SF 424 R&R and PHS-398
SpecificTable of Contents

PI: **Sun, Wei**	Title: Estimation and association analysis of biomarkers for tumor immune microenvironment	
	FOA: PA16-175	
	FOA Title: Exploratory Grants in Cancer Epidemiology and Genomics Research (R21)	
	Organization: FRED HUTCHINSON CANCER RESEARCH CENTER	
	Department: Biostatistics & Biomathematics	
Senior/Key Personnel:	*Organization:*	*Role Category:*
Wei Sun Ph.D	Fred Hutchinson Cancer Research Center	PD/PI
Michael Wu Ph.D	Fred Hutchinson Cancer Research Center	Co-Investigator

OMB Number: 4040-0001
Expiration Date: 10/31/2019

RESEARCH & RELATED OTHER PROJECT INFORMATION

1. Are Human Subjects Involved?* ● **Yes** ❍ **No**

1.a. If YES to Human Subjects

Is the Project Exempt from Federal regulations? ● Yes ❍ No

If YES, check appropriate exemption number: 1 __ 2 __ 3 ✔ 4 __ 5 __ 6

If NO, is the IRB review Pending? ❍ **Yes** ❍ **No**

IRB Approval Date:

Human Subject Assurance Number 00001920

2. Are Vertebrate Animals Used?* ❍ **Yes** ● **No**

2.a. If YES to Vertebrate Animals

Is the IACUC review Pending? ❍ Yes ❍ No

IACUC Approval Date:

Animal Welfare Assurance Number

3. Is proprietary/privileged information included in the application?* ❍ **Yes** ● **No**

4.a. Does this project have an actual or potential impact - positive or negative - on the environment?* ❍ **Yes** ● **No**

4.b. If yes, please explain:

4.c. If this project has an actual or potential impact on the environment, has an exemption been authorized or an environmental assessment (EA) or environmental impact statement (EIS) been performed? ❍ Yes ❍ No

4.d. If yes, please explain:

5. Is the research performance site designated, or eligible to be designated, as a historic place?* ❍ **Yes** ● **No**

5.a. If yes, please explain:

6. Does this project involve activities outside the United States or partnership with international collaborators?* ❍ **Yes** ● **No**

6.a. If yes, identify countries:

6.b. Optional Explanation:

PROJECT SUMMARY

Molecular features derived from tumor samples (e.g., somatic mutations, gene expression, or DNA methylation) can be very useful biomarkers for epidemiology studies. Recent success of immunotherapy demonstrated that tumor immune microenvironment plays a crucial role for tumor growth and inhibition. Therefore, biomarkers derived from tumor immune microenvironment are great additions to many large epidemiology studies that have access to tumor samples. In this project, we propose to develop a set of statistical methods and computational tools to study biomarkers in tumor immune microenvironment, and as a demonstration, apply them to analyze the omic data from The Cancer Genome Atlas (TCGA). Specifically, we will estimate immune cell composition in the TCGA samples using gene expression and/or DNA methylation data, which can be collected from either fresh frozen or formalin-fixed paraffin-embedded (FFPE) samples. Next we will use immune cell composition to construct prognostic signatures of patient survival time. Our methods and software packages will provide important resources that will enable new epidemiology studies, such as association of immune features with environmental/genetic factors, or cancer risk prediction for cancer subtypes defined/refined by immune biomarkers.

PROJECT NARRATIVE

We propose to develop statistical methods and software packages to study tumor immune microenvironment, and associate immune cell composition with survival time. Our project will break new ground for epidemiology studies, for example, to enable stratified association analysis for patients with particular tumor immune microenvironment.

FACILITIES AND RESOURCES

FRED HUTCHINSON CANCER RESEARCH CENTER

The Fred Hutchinson Cancer Research Center (Fred Hutch), home of three Nobel Laureates, is an independent, nonprofit research institution dedicated to the development and advancement of biomedical research. Fred Hutch's organizational mission is the elimination of cancer as a cause of human suffering and death. In recent years, Fred Hutch took the lead role in forming the Fred Hutch/University of Washington Cancer Consortium. This combined comprehensive cancer center includes Fred Hutch, the University of Washington, and Children's Hospital and Regional Medical Center. These institutions have a long history of collaboration across the disciplines of basic, clinical and public health sciences. The Consortium, one of 40 NCI-designated comprehensive cancer research centers nationwide, offers new opportunities to reduce suffering and mortality from cancer through knowledge gained from increased interdisciplinary research. In addition, the Hutchinson Center enjoys productive industry collaborations that enhance its ability to achieve advances in cancer research and treatment. The Hutchinson Center is organized into six divisions: the five scientific divisions of Basic Sciences, Clinical Research, Public Health Sciences, Human Biology, and Vaccine and Infectious Diseases; and the divisions of Administration and Development.

The Division of Public Health Sciences is home to the nation's oldest and largest program devoted to cancer-prevention research - an important endeavor, considering that many cancers may be avoidable by changes in lifestyle. The Division was originally established within Fred Hutch in 1975 as the Program in Epidemiology and Biostatistics. In 1983, it gained Division status coincident with the creation of the Cancer Prevention Research Program – the first NCI funded cancer prevention research unit. Since its inception, the Division has grown to be the largest of Fred Hutch's five scientific Divisions with over 141 faculty and 525 staff working towards the Division's mission of identifying strategies that will ultimately reduce the incidence of and mortality from cancer and other diseases.

The PHS Division is organized into five administrative programs. Each of these programs has a faculty with a wide range of interests and an interactive and interdisciplinary research orientation.

Faculty in the PHS Division also actively participate in several of Fred Hutch's interdisciplinary scientific programs, including leadership of the Center's gastroenterology, prostate, and breast cancer initiatives. Research in the PHS Division focuses on determining causes of cancer, helping to identify and assess effective screening and treatment methods, developing prevention strategies that reduce the risk of cancer, and developing research strategies to assist people to change behaviors toward healthier lifestyles. This prevention-oriented research also extends to other diseases, including HIV/AIDS, cardiovascular disease, diabetes, and fractures.

The Program in Biostatistics and Biomathematics is within the Public Health Sciences Division. Broadly, projects and faculty within the Program employ statistics and mathematical principles to analyze biological and genetic data and processes and evaluate diagnostic, therapeutic and preventive medical methods and practices. Individually, projects and faculty might provide statistical collaboration and coordination for research programs within and outside the center, develop and evaluate new quantitative methods for the efficient design and analysis of a broad range of biomedical studies and construct biomathematical models of carcinogenesis and other biological processes.

Office space and furnishings are available within the Robert M. Arnold PHS Building, a state-of-the-art facility. Core program facilities available are located on Levels D and 1-5, and include offices for scientists, research associates and research support staff. In addition, research projects within the PHS Division can make use of specialized spaces for conferences, telephone and in-person interviewing, exercise studies, feeding studies and medical examination functions. The on-line Event Management System allows for scheduling appropriately-sized conference rooms and support services, including individualized room set-up, computerized projection, video conference connections, and catering for meetings held within the Arnold Building and across the Day Campus. PHS researchers are located in close proximity of the Consortium shared resources, other Fred Hutch scientific divisions and facilities, and the PHS Laboratories.

