UNIVERSITY OF CALIFORNIA

Los Angeles

Integrating Multimodal Data for Personalized Models of Cancer

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

by

Amy Lauren Cummings

2022

ABSTRACT OF THE DISSERTATION

Integrating Multimodal Data for Personalized Models of Cancer

by

Amy Lauren Cummings

Doctor of Philosophy in Bioengineering

University of California, Los Angeles, 2022

Professor Alex Ahn-Tuan Bui, Chair

Advances in next-generation sequencing coupled with comparative analyses have had tremendous implication for oncology. Differentially expressed genomic features have revealed molecular pathways, oncogenic drivers, and resistance patterns targeted by drug development leading to significant decreases in cancer mortality. The introduction of high throughput multi-omics, including transcriptomics, epigenomics, and proteomics, promised to scale translational innovation. Increasing feature complexity, however, cannot be comparatively resolved, and many efforts in this space have failed due to inconclusive or conflicting results. Computational systems biology and functional genomics have proposed dynamic integration of multiple molecular pathway models for data harmonization, yet the requirement for complete information has biased discovery. Similarly, the incorporation of probabilistic approaches that constrain features may obscure incremental biologic effects. This problem is exemplified by the several dozen cancer genomic biomarkers and models from peer-reviewed high impact publications that do not meet statistical significance when applied beyond their training and validation datasets.

This dissertation seeks to employ a scientifically rigorous process of evaluation using iterative modeling to benchmark multiomic comparisons. First, this research develops a framework for characterizing multimodal data based on structural and functional information. Second, this research benchmarks similarity metrics using varied data structures, offering techniques to reveal key biologic differences using probabilistic modeling (hierarchical clustering). This framework is then applied to neoepitope prediction incorporating human leukocyte antigen-B supertypes and used to resolve previously inconclusive and conflicting results, including the difference in survival based on B44 supertype in patients with non-small cell lung cancer (NSCLC) and melanoma treated with immune checkpoint blockade (ICB).

Ultimately, this dissertation advances our understanding of the interaction between feature selection and power in multiomic analyses and offers recommendations to enhance the reliability of these investigations.

The dissertation of Amy Lauren Cummings is approved.

Denise R. Aberle

Steven M. Dubinett

Edward B. Garon

William Hsu

Arash Naeim

Alex Ahn-Tuan Bui, Committee Chair

University of California, Los Angeles

2022

To my son, Declan Patrick Conwell Cummings, for what I cherish about the future

and to my grandmother, Gloria Jean Delaney, for what I cherish about the past

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

ACKNOWLEDGMENTS

“You have the ability to not just make a difference in your patient’s life, but also everyone else in that patient’s life. It is the most amazing gift.” Alexandra Levine’s talk on humanism in medicine was formative in my medical training, and one I looked to often on long days. Over the past several years, however, I have realized that the gift is more than the difference I make as the life most touched has been my own. Aptly, this dissertation is built upon the generosity of many. Patients who donated their stories, tissue, time, and bravery in their fight against cancer, my family who did it all while I was figuring out “just one more thing”, and all of the faculty, colleagues, and administrators who created time in overpacked schedules to support me along the way.

Thanks are not nearly enough, and among the many skills and insights I’ve gained, here is a short list of what I’ve found most meaningful: thank you to Dr. Aberle for showing me how to be a trail blazer; Dr. Dubinett for how to think comprehensively; Dr. Garon for how to create a data-driven narrative; Dr. Hsu for how to embed consistency; Dr. Naiem for how to strategize; and Dr. Bui for how to innovate. I would not be here without Drs. Demer and Wong and the STAR program, who saw something in me I did not see in myself; Drs. Slamon and Larson, who figured out how to make it work logistically; the faculty and students at BE MII and UCLA QCB, Clifford Kravitz and the DGIT team, and Di Wu and the Falcon computing team who all helped me learn how to code and build a pipeline in the cloud; and Drs. Garon, Goldman, Lisberg, Fares, Velez, Li, and the entire Lung Correlative Team (including Jackie Gukasyan, Debory Li, Paige Brodrick, Henry Lu, and many others), who have been my work family, sounding board, and main collaborators.

To my family, including my wife Siobhan Conwell, son Declan, parents Glen and Debra Cummings, and the entire Nasser Family (Jordan, Allison, Amelia, and Crosby), thank you for always believing in me and making time whenever I needed it. I am who I am because of you, and I am humbled and grateful to have had your support for all these years.

I promise I won’t get any more degrees.

Chapter 5 is a version of **Cummings AL**, Gukasyan J, Lu HY, Grogan T, Sunga G, Fares CM, Hornstein N, Zaretsky J, Carroll J, Bachrach B, Akingbemi WO, Li D, Noor Z, Lisberg A, Goldman JW, Elashoff D, Bui AAT, Ribas A, Dubinett SM, Rossetti M, Garon EB. Mutational landscape influences immunotherapy outcomes among patients with non-small cell lung cancer with human leukocyte antigen supertype B44. *Nat Cancer*. 2020 Dec;1(12):1167-1175. DOI: 10.1038/s43018-020-00140-1. JG, HY, GS, CMF, BB, and DL collected data and performed analyses; NH, JZ, ZN, AL, AATB, AR, and SMD contributed data and analysis tools; JC, BB, WOA, JWG contributed data; TG, DE, and MR conceived and designed part of the analyses, ALC and EBG conceived and designed the analysis, collected the data, contributed data and analysis tools, performed the analysis, and wrote the paper. The results are in part based on data generated by the TCGA Research and CPTAC Networks: https://www.cancer.gov/tcga; https://proteomics.cancer.gov. Experiments in this dissertation were funded by EBG's NIH-NCI R01 CA28403, the Cancer Center Support grant P30 CA016042, and NIH-NCAT UL1TR001881.

