inclusion of all (j = 7) predictors (sex, KPS, total number of comorbidities, and all four principal components). Additionally, let N be set as the number of given individuals and x_i as the vector of predictors assigned to subject i. With the intercept defined as predictor "0" ($x_{i0} = 1\ \forall$ i), the total number of predictors moves to j = 8. It is then possible to construct a matrix X with elements x_{ij} each describing each subject's predictor values.

Additionally, the effect sizes for all predictors and phenotypes can be set as an "effect size" matrix B, where $B_{jd} = \beta_{jd}$ is the effect size for predictor j on phenotype d. Thus, let

β_d be the vector of effect sizes of all predictors on phenotype d, and β_j as the vector of effect sizes of predictor j on all phenotypes. Equation 2.28 defines the probability that a subject is in category d given predictors x_i and matrix B.

$$Pr\left(w_i = d|B, x_i\right) = \begin{cases} \dfrac{e^{\beta_d^T x_d}}{1 + \sum_{v=1}^{D} e^{\beta_v^T x_i}}, \text{if } w_i > 0 \\ \dfrac{1}{1 + \sum_{v=1}^{D} e^{\beta_v^T x_i}}, \text{if } w_i = 0 \end{cases} \tag{2.28}$$

Finally, omnibus testing can be performed by using a likelihood ratio test between the full model and null model which differentiate based on the inclusion or exclusion of the genotypic effect respectively. Having a full model demonstrating a statistically significant difference when compared to the null model will indicate that the variant is significantly associated with any of the phenotypes [Jostins and McVean, 2016].

2.4.4 META-ANALYSIS

For the meta-analysis, the Z-Transform Test was utilized to combine the p-values from the Group A and Group B UCSF analyses into an overall $UCSF_{total}$ p-value. As the multinomial logistic regression omnibus test relies on the likelihood ratio test which follows a χ^2 distribution, it can be considered a series of one-tailed tests which do not include directional effect. As a standard normal deviate Z can range between -inf to inf and p_i can take any value from 0 to 1, any value of p_i will be bijective with a value of Z. Dubbed "Stouffer's Method" (Z_s), this Z-Transform test is capable of transforming one-tailed p-values (p_i) from k independent tests into standard normal deviates Z_i [Whitlock, 2005]. Furthermore, the sum of the standard normal deviates (Z_i's) divided by the square root of the total number of tests (k) will follow a standard normal distribution under the null hypothesis.

$$Z_s = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \tag{2.29}$$

As opposed to using a Fisher's combined probability test approach, the Z-Transform Test was chosen as it provides increased power and precision. Fisher's method is asymmetrically sensitive to small p-values when compared to larger p-values which can bias results and incorrectly reject the null hypothesis [Whitlock, 2005]. It was utilized to combine the

multinomial logistic regression (TRINCULO) results from both UCSF$_A$ & UCSF$_B$ in order to fully represent the UCSF population.

2.4.5 GWAS COMPARISON

From both the Multinomial Logistic Regression and Dirichlet Regression methods, SNPs of interest were evaluated at three different levels: 1) 5×10^{-8}, 2) 1×10^{-5}, and 3) a conservative Bonferroni-corrected threshold. In a genome-wide association setting, the initial value of 5×10^{-8} is commonly indicative of significant and replicable results, while 1×10^{-5} is "suggestive" [Panagiotou et al., 2011]. For the Multinomial Logistic Regression pipeline, the Bonferroni-corrected values for $UCSF_{total}$ and TOR_1 were 2.651e-06 and 3.607e-07 respectively. For the Dirichlet Regression, the $UCSF_{total}$ cohort was split into $UCSF_A$ and $UCSF_B$; $UCSF_A$, $UCSF_B$, and TOR_1 had Bonferroni-corrected values of 1.973e-07, 2.703e-06, 3.667e-07 respectively. In lieu of a more rigorous comparison which would involve running a simulation to test the power of the Dirichlet Regression method when compared to that of the Multinomial Logistic Regression method, we will solely compare the results and request further study.

3.0 RESULTS

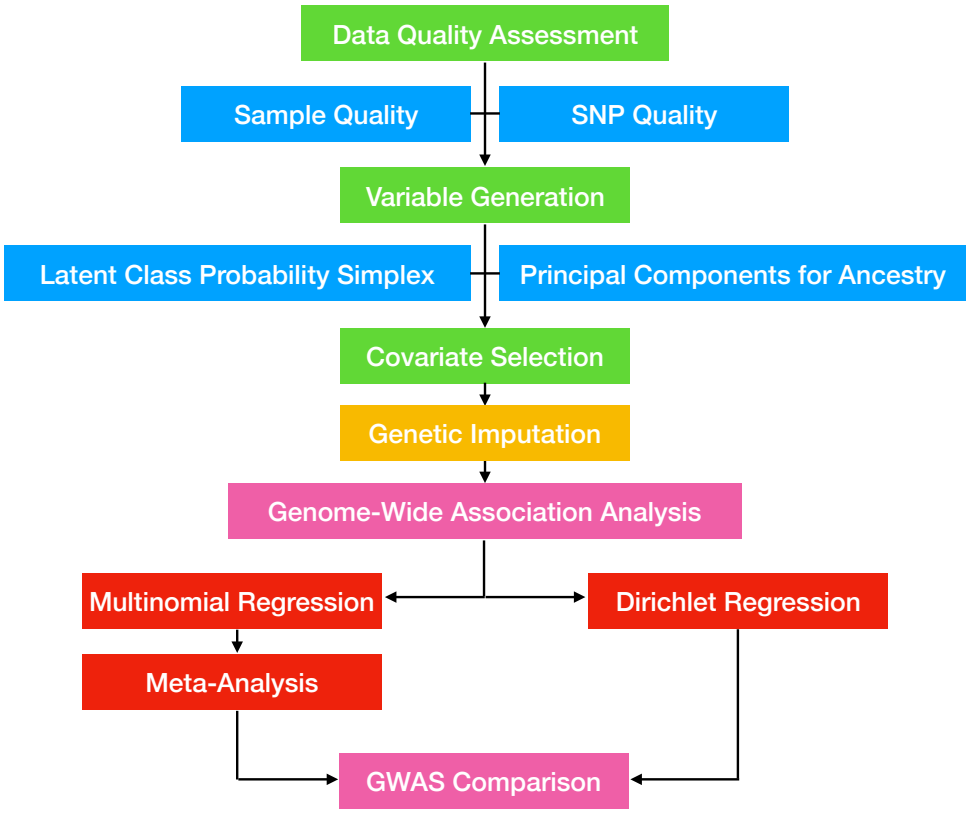

Figure 1: **Analysis Roadmap**

Dr. Cooper's LCA found that a total of four separate latent classes was most appropriate for both the $UCSF_{total}$ and TOR_1 cohorts. As shown in Figure 2, both the $UCSF_{total}$ and TOR_1 cohorts encounter similar situations. Latent Classes 1 (in blue) and 4 (in red) show

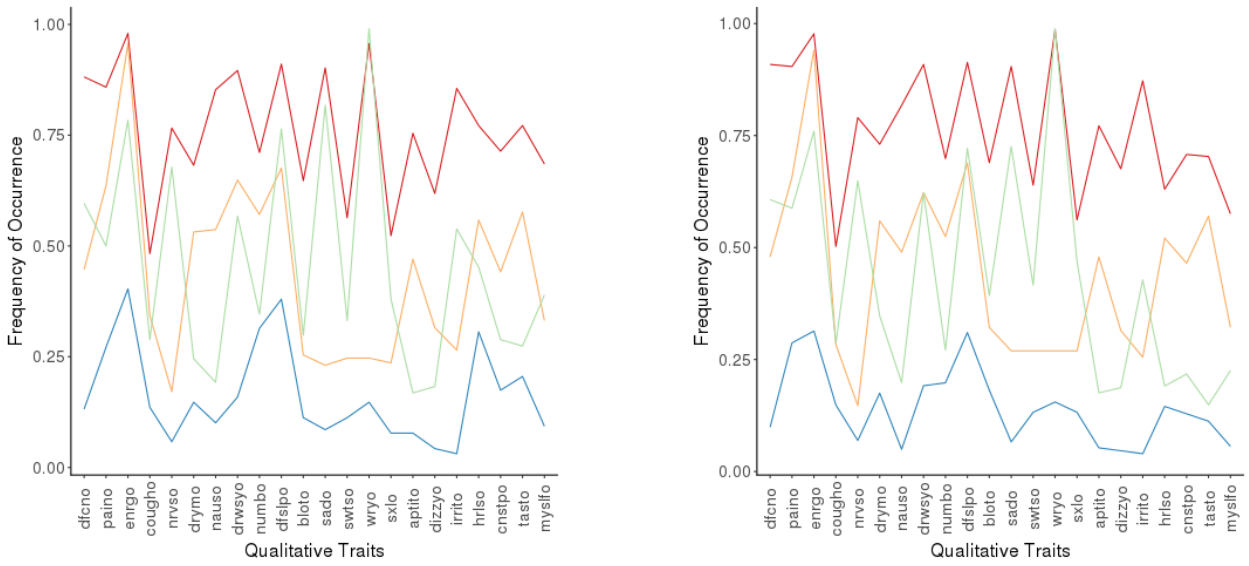

Figure 2: **Frequency of Occurrence for the 22 Qualitative Traits as Separated by Latent Class for UCSF (Left) and TOR1 (Right) Cohorts**

In both cases, Latent Class 4 individuals (blue) appear to have the lowest overall frequency of traits, while Latent Class 1 subjects (red) have with the highest overall frequency of traits. Subjects from Latent Class 2 (green) and 3 (yellow) appear to be less distinguishable.

the extremes of the symptom cluster burden, while latent classes 2 and 3 (in green and yellow) are more difficult to differentiate. Some qualitative traits, such as the experience of "lack of energy" (enrgo) and "difficulty sleeping" (dfslpo), that appear to be two of the most frequently occurring qualitative traits comparatively within all groups. Additionally, it may not be too far of a stretch to imagine how all three may possible be related to increases in overall "stress" for any surviving cancer patient.