Computing Resources are particularly crucial for executing GDAC@FH, because of 1) heavy computing, 2) huge amount of genomic and clinical data, 3) need for real-time exchanges with private and public clouds, and 4) data security. All resources and services are complimentary to grant funded projects of Fred Hutch faculty, except for “extended services” which involve charges. In the following, we provide in-depth introduction on our current computing resources:

Scientific Computing (Center IT)
The Scientific Computing group is staffed with 5 FTE and provides the following services to Fred Hutch research groups and shared resources:

- High Performance Computing (access to the Gizmo HPC cluster including large memory machines, scratch spaces and web gateways for job submission through web browsers)
- General Linux/Unix support (Linux Desktop Support, managing applications on departmental Linux servers)
- Software Development Support (SCM/Subversion source code management, code evaluation for R, Python, shell scripting support, software packaging, performance evaluation)
- HPC and Linux/Unix Training
- File and Data management assistance / Data archiving

Local Computing Resources
The 'Gizmo' cluster is currently equipped with 495 compute nodes / 2520 cpu cores and more than 24 TB of main memory (RAM) and is connected to a fully redundant Isilon high performance storage cluster via Cisco Nexus 7000 series 10G networking equipment. Gizmo can directly access a storage capacity of more than 1 Petabyte and has dedicated high performance scratch space of 150TB with a maximum throughput of 4GB/s. Gizmo consists of:

- 3 Intel Xeon (E5-2697v3, 2.6 Ghz) nodes, each with 28 cores / 384 GB RAM / 10G network for a total of 84 cores / 1152 GB RAM (head / development nodes)
- 456 Intel Xeon (E3-1270/1241 v3, 3.5 Ghz) nodes, each with 4 cores / 32 GB RAM for a total of 1824 cores / 14592 GB RAM
- 10 Intel (E5-2667v3, 3.2 Ghz) nodes, each with 16 cores / 256 GB RAM / 10G network for a total of 160 cores / 2560 GB RAM
- 3 Intel Xeon (E5-2697v3, 2.6 Ghz) nodes, each with 28 cores / 786 GB RAM / 10G network for a total of 84 cores / 2304 GB RAM
- 23 Intel Xeon (E5-2670, 2.6 Ghz) nodes, each with 16 cores / 64 GB RAM for a total of 368 cores / 1472 GB RAM

Each cluster node contains 1-6 TB local disk space which can be used for non-shared temporary data. 16 Nodes are equipped with fast 400GB NVMe flash disks for scratch space. All of the cluster nodes are 64-bit systems running the current version of the Ubuntu LTS Linux operating system. A variety of standard software is installed on the head nodes and on each work node. This includes the current version of R which is updated on a monthly basis as well as several MPI / parallel processing frameworks.

A total of 352 nodes or 1672 cpu cores are equally shared by all Fred Hutch research groups. The Slurm HPC scheduler is used to implement fair sharing of HPC resources and ensures that all research groups can access a minimum number of resources at all times. In this cluster segment all Principal Investigators are given the same priority. The remainder of the resource is owned by individual PIs and managed in a condominium model. 50% of these “private nodes” in the condominium are lightly used. When systems are not used by their owners, other research groups are permitted to use them temporarily to maximize the return of the investment.

Cloud Computing Resources
The cloud team is currently staffed with 6 individuals from the Scientific Computing and Information Security teams. The Center is currently leveraging Amazon Web Services (AWS) for its Infrastructure as a Service (IaaS) cloud platform. Our internal network has been extended into a secure AWS Virtual Private Cloud (VPC) that is covered under a HIPAA Business Associates Agreement (BAA) with Amazon. This environment is fully audited and integrated with our Security Information and Event Management (SIEM) system.

This environment allows Researchers to quickly provision computing, storage and database resources with the ability to scale systems both vertically (scale up) and horizontally (scale out) easily to meet just about any performance or capacity requirement.

In addition to the secure VPC, we also have a public VPC for external collaborations.

Data Storage Service (Center IT)
Storage Services are staffed with 3 FTE (24/7 on call support)

The Fast File storage service provides a high performance Posix file system. The Service is based on Isilon X Series scale out network attached storage technology and is mirrored to Isilon Nearline (NL) Series located in a different building on campus. The system can be accessed via CIFS and NFS protocols. Data protection is provided by file system snapshots which are stored on the system for 7 days. In addition 2 copies of all data is stored on tape backup of which one copy is kept at an offsite location.

The Economy File storage service provides a high capacity object storage system that can scale to sizes > 50 Petabyte at low cost. The Service is based on commodity storage hardware, open source cloud storage technology (Openstack Swift) and is commercially supported by swiftstack.com. The system can be accessed via Swift or S3 protocols. 3 replicas (copies) of the data are stored in 3 different buildings on campus providing high resiliency and data protection. The system is equipped with a recycle bin that protects against accidental deletions of files and currently allows for data restoration within 60 days after deletion.

Extended storage services:
Since Nov 1, 2014 Fred Hutch started charging for data storage usage:

- Every Principal Investigator receives an allocation of 10 TB (5 TB of high performance storage “Fast File” and 5 TB of slower storage “Economy File”) as complimentary service. Usage beyond that amount will incur a monthly cost per the published storage price list.
  (current pricing: $40 per TB/month for “Fast File” and $3 per TB/month for “Economy File”)

Information Security Office (Center IT)
The Information Security Office (ISO) is staffed with 5 FTE. The ISO maintains a highly available Intrusion prevention system and high performance firewall and offers services such as encryption, forensics and consulting

Fred Hutch Information Technology Core Services (Center IT)
Center IT is staffed with nearly 120 FTE (including FTE mentioned above) and provides support for a complex heterogeneous information technology environment.

Center IT manages the Fred Hutch storage services which consist of a Nimble Storage Area Network (SAN), a NetApp Enterprise storage cluster for enterprise file services and an Isilon Storage cluster to provide high throughput data access for high performance computing. The total networked storage capacity is currently 2 PetaByte. Data protection is implemented by DataDomain appliances in conjunction with Commvault Simpana backup software and IBM Tivoli Storage Manager.

Many of the server services are provided by virtual systems using the advanced VMWare VSphere technology. The virtual systems have the added protection of being recoverable through snapshots which are taken and stored on a daily basis. The vSphere environment is configured as self service Enterprise cloud and currently hosts more than 800 virtual machines.

The network infrastructure consists of multiple Local Area Networks (LANs) connected to the Fred Hutch network. Intel personal computers, Macintosh systems, and LINUX/UNIX workstations are connected to servers running different operating systems. The Fred Hutch network is connected to the Internet through the Pacific Northwest GigaPop Network (PNWGP) with 1 Gbit/s. The Network is protected by redundant firewalls and a redunant intrusion prevention system. There are over 5,000 local workstations, printers, and servers connected to the Fred Hutch Network.

Center IT supports both UNIX/IMAP (Zimbra) and Microsoft Exchange for e-mail services in a highly available configuration.

Center IT also provides support for over 60 applications that are made available to the entire Center including an enterprise level SharePoint Collaboration platform.

Biostatistics Share Resource is a Center-wide biostatistics resource, jointly managed by the PHS Program in Biostatistics and Biomathematics along with Clinical Statistics. This shared resource has two primary functions. The first function is to provide statistical computing supports to computational faculty members who need to have following supports: 1) implementing novel methodologies in R, MATALB, C/C++, Java, Python, or any computer languages, 2) conducting simulation studies to test new methods, 3) processing various high dimensional data (RNAseq, microbiome, proteomics, metabolomics, etc.), 4) managing processed data (omics and clinical data), 5) performing customized data analyses in R, MATLAB, SAS, and SPSS. The second function is to provide biostatistical consultations to faculty members who are affiliated with public health sciences, clinical sciences, cancer biology or basic sciences, and to assist them with statistical analyses with various software packages.

Collaborative Data Services (CDS):
Collaborative Data Services (CDS), based at Fred Hutch, provides a series of core services to support investigators and projects at the Hutch and across the Seattle Cancer Consortium. CDS is part of Fred Hutch's Center-wide Shared Resources, and is partially supported by a P30 Cancer Center Support Grant award. CDS is invested in supporting research infrastructure that enhances collaborative, transdisciplinary research. We are comprised of three cores: Programming, Data Control, and Interviewing, and ancillary services such as tracking participants who are lost to follow-up. CDS has many years of working with Fred Hutch investigators, and the approach to designing and implementing data collection, interviewing, and programming is rigorous, cost-effective and of high quality.

Shared Resources:
The Shared Resources at the Fred Hutchinson Cancer Research Center consist of facilities and/or laboratories which are available for use by Fred Hutch investigators, Cancer Consortium members, and the external academic and biotechnology community. These centralized resources provide support for basic scientific and clinical research and projects within the public health sciences. The facilities give investigators the opportunity to augment their research with resources that would not otherwise be convenient or cost effective in each individual laboratory. Shared Resources also partners extensively with investigators, assisting with the initial experimental design and follow-on data analysis.