VITA

2000-2004	B.A., Communications Studies, English Minor, Honors Program; UCLA, Los Angeles, California.
2007-2008	Postbaccalaureate Premedical Program; University of Southern California, Los Angeles, California.
2009-2013	M.D.; Keck School of Medicine at the University of Southern California, Los Angeles, California.
2013-2016	Internal Medicine Internship and Residency; Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California.
2016-2019	Hematology and Oncology Fellowship; Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California.
2018-2019	Chief Fellow; Division of Hematology and Oncology, Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California.
2019-2020	Chief Fellow; Specialized Training in Advanced Research (STAR), Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California.
2020-2022	Health Sciences Clinical Instructor of Medicine; Division of Hematology and Oncology, Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California.
2021-	Co-Director; Preliminary Resident-Oriented (pro)-STAR Program, Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California.
2022-	Director; Justice, Equity, Diversity, and Inclusion (JEDI), Jonsson Comprehensive Cancer Center (JCCC), Los Angeles, California.
2022-	Assistant Clinical Professor of Medicine; Division of Hematology and Oncology, Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California.

PUBLICATIONS

Cummings AL, Garon EB. The ascent of immune checkpoint inhibitors: is the understudy ready for a leading role? *Cancer Biol Med*. 2017 Nov;14(4):341-347. PMID: 29372100.

Lisberg A, **Cummings AL**, Goldman JW, Bornazyan K, Reese N, Wang T, Coluzzi P, Ledesma B, Mendenhall M, Jones B, Madrigal J, Carroll J, Gukasyan J, Williams T, Sauer L, Wells C, Hardy A, Linares P, Adame C, Garon EB. A phase II study of pembrolizumab in EGFR-mutant, PD-L1+, tyrosine kinase inhibitor naïve patients with advanced NSCLC. *J Thorac Oncol*. 2018 Aug13(8):1138-1145. PMID: 29874546.

Hu-Lieskovan S, Lisberg A, Zaretsky JM, Grogan TR, Rizvi H, Wells DK, Carroll J, **Cummings AL**, Madrigal J, Jones B, Gukasyan J, Shintaku P, Slamon D, Dubinett S, Goldman JW, Elashoff D, Hellmann MD, Ribas A, Garon EB. PD-L1 expression associates with long-term clinical benefit to pembrolizumab in advanced non-small cell llung cancer. *Clin Cancer Res*. 2019 Aug15;25(16):5061-5068. PMID: 31113840.

Noor Z, **Cummings AL**, Johnson MA, Spiegel M, Goldman JW. Targeted therapy for non-small cell lung cancer. *Semin Respir Crit Care Med*. 2020 Jun;41(3):409-434. PMID: 32450595.

Cummings AL, Gukasyan J, Lu HY, Grogan T, Sunga G, Fares CM, Hornstein N, Zaretsky J, Carroll J, Bachrach B, Akingbemi WO, Li D, Noor Z, Lisberg A, Goldman JW, Elashoff D, Bui AAT, Ribas A, Dubinett SM, Rossetti M, Garon EB. Mutational landscape influences immunotherapy outcomes among patients with non-small cell lung cancer with human leukocyte antigen supertype B44. *Nat Cancer*. 2020 Dec;1(12):1167-1175. PMID: 35121931.

Garon EB, Spira AI, Johnson M, Bazhenova L, Leach J, **Cummings AL**, Candia A, Coffman RL, Janatpour MJ, Janssen R, Gamelin E, Chow LQM. A phase 1b open-label, multicenter study of inhaled DV281, a TLR9 agonist, in combination with nivolumab in patients with advanced or metastatic non-small cell lung cancer. *Clin Cancer Res*. 2021 Aug 15;27(16):4566-4573. PMID: 34108179.

Zhou N, Velez MA, Bachrach B, Gukasyan J, Fares CM, **Cummings AL**, Lind-Lebuffe JP, Akingbemi WO, Li DY, Brodrick PM, Tessuf NM, Rettinger S, Grogan T, Rochigneux P, Goldman JW, Garon EB, Lisberg A. Immune checkpoint inhibitor induced thyroid dysfunction is a frequent event post-treatment in NSCLC. *Lung Cancer*. 2021 Nov;161:34-41. PMID: 34507111.

Gao Y, Stein MM, Kase M, **Cummings AL**, Bharanikumar R, Lau D, Garon EB, Patel S. Comparison of the tumor immune microenvironment and checkpoint blockade biomarkers between stage III and IV non-small cell lung cancer. *Cancer Immunol Immunother*. 2022 [in press].

CHAPTER 1

Introduction

1.1 Motivation

The earliest comparative studies of cancer genomics were unintentional. Using gel electrophoreses, the pan-cancer tumor suppressor p53, the most commonly mutated gene in cancer, was found when untreated control cancer cells exhibited the same size protein induced by an oncogenic virus.[1] Ever since, differential gene expression has been the most used analytic methodology in cancer genomics. Early successes established signaling pathways/co-expression, often using unsupervised clustering,[2-8] and identified multiple oncogenic drivers, or key mutations responsible for the initiation and maintenance of cancer, including alterations to human epidermal growth factor receptor 2 (HER2) and other tyrosine kinase growth factor receptors.[9-11] Targeting this low-hanging fruit resulted in incredibly meaningful improvements in cancer survival,[12-14] but most cancer patients do not exhibit targetable mutations,[15] and of those who do, only a proportion will respond and only for a limited time.