Regarding, other characteristics about the subjects, it was worth noting the demographics of the final groups post-filtering. For the $UCSF_{total}$ group (Table 4), females outnumbered males overall by roughly 3.6:1, while "white" was the most common self-reported race. For the TOR_1 group (Table 5), females, again, outnumbered males overall by about 2.6:1. However, "white" was the overwhelming majority for self-reported race with only one "black" and three "Asian/Pacific Islander" subjects in comparison.

Table 4: **Distribution of Sex and Self-Reported Race as Separated by Latent Class for the Filtered $UCSF_{total}$ Cohort**

LC	Female	Male	Asian/Pacific Islander	Black	Hispanic/Mixed/Other	White
1	283	36	38	22	44	211
2	386	124	67	43	48	343
3	163	35	14	10	17	157
4	164	81	38	17	24	162

The distributions of the main covariates, Karnofsky Performance Status (KPS) and the total number of comorbidities, were also assessed for each of the different groups. A KPS value of 100 indicates "perfect health", while a score of 0 represents a deceased individual. For both UCSF and TOR1, subjects within Latent Class 1 were shown to have a lower "overall wellbeing" especially when compared to those within Latent Class 4. Additionally, Latent Class 1 appeared to have a higher mean value of comorbidities when compared to Latent Class 4.

3.1 QUALITY CONTROL FILTERING

3.1.1 SUBJECT QUALITY CONTROL FILTERING

From an initial count of 1613 samples (include technical duplicates), the $UCSF_{total}$ cohort retained 1272 unique individuals after the filtering process, while the TOR_1 cohort retained 415 (Table 6).

Table 5: **Distribution of Sex and Self-Reported Race as Separated by Latent Class for the Filtered TOR_1 Cohort**

LC	Female	Male	Asian/Pacific Islander	Black	White
1	43	15	2	0	56
2	29	9	0	0	38
3	116	36	1	0	151
4	111	56	0	1	166

3.1.2 SNP QUALITY CONTROL FILTERING

From Group A, there were 527182 SNPs remaining after the filtering process. From Group B, there were 239101 SNPs post exclusion (Table 7).

3.1.3 RETAINED COVARIATES

From the 11 total variables possibly available as covariates (Section 2.4.1), three were chosen for the final set of covariates within the model: "sex", "KPS", and "Number of Comorbidities." While, "KPS" and "Number of Comorbidities" were highlighted as common variables from the $UCSF_{total}$ and TOR_1 LASSO regression model outputs, "sex" was retained due to its biological significance and relation to cancer susceptibility. [Dorak and Karpuzoglu, 2012, Edgren et al., 2012]. Additionally, the four principal components were retained to account for existing population substructures produced by genetic ancestry especially at the case-control level. In total, these four covariates were utilized in both the Trinculo and Dirichlet Regression methodologies.

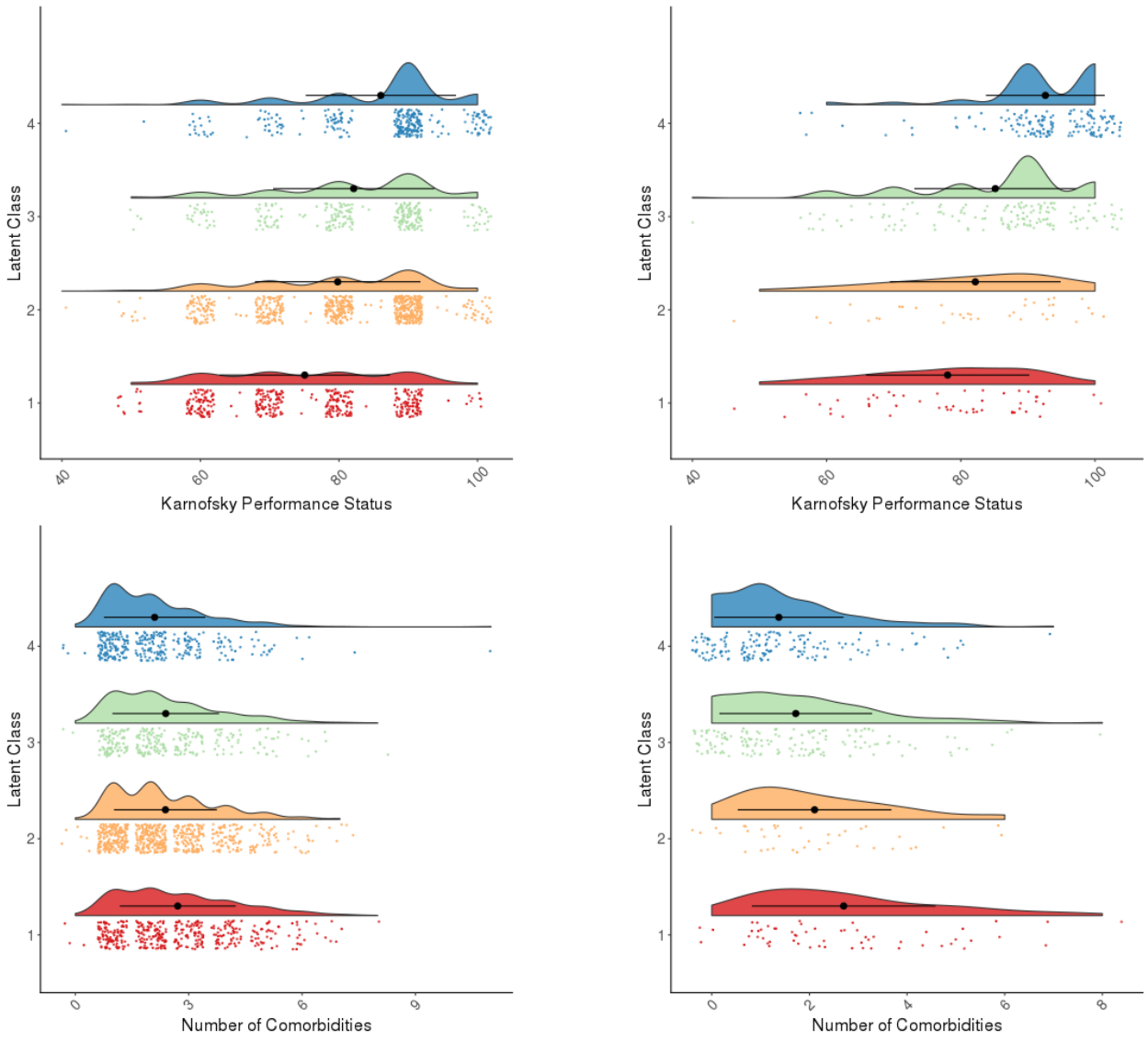

Figure 3: **Distribution of Karnofsky Performance Status (KPS) and Comorbidities Count per Latent Class for $UCSF_{total}$ (Left) and TOR_1 (Right) Cohorts**

The center dot represents the mean, while its intersecting line indicates the interquartile range. The values are jittered in order to more easily highlight the density at certain points

Table 6: **Subject Quality Control for $UCSF_{total}$ and TOR_1 Cohorts**

	$UCSF_{total}$		TOR_1	
Filtering Criteria	Retained	Excluded	Retained	Excluded
Starting Count	1613	NA	498	NA
Missing.e2 $\geq$ 0.03	1610	3	498	0
Females with low X intensities	1602	8	497	1
Males with low Y intensities	1597	5	495	2
Genotype quality scores	1595	2	493	2
Sibling	1594	1	493	0
Unexpected Duplicates	1592	2	493	0
Expected Duplicates	1469	123	446	47
High BAF sd	1461	5	443	3
Missing Gender	1296	165	419	24
Base distance > 5mb	1285	11	415	4
Missing Latent Class	1272	13	415	0

Table 7: **SNP Quality Control for Group A and Group B Cohorts**

	Group A		Group B	
Filtering Criteria	Retained	Excluded	Retained	Excluded
Starting Count	547644	NA	242901	NA
Missing.n2 $\geq$ 0.03	535198	12446	239880	3021
Unknown Chromosome Position SNPs	534505	693	239880	0
Duplicate SNPs	527182	7323	239101	779

3.2 MULTINOMIAL LOGISTIC REGRESSION RESULTS

After applying the multinomial logistic regression and meta-analysis on the $UCSF_{total}$ cohort to test for genome-wide association, we found that there were no SNPs that achieved genome-