Arnold Library:
The Arnold Library provides high quality, responsive services and resources in support of Fred Hutch's research, education and patient care programs. Our physical space houses study carrels with wireless Internet access, patron computers and the Shared Resources Computer Lab. The digital side of our operation encompasses subscription management for more than 25,000 ebooks and over 32,000 online journals and a variety of databases and web services. Librarians curate Fred Hutch researchers profiles, provide center-wide tracking of scholarly publishing, support Center authors with NIH Public Access Policy compliance, manage the Shared Resources website, provide training and support for citation management tools like EndNote, provide reports and consultation on publication metrics, host a course guides system to support faculty instructors, manage the Fred Hutch history archive and administer several institutional repositories.

EQUIPMENT

FRED HUTCHINSON CANCER RESEARCH CENTER (Fred Hutch)

Computer Resources
The staff members of this project have desktop PCs running Windows XP or Linux- all of which are connected to a shared high performance computing resource, the 'Gizmo' cluster which is currently equipped with 389 compute nodes / 3004 cpu cores and more than 13 TB of main memory (RAM) and is connected to a fully redundant Isilon high performance storage cluster via Cisco Nexus 7000 series 10G networking equipment. Gizmo can directly access a storage capacity of more than 1 Petabyte and has dedicated high performance scratch space of 180TB with a maximum throughput of 4GB/s. Gizmo consists of:

- 4 Intel Xeon (E5-2690) nodes, each with 16 cores/384 GB RAM for a total of 64 cores/1536 GB RAM (Head- and development nodes)
- 1 Intel Xeon (7560) node with 64 cores and 512GB RAM
- 228 Intel Xeon (E3-1270 v3) nodes, each with 4 cores/32 GB RAM for a total of 912 cores / 3648 GB RAM
- 23 Intel Xeon (E5-2670) nodes, each with 16 cores/64 GB RAM for a total of 368 cores / 1472 GB RAM
- 85 Intel Xeon (X5650) nodes, each with 12 cores/48 GB RAM for a total of 1020 cores/4080 GB RAM

48 AMD “Istanbul” nodes, each with 12 cores/32 GB RAM for a total of 576 cores/1536 GB RAM

Fred Hutch recently received an award to expand the high-performance computing (HPC) cluster to provide the capacity for the growing computational needs in a broad range of biomedical research studies at Fred Hutch.

Computing Cluster Expansion with 992 Computing cores, 5248GB memory, 272 TB scratch space:
12 Standard Nodes Units (12@4 cores and 32GB & 1gbit)
20 High Core Nodes (16 cores & 128GB & 10gbit)
4 Large Memory Nodes (24 cores & 384GB & 10gbit)
4 Scratch Storage Nodes
1 10G Network Switch Appliance

Hardware
Hardware is chosen to include redundancy features such as RAID disk setups, dual power supplies and Network Interface Cards to take advantage of the power and network systems present Center-wide. The major hardware supplier is Dell, with whom we have a close working relationship and a very responsive support system.

Software
Analysis tasks are run with a range of statistical packages such as R, Matlab, STATA and SAS, standard programming languages as well as custom applications developed in-house. Databases at the Center are currently run with a range of technologies including Oracle, MS SQL Server and PostgreSQL. A range of Windows and Linux-based applications are available for use by the staff. Software Development Support (SCM/Subversion source code management, code evaluation for R, C, Python, shell scripting support, software packaging, performance evaluation) is also provided.

Servers
Servers are situated in restricted-access; environmentally-controlled server rooms equipped with locked server cabinets and FM200 fire suppression systems. The Center’s main infrastructure servers are Windows-based systems running many of the domain, file-serving, e-mail and backup systems. In addition there are Linux based computing servers and clusters used for research and analysis projects.

Network Infrastructure
The network infrastructure consists of multiple Local Area Networks (LANs) connected to the Fred Hutch Network. Intel personal computers, Macintosh systems, and LINUX/UNIX workstations are connected to servers running different operating systems. The Fred Hutch network is connected to the Internet through the Pacific Northwest GigaPop Network (PNWGP) with 1 Gbit/s. The Fred Hutch Network is protected by

redundant firewalls and a redundant intrusion prevention system. There are over 3,000 local workstations, printers, and servers connected to the Fred Hutch Network. Fred Hutch supports both UNIX/IMAP and Microsoft Exchange for e-mail services. Fax and photocopier machines are also available throughout the Center. Center IT also provides support for over 60 applications that are made available to the entire Center including an enterprise level SharePoint Collaboration platform.

OMB Number: 0925-0001
Expiration Date: 10/31/2018

PHS 398 COVER PAGE SUPPLEMENT

1. Human Subjects Section

Clinical Trial? ❍ Yes ● No

*Agency-Defined Phase III Clinical Trial? ❍ Yes ❍ No

2. Vertebrate Animals Section

Are vertebrate animals euthanized? ❍ Yes ● No

If "Yes" to euthanasia

Is the method consistent with American Veterinary Medical Association (AVMA) guidelines?

❍ Yes ❍ No

If "No" to AVMA guidelines, describe method and proved scientific justification

3. *Program Income Section

*Is program income anticipated during the periods for which the grant support is requested?

❍ Yes ● No

If you checked "yes" above (indicating that program income is anticipated), then use the format below to reflect the amount and source(s). Otherwise, leave this section blank.

*Budget Period *Anticipated Amount ($) *Source(s)

4. Human Embryonic Stem Cells Section

*Does the proposed project involve human embryonic stem cells? ❍ Yes ● No

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://grants.nih.gov/stem_cells/registry/current.htm. Or, if a specific stem cell line cannot be referenced at this time, please check the box indicating that one from the registry will be used:

❑ Specific stem cell line cannot be referenced at this time. One from the registry will be used.

Cell Line(s) (Example: 0004):

5. Inventions and Patents Section (RENEWAL)

*Inventions and Patents: ❍ Yes ● No

If the answer is "Yes" then please answer the following:

*Previously Reported: ❍ Yes ❍ No

6. Change of Investigator / Change of Institution Section

❑ Change of Project Director / Principal Investigator

Name of former Project Director / Principal Investigator

Prefix:

*First Name:

Middle Name:

*Last Name:

Suffix:

❑ Change of Grantee Institution

*Name of former institution:

PHS 398 RESEARCH PLAN

OMB Number: 0925-0001
Expiration Date: 10/31/2018

Introduction	
1. Introduction to Application (Resubmission and Revision)	
Research Plan Section	
2. Specific Aims	wsun_SpecAim_2017_02_16.pdf
3. Research Strategy*	wsun_ResStrat_2017_02_16.pdf
4. Progress Report Publication List	
Human Subjects Section	
5. Protection of Human Subjects	wsun_ProtHum_2017_02_16.pdf
6. Data Safety Monitoring Plan	wsun_DataSafe_2017_02_16.pdf
7. Inclusion of Women and Minorities	wsun_InclWM_2017_02_16.pdf
8. Inclusion of Children	wsun_InclChild_2017_02_16.pdf
Other Research Plan Section	
9. Vertebrate Animals	wsun_VertAnim_2017_02_16.pdf
10. Select Agent Research	wsun_SelAgent_2017_02_16.pdf
11. Multiple PD/PI Leadership Plan	
12. Consortium/Contractual Arrangements	wsun_ConsContr_2017_02_16.pdf
13. Letters of Support	wsun_LettSupp_2017_02_16.pdf
14. Resource Sharing Plan(s)	wsun_ResShar_2017_02_16.pdf
15. Authentication of Key Biological and/or Chemical Resources	
Appendix	
16. Appendix	

Title: Estimation and association analysis of biomarkers for tumor immune microenvironment

SPECIFIC AIMS

The long term goal of this project is to construct biomarkers characterizing **tumor immune microenvironment (TIME)**, and assess associations between TIME biomarkers and epidemiology variables, molecular features, or clinical outcomes. **In this exploratory grant, we focus on computational methods to estimate TIME biomarkers from gene expression and/or DNA methylation data, and to assess their roles as prognostic biomarkers for patient survival time. Our project, if proved to be successful, will break new ground for many potential usages of TIME biomarkers in epidemiology or translational studies.** For example, use them as predictive biomarkers for cancer immunotherpay, or use them to define cancer subtypes, and thus enable association studies between environmental/genetic factors and cancer risks within each TIME subtype.