Advances in next-generation sequencing, multiplexing, high-throughput multiomics, and the public availability and standardization of bioinformatics workflows[16-24] promised to inspire and widen translational efforts, but treatment strategies leveraging these data have yielded disappointing outcomes. The National Cancer Institute Molecular Analysis for Therapy Choice (NCI-MATCH) trial, one of the largest endeavors using varied molecular biomarkers, assigned 875 participants to 37 genotype-matched treatment cohorts and showed an overall response rate of 7.5%, no different than standard second-line chemotherapy.[25, 26] Moreover, the recent investigation of several dozen multiomic biomarkers identified in peer-reviewed high impact publications showed that most failed to reach statistical significance and exhibited inconsistent findings when applied across cancer histology.[27]

The complexity of genomic data, inability to account for multimodal multivariable interactions, and large number of redundant variables in relatively small sample sizes have been identified as key determinants in multiomic analytic shortcomings.[28-33] Computational systems biology and functional genomics tackled these issues with dynamic integration of multiple molecular pathway models, leveraging unsupervised machine learning techniques to make key progress in resolving inconsistent findings.[34-37] The Pathway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM), for example, was able to successfully delineate subtypes of glioblastoma based on transcriptomic information, but only reaffirmed known clinically relevant subgroups, suggesting a tendency of these approaches to bias discovery and recapitulate what is already known.[36] Probabilistic approaches using Bayesian and Markov chain Monte Carlo (MCMC) sampling similarly have integrated multiomic data successfully and have been proven capable of generating novel insight.[38-42] Onco-Multi-OMICS, for instance, showed cancer-specific histone marks associate with transcriptional changes in driver genes in head and neck squamous cell carcinoma, uniquely marrying the concepts of epigenetics and driver mutations in cancer evolution.[41] This model has not been applied to other histology types, however, and a review of a dozen probabilistic models integrating multiomic analyses did not identify any approach that was successful in more than one clinical scenario,[43] suggesting these approaches may lack flexibility and cannot be broadly applied.

Through this lens, challenges in multiomic analyses are explored throughout this dissertation to evaluate systematic strategies to identify meaningful change and relative value in cancer multiomic models, particularly with respect to managing noise in small sample sizes and enabling model flexibility.

1.2 Contributions

This dissertation advances our understanding of practical applications of multiomic-based model development and establishes techniques to improve reproducibility and iteration of multiomic analyses across cancer histology by fulfilling the following two aims:

- Aim 1: Developing a framework for characterizing multimodal data. This dissertation supports an expanded role for germline and normal tissue in cancer multiomic analyses. While the inclusion of a germline sample is routine in DNA-based variant analyses and done for the purpose of decreasing false positive variant calls,[44] normal tissue and cancer multiomic analyses are typically performed separately and compared as a final step (when compared at all). This dissertation supports using germline and normal tissue features to identify expected or probabilistic findings at each step of analysis, enabling identification not only of features that are statistically different, but also those that are expectedly different and unexpectedly similar. Using this framework, these experiments reveal technical artifacts and associations that otherwise could lead to spurious results as well as demonstrate the relationship of probabilistic distributions to numbers of observed events, thereby enabling discernment of nuanced subgroups.
- Aim 2: Benchmarking similarity metrics using genomic data. Experiments performed as part of this aim suggest the introduction of error and noise to dissimilar datasets leads to shifts in data structure that falsely enhance measured similarity. To protect against this source of error, this dissertation demonstrates repeat measurements with similarity metrics with distinct strengths (i.e., distance metrics more and less sensitive to scale and non-distance metrics) increases reliability and helps maintain inherent data structure. This idea is demonstrated in two methodologies commonly used in multiomic analyses: (1) comparisons at an individual level for the purpose of showing meaningful change, and (2) comparisons across a cohort for subgroup discovery. Using standard single variable and regression models, subsequent

experiments demonstrate how multiomic analyses with small sized cohorts tend to lack power, particularly with queries requiring higher feature complexity. Using the techniques and known and uncovered relationships from Aims 1 and 2, probabilistic, iterated model experiments demonstrate utility in resolving increasingly complex and varied queries, such as the role of mutational signature in cancer-immune interactions and transcriptomic and epigenetic features associated with protein expression.

Motivating scenario. Immune checkpoint blockade (ICB) is arguably the most important therapeutic advance in cancer in the past decade, yet prediction of clinical benefit remains challenging. Most advanced cancer patients receive ICB at some point in their disease although only a third will benefit clinically. This approach is at great cost to healthcare systems at over $30B worldwide in 2021 (projected $1.4T by 2030)[45] – and not without toxicity, including fatal toxicities occurring at a rate of 0.6%.[46, 47] The significance of this clinical problem and the underlying conflicting evidence of relevant biomarkers provides a foundation for exploring and developing methods in this dissertation, particularly with respect to cancer-immune interactions.

1.3 Organization

The remaining chapters of this dissertation are organized as follows:

- Chapter 2 describes the technical background of major aspects of this dissertation, including techniques used in dataset creation, methods for structuring and interrogating relationships among multiomic data, and approaches to computational analyses.
- Chapter 3 presents the framework for multimodal data characterization, showing that incorporating germline and normal tissue comparisons helps limit redundant variables and reveals technical artifacts and associations, improving model robustness against spurious results.

- Chapter 4 discusses the impact of noise in multiomic analyses, benchmarking similarity metrics commonly used for characterization and demonstrating the value of repeat measurements with metrics leveraging distinct techniques.
- Chapter 5 applies these principles in multiomic analyses focusing on cancer-immune interactions, demonstrating limitations of feature selection and power with standard approaches and the value of personalized, iterative modeling in resolving discrepancies and answering complex queries in small sample sizes.

Finally, Chapter 6 summarizes the contributions and findings from this dissertation, and then provides possible directions to build on these developments to serve the goal of enhancing reliability in multiomic investigations.