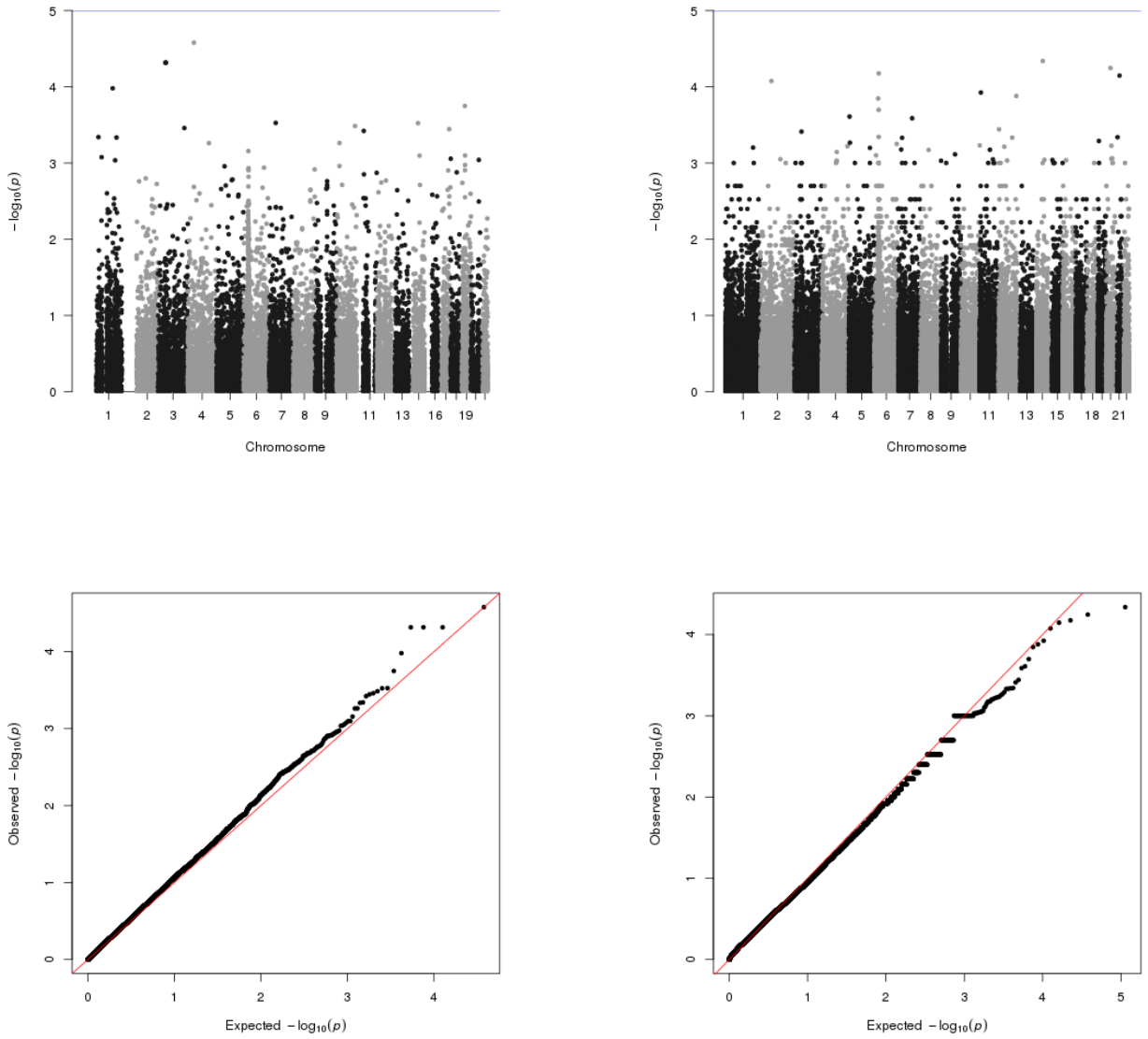

Figure 4: **Manhattan Plot and Q-Q Plot of Single Nucleotide Polymorphisms for $UCSF_{total}$ (Left) and TOR_1 (Right) Cohorts from TRINCULO.**

"Suggestive Association" and "Bonferroni-corrected" Significance are represented via the blue and green lines respectively.

wide significance ($p < 5 \times 10^{-8}$), Bonferroni-corrected significance ($p < 2.651 \times 10^{-6}$), or "suggestive association" significance ($p < 10^{-5}$). However, the top three SNPs that were closest to the "suggestive association" threshold can be found in Table 8. Additionally, the Q-Q plot of the expected vs observed p-values additionally reflects this notion; it's stark linearity suggests a lack of associated SNPs with the latent class outcome. Referencing Figure 4, all of the SNPs in the TOR_1 cohort failed to achieve genome-wide significance ($p < 5 \times 10^{-8}$), Bonferroni-corrected significance ($p < 8.792136 \times 10^{-7}$), and "suggestive association" significance ($p < 10^{-5}$). Its Q-Q plot also demonstrated this lack of association.

3.3 DIRICHLET REGRESSION

When we applied the Dirichlet regression to the probability simplexes, there were a few key observations. Initially, there were a marked excess of extremely small p-values as demonstrated in both the Manhattan and Q-Q plots (left-half of Figure 5). Further investigation revealed that all SNPs with p-values that exceeded the least restrictive "suggestive association" threshold, had minor allele frequencies (MAF) less than 0.001. Possibly, this may be indicative that the test statistic is not robust to very low MAF conditions. Relatedly, Wald tests in regular association test of genetic variants with low MAFs have also encountered this issue and had led to application of correction methods such as Firth's bias correction [Zhou et al., 2017]. Additionally, when we filter out the results when the minor allele frequency is less than 5 %, we are left with SNPs with p-values $\geq 1.917025 \times 10^{-4}$. As demonstrated in the right-half of Figures 5, 6, and 7, all SNPs from the $UCSF_A$, $UCSF_B$, and TOR_1 cohorts with MAF ≥ 0.05 have non-significant p-values $> 1 \times 10^{-5}$ (the least restrictive "suggestive association" threshold). Additionally, when evaluating the Manhattan Plots before the $< 5\%$ MAF Exclusion, it is possible to see a horizontal "chunks" of "white space." After further evaluation, we postulate that this phenomena may be related to the extremely low MAFs associated with certain SNPs.

When we graph the $-log_{10}$(p-values) from the Multinomial Meta-Analysis vs Dirichlet Regression results from the UCSF cohort after removing the low MAF SNPs, we fail to

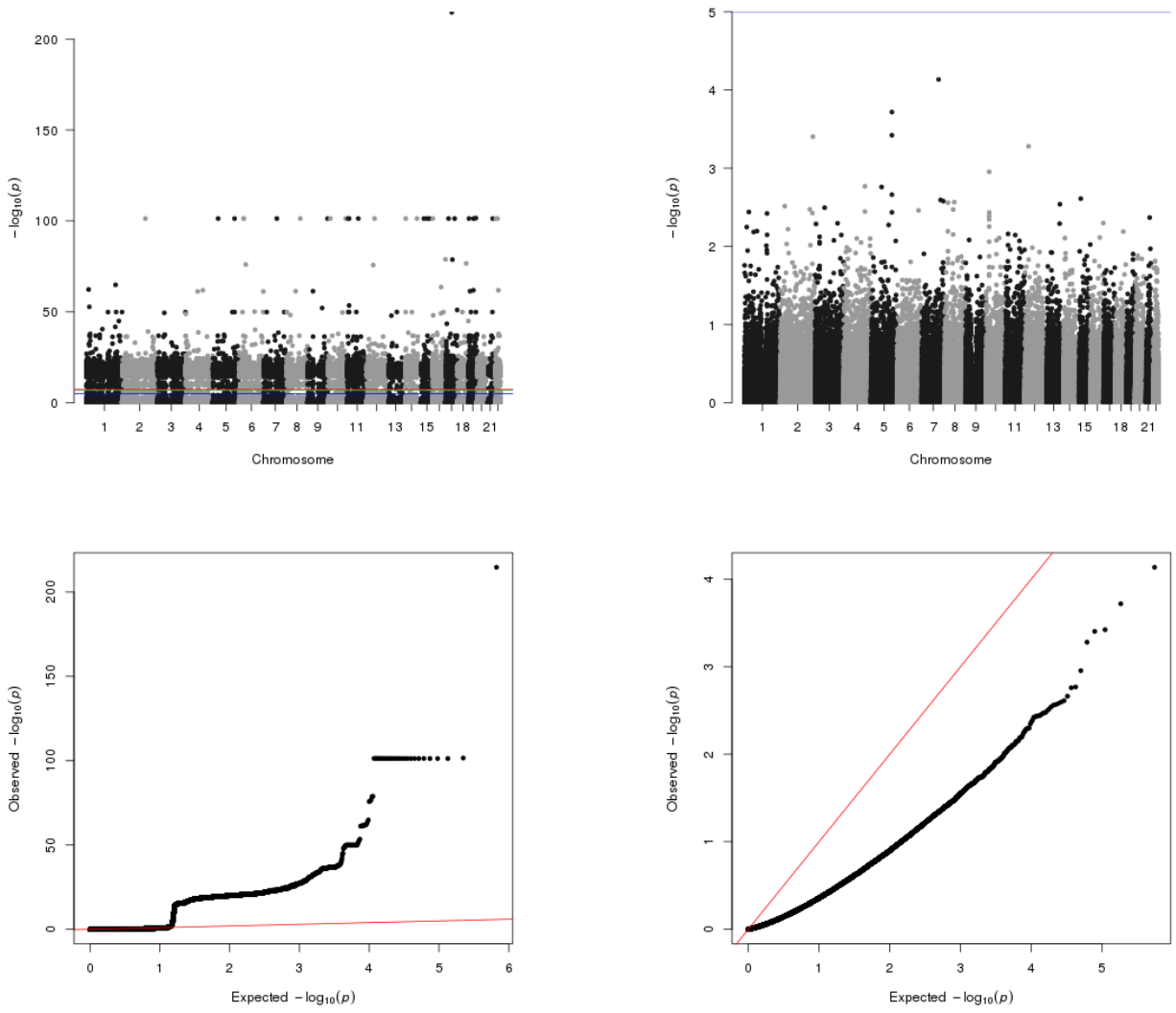

Figure 5: **Manhattan and Q-Q Plots of Single Nucleotide Polymorphisms for $UCSF_A$ Cohort Before (Left) and After (Right) less than 5% Minor Allele Frequency Exclusion from Dirichlet Regression**

"Genome-Wide Association", "Suggestive Association", and "Bonferroni-corrected" Significance are represented via the red, blue, and green lines respectively.