Cancer immunotherapy (e.g., immune checkpoint therapy [1] or adoptive cell therapy [2]) prompts the immune system to identify and kill cancer cells. The phenomenal successes of cancer immunotherapy in a subset of cancer patients underscores the importance of TIME for tumor growth/inhibition [3]. However, it is very challenging to estimate immune cell composition in TIME using omic data (e.g., gene expression or DNA methylation) for at least three reasons. First, different cell types may have very similar gene expression and/or DNA methylation profiles and thus are difficult to distinguish. Second, cell type decomposition relies on cell-type-specific omic data, which are often collected from blood; and gene expression or DNA methylation of certain cell type may vary between blood and TIME. Third, different types of immune cells have variable cell sizes and transcriptional activity that affect the total amount of transcripts. Furthermore, it is also very important to account for the uncertainty of cell type decomposition in the following association analysis.

A few groups have studied TIME using gene expression data from The Cancer Genome Atlas (TCGA) project [4–8]. These pioneering works have demonstrated promising results to study TIME, but also bear some limitations. For example, some of these works estimate immune cell presence using the expression of only one or a few genes [4,5], or identify immune cell types that are over-represented in TIME, instead of quantitatively estimating immune cell composition [6]. Newman et al. 2015 [7] and Li et al. 2016 [8] estimated immune cell composition using support-vector regression or non-negative least squares, respectively. Their approaches rely on the assumption that cell-type specific gene expression measured in blood are the same as in TIME and they ignore the fact that different types of cells have different sizes and/or transcriptional activity. We propose a statistical framework to estimate immune cell composition using both gene expression and DNA methylation data that can overcome all of the aforementioned limitations. In addition, we will also develop statistical methods to use immune cell composition to predict survival time while accounting for the uncertainty of immune cell decomposition.

Specific Aim 1: Estimate immune cell composition. We will develop a likelihood-based framework to combine gene expression and DNA methylation data to estimate immune cell composition. Inclusion of DNA methylation helps correct the bias due to cell sizes or transcriptional activities since there are always two copies of DNA regardless of cell type. Combining these two types of data helps to better separate closely related cell types. In addition, our model allows a subset of molecular features (gene expression or DNA methylation) to be different between blood and TIME. This likelihood-based framework also allows us to quantify estimation uncertainty, which will be very useful for down-stream association analysis.

Specific Aim 2: Construct prognostic immune signature of patient survival time. Previous studies have shown that immune cell type proportions in tumor microenvironment can predict patient survival time [9–11]. Building on these works, we will jointly model patient survival time and DNA methylation/gene expression with immune cell proportions as latent variables, so that we can assess the association between immune cell type signature and survival time while accounting for the uncertainty of cell type deconvolution.

All the methods developed in this R21 grant will be implemented in open source R packages. The computational intensive parts will be implemented using C or C++ and we will use R package Rcpp for R and C++ integration. We will apply our method to study TCGA data as well as other datasets [12, 13]. Our methods/software and data analysis results will help build a statistically rigorous foundation to use TIME biomarkers in epidemiological or clinical studies.

RESEARCH STRATEGY

(a) Significance

We will develop statistical methods to profile the immune landscape of tumor immune microenvironment (TIME), and to identify immune cell signatures that are associated with patient survival time from The Cancer Genome Atlas (TCGA) samples. The deliverables of our research projects, including statistical methods, software packages, and results from TCGA data analysis, will help build a statistically rigorous foundation to use TIME biomarkers in epidemiological or clinical studies.

In **Specific Aim 1**, we will develop a computational approach to estimate immune cell composition of tumor samples using gene expression and/or DNA methylation data. We will integrate both types of data in a likelihood-based framework to estimate model parameters. Then our model can be applied for cell type deconvolution with one type or both types of data available. We will apply our methods to estimate immune cell composition in all cancer types of TCGA. Though we will start with colon cancer and lung cancer, based on local expertise in Fred Hutch. Our collaborator Dr. Ulrike Peters is the PI of Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO). Another collaborator, Dr. McGarry Houghton, is a pulmonologist who had conducted comprehensive studies of immune cell composition in lung cancer samples. **Our method will significantly improve the accuracy of immune cell composition estimation, and thus improve the sensitivity and specificity for down-stream association analysis.** The improved performance are due to the fact that we borrow information between gene expression and DNA methylation, and the flexibility of our method to allow cell-type-specific gene expression/DNA methylation to vary between blood and TIME. Our likelihood-based framework also provides estimation uncertainty of cell type composition, which is crucial for the next step association analysis.

Flow cytometry or single cell sequencing data are two popular alternative approaches to estimate cell type composition. However, both approaches are not practical for clinical use. Flow cytometry is labor intensive and it requires large amount of fresh processed tumor tissue. Single cell sequencing also has very high requirement on sample quality and it is cost-prohibitive to sequence most cells within a tumor sample. Even tens of thousands of cells still constitute a tiny proportion of the cells within a tumor sample, and thus may not give representative estimates of cell type composition. Therefore a computational method that can estimate immune cell composition from bulk tumor sample is of great value for many down-stream analysis, for example, to identify biomarkers for immunotherapy efficacy, or to identify tumor subtypes based on immune microenvironment.

In **Specific Aim 2**, we will use immune cell composition to construct prognostic signatures for survival time. Previous studies have shown that immune cell composition within human cancers have prognostic value in multiple types of cancers [3, 9, 11, 14]. Two types of immune cells may have very similar gene expression or DNA methylation, and thus cell type proportion estimates of these two cell types will have a large degree of uncertainty, which should not be ignored for the purpose of survival time prediction. By jointly modeling survival time and omic data while treating immune cell composition as latent variables, **we provide accurate estimate of survival time association while accounting for the uncertainty of immune cell decomposition. In addition, we also propose an effective penalized estimation approach to borrow information across cell types with similar gene expression and DNA methylation, and thus improve the stability of our prediction.** We will apply our method to analyze TCGA data, and our results will provide benchmarks of survival time prediction using cell type composition across all major cancer types.

Research Team. The PI, Dr. Wei Sun, has extensive experience working on different types of omic data including somatic copy number aberration [15, 16], epigenetic marks [17, 18], and gene expression [19, 20], as well as software development. Dr. Sun was an Associate Professor at UNC Chapel Hill with a joint appointment in the Department of Biostatistics and the Department of Genetics. He moved to Fred Hutchinson Cancer Research Center (FHCRC) at 2015 fall as an Associate Member in the Public Health Sciences division. Dr. Michael Wu is an Associate Member at FHCRC. He moved from UNC to FHCRC at 2013. Dr. Wu's work for rare variant association using SKAT (Sequence Kernel Association Test) has made huge impact in genetic association studies. He is also an expert on DNA methylation data analysis [21–23], and he will mainly contribute on the DNA methylation part of this project. Drs. Sun and Wu knew each other very well since their overlap at UNC. Their expertise complement with each other, and they form a strong team to carry out the proposed research projects.