CHAPTER 2

Background

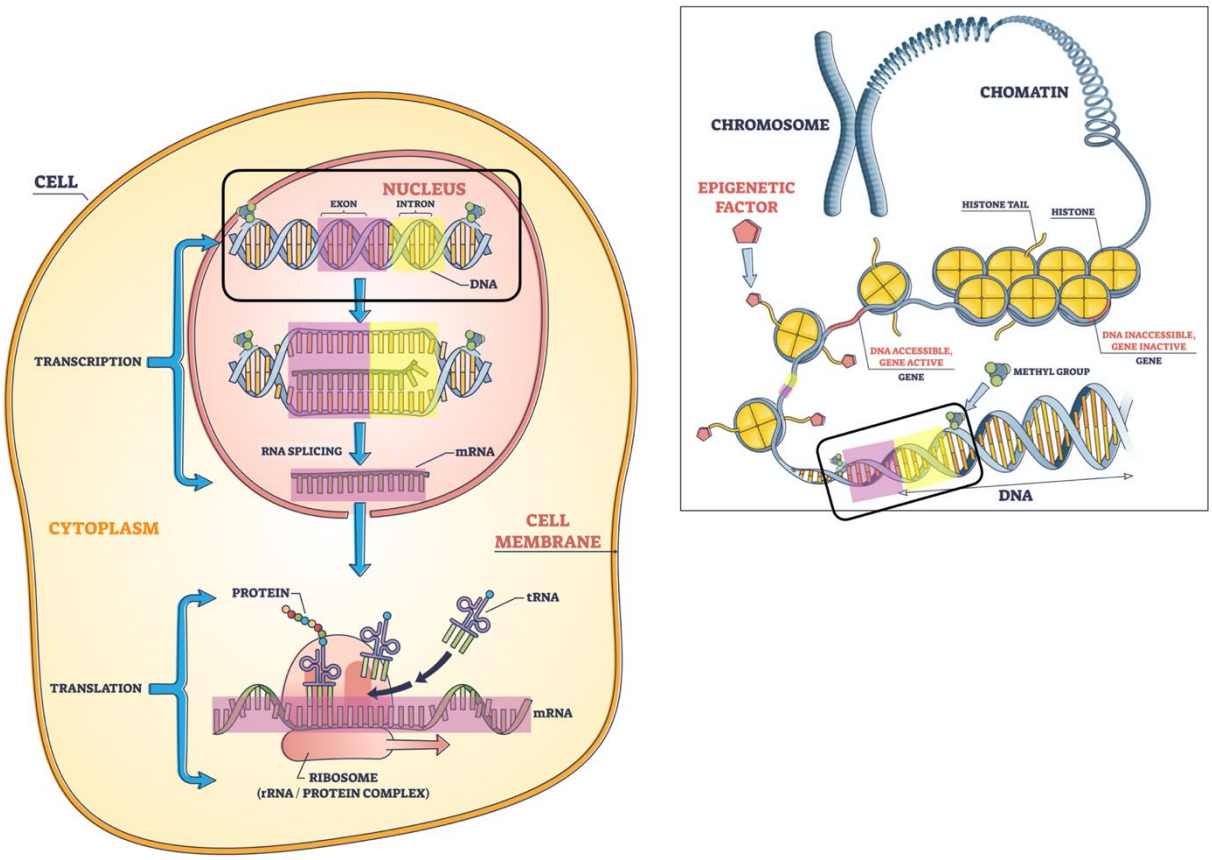
This chapter provides an overview of current methods pertinent to the methodological developments in this dissertation. We first discuss the challenges of multiomic analyses, focusing on cancer-specific models, and current approaches to solving these problems. Two specific areas are then described, highlighting state-of-the-art methods and approaches: (1) strategies for managing conflicting information; and (2) techniques for assessing data similarity. Details of comparable biologically driven multiomic models are presented in subsequent chapters.

2.1 Challenges in Multiomic Analyses

Genomic data complexity. Cancer is, by definition, genomic instability that results in unrestrained growth and/or inappropriate persistence of abnormal cells.[48] Based on the central dogma of molecular biology, intermediary/regulatory elements (i.e., epigenetics) transform an individual's blueprint of three billion DNA base pairs into RNA and amino acid sequences that form all protein responsible for life processes,[49] as illustrated in Fig. 2.1. Alterations to this blueprint may hijack any cellular process and/or change the behavior, expression, or regulation of any of these components or their networks, as well as vary over space, time, and impact.[50-63] Multiomic analyses thereby require interpretation of molecular variation at each level of interaction, balancing techniques for robust data structure/integration and dimensionality reduction with appropriately powered cohorts. Key advances have drawn extensively from the fields of statistics, information theory, and machine learning (ML) for these purposes.

Multiomic data structure/integration. Currently, there are two main approaches to integrating omics data: sequential and simultaneous.[64] Sequential approaches allow for the confirmation or refinement of findings based on one data type, with additional analyses of further omics data

Figure 2.1: Central dogma of molecular biology.



Adapted from stock images with Creative Commons (CC) licenses available upon request.
Not drawn to scale.

obtained from the same set of samples.[65] Simultaneous approaches combine omics data into single probabilistic models, often using a two-step optimization procedure.[66] Sequential approaches tend to confer hierarchical relationships among omics layers while simultaneous approaches are less biased in structure. Both model types require computational approaches that apply several methods to carry out a sequence of tasks in a "pipeline".[64] Those methods that use graphs to model the interactions among variables are "network-based" and consider currently known or predicted relationships between biologic variables.[67] Those techniques that are Bayesian use statistical models which, starting from an *a priori* assumption about data probability distribution, compute updated posterior probability distributions[68] [note (un)supervised learning algorithms are considered network-free and non-Bayesian]. These approaches and techniques

enable the creation of varied data structures, each with their benefits and limitations, which are explored in subsequent components of this dissertation.

Dimensionality reduction. Motivating scenarios for multiomic analyses drive the selection of statistical tools embedded in both sequential and simultaneous multiomic model pipelines. These scenarios are typically described as: (1) the characterization of molecular behaviors, mechanisms, and/or relationships, (2) subgroup classification, or (3) prediction of an outcome or phenotype.[69] The approach to dimensionality reduction given these distinct purposes is variable, contributing to limitations in cross-model comparisons and flexible application.