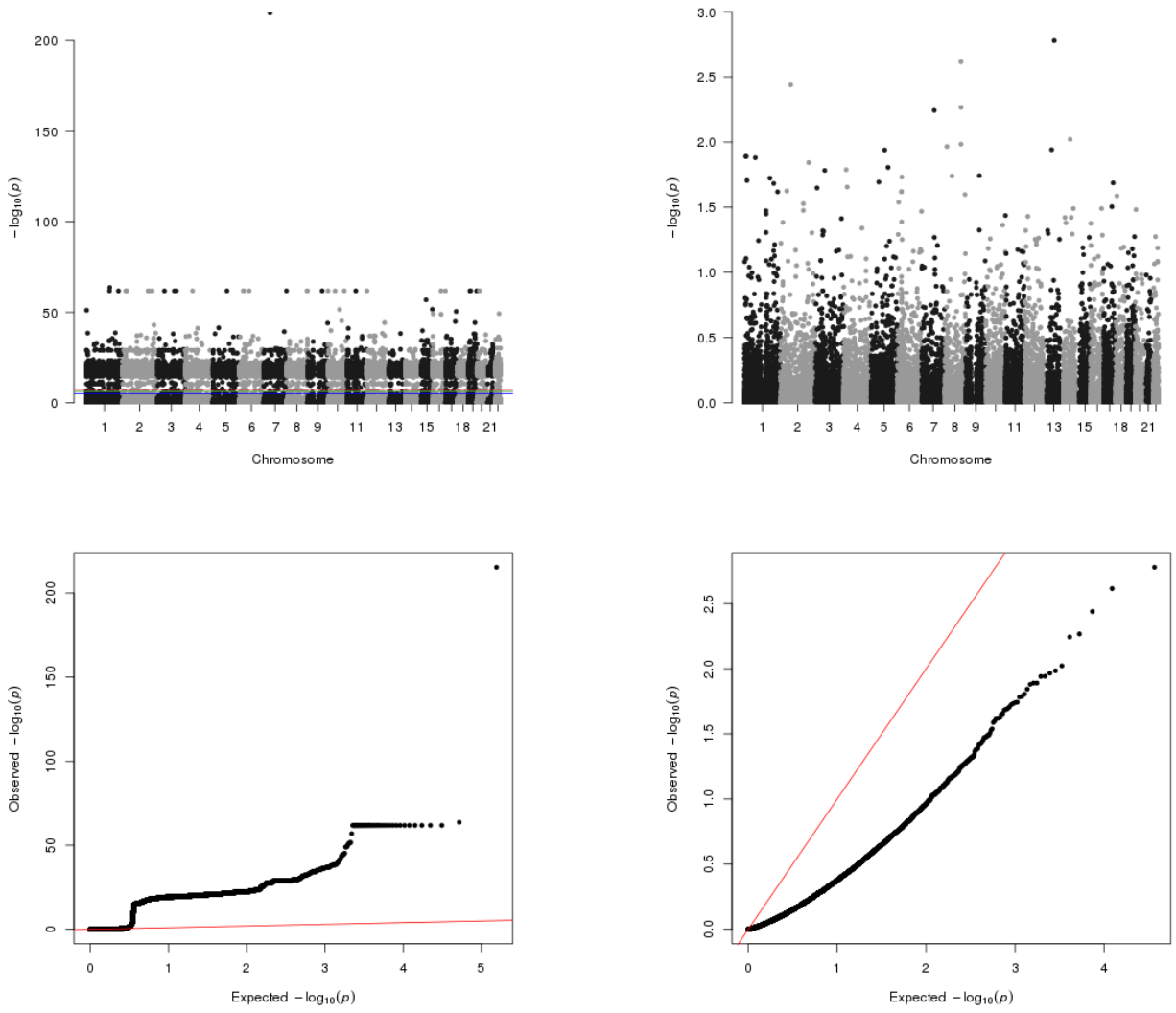

Figure 6: **Manhattan and Q-Q Plots of Single Nucleotide Polymorphisms for $UCSF_B$ Cohort Before (Left) and After (Right) less than 5% Minor Allele Frequency Exclusion from Dirichlet Regression**

"Genome-Wide Association", "Suggestive Association" and "Bonferroni-corrected" Significance are represented via the red, blue, and green lines respectively.

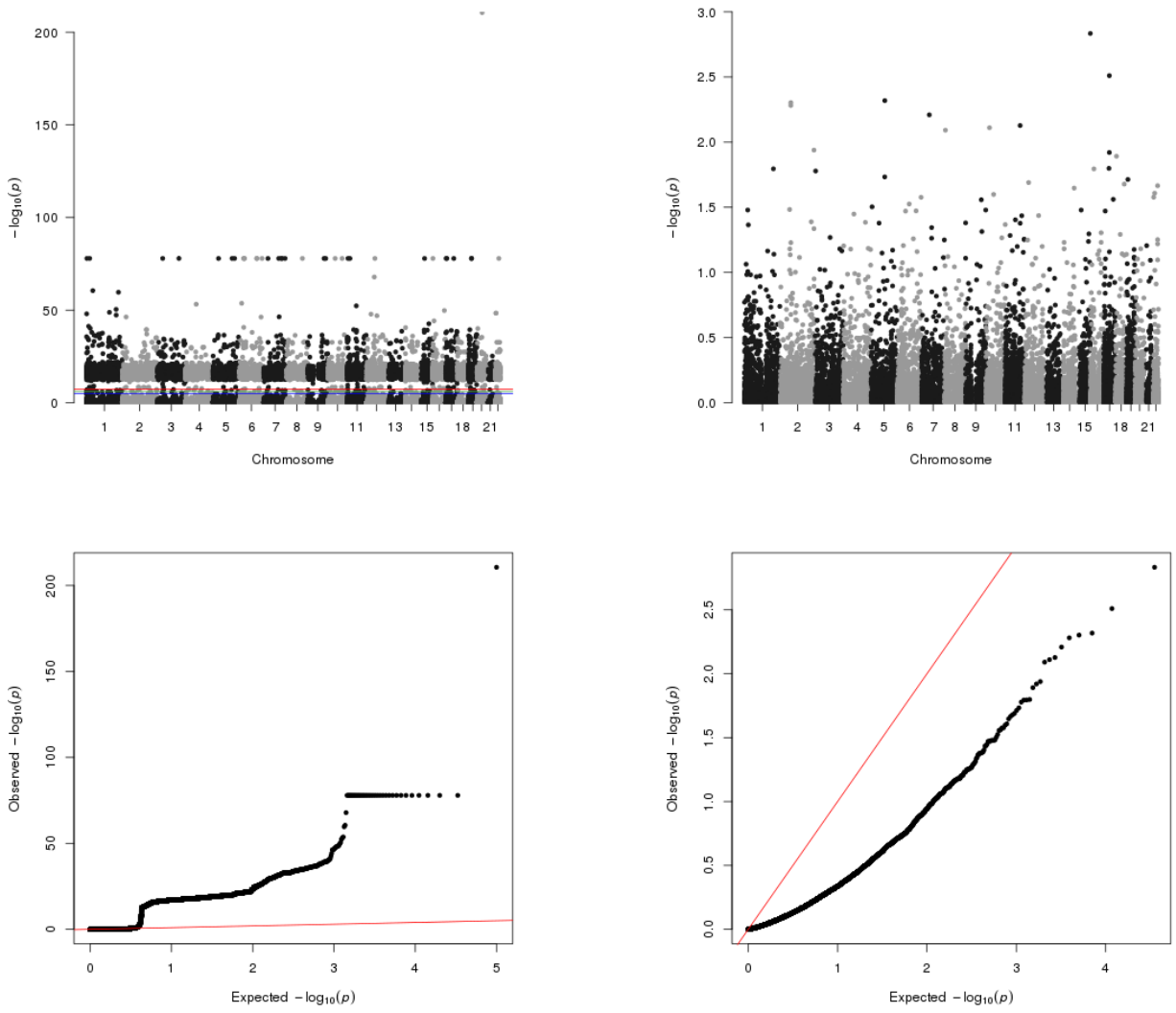

Figure 7: **Manhattan and Q-Q Plots of Single Nucleotide Polymorphisms for TOR_1 Cohort Before (Left) and After (Right) less than 5% Minor Allele Frequency Exclusion from Dirichlet Regression**

"Genome-Wide Association", "Suggestive Association" and "Bonferroni-corrected" Significance are represented via the red, blue, and green lines respectively.