(b) Innovation

Estimation of immune cell composition. The vast majority of TCGA studies only focus on tumor cells and consider non-tumor cells within tumor samples as "contamination". However, such "contamination" allows us to study immune landscape in tumor microenvironment. From this perspective, our methods and results are novel and different from most published works on TCGA data analysis. Some pioneering studies of immune cell composition in TCGA tumor samples have adopted some ad hoc approaches, e.g., capture the immune system activity by the expression of two or three genes [4, 5] or assess whether genes annotated to a specific immune cell type are over-expressed using gene set enrichment analysis (GSEA) [6]. Newman et al. 2015 [7] and Li et al. 2016 [8] estimated immune cell composition using support-vector regression or non-negative least squares, respectively. Their approaches rely on the assumption that the immune cell-type specific expression measured in blood are the same as in TIME. In addition, they ignore difference in cell sizes and do not provide cell type deconvolution uncertainty, which is very important for association analysis of immune cell composition. **In contrast to these methods, our method has the following novelties. (1) Borrowing information between gene expression and DNA methylation data. (2) Alleviate bias due to cell size effects on gene expression by exploiting DNA methylation data. (3) Allow gene expression or DNA methylation of one cell type to be different between the cells in TME and the cells in reference samples, which are usually collected from blood.**

Associate immune cell composition with survival time. Previous works have assessed the association between immune cell composition and survival time with a two-step approach: first estimate immune cell composition and then associate such estimates with survival time while ignoring estimation uncertainty in the first step. Failure to account for estimation uncertainty is equivalent to regression ignoring measurement error, and it is well known that such an approach can lead to estimation bias, reduced power, and/or inflated type I error [24, 25]. **The novelty of our method is that we overcome the limitation of the two-step approach by jointly modeling survival time and omic data while including immune cell composition as a latent variable. In addition, we employ penalized estimation on cell-type specific effect to obtain more stable predictions**.

(c) Approach

Preliminary results on TCGA data analysis. We estimated immune cell composition in TCGA colon cancer patients to demonstrate that gene expression and DNA methylation indeed provide consistent, but non-redundant information for cell type deconvolution. DNA methylation of individual immune cell types were obtained from six previous studies [26, 34–38]; and cell-type specific gene expression data were from [7]. We conducted cell type decomposition using gene expression and DNA methylation separately. We employed a generalized linear model (glm) for log-normal distribution to decompose gene expression data and a maximum likelihood estimate to decompose DNA methylation data (see the Approach section for more details). As a preliminary analysis, we only included a few cell types in our studies, and we estimated cell type composition for each sample separately, instead of jointly analyzing all samples to refine the estimates of cell type-specific expression or methylation. Nevertheless, we still observed significant associations between cell type proportion estimates from methylation and gene expression (Figure 1(A-B)). We also demonstrated that such crude estimates of cell type composition can be associated with tumor subtypes. Specifically, colon cancer patients can be classified into two subtypes: hyper-mutated versus non-hyper-mutated. We obtained the somatic mutation calls of these patients from file PANCAN12.mutation.whitelist.maf (https://www.synapse.org/#!Synapse:syn1710431), and based on the number of mutations, we classified the patients with more than 750 mutations as hyper-mutated ones.

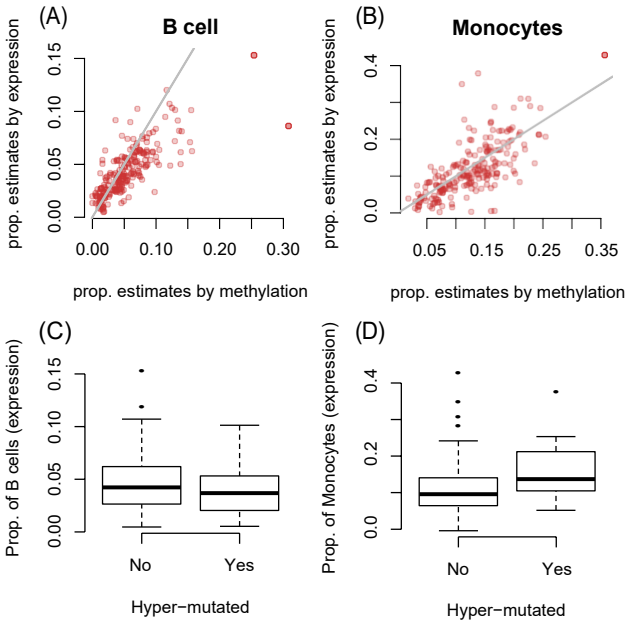

Figure 1: (A-B) Proportions of B cells or Monocytes estimated by methylation versus expression. The grey line is $y = x$. (C-D) Association between hypermutation status and the proportion of B cells (Wilcox test p-value 0.046) or Monocytes (Wilcox test p-value 0.005).

Those hypermutated samples tend to have lower proportion of B cells and higher proportion of monocytes (Figure 1(C-D)). These results are interesting and relevant for immunotherapy because it has been shown that those hyper-mutated colorectal cancer patients have higher response rate to immunotherapy [28].

Specific Aim 1: Estimate immune cell composition.

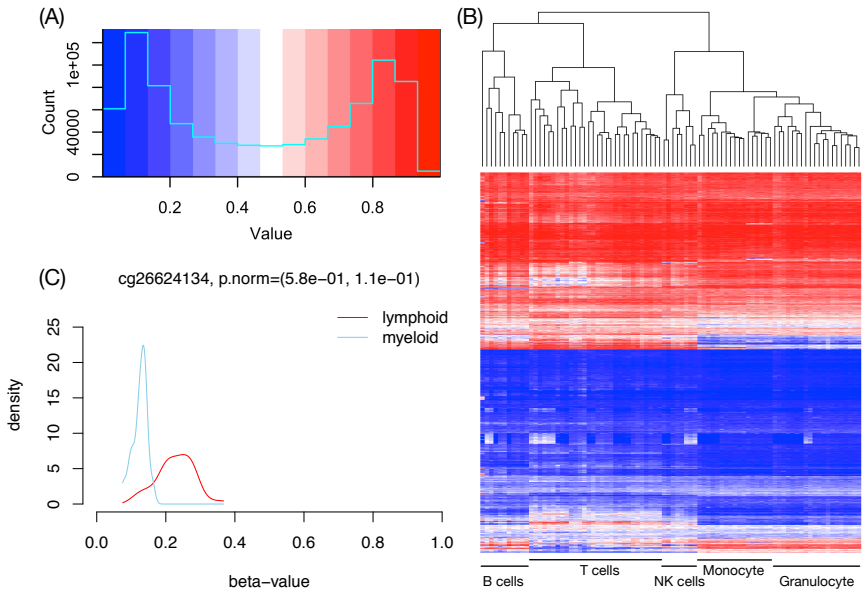

Figure 2: (A) The distribution (and color coding) of the methylation of 9,401 CpGs that are shared between two studies (Illumina 27k array for [27] and Illumina 450k array for [26]) and with standard deviation > 0.02 in both studies. (B) A heatmap of the methylation of 9,401 CpGs (rows) in 86 samples from 10 cell types (columns). These 10 cell types can be further grouped into 5 categories labeled at the bottom. (C) Distribution of the methylation of a CpG in lymphoid cells or myeloid cells, where p.norm are p-values for normality test within each group.

We assume that external gene expression and DNA methylation data from pure cell types are available, and tumor purities have been estimated, e.g., from copy number data [16, 29]. For the deconvolution problem, it is important to ensure the omic data are measured at appropriate scale so that the deconvolution is in linear space [30]. With this in mind, for expression data from RNA-seq, we use read-depth corrected total read count per gene, and for expression data from microarray, we take the intensity values before applying log transformation [30]. For DNA methylation data, we choose the scale of beta-value: $\beta = M/(M+U+100)$ where M and U are methylation and unmethylation probe intensities, respectively. In these scales, we model gene expression data by log-normal distribution and DNA methylation data by normal distribution. As shown in Figure 2, although methylation data across all CpG's form a bimodal distribution, the methylation of each CpG can be reasonably approximated by a normal distribution. Similar checking has shown that log-normal distribution assumption is reasonable for gene expression data. In contrast to more complicated distributions such as a negative binomial distribution for RNA-seq data or a beta distribution for DNA methylation data, the log-normal and normal distributions provide good enough fit of real data for the purpose cell type deconvolution and allow us to develop computationally very efficient implementation.

Next we introduce a few notations to formally define our likelihood model.