Partial least squares (PLS) and principal component analysis (PCA), in addition to variations of both these techniques, are the most used to identify predictors of a response dataset. PLS leverages linear combinations of predictors (enabling supervision) to define a new set of features chosen so that the highest weight is assigned to variables most strongly correlated to an outcome.[70] PCA, in contrast, is unsupervised, using a set of p variables represented as unit vectors (jointly normally distributed), where the i-th vector is the direction of a line that best fits the data while being orthogonal to the first $i-1$ vector, explaining variance through p iterations, reducing (even highly correlated) variables to an independent set. Other commonly used strategies for dimensionality reduction involve (un)supervised clustering using expectation-maximization algorithms,[71] Lasso penalized regression,[72] gene set analysis,[73] and pairwise correlation analysis,[74] many of which employ false discovery rate (FDR), or Benjamini-Hochberg corrections, to correct for multiple hypothesis testing.[75]

Power

Calculating power, or the probability that a test of significance will detect a deviation from a null hypothesis should such a deviation exist, is limited in multiomic models due both to a lack of comparative performance parameters across omic technologies and the presence of redundant

variables in relatively small sample sizes.[76] Central to these issues are the unique and often conflicting characteristics inherent to each omic technology.

Relationship mapping

At a practical level, multiomic data structure relies on relationships of primary measurement layers that require mapping of variables or leverage of a common unit. Each omic feature and how it is measured, however, has been independently developed, leading to distinct methodologies and ontologies summarized in Table 2.1. As RNA expression data is the most common multiomic data and resolved at the level of the gene,[76] data are classically mapped based on gene identifiers, which leads to inherent dataset shift and feature loss as some multiomic data have more refined identifiers based on transcript or reference sequence.[77] Methylation data, for instance, are measured with probes that assess CpG islands (areas of methylation near promoter regions of genes) annotated with transcript identifiers – each gene has multiple probes and each probe can cover multiple genes. Protein data are annotated with reference sequences that are unique to both gene and transcript identifiers.[78] Notably, microRNA (miRNA) notation is sequential based on discovery date, with "*" notation used to differentiate predominantly expressed miRNA sequences from other sequences that originate from the same predicted precursor (* is used for the less expressed sequence).[79] A reliable mapping of the miRNA interactome, including how miRNA interact with target messenger RNA (mRNA) or partner with long, noncoding RNAs and pseudogenes, has yet to be elucidated, limiting the applicability of miRNA data to multiomic analyses.[80] For this reason, miRNA are not included in this dissertation's experiments (see Section 2.2, systematic exclusion as a method for managing conflicting information).

Technical limitations

Beyond relational data structure limitations, each multiomic assessment also contributes unique technical limitations at both data creation and normalization steps that affect overall data structure and reliability. The vast majority of multiomic data (currently) is produced by bulk analyses, which

Table 2.1: Distinct methodologies and ontologies in multiomic data.

Source	Experiment	Platform	Processing	Output	Reference	Match
CPTAC3	whole exome seq	Illumina NovaSeq	Mutect2, VEP	SNP, SNV, aa substitutions	GRCh38	ENSG
CPTAC3	methylation	Illumina HM450	SeSAMe	beta value	cg probe to gene symbol	gene symbol
CPTAC3	gene expression	Illumina HiSeq	STAR 2 TwoPass	htseq counts	GRCh38	ENSG
CPTAC3	mass spectrometry	Hfx Kelsey MS	ReAdW4Mascot2, MSGF+, ProMS	precursor area ion counts, E-value	gene symbol to gene name to NP	NCBI RefSeq NP
CPTAC3	whole genome seq	Illumina Genome Analyzer IIx	ASCAT2	copy number	GRCh38	ENSG
TCGA	whole exome seq	Illumina HiSeq	Mutect2, VEP	SNP, SNV, aa substitutions	GRCh38	ENSG
TCGA	methylation	Illumina HM27	SeSAMe	beta value	cg probe to gene symbol	gene symbol
TCGA	gene expression	Illumina HiSeq	STAR 2 TwoPass	htseq counts	GRCh38	ENSG

Aa – amino acid, ASCAT2 – Allele-Specific Copy number Analysis of Tumors 2,[24] CPTAC – Clinical Proteomic Tumor Analysis Consortium,[81] ENSG – Ensembl gene identification,[82] MS – mass spectrometry,[83, 84] NCBI – National Center for Biotechnology Information, NP – nomenclature protein,[79] seq – sequencing, SNP – single nucleotide polymorphism, SNV – single nucleotide variant, SeSAMe – SEnsible Step-wise Analysis of Methylation data,[85] STAR – Spliced Transcripts Alignment to a Reference,[86] TCGA – The Cancer Genome Atlas,[87] VEP – Variant Effect Predictor.[82]

leads to sources of error as well as architectural feature loss due to multiplexing techniques and raw material limitations.[33] For example, whole exome sequencing (WES) provides comprehensive coverage of a targeted space (i.e., DNA that leads to protein creation), but coverage of the reference genome varies based on sample library preparation.[88] Samples that have features leading to poor library preparations, such as small biopsy size, will systematically be excluded from multiomic analyses, leading to discovery bias for specimens that are easier to obtain (e.g., melanoma), and/or have larger volumes, a key limitation in the interpretation of The Cancer Genome Atlas (TCGA) cancer specimens.[89]

Additionally, WES Illumina platforms, the most used NGS technology, rely on cyclic reversible termination. Even with an accuracy rate of >99.5%, this underrepresents adenine-thymine and guanine-cytosine-rich regions, exhibiting a tendency towards substitution errors, leading to a false-positive rate of 2.5%.[90] The Broad Institute's MuTect2 algorithm, the most used approach for variant calling, has been developed to address this issue by pairing a matched germline sample to cancer samples to improve somatic, or mutated, single nucleotide variant calls (SNVs), requiring a prespecified number of observations to make a variant call.[44] While helpful for specificity, poorly processed specimens, rare events, and/or unequal coverage of regions can cause dataset shift and contribute to data missingness.