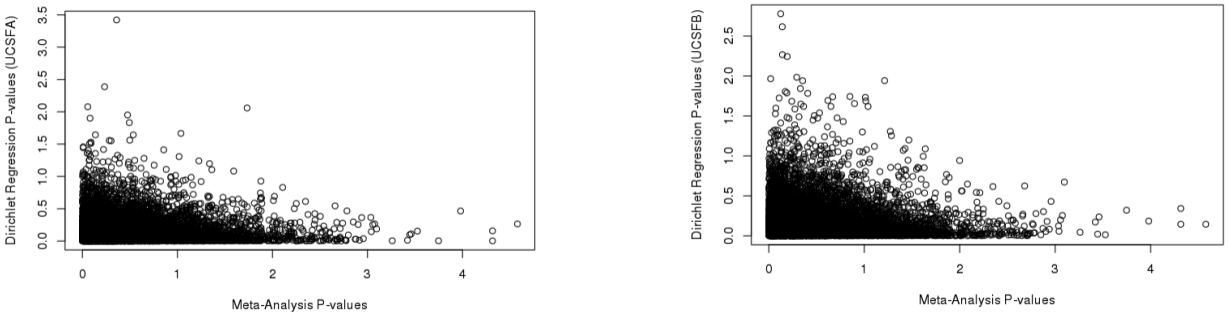

Figure 8: $-log_{10}$ **P-values for Meta-Analysis vs Dirichlet for** $UCSF_A$ **(Left) &** $UCSF_B$ **(Right)**

see consistent results and good agreement (Figure 8). This may be indicative of greater issues such as mismatched assumptions underlying each model that would allow for such discordance.

3.4 $UCSF_{TOTAL}$ MULTINOMIAL LOGISTIC REGRESSION TOP HITS

After running the meta-analysis for the $UCSF_{total}$ individuals from the multinomial logistic regression results, the top three SNPs after the < 0.05 MAF filtering were rs278981, rs72932959, and rs35455589 as shown in Table 8. Each SNP appeared to fall on 7 transcripts in 2 genes, 21 transcripts in 3 genes, and 11 transcripts in 2 genes respectively. Both rs278981 and rs35455589 appear to have relatively common allele frequency in many populations; however, rs72932959 appears to have a much higher allele frequency in Asian and Latino populations. Additionally, both rs278981 and rs35455589 are missense variants inferring that they may be functional [Lek et al., 2016]. RBM47, HYAL3, and HYAL2 each represent a single gene that has been associated with one of the SNPs, and each of them have been investigated in literature reviews and associated with breast cancer progression or lung cancer tumors [Vanharanta et al., 2014, Rai et al., 2001, Udabage et al., 2005].

Table 8: **Meta-Analysis of Multinomial Regression Results - Top Hits**

SNP	Chr:BP	P.Value	Lit Review
rs278981	4:40,428,010	2.642e-05	RBM47: suppressor of breast cancer progression and metastasis [Vanharanta et al., 2014]
rs72932959	3:50,336,292	4.833e-05	HYAL3: candidate lung cancer tumor suppressor [Rai et al., 2001]
rs35455589	3:50,355,730	4.833e-05	HYAL2: over-expression implicated in invasiveness of breast cancer [Udabage et al., 2005]

When assessing the p-values of these SNPs from the $UCSF_A$ and $UCSF_B$ Dirichlet Regression results, we found that all values were significantly larger and none were as close to the "associated significance" threshold (Table 9). Both the Multinomial and Dirichlet Regression results for the TOR_1 cohort (Table 10) repeated these findings. Lastly, when assessing the odd ratios of each latent class against the latent class with the highest "symptom burden" (LC1) (Table 11), we found that none of them were unnaturally inflated, yet almost all groups (except for LC2) appeared to demonstrate that as the load of minor alleles increases (i.e. as the genotype increases from 0 to 1 to 2, where the number represents the number of rare alleles the individual has), the odds of being in a class other than LC1 increases. When looking at the Counts and Percentages of each SNP more closely (Tables 12 - 17), we can see that the trend holds in the $UCSF_A$ but not in the $UCSF_B$ tables. Possibly, each SNP, even those designated as "missense" and therefore functional, may play a role in actually alleviating aspects of the "symptom burden" for the subjects. From Tables 13 & 14 and 16 & 17, we can see that rs729392959 and rs35455589 are generating very similar counts. As these two SNPs are on the same chromosome and relatively close to one another (Table 8), this may be indicative that the markers are in linkage disequilibrium.

Table 9: $UCSF_{total}$ **Dirichlet Regression P-Values - Top Hits**

	rs278981	**rs72932959**	**rs35455589**
$UCSF_A$ P-value	0.0112	0.477	0.477
$UCSF_B$ P-value	0.632	0.632	0.877

Table 10: TOR_1 **Regression P-Values - Top Hits**

	rs278981	**rs72932959**	**rs35455589**
Multinomial P-value	0.643	0.129	0.129
Dirichlet P-value	0.904	0.961	0.961

Table 11: **Odds Ratio Comparisions for Top Hit SNPs from Multinomial Regression (TRINCULO)**

	rs278981	**rs72932959**	**rs35455589**
$UCSF_A$ OR (LC2 vs LC1)	0.200	2.323	0.578
$UCSF_A$ OR (LC3 vs LC1)	1.419	1.448	2.323
$UCSF_A$ OR (LC4 vs LC1)	1.419	1.416	0.869
$UCSF_B$ OR (LC2 vs LC1)	1.458	17.940	2.256
$UCSF_B$ OR (LC3 vs LC1)	1.458	17.940	2.256
$UCSF_B$ OR (LC4 vs LC1)	1.255	0.683	0.860

Table 12: **Genotype (rs278981) by Latent Class, Counts (Left) and Row Percentages (Right), $UCSF_A$ Multinomial Regression**

	LC1	LC2	LC3	LC4
0	108	137	79	75
1	69	127	36	58
2	4	28	3	6

	LC1	LC2	LC3	LC4
0	0.271	0.343	0.198	0.188
1	0.238	0.438	0.124	0.200
2	0.098	0.683	0.073	0.146

Table 13: **Genotype (rs72932959) by Latent Class, Counts (Left) and Row Percentages (Right), $UCSF_A$ Multinomial Regression**

	LC1	LC2	LC3	LC4
0	153	236	102	115
1	18	40	12	14
2	9	16	4	9

	LC1	LC2	LC3	LC4
0	0.252	0.389	0.168	0.190
1	0.214	0.476	0.143	0.167
2	0.237	0.421	0.105	0.237

Table 14: **Genotype (rs35455589) by Latent Class, Counts (Left) and Row Percentages (Right), $UCSF_A$ Multinomial Regression**

	LC1	LC2	LC3	LC4
0	154	235	101	114
1	18	40	12	15
2	9	16	4	9

	LC1	LC2	LC3	LC4
0	0.255	0.389	0.167	0.189
1	0.212	0.471	0.141	0.176
2	0.237	0.421	0.105	0.237

Table 15: **Genotype (rs278981) by Latent Class, Counts (Left) and Row Percentages (Right), $UCSF_B$ Multinomial Regression**

	LC1	LC2	LC3	LC4
0	20	35	22	22
1	17	28	15	11
2	2	2	2	1

	LC1	LC2	LC3	LC4
0	0.202	0.354	0.222	0.222
1	0.239	0.394	0.211	0.155
2	0.286	0.286	0.286	0.143

Table 16: **Genotype (rs72932959) by Latent Class, Counts (Left) and Row Percentages (Right), $UCSF_B$ Multinomial Regression**

	LC1	LC2	LC3	LC4
0	27	48	33	29
1	11	12	4	3
2	1	5	2	2

	LC1	LC2	LC3	LC4
0	0.197	0.350	0.241	0.212
1	0.367	0.400	0.133	0.100
2	0.100	0.500	0.200	0.200

Table 17: **Genotype (rs35455589) by Latent Class, Counts (Left) and Row Percentages (Right), $UCSF_B$ Multinomial Regression**

	LC1	LC2	LC3	LC4
0	27	47	33	28
1	11	12	4	3
2	1	5	2	2

	LC1	LC2	LC3	LC4
0	0.200	0.348	0.244	0.207
1	0.367	0.400	0.133	0.100
2	0.100	0.500	0.200	0.200

4.0 DISCUSSION

As demonstrated in Figure 1, the project was broken down into several phases: data quality assessment, variable generation, covariate selection, genetic imputation, and the genome-wide association analyses. Starting with the first phase, recall that there were two separate cohorts, UCSF and TOR1, which involves individuals both from the San Francisco Bay Area as well as Norway. In total there initially were 1613 and 498 individuals in each respective group; 1272 and 415 respectively remained after the sample filtering process. Additionally, the genotype calling was split into two separate groups, Group A and Group B, based on the different arrays used; the UCSF cohort was split between Group A and Group B, while TOR1 was only called in Group B. Post SNP quality control, there were 527,182 and 239,101 SNPs evaluated within Group A and Group B respectively.