- I=number of samples, J=number of genes, K=number of CpGs, and Q=number of non-tumor cell types.
- $\mathbf{Y} = (y_{ji})_{J\times I}$: gene expression data matrix. y_{ji} is the expression of the j-th gene in the i-th tumor sample.
- $\mathbf{Z} = (z_{ki})_{K\times I}$: DNA methylation data matrix. z_{ki} is the methylation of the k-th CpG in the i-th tumor sample.
- $\mathbf{F}_q = (f_{jqr})_{J\times R_q}$: the gene expression data from the q-th cell type with $1 \le q \le Q$. f_{jqr} is the expression of the j-th gene in the r-th replicate, $1 \le r \le R_q$, and R_q is the number of replicates of the q-th cell type.
- $\mathbf{H}_q = (h_{kqr})_{K\times T_q}$: the DNA methylation data from the q-th cell type. h_{kqr} is the DNA methylation of the k-th CpG in the r-th replicate, $1 \le r \le T_q$, and T_q is the number of replicates of the q-th cell type.

Denote normal and log-normal distributions by $\mathcal{N}(\nu, \tau^2)$ and $\mathcal{LN}(\mu, \sigma^2)$, respectively. If $Y \sim \mathcal{LN}(\mu, \sigma^2)$, $\log(Y) \sim \mathcal{N}(\mu, \sigma^2)$, and $E(Y) = \exp(\mu + \sigma^2/2)$. We assume $f_{jqr} \sim \mathcal{LN}(\mu_{jq}, \sigma^2_{jq})$, $h_{kqr} \sim \mathcal{N}(\nu_{kq}, \tau^2_{kq})$, $y_{ji} \sim \mathcal{LN}(\tilde{\mu}_{ji}, \tilde{\sigma}^2_{ji})$, and $z_{ki} \sim \mathcal{N}(\tilde{\nu}_{ki}, \tilde{\tau}^2_{ki})$. Let ρ_{qi} be the proportion of the q-th cell type in the i-th sample. Denote the purity of i-th sample by η_i. Then $\sum_{q=1}^{Q} \rho_{qi} = 1 - \eta_i$. We further assume that in tumor cells, the expression of the j-th gene

follows a log normal distribution $\mathcal{LN}(\mu_{j0}, \sigma^2_{j0})$, and the methylation of the k-th CpG follows a normal distribution $\mathcal{N}(\nu_{k0}, \tau^2_{k0})$. Since summation of normal random variables follows a normal distribution, for DNA methylation data, we have $\tilde{\nu}_{ki} = \eta_i \nu_{k0} + \sum_{q=1}^{Q} \rho_{qi} \nu_{kq}$ and $\tilde{\tau}^2_{ki} = \eta_i^2 \tau^2_{k0} + \sum_q \rho^2_{qi} \tau^2_{kq}$. For expression data, the summation of log-normal random variables can be approximated by a log-normal distribution [31,32]. Let $\gamma_{jq} = E(f_{jqr}) = \exp(\mu_{jq} + \sigma^2_{jq}/2)$. Then $\tilde{\mu}_{ji} + \tilde{\sigma}^2_{ji}/2 = \log(\sum_q \rho_{qi} \gamma_{jq})$, and $\tilde{\sigma}^2_{ji} = \log\left\{\sum_q [\rho^2_{qi} \gamma^2_{jq} \exp(\sigma^2_{jq} - 1)] / (\sum_q \rho_{qi} \gamma_{jq})^2 + 1\right\}$. We have validated in extensive simulations that such log-normal distribution approximation allows us to obtain accurate estimates of immune cell composition. For example, results of one set of simulation is shown in Figure 3. Here we treat the problem of estimating ρ_{qi}'s as a regression problem with observed values of γ_{jq} as covariates. We either fit a linear regression model or a generalized linear model for log-normal distribution, and the latter leads to more accurate estimates.

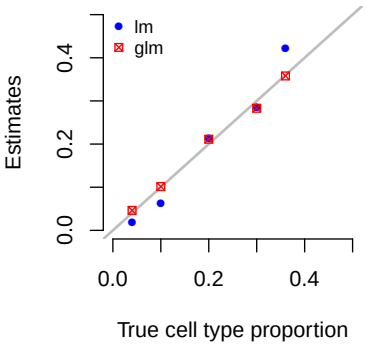

Figure 3: A simulation to estimate the mixture proportion of 5 cell types. Gene expression of the tissue sample is simulated by the summation of 5 random variables, each following a log-normal distribution. lm: linear regression assuming normal distribution. glm: ageneralized linear model assuming log-normal distribution.

Denote the parameters for the q-th cell type to be $\Theta_{qF} = \{\mu_{jq}, \sigma^2_{jq}\}$ and $\Theta_{qH} = \{\nu_{kq}, \tau^2_{kq}\}$. Similarly, denote the parameters for tumor cells to be $\Theta_{0F} = \{\mu_{j0}, \sigma^2_{j0}\}$ and $\Theta_{0H} = \{\nu_{k0}, \tau^2_{k0}\}$. Let Θ be all the parameters, the likelihood function, denoted by $\mathcal{L}(\Theta|\mathbf{Y}, \mathbf{Z}, \{\mathbf{F}_q\}, \{\mathbf{H}_q\})$, can be written as

$$p(\mathbf{Y}|\Theta_{0F}, \{\Theta_{qF}\}, \{\rho_{qi}\}) p(\mathbf{Z}|\Theta_{0H}, \{\Theta_{qH}\}, \{\rho_{qi}\}) \prod_{q=1}^{Q} p(\mathbf{F}_q|\Theta_{qF}) \prod_{q=1}^{Q} p(\mathbf{H}_q|\Theta_{qH}). \quad (1)$$

Maximizing this likelihood function to estimate the parameters is not trivial because of the large number of parameters. We have designed the following coordinate ascend algorithm.

1. Obtain initial estimates of cell-type specific parameters using the data from each cell type: $\hat{\Theta}_{qF} = \arg\max_{\Theta_{qF}} \left[p(\mathbf{F}_q|\Theta_{qF})\right]$, and $\hat{\Theta}_{qH} = \arg\max_{\Theta_{qH}} \left[p(\mathbf{H}_q|\Theta_{qH})\right]$.
2. Obtain initial estimates of the parameters of gene expression or DNA methylation in tumor cells (i.e, $\hat{\Theta}_{0F}$ and $\hat{\Theta}_{0H}$) using the tumor samples with relatively high purity, while ignoring non-tumor cell types.
3. Given $\hat{\Theta}_{qF}$, $\hat{\Theta}_{qH}$, $\hat{\Theta}_{0F}$, and $\hat{\Theta}_{0H}$, estimate $\{\rho_{qi}\}$.
4. Given $\{\hat{\rho}_{qi}\}$, re-estimate $\hat{\Theta}_{qF}$, $\hat{\Theta}_{qH}$, $\hat{\Theta}_{0F}$, and $\hat{\Theta}_{0H}$.

Then we iterate between steps 3 and 4 until convergence to obtain the final estimates. The above estimation procedure can be applied to the situation that cell-type-specific reference is available only for gene expression or DNA methylation data or be extended to include other data types, such as other types of epigenetic data Systematic difference of immune cell composition estimates from gene expression and DNA methylation data implies bias on of gene expression estimates due to variation of cell sizes or transcriptional activity [33]. For example, if the estimated immune cell composition of Neutrophil from gene expression data is systematically lower than the estimates from DNA methylation data, the most likely reason is that a Neutrophil cell has lower total number of transcripts than other types of cells. **We will multiply the cell type specific expression by a "cell size factor" to remove such bias. Such "cell size factor" are identifiable as long as the number of tumor samples is larger than the number immune cell types.**

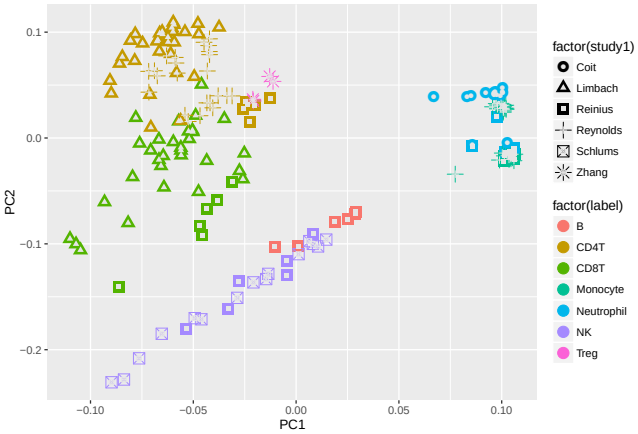

Figure 4: PC1 vs. PC2 for DNA methylation data after cell-type-specific quantile normalization. The shapes of points indicate studies and the colors indicate cell types.