In juxtaposition, mass spectrometry (MS) experiments are strongly biased toward abundant proteins.[91] MS experiments are known for creating multiple spectra of the same peptide ion and multiple peptides of the same protein, with varying quantitative patterns within and between MS runs.[92, 93] These intensities can further be compromised by ion interference due to co-isolation and co-fragmenting of isobaric ions, with technical variation coming from sample preparation and instrument fluctuation, limiting the ability to discern reliable differences in protein quantification.[94] Including a common reference increases precision by facilitating normalization

across batches, although other approaches pool samples to create a reference to dampen individual effects.[95]

To address these technical limitations, MS-GF+, software that analyzes tandem mass spectra, relies on E-value cutoffs that reflect the number of times a peptide matches incorrectly based on a target-decoy approach.[84] This target decoy approach substitutes other amino acids in the last position of the observed peptide, which are then matched to observed data in ProteinDB to estimate FDRs with a dot-product.[84] With prespecified E-values typically set at <0.001 (one erroneous match out of 1000), there is high fidelity in peptide spectral matches with adequate E-values, but only about half of peptides are identified, contributing to data missingness and bias against abnormal proteins.[96] As such, the order in which WES and MS layers are incorporated can dramatically alter data structure due to mismatches in data missingness, with irregularities in the technical processing compounding spurious results and/or leading to experiment failure.[97]

Performance evaluation

While most omic technologies include Figures of Merit (FoM) such as accuracy, reproducibility, sensitivity, and dynamic range, the definition of each is different depending on the technology considered, and each omic platform incorporates different critical FoM.[76] Different omics platforms present distinct noise levels in dynamic ranges, suggesting analytic methods that address noise within an omic layer might not be equally applicable in a multiomic structure. For example, Infinium, the Illumina methylation platform, measures methylation at single base resolution using two bead types per CpG, one that reads methylation (indicated by a green signal) and one unmethylation (indicated by a red signal).[98] Methylation data is thus reported as either beta- (Eq. 2.1.) or M-values (Eq 2.2.), which are defined based on M representing the signal from the methylated allele and U representing the unmethylated signal:

$$\beta = \frac{M}{(M+U)} \tag{2.1}$$

$$\mathrm{M} = log_2 \frac{M}{U} \quad (2.2)$$

β intensity values determine the proportion of methylation at each CpG locus while the transformation of M leads to distributions more appropriate for statistical testing.[99] With these representations, there are issues with signal bleed-through when measurement from one channel affects the measurement in the other channel, as well as dye bias, referring to the difference in signal intensity between the two color channels.[100] SeSAMe, a landmark approach in methylation data normalization, implements background subtraction based on normal-exponential deconvolution and experiment-dependent masking, using signal detection p-values to help account for experimental noise to reduce these artifacts.[85] However, the impact of these transformations in multiomic data structures, particularly with respect to averaged beta- or M-values given their redundance at the gene level, remains unknown.[101]

Many technical advances addressing false positives (specificity) in multiomic discovery have been reviewed, but there are far fewer techniques to address false negatives (sensitivity). Returning to TCGA, which are presently the most influential datasets in multiomic analyses given their use in multiomic benchmarking,[102, 103] Huang and colleagues showed the role of *ERF* mutations was missed in TCGA prostate cancer assessments because there were a limited number of African American men included in the TCGA dataset.[104, 105] There since have been efforts to harmonize quality metrics, including MultiPower and MultiML algorithms that relate sensitivity to reproducibility,[76] yet these have not been broadly applied or validated. Model benchmarking, particularly with respect to bias, is often an implied, assumed, and/or overlooked step in multiomic analyses, which serves as fundamental inspiration for this dissertation.

2.2 Management of Conflicting Information

Conflicting information is a key challenge in multiomic data interpretation. Standard statistical methodology includes three main approaches: (1) systematic exclusion, (2) normalization and/or

stratification, and (3) selection of a ground truth. Conflicting information in multiomic analyses, as evidenced by the review above, however, may come from underpowered experimental design, noisy measurements, and inappropriate integration methods, but also may reflect important biologic themes. Thus, an understanding of the biologic relevance of conflicting information is required to select appropriate techniques. Recent advances in multiomic analyses incorporate statistical techniques that iteratively update data structure and/or consider biologic domain mapping, representing central themes built upon in Chapter 3.

Altering data structure without mapping to biologic domain. Outlier exclusion often enhances statistical interpretation of biologic data, yet outliers can be difficult to detect in multiomic data structures.[106] Strategies assessing variability within and between variables have been shown to help with outlier identification, and the Patient-Specific Data Fusion (PSDF) algorithm represents a key innovation in this regard.[39] PSDF applies metrics (i.e., noise structures) specific to each data type using common latent labels across all data types, then aggregates latent variable labels at the patient level to systematically exclude (or minimize) contradictory samples.[39] By converging on patients with multiple latent labels, these techniques can be extended to supervised clustering and accommodate several data types, although the number of clusters can be difficult to determine, often requiring cross validation methods.[69, 107] Pairwise correlation analysis similarly can be used to adjust for potential confounding observations/variables using statistically defined data relationships to alter data structure leveraging variable weights.[108] Using correlation matrices to reflect association strengths, pairwise analyses can identify more influential observations, which can then be isolated and evaluated using standard techniques.[74, 109]

Altering data structure by mapping to biologic domain. Gene set analysis or pathway analysis offers a strategy that alters data structure based on relationships to biologic domains.[73] These approaches have two starting points: (1) an identified set of observations of genes of interest, or (2) comparatively identified genes with high and low differential multiomic expression.[110] Both

options use publicly available knowledgebases (KBs) to map experimental data to known pathways to create scores used to confer membership in subgroups.[111] These approaches have been particularly helpful to identify gene modules expressed in non-malignant bystander cells, limiting both conflicting information and noise in multiomic datasets.[112] One limitation of these approaches, however, is that without careful application and biologic insight, pathway analyses may exacerbate conflicting information, making data structures uninterpretable.