4.1 PHENOTYPIC COMPARISONS

When looking at the frequency of the occurrence for the 22 qualitative traits as separated by latent class (Figure 2), it initially is worth noting how strikingly similar the $UCSF_{total}$ and TOR_1 profiles are especially given the inherent differences in the populations. However, as the latent classes were generated using a combination of both the UCSF and TOR1 datasets (while adjusting for site), this should come as no surprise. It is notable that the middle latent classes (green & yellow) are a bit indistinguishable from each other, while the lowest and highest latent classes (blue & red) more easily visualize the individuals who had the least and greatest frequency of occurrence for these qualitative traits. In Figure 3, which describes the distributions of the covariates separated by latent class, we can see something similar:

the latent class with the lowest presentation of the 22 qualitative traits (blue) tends to have the highest KPS and lowest number of comorbidities, while the highest variable presenting latent class (red) appears to have the lowest KPS and highest number of comorbidities. With regards to self-reported ethnicities, "white" remained the most common self-reported race in both groups (Tables 4 & 5), yet the overall distributions in each population were different. TOR_1 reflected a nearly homogeneous "white" population (as would be expected for a Norwegian sample), while $UCSF_{total}$ had significantly more variation in comparison.

A point of interest to make is that the latent classes are determined from the dataset as a whole and are not an attribute directly measurable on a single individual. This may lead to issues with regards to reproducibility unless future analyses utilize the exact same models created here to maintain the same definitions for the latent classes. However, from a public health perspective, this approach may be worth developing further as its ability to draw inferences from observed variables may allow the creation of models capable of indicating certain latent characteristics of the patient. For instance, by setting a new subject's information (defined by the exact same set of variables used for the LCA), into the model, it would be possible to assign her into a specific "treatment group" as extrapolated from the exact latent class assignment. Thereby, once it is know that the patient is in that specific treatment group, she may be given a specifically tailored set of special care which may be more appropriate than the usual standard-of-care.

4.1.1 IMPUTATION

Recall that the $UCSF_{total}$ group was split into Group A and Group B for genotype calling, and a meta-analysis was performed on the common set of SNPs between $UCSF_A$ and $UCSF_B$. By utilizing panel references such as the Haplotype Reference Consortium (HRC), shared haplotypes can be identified amongst subjects and the missing genotypes from each individual can be replaced by the observed alleles from the reference [Li et al., 2009]. Through imputation, we were able to increase the number of commonly shared SNPs found in both groups thereby allowing a larger number of SNPs to be analyzed in the meta-analysis for the $UCSF_{total}$ cohort. Unfortunately due to time constraints, the imputed values were only

utilized for the $UCSF_{total}$ multinomial logistic regression analyses. More specifically, we were able to conduct a meta-analysis on 18,862 commonly shared $UCSF_{total}$ variants; the TOR_1 subjects did not have any imputed values assessed.

4.1.2 COVARIATE SELECTION: SIMPSON'S PARADOX

When performing the covariate selection for the regression models, it is important to highlight the potential issue of Simpson's Paradox, which describes events where an association may be present within individual groups but may not appear to be the case when all groups are combined. Recall that in the analysis, the $UCSF_{total}$ and TOR_1 subjects each underwent a LASSO multinomial regression in order to determine which variables would be associated with the latent classes. There were differences in the wording and options that the $UCSF_{total}$ and TOR_1 groups received as part of questionnaire. Condensing down the TOR_1 group's definition of "partnered" to match that of the $UCSF_{total}$ group's variable (e.g. Umarried; Married/Living Together; Divorced; Widow; Separated into Partnered:Yes/no) would potentially lead to a misrepresentation of the data. As mentioned previously, it is worth noting that the dataset used at the time of analysis has been replaced by a newer version. In its up-to-date form, the discordant variables have now been harmonized thus giving any future analysis a direct way to circumvent the issues listed above.

4.1.3 COVARIATE SELECTION: KPS

While Karnofsky Performance Status (KPS) was included in the regression models as a covariate due to its performance within the LASSO selection process, it is important to highlight the accompanying set of issues. KPS describes a measure of overall "wellness of being/function" for an individual as observed by a separate party; having certain symptoms of disease which could manifest in the individual as "diarrhea", "nausea", or "pain" could lead to an overall lower KPS score or lower "wellness of being." Perhaps it may be repetitive then to include the KPS score as a covariate within the model as there is arguably substantial overlap with some of the 22 qualitative traits utilized in the latent class analysis to subcategorize the individuals into different levels of "quality-of-life." Additionally, as pa-

tients in this study had already undergone chemotherapy, their collective KPS values would have been skewed towards a generally lower overall "wellness of being."

4.1.4 COVARIATE SELECTION: INCLUSION

The current approach of selecting common covariates from both the $UCSF_{total}$ and TOR_1 analyses provided the opportunity to evaluate the outcome variable as a function of the same set of variables. When variables are added to the model, the proportion of variance in the dependent variable that can be explained by the set of independent variables or coefficient of determination (r^2) increases. Therefore, it may be possible to view that the way one might define the regression residuals may be different depending on which variables were selected to be in the model. It is possible to see this in genetics, where the inclusion of covariates within the regression model changes the context of the genetic effects one may be assessing. More specifically, a confounder, a risk factor which will skew the relationship between exposure of interest and disease manifestation, will have differing implications if it is included in the model versus if it is not. For instance, inclusion of BMI in a genetic association test for Type 2 Diabetes (T2D) would demonstrate an entirely different context as BMI's nature as a strong risk factor would influence the occurrence of cases within the population. It would not be possible to conclude that, given the statistic passing the set significance threshold, there is an association just between this particular SNP and T2D; BMI, played a major role as a confounder, and the interpretation of the the association would need to include that aspect as well. In other words, one would be assessing the association between the SNP and phenotype conditional on the covariate by using this approach [Vansteelandt et al., 2009].

Additionally, covariate presence in genetic studies has been met with caution when testing for novel associations especially if they are not confounders. In case-control studies, the ascertainment process for subjects can bias cases to be individuals with risk genotypes as well as high-risk covariate levels; overall this can lead to a loss of power due to increased standard error of the genetic association due to the correlation [Mefford and Witte, 2012]. [Kuo and Feingold, 2010] assessed the effect of covariate inclusion via simulation and found that the regression model with just the genotype alone was more powerful in determining a

genetic effect when compared to the model including a covariate; this indicates that it may be detrimental to include covariates within the model when performing genetic association. Finally, including covariates within a genetic model relies on the assumption that there is no association between the covariate and the marker locus. This may not hold up in a genome-wide context especially when the covariate of interest may have a high heritability [Vansteelandt et al., 2009].

4.2 INFLUENCE OF LOW MINOR ALLELE FREQUENCY

With regards to the results of the analysis pipelines, we did not find any genome-wide significant signals after excluding for extremely low MAF. Though rare and low frequency variants with MAFs $< 0.5 - 1\%$ are thought to play a major role in heritability for common, complex diseases, it is important to highlight that low MAF has been associated with an increased false-positive rate [Tennessen et al., 2012, Tabangin et al., 2009]. Given the number of latent class and genotype assignments possible, it should be understood that low MAFs may be very possible within our dataset. Recall that in the original multinomial regression approach, there were four latent class assignments; moreover, there were three separate genotypes possible depending on the mixture of major (A) and minor (B) alleles for each individual: 1) AA, 2) AB, and 3) BB. In a situation where the minor allele frequency or overall number of minor alleles has been determined to be quite low within the sample, it should come as no surprise that after separating that sample into four latent classes, the number of minor alleles in each group may possible be extremely low or non-existent. In fact, Figures 5, 6, & 7 demonstrated excessively small p-values previous to any MAF filtering which could be seen as a result of the increased false-positive rate from having SNPs with low MAF.

4.3 MULTINOMIAL LOGISTIC REGRESSION

In our data, it has been shown that there are many markers with very low MAFs; subsequently, it is the case that their latent class/genotype contingency tables have empty cells. In general, it is important to note that having low or empty counts in the marker's latent class/genotype contingency table can cause high standard errors for the coefficient associated with that category [Menard, 2002]. In logistic regression, having a similar scenario of a low number of event-per-variable (EPV) can lead to biased results; this has led to the application of Firth's bias correction as a correction to apply when dealing with small sample sizes [Peduzzi et al., 1996, Maiti and Pradhan, 2008]. With regards to the Multinomial Logistic Regression, it is also possible to extend Firth's bias correction. In the genetic context, previous research has demonstrated the use of Firth's bias correction in rare variant association tests when dealing with single low-count variants [Bull et al., 2002, Wang, 2014, Ma et al., 2013]. Unfortunately, the TRINCULO software utilized for the multinomial logistic regression analysis did not appear to use this method, and this may be related to the abnormal Q-Q plots observed within our results (Figure 4).