Data analyses We will apply our method to analyze all the TCGA cancer types. As for reference gene expression data, we will employ immune cell expression data used by [6] and [7], which include more than 800 samples for 28 cell types. For methylation, we will combine immune cell methylation data of Illumina 450k arrays from six previous studies [26, 34–38]. For studies comparing methylations

between case and controls, we only take control samples. An initial data harmonization by cell-type-specific quantile normalization show that we can remove the batch effects across studies (Figure 4). Our results of TCGA data analyses will provide not only the immune cell type composition across multiple types of cancer, but also estimates of cell-type-specific gene expression and DNA methylation ($\hat{\Theta}_{qF}$ and $\hat{\Theta}_{qH}$) within tumor microenvironment for each cancer type. **We will use our estimates of cell type composition to cluster cancer samples to identify possible subtypes.** The distance between two samples can be computed by a weighted Kullback-Leibler divergence while the weight for each cell type proportion is inversely proportional to the variance of the proportion estimate. **Given the estimates of $\hat{\Theta}_{qF}$ and/or $\hat{\Theta}_{qH}$, we can estimate cell type composition of any other new samples using either gene expression and/or DNA methylation data.** A caveat of TCGA samples is that they may not be representative of all cancer patients. For example, some samples with very high percentage of immune cell infiltration have been discarded during sample preparation. The potential biased sampling of TCGA does not affect our method development, and we will interpret our results with caution.

Challenges, limitation, and alternative strategies. Most cell-type-specific expression/methylation data were collected from blood, which may not be the same as the expression/methylation of this cell type in TIME. In our model, the omic data from individual cell types provide valuable initial estimates and the omic data from large number of TCGA samples can correct potential bias of those initial estimates. **To further improve the robustness of our method, we refine our model to allow for some genes or some CpGs to be "noise" features that do not follow the decomposition model.** This feature selection step can be implemented by assuming genes (or CpGs) follow a mixture distribution of two components, with one component for "informative" features for cell type decomposition, and the other component for "noise" features. To demonstrate the efficacy of this approach, we performed a preliminary analysis to estimate cell type composition using a methylation dataset with known cell type composition [27]. We first selected 362 informative CpGs that are differentially methylated across cell types, and then estimated cell type composition using non-negative least squares (NNLS) or or maximum likelihood estimate (MLE) based on our mixture model. MLE has better performance than NNLS (Figure 5(A-B)) because MLE can account for the fact that methylation across CpGs have different variances. Next we randomly selected 10% of these 362 CpGs, and purterbed their methylation values. In this situation, our mixture-model-based MLE has much better performance than NNLS method (Figure 5(C-D)).

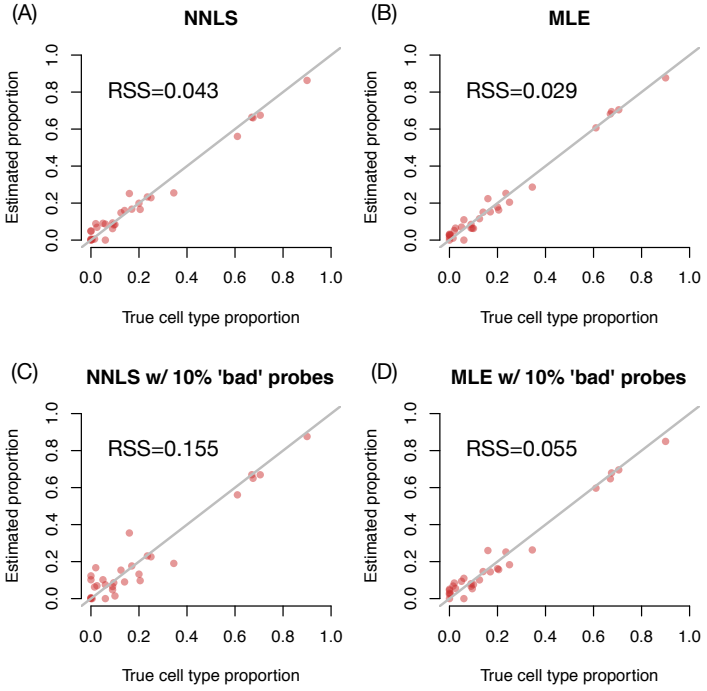

Figure 5: Comparison of the known cell type proportion for 5 cell types in 6 samples (hence 30 points per plot) versus the estimates obtained by non-negative least squares (NNLS) or maximum likelihood estimates (MLE). RSS stands for residual sum squares, which equals to the summation of squared difference between estimated and true cell type proportions.

Results validation. We will carefully evaluate our method using both simulated data and published experimental data. Following previous work [40], gene expression and methylation data of complex tissues will be simulated from a weighted mixture of gene expression and methylation measured in all constituting cell types. Using such simulated data, we will evaluate our method and other methods [7, 11, 40–44]. We expect that our method has better performance because we combine the information of gene expression and DNA methylation. We will also evaluate our cell type composition estimates using the samples for which immune-cell compositions have been established by flow cytometry [7, 42]. Some studies have measured gene expression of multiple types of immune cells in TIME [45, 46], and such data can be used to evaluate our estimates of cell-type-specific expression in TIME.

Specific Aim 2: Construct prognostic immune signature of patient survival time.

Let T_i be the observed survival/censoring time for the i-th sample, and let δ_i be an indictor for actual death. For the i-th sample, denote the other covariates that may be included in the prognostic signature by X_i, and denote the non-tumor cell proportions by $\boldsymbol{\rho}_i = (\rho_{1i}, ..., \rho_{Qi})^T$. Recalling that η_i denotes the pre-estimated tumor purity of the i-th sample, we assume a linear predictor of survival time: $h(\boldsymbol{\rho}_i, \eta_i) = \beta_0 + \sum_q \rho_{iq}\beta_q + \eta_i\gamma$ and we will also explore other forms of h if needed. Let Θ_T be the set of parameters used to model survival time. The joint likelihood function of survival time, DNA methylation, and gene expression is

$$\mathcal{L}(\Theta, \Theta_T | \{T_i\}, \mathbf{Y}, \mathbf{Z}, \{\mathbf{F}_q\}, \{\mathbf{H}_q\}, \{X_i\}) \quad = \quad p(\mathbf{Y}, \mathbf{Z}, \{\mathbf{F}_q\}, \{\mathbf{H}_q\} | \Theta) \prod_{i=1}^{I} p[T_i | h(\boldsymbol{\rho}_i, \eta_i), X_i, \Theta_T]. \tag{2}$$

where $p(\mathbf{Y}, \mathbf{Z}, \{\mathbf{F}_q\}, \{\mathbf{H}_q\}|\Theta)$ has been defined in equation (1). Let $\lambda[T_i|h(\boldsymbol{\rho}_i, \eta_i), X_i]$ be the hazard function, then $p(T_i|h(\boldsymbol{\rho}_i, \eta_i), X_i, \Theta_T) = \lambda[T_i|h(\boldsymbol{\rho}_i, \eta_i), X_i]^{\delta_i} \exp\left\{-\int_0^{T_i} \lambda[t|h(\boldsymbol{\rho}_i, \eta_i), X_i]dt\right\}$. It is well known that one may specify a parametric or semi-parametric form for hazard function. We will employ a parametric form to reduce the computational cost, though change to a semi-parametric form is straightforward if needed. For example, assuming exponential distribution, $\lambda[T_i|h(\boldsymbol{\rho}_i, \eta_i), X_i] = \exp\{h(\boldsymbol{\rho}_i, \eta_i) + X_i\kappa\}$, where κ are regression coefficients for X_i. The likelihood function can be divided into two parts for gene expression/DNA methylation data and survival time. Each part has their own parameters, two parts share $\boldsymbol{\rho}_i$'s. We will employ a block-wise co-ordinate ascend algorithm to obtain MLE, by estimating parameters for each part and $\boldsymbol{\rho}_i$'s iteratively, until convergence.