Mapping to biologic domain without altering data structure. The creation of ground truth datasets used for benchmarking have had tremendous impact on technical advancements within single omic layers[113] and represent a key opportunity in cancer multiomics.[114] The main purpose of multiomic assessments is to demonstrate information flow from genotype to phenotype.[115] Genetic variations, or missense mutations, resulting in a change of amino acid sequence can have a dramatic effect on stability, hydrogen bond networks, conformational dynamics activity, and many other physiologically important properties of proteins.[116] Amino acid substitutions thus provide structural and functional information that connect genomic alterations to protein-level consequences. Recently, advances in Tandem Mass Tagged (TMT)-multiplexed MS experiments have significantly improved proteomic quality, depth, availability, and reproducibility.[23, 117, 118] While not without limitations, the use of protein-expression as ground truth to validate conflicting intermediary multiomic relationships represents a novel approach to benchmarking multiomic assessments, which is the focus of Chapters 3 and 4.

2.3 Techniques for Assessing Data Similarity

To date, multiomic analyses have predominantly used hierarchical clustering algorithms leveraging distance similarity metrics to identify dataset differences.[2, 103, 119-121] These metrics are additionally used in ML, with some preference for scale sensitive metrics, such as Euclidean distance and Manhattan or city block distance, given more robust results.[122] Building on this history, there are two innovations that contextualize recent advances in data similarity

assessments: (1) repeatedly concordant similarity metrics that demonstrate model robustness,[123] and (2) iterated patient connectivity models that converge based on similarity.[124]

Repeatedly concordant similarity metrics. Clustering and cosine similarity have been used in genomic analyses for over 20 years, initially with RNA microarray results.[2] While potentially appropriate for single omic analyses, these metrics rely on linear relationships and inherently cannot account for the background distribution of genomic data, and to date, have not been thoroughly interrogated in multiomic modeling.[103] More recently, the introduction of a "mass-distance" measure has offered methodology to adjust to background distribution by estimating the probability to observe by chance a vector inside the volume delimited by the expression profiles (i.e., smaller volumes suggest more similar profiles).[125] In an assessment of gene relationships of yeast, mass-distance outperformed other linear similarity and statistical metrics, including Pearson, Spearman, and Euclidean correlations.[125] Given recent advances in imaging assessment and signal processing, nonlinear similarity metrics, such as mutual information,[126, 127] additionally may offer value. While these have not been classically employed in genomic analysis, the work of Yu and colleagues in statistical stability suggests repeatedly concordant similarity metrics may be used to suggest robustness, offering additional rationale for benchmarking linear and non-linear approaches. Similarity metrics included in experiments in this dissertation are depicted in Table 2.2 and explored in Chapter 4.

Patient connectivity models. Avoiding overfitting, especially given the multiple sources of data heterogeneity previously discussed, is a key hurdle for multiomic analyses.[128, 129] One strategy, as opposed to developing a pipeline to identify differences, is to develop one to detect similarity. A novel approach in this vein with increasingly favorable reviews is that of patient connectivity networks (PCN) based on message-passing theory.[64, 114, 124] Message-passing theory iteratively updates a network to make it more like the others with each iteration – weak connections (i.e., noise) disappear with iterations, whereas strong connections are propagated through

Table 2.2: Similarity metrics.

Feature	Linear	Definition	Key equation	Program	Interpretation	Advantage
City block distance	yes	distance between points a and b with dimensions k	$\sum_{j=1}^{k} \lvert a_j - b_j \rvert$	SciPy	Range 0-∞, 0 identical	Effect of large difference dampened (helpful for count data)
Cohen's kappa	no	measure of inter-annotator agreement	$\kappa = \frac{(p_o - p_e)}{(1 - p_e)}$	sklearn	Range -1-1, >0.6 excellent, >0.2 adequate	Categorical comparisons
Cosine distance	yes	angle difference in vectors A, B	$1 - \cos\theta$, $\cos\theta = \frac{A \cdot B}{\lVert A \rVert \lVert B \rVert} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$	SciPy	Range 0-1, 0 identical	Disregards effect of size (helpful for count data)
Euclidean distance	yes	Length between points p and q	$\sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$	SciPy	Range 0-∞, 0 identical	Includes effect of size (helpful for proportion data)
Jaccard distance	yes	dissimilarity of sample sets, difference over size of union	$d_j(A,B) = 1 - J(A,B) = \frac{\lvert A \cup B \rvert - \lvert A \cap B \rvert}{A \cup B}$	SciPy	Range 0-1, 0 identical	Incorporates multidimensional scaling
Kendell's Tau	no	measure of n ordinal associations of x, y	$\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j)$	SciPy	Range -1-1, >0.6 excellent, >0.2 adequate	Non-parametric rank comparisons
Mutual information	no	Difference of joint distribution of pair i, j	$MI(i,j) = \sum_{a.b} P(i_i, j) \cdot \log\left(\frac{P(i_i, j_i)}{P(i_i) \cdot P(j_i)}\right)$	sklearn	Range 0-1, 1 identical	High-dimensional generalization scheme without linear dependence
Pearson correlation	yes	correlation of variable X, Y in population ρ	$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$	SciPy	Range 0-1, 1 identical	Evaluates linear relationship
Spearman correlation	no	correlation of variable X, Y in population ρ based on rank	$r_s = \rho_{R(X),R(Y)} = \frac{cov(R(X),R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}}$	SciPy	Range 0-1, 1 identical	Evaluates monotonic relationship

convergence.[130] PCNs are graphs where patients are represented as nodes and connectivity as edges, leveraging similarity matrices generated by merging connectivity from all data types.[124] PINSPlus, or perturbation clustering for data integration and disease subtyping, leverages this approach by identifying how often patients are grouped together in three scenarios: (1) when the data are perturbed, (2) when using different types of omics data, and (3) when a different clustering technique is used.[131] These themes, including exploration of data perturbation, use of similarity matrices, employment of repeat measurements, and iterative models of increasing similarity to convergence, form the methodology backbone of Chapters 3 and 4.