4.4 DIRICHLET REGRESSION

In the Dirichlet Regression approach, the phenotype (e.g. the independent variable) is not the categorical latent class assignment; rather, it is the probability simplex, the vector of four posterior probabilities associated with each latent class. Figure 9 visually demonstrates the distribution of those simplex probabilities. In this context, each vertex represents one of the four probabilities (p_1, p_2, p_3, p_4) that make up the probability simplex, and the combined probability simplex sums to a total of 1 $\left(\sum_{i=1}^{4} p_i = 1\right)$. Having a point within Figure 9 be near top of the pyramid, for instance, would indicate that a subject's p_1 value would be close to 1 and subsequently her other probability values (p_2, p_3, p_4) would be near or at 0. The Dirichlet Regression approach allows us to utilize the entirety of the probability simplex space as containing possible phenotypes; however, it is important to note that each subject's

probability simplexes appear to be stratified to the area closely surrounding the vertices. As the Dirichlet Regression method is capable of handling a more heterogeneous spread of probability simplexes but the DirichletReg R package utilized to perform the analysis did not include an option to account for skewness nor heteroscedasticity, perhaps this lack of adjustment could be responsible for the abnormal Q-Q plots (Figures 5, 6, & 7) [Maier, 2014].

4.5 COMPARING PHENOTYPES FOR ASSOCIATION

The goal of this analysis was to assess how probability simplex phenotypes perform compared to those of the most-likely latent class phenotypes when evaluating association. Overall, as shown in Figure 8, the p-values showed very poor agreement. Moreover, while the top three SNPs from the $UCSF_{total}$ Multinomial Logistic Regression were also present within the Dirichlet Regression results, comparatively, the p-values were significantly larger when compared to the 1×10^{-5} "associated significant" threshold (Figures 8 & 9).

Additionally, consider a situation where there exists a "perfect" three class model where every observation is predicted perfectly by the genotype. Plotting the probabilities would result in a triangular graph with all points centered at the three separate vertices (Figure 9). We created a three class artificial dataset with 3,000 individuals per vertex, where everyone in vertex i had genotype i-1 for $i \in (1...3)$, so genotype perfectly predicted each person's probability simplex. Running the Multinomial Logistic Regression on these data led to predicted probabilities close to the expected vertices (Table 18). However, when these data were analyzed with the Dirichlet Regression method, we found that the predicted, fitted values for each genotype were much farther away from the vertices (Table 19). Inherently, the Multinomial Logistic Regression and Dirichlet Regression approach are modeling the data in completely different manners. Such differences may be another reason for the discrepancies observed between the p-values from the actual analyses.

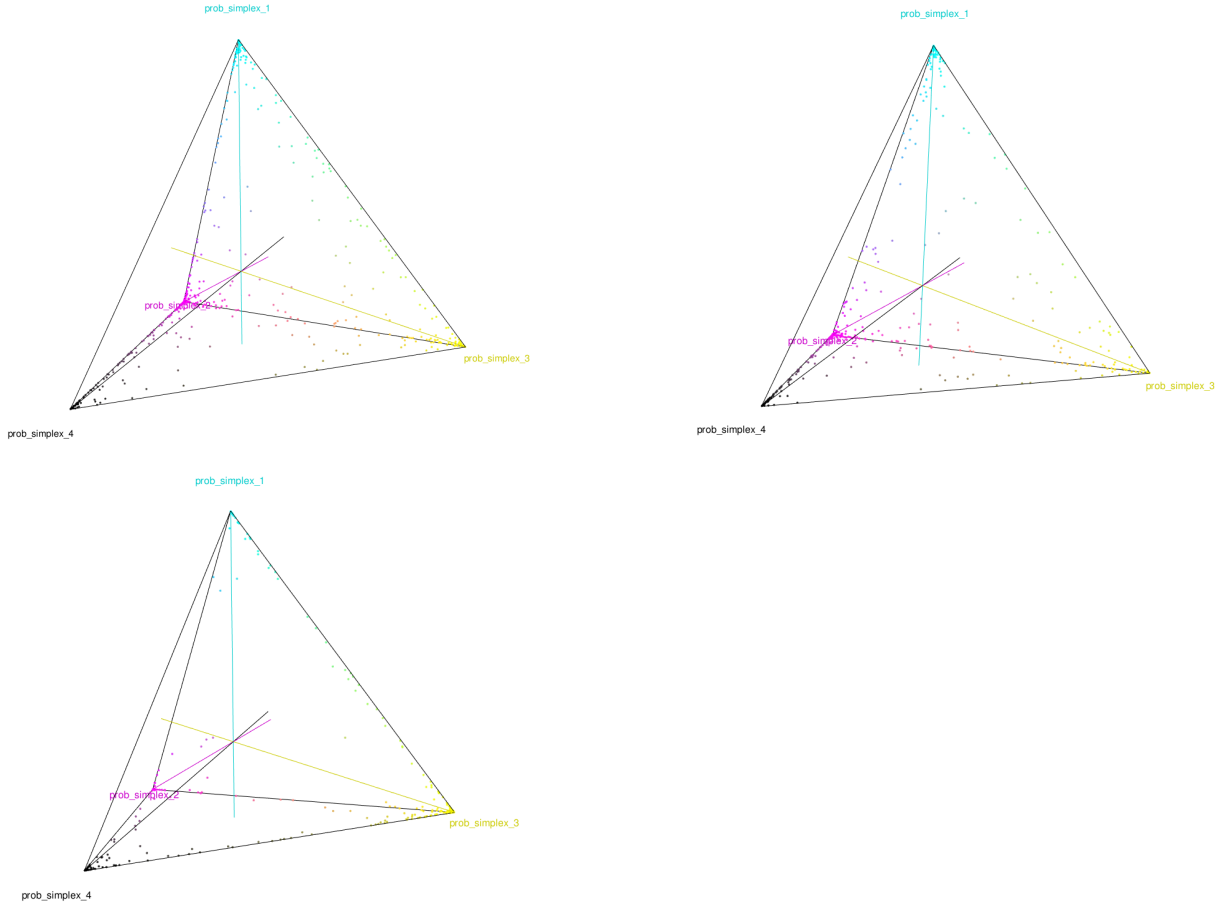

Figure 9: **Distribution of Dirichlet Probability Simplex for $UCSF_A$, $UCSF_B$, and TOR_1 (clockwise from upper left)**

The four probabilities of the simplex are shown in teal, pink, yellow, and black respectively.

Table 18: **"Perfect" Expected Probabilities (Left) vs Multinomial-Derived Predicted Probabilities (Right)**

	1	2	3
0	1.000	0.000	0.000
1	0.000	1.000	0.000
2	0.000	0.000	1.000

	1	2	3
0	1.000	1.268e-32	4.729e-09
1	0.000	1.000	2.243e-08
2	0.000	5.327e-09	1.000

Table 19: **"Perfect" Expected Probabilities (Left) vs Dirichlet-Derived Predicted Probabilities (Right)**

	1	2	3
0	1.000	0.000	0.000
1	0.000	1.000	0.000
2	0.000	0.000	1.000

	1	2	3
0	0.683	0.218	0.099
1	0.353	0.295	0.353
2	0.099	0.218	0.683

4.6 FUTURE WORK

There are a number of ways to move forward with this project. One avenue may involve recruiting more individuals. If focusing on bolstering the $UCSF_{total}$ and TOR_1 cohorts, a boost in sample size could lead to an increase in power for the GWAS [Spencer et al., 2009]. TOR_1 was chosen to act as a dataset capable of validating any potential findings within the $UCSF_{total}$ cohort. While, in this analyses, no findings could be replicated, perhaps having the funds to recruit heavily from a separate population may provide that opportunity in the future. With regards to the existing dataset, utilizing the newest updated dataset with the harmonized variables would be necessary and would allow us to run the LASSO variable selection process with both cohorts together with site as a covariate. Additionally, imputing the TOR_1 subjects and using the imputed data in the Dirichlet Regression analyses would be the ideal choice. Alternatively, we could follow through with whole genome sequencing which would provide a significantly larger number of available variants with compared to our current approach of focusing on the exome. However, in order to determine if the Dirichlet Regression approach is comparable to the Multinomial Logistic Regression method at hand, a simulation study including power calculations should be run.

BIBLIOGRAPHY

Astrup, G. L., Hofsø, K., Bjordal, K., Guren, M. G., Vistad, I., Cooper, B., Miaskowski, C., and Rustøen, T. (2017). Patient factors and quality of life outcomes differ among four subgroups of oncology patients based on symptom occurrence. *Acta Oncologica*, 56(3):462–470.

Bull, S. B., Mak, C., and Greenwood, C. M. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics & Data Analysis*, 39(1):57–74.

Carr, D., Goudas, L., Lawrence, D., Pirl, W., Lau, J., DeVine, D., Kupelnick, B., and Miller, K. (2002). Management of cancer symptoms: pain, depression, and fatigue. *Evidence report/technology assessment*, 61:368–374.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284.

De Moor, J. S., Mariotto, A. B., Parry, C., Alfano, C. M., Padgett, L., Kent, E. E., Forsythe, L., Scoppa, S., Hachey, M., and Rowland, J. H. (2013). Cancer survivors in the united states: prevalence across the survivorship trajectory and implications for care. *Cancer Epidemiology and Prevention Biomarkers*, 22(4):561–570.

Deshields, T. L., Potter, P., Olsen, S., and Liu, J. (2014). The persistence of symptom burden: symptom experience and quality of life of cancer patients across one year. *Supportive Care in Cancer*, 22(4):1089–1096.

Dorak, M. T. and Karpuzoglu, E. (2012). Gender differences in cancer susceptibility: an inadequately addressed issue. *Frontiers in genetics*, 3:268.