In the above procedure, one step that needs extra attention is to estimate the coefficients in $h(\boldsymbol{\rho}_i, \eta_i) = \beta_0 + \sum_q \rho_{iq}\beta_q + \eta_i\gamma$ given $\hat{\rho}_{iq}$ and pre-specified η_i. The covariates in this link function (i.e., ρ_{iq} and η_i) must satisfy a constraint that $\eta_i + \sum_{q=1}^{Q} \rho_{iq} = 1$. This is the so-called compositional data, which consists of the proportions that add up to 1. The constraint that the summation of compositions equals to 1 renders many standard statistical estimation methods inappropriate [47]. Following Lin et al. (2014) [48], we adopt the solution to use log transformed compositions: $h(\boldsymbol{\rho}_i, \eta_i) = \beta_0 + \sum_q \log(\rho_{iq})\beta_q + \gamma \log(\eta_i)$, with extra constraint that $\sum_{q=1}^{Q} \beta_q + \gamma = 0$.

Results validation. We will evaluate our method by analyzing TCGA data. Our goal is to make best prediction instead of hypothesis testing and we will evaluate our results using prediction accuracy measured by C-index [49] calculated through cross-validation. We will compare our method versus methods that use gene expression and methylation data separately, as well as the two-step approach that first estimate cell type composition and then predict survival time assuming cell type composition is known.

Challenges, limitations, and alternative strategies. Some immune cell types may have highly similar gene expression and DNA methylation profiles and thus they are very difficult to distinguish. For the purpose of predicting survival time, it is not necessary to distinguish such highly similar cell types. To address this challenge, we impose a penalty in the likelihood function to encourage similar cell types to have similar effect sizes on survival time. Specifically, we will first run hierarchical clustering of individual cell types using both gene expression and DNA methylation data and record their order in the resulting dendrogram. This order is a one-dimensional projection that reflects the similarities across Q cell types such that two cell types that are next to each other tend to (though not necessarily) be more similar. Denote this order by $(\pi_1, ..., \pi_Q)$, where π_t takes values from 1 to Q and it is the index of the cell type that is located at the t-th position. Now the regression model can be written as $h(\boldsymbol{\rho}_i, \eta_i) = \beta_0 + \sum_{t=1}^{Q} \log(\rho_{i,\pi_t})\beta_t + \gamma \log(\eta_i)$. We add a fused Lasso penalty [50] to the log likelihood function in the form of $\lambda \sum_{t=2}^{Q} |\beta_t - \beta_{t-1}|$, where λ is a tuning parameter that can be selected by BIC. This penalty encourages adjacent cell types to have similar regression coefficients and thus improves the robustness of prediction. The penalized log likelihood can be maximized following the approach proposed by Lin et al. (2014) [48]. We have been working on survival time prediction using TCGA data in other projects [51] and we are aware that the survival time in TCGA has limited quality (e.g., non-cancer-specific survival) and often has high percentage of censoring. Nevertheless, numerous studies have demonstrated that patients of different stages or subtypes can have different survival time distributions. Therefore it is valuable to predict survival time using TCGA data, and we will also explore other datasets with both omic data and survival time [12].

Timeline. We plan to finish the works of Specific Aims 1 and 2 in years 1 and 2, respectively. We will publish our methods and results as well as our software packages.

PROTECTION OF HUMAN SUBJECTS

This project aims to develop statistical methods and apply these methods to analyze publicly available and de-identified omic dataset from human subjects. No personnel identification information will be used and no individual will be recruited in this project. Therefore, our project fits the Exemption 4 for human subject study: “research involving the collection or study of existing data or specimens if publicly available or information recorded such that subjects cannot be identified.”

Risks to the Subjects

a. Human Subjects Involvement and Characteristics:

This project involves no direct human interaction with a subject and no human subject recruitment. We will use publicly available and de-identified omic data from The Cancer Genome Atlas (TCGA), Encyclopedia of DNA Elements (ENCODE), as well as the following published studies:

De Simone, et al. (2016). Transcriptional landscape of human tissue lymphocytes unveils uniqueness of tumor-infiltrating T regulatory cells. Immunity, 45(5), pp.1135-1147.

Farlik et al. (2016). DNA methylation dynamics of human hematopoietic stem cell differentiation. Cell stem cell, 19(6), pp.808-822.

Angelova et al. (2015) Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. Genome biology, 16(1)

Newman et al. (2015) Robust enumeration of cell subsets from tissue expression profiles. Nature methods, 12(5), 453–457

Reinius et al. (2012) Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. PloS one, 7(7), p.e41361.

Accomand et al. (2014) Quantitative recon- struction of leukocyte subsets using DNA methylation. Genome Biology, 15(3)

Tirosh et al. (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science, 352(6282), pp.189-196.

Limbach et al. (2016) Epigenetic profiling in CD4+ and CD8+ T cells from Graves' disease patients reveals changes in genes associated with T cell receptor signaling. *Journal of autoimmunity*, *67*, pp.46-56.

Sallustio et al. (2016) Aberrantly methylated DNA regions lead to low activation of CD4+ T-cells in IgA nephropathy. Clinical Science, 130(9), pp.733-746.

Reynolds et al. (2014) Age-related variations in the methylome associated with gene expression in human monocytes and T cells. *Nature communications*, *5*, p.5366.

Vento-Tormo et al (2016). IL-4 orchestrates STAT6-mediated DNA demethylation leading to dendritic cell differentiation. *Genome biology*, *17*(1), p.4.

Coit et al. (2015) Epigenome profiling reveals significant DNA demethylation of interferon signature genes in lupus neutrophils. *Journal of autoimmunity*, *58*, pp.59-66.

Schlums et al. (2015) Cytomegalovirus infection drives adaptive epigenetic diversification of NK cells with altered signaling and effector function. *Immunity*, *42*(3), pp.443-456.

Zhang et al. (2013) Genome-wide DNA methylation analysis identifies hypomethylated genes regulated by FOXP3 in human regulatory T cells. *Blood*, *122*(16), pp.2823-2836.

Kulis et al. (2015) Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nature genetics*, *47*(7), pp.746-756.

b. Sources of Materials:

De-identified omic data will be used for our study. Limited demographic or clinical variables will also be used if available, such as gender, age, and tumor stage; they are publicly available and de-identified as well.

c. Potential Risks:

The major foreseeable risk is loss of privacy through identification of an individual subject from omic data. These risks are highly unlikely. First, all the data used for this project are de-identified. Secondly, identifying a person using these omic data is unlikely because there is no database that links gene expression or DNA methylation to patient identification. Finally, if funded, the data will be stored in a designated and secured machine with no access except for the personnel directly involved in this project.

Adequacy of Protection Against Risks

a. Recruitment and Informed Consent:

Not applicable because no human subjects will be recruited in this project.

b. Protections Against Risk:

We will not pursue the human subject identification information at any stage of this study and all the gene expression and DNA methylation data will be stored in a designated and secured machine with no access except the personnel directly involved in this project. Aside from the remote risk for breach of confidentiality using gene expression or DNA methylation data, this study poses no personal risk to any individuals.

Potential Benefit of the Research to Subjects and Others

There will be no benefit to the individual subjects whose samples are studied. These subjects will not be contacted or made aware of findings. The potential benefit to society pertains to knowledge gained and potential benefit to future research or clinical practice.

Importance of the Knowledge to be gained

The statistical methods developed in this project may enable a better understanding of the immune microenvironment in tumor. Considering that it is very unlikely to lose privacy of the human subjects in our study, the risks are reasonable in relation to the importance of the knowledge to be gained.

INCLUSION OF WOMEN AND MINORITIES

Not applicable because this project only carries out analysis on publicly available and de-identified data, and we do not recruit any individuals (Exemption 4 of human subject study).

INCLUSION OF CHILDREN

Not applicable because this project only carries out analysis on publicly available and de-identified data, and we do not recruit any individuals (Exemption 4 of human subject study.