CHAPTER 3

Developing a Framework for Characterizing Multimodal Data

Structural approaches to data integration are foundational to multiomic experiments. In this chapter, we explore the use of amino acid substitutions as an organizational schema for multiomic integration. The impact of varied matrix structures and representations (e.g., count versus proportion data) are examined with respect to subgroup identification and biologically-driven cohorts. Iterative hierarchical clustering comparing normal (germline) and tumor (somatic) similarity across cohorts suggests the value of including normal multiomic comparisons, enabling high variance features to be mapped to biologic correlates while also identifying technical artifacts and outliers.

3.1 Introduction

A multiomic data framework architecture requires identification of purpose.[69] Historically, motivating scenarios have included characterization, subgroup classification, and prediction – partitions that can contribute to model rigidity, limiting a model's ability to be applied to more than one scenario. Using a concept-driven as opposed to task-driven architecture, however, may lead to increased model flexibility.[132] There are two observations that contribute to conceptualization of this dissertation's experimental architecture: (1) clinical interpretation of multiomic reports is limited by an inability to know whether findings are functionally relevant (i.e., likely to influence cellular function at the protein level), and (2) multiplexed TMT-MS experiments can enable protein confirmation of multiomic observations at scale.

In this light, an integration schema that follows multiomic data flow from genomic alteration to protein expression can help to answer biological domain questions about the likelihood of detectable protein expression, making characterization, subgrouping, and prediction tasks possible. To derive this data structure, a hybridized sequential/simultaneous approach is used. A DNA-based layer (whole exome sequencing, WES) provides the foundation of the data structure

to which transcriptomic, epigenetic, and proteomic data are mapped sequentially and then analyzed as a whole with the proteomic layer considered ground truth. As DNA-based assessments are the most complete, this structure avoids excessive data loss related to incorporation of a proteomic layer. Additionally, it enables addition of normal/germline multiomic data, which can be compared separately or added to the data structure. The key contributions of normal multiomic inclusion are threefold: (1) it provides a biologic control to multiomic experiments as variations in normal tissue are less likely to be pathogenic compared to variations in cancer tissue (i.e., what is expectedly different), (2) it provides a technical control as similar artifacts may be seen in both normal and tumor (i.e., what is unexpectedly similar), and (3) it provides a way to compare data structure relationships (i.e., what is unexpectedly different). Further, it requires no prior knowledge as the data are observed.

3.2 Related Work

The prior chapter reviewed multiomic data structure and architecture, particularly with reference to management of conflicting information and techniques for assessing data similarity. Other work pertinent to the experiments in this chapter involve visualization as a discovery technique, dimensionality reduction, and representation of biological domain knowledge, particularly with respect to amino acid substitutions.

Visualization for discovery. At its core, this work seeks to provide insight into the feasibility and potential development of artificial intelligence-based clinical decision support in oncology, which relies on developing techniques that can improve classification.[133] Data visualization, through this lens, is an indispensable part of exploratory data analysis.[134] Requiring dimension reduction and exploiting humans' pattern recognition abilities, visualizations can be an effective way of interpreting complex information, although they require appropriate technique to prevent misinterpretation.[135] Best practices include applying appropriate geometry, including bar graphs or boxplots for histograms/distributions evaluated by count and by proportion, and a two-

dimensional topology of rows and columns with either a rectangular or hexagonal lattice to evaluate more complex features such as heatmaps or clustering.[136]

More advanced theories in visualization for discovery involve depicting error,[137] small multiples,[138] and leveraging visualization of probabilistic models (e.g., hierarchical clustering) as experimental methodology.[139] Standard deviation, which is based on the spread of data; standard error, which depicts uncertainty in the mean based on sample size (and mostly fallen out of favor for this reason);[140] confidence intervals, which display the reliability of a measurement; and credible intervals, exclusively associated with Bayesian methods, are all common metrics of uncertainty.[135] Krzywinksi and Altman, after an extensive review, recommend matching error depictions to the data to highlight what may be perceived limitations but concede that any error representation is better than none.[137]

The concept of small multiples, or paneling or faceting, describes the technique of iterating a figure with common axes, axes scales, and geometry to highlight differences.[138] This approach changes one variable per panel to highlight differences among the panels. Advanced visualizations, such as hierarchical cluster visualizations, can also be used to compare forms of discovery to provide additional insight.[139] We blend both techniques to iteratively examine small multiples of hierarchical clustering, leading to the identification of both unexpected similarities and differences that fuel further investigation.

Dimensionality reduction. In 2013, the National Cancer Institute (NCI) set standardized criteria for omics-based model development focusing on rigorous data, statistics, and model analyses to aid in clinical translation.[129] While there is no consensus on measures to include in omics evaluations, copy number, DNA methylation, gene expression, miRNA expression, metabolites, clinical information, and epidemiologic data have been recommended as features that should be subjected to outcome analysis and model building.[129] While multiomic investigations are conventionally limited to one histology and type of approach for feasibility, the value of this

proposed methodology is that it seeks to provide a generalized framework for multimodal cancer genomic analyses across histology by using multiple probabilistic approaches. As such, alternative dimensionality reduction is required for feasibility. A first strategy is to limit genomic observations to protein-coding regions of the genome (WES). This limitation decreases evaluated regions to 1% of the genome (~38 million base pairs