Edgren, G., Liang, L., Adami, H.-O., and Chang, E. T. (2012). Enigmatic sex disparities in cancer incidence. *European journal of epidemiology*, 27(3):187–196.

Espin-Garcia, O., Shen, X., Qiu, X., Brhane, Y., Liu, G., and Xu, W. (2014). Genetic association analysis for common variants in the genetic analysis workshop 18 data: a dirichlet regression approach. In *BMC proceedings*, volume 8, page S70. BioMed Central.

Ferrell, B. R., Wisdom, C., and Wenzl, C. (1989). Quality of life as an outcome variable in the management of cancer pain. *Cancer*, 63(11):2321–2327.

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One*, 8(11):e79667.

Gill, A., Chakraborty, A., and Selby, D. (2012). What is symptom burden: a qualitative exploration of patient definitions. *Journal of palliative care*, 28(2):83.

Ginsburg, G. S. and McCarthy, J. J. (2001). Personalized medicine: revolutionizing drug discovery and patient care. *TRENDS in Biotechnology*, 19(12):491–496.

Gogarten, S. M., Bhangale, T., Conomos, M. P., Laurie, C. A., McHugh, C. P., Painter, I., Zheng, X., Crosslin, D. R., Levine, D., Lumley, T., Nelson, S. C., Rice, K., Shen, J., Swarnkar, R., Weir, B. S., and Laurie, C. C. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, 28(24):3329–3331.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Harrington, C. B., Hansen, J. A., Moskowitz, M., Todd, B. L., and Feuerstein, M. (2010). It's not over when it's over: long-term symptoms in cancer survivorsa systematic review. *The International Journal of Psychiatry in Medicine*, 40(2):163–181.

Henry, D. H., Viswanathan, H. N., Elkin, E. P., Traina, S., Wade, S., and Cella, D. (2008). Symptoms and treatment burden associated with cancer treatment: results from a cross-sectional national survey in the us. *Supportive care in cancer*, 16(7):791–801.

Illi, J., Miaskowski, C., Cooper, B., Levine, J. D., Dunn, L., West, C., Dodd, M., Dhruva, A., Paul, S. M., Baggott, C., et al. (2012). Association between pro- and anti-inflammatory cytokine genes and a symptom cluster of pain, fatigue, sleep disturbance, and depression. *Cytokine*, 58(3):437–447.

Jostins, L. and McVean, G. (2016). Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. *Bioinformatics*, 32(12):1898–1900.

Kuo, C.-L. and Feingold, E. (2010). What's the best statistic for a simple test of genetic association in a case-control study? *Genetic epidemiology*, 34(3):246–253.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ODonnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annual review of genomics and human genetics*, 10:387–406.

Linzer, D. A. and Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10):1–29.

Lo, Y., Mendell, N. R., and Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3):767–778.

Ma, C., Blackwell, T., Boehnke, M., Scott, L. J., and Investigators, G. (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic epidemiology*, 37(6):539–550.

Maier, M. J. (2014). Dirichletreg: Dirichlet regression for compositional data in r.

Maiti, T. and Pradhan, V. (2008). A comparative study of the bias corrected estimates in logistic regression. *Statistical Methods in Medical Research*, 17(6):621–634.

McLachlan, G. and Peel, D. (2000). Multivariate normal mixtures. *Finite Mixture Models*, pages 81–116.

Mefford, J. and Witte, J. S. (2012). The covariate's dilemma. *PLoS genetics*, 8(11):e1003096.

Menard, S. (2002). *Applied logistic regression analysis*, volume 106. Sage.

Miaskowski, C., Cooper, B. A., Aouizerat, B., Melisko, M., Chen, L.-M., Dunn, L., Hu, X., Kober, K., Mastick, J., Levine, J., et al. (2017). The symptom phenotype of oncology outpatients remains relatively stable from prior to through 1 week following chemotherapy. *European journal of cancer care*, 26(3).

Miaskowski, C., Cooper, B. A., Melisko, M., Chen, L.-M., Mastick, J., West, C., Paul, S. M., Dunn, L. B., Schmidt, B. L., Hammer, M., et al. (2014). Disease and treatment characteristics do not predict symptom occurrence profiles in oncology outpatients receiving chemotherapy. *Cancer*, 120(15):2371–2378.

Miaskowski, C., Dunn, L., Ritchie, C., Paul, S. M., Cooper, B., Aouizerat, B. E., Alexander, K., Skerman, H., and Yates, P. (2015). Latent class analysis reveals distinct subgroups of patients based on symptom occurrence and demographic and clinical characteristics. *Journal of pain and symptom management*, 50(1):28–37.

Mor, V., Laliberte, L., Morris, J. N., and Wiemann, M. (1984). The karnofsky performance status scale: an examination of its reliability and validity in a research setting. *Cancer*, 53(9):2002–2007.

Muthén, L. K. and Muthén, B. O. (2012). Mplus version 7 users guide. *Los Angeles, CA: Muthén & Muthén*.

Nipp, R. D., El-Jawahri, A., Moran, S. M., D'arpino, S. M., Johnson, P. C., Lage, D. E., Wong, R. L., Pirl, W. F., Traeger, L., Lennes, I. T., et al. (2017). The relation-

ship between physical and psychological symptoms and health care utilization in hospitalized patients with advanced cancer. *Cancer*, 123(23):4720–4727.

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural equation modeling*, 14(4):535–569.

Panagiotou, O. A., Ioannidis, J. P., and Project, G.-W. S. (2011). What should the genome-wide significance threshold be? empirical replication of borderline genetic associations. *International journal of epidemiology*, 41(1):273–286.

Patrick, D., Ferketich, S., Frame, P., Harris, J., Hendricks, C., Levin, B., Link, M., Lustig, C., McLaughlin, J., Ried, L., et al. (2003). National institutes of health state-of-the-science conference statement: symptom management in cancer: pain, depression, and fatigue, july 15-17, 2002. *Journal of the National Cancer Institute*, 95(15):1110–1117.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12):1373–1379.

Portenoy, R. K., Thaler, H. T., Kornblith, A. B., Lepore, J. M., Friedlander-Klar, H., Kiyasu, E., Sobel, K., Coyle, N., Kemeny, N., Norton, L., et al. (1994). The memorial symptom assessment scale: an instrument for the evaluation of symptom prevalence, characteristics and distress. *European Journal of Cancer*, 30(9):1326–1336.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904.

Rai, S. K., Duh, F.-M., Vigdorovich, V., Danilkovitch-Miagkova, A., Lerman, M. I., and Miller, A. D. (2001). Candidate tumor suppressor hyal2 is a glycosylphosphatidylinositol (gpi)-anchored cell-surface receptor for jaagsiekte sheep retrovirus, the envelope protein of which mediates oncogenic transformation. *Proceedings of the National Academy of Sciences*, 98(8):4443–4448.

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics, 2018. *CA: a cancer journal for clinicians*, 68(1):7–30.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Smith, B. D., Smith, G. L., Hurria, A., Hortobagyi, G. N., and Buchholz, T. A. (2009). Future of cancer incidence in the united states: burdens upon an aging, changing nation. *Journal of clinical oncology*, 27(17):2758–2765.

Spencer, C. C., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS genetics*, 5(5):e1000477.

Tabangin, M. E., Woo, J. G., and Martin, L. J. (2009). The effect of minor allele frequency on the likelihood of obtaining false positives. In *BMC proceedings*, volume 3, page S41. BioMed Central.

Tennessen, J. A., Bigham, A. W., OConnor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *science*, 337(6090):64–69.

Udabage, L., Brownlee, G. R., Nilsson, S. K., and Brown, T. J. (2005). The overexpression of has2, hyal-2 and cd44 is implicated in the invasiveness of breast cancer. *Experimental cell research*, 310(1):205–217.

Vanharanta, S., Marney, C. B., Shu, W., Valiente, M., Zou, Y., Mele, A., Darnell, R. B., and Massagué, J. (2014). Loss of the multifunctional rna-binding protein rbm47 as a source of selectable metastatic traits in breast cancer. *Elife*, 3.

Vansteelandt, S., Goetgeluk, S., Lutz, S., Waldman, I., Lyon, H., Schadt, E. E., Weiss, S. T., and Lange, C. (2009). On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects. *Genetic epidemiology*, 33(5):394–405.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333.

Wang, X. (2014). Firth logistic regression for rare variant association tests. *Frontiers in Genetics*, 5:187.

Ware Jr, J. E., Kosinski, M., and Keller, S. D. (1996). A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Medical care*, 34(3):220–233.

Whitlock, M. C. (2005). Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *Journal of evolutionary biology*, 18(5):1368–1373.

Yang, P., Cheville, A. L., Wampfler, J. A., Garces, Y. I., Jatoi, A., Clark, M. M., Cassivi, S. D., Midthun, D. E., Marks, R. S., Aubry, M.-C., et al. (2012). Quality of life and symptom burden among long-term lung cancer survivors. *Journal of Thoracic Oncology*, 7(1):64–70.

Yates, J. W., Chalmer, B., McKegney, F. P., et al. (1980). Evaluation of patients with advanced cancer using the karnofsky performance status. *Cancer*, 45(8):2220–2224.

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Elvestad, M. B., Wolford, B. N., LeFaive, J., VandeHaar, P., Gifford, A., Bastarache, L. A., Wei, W.-Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., and Lee, S. (2017). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *bioRxiv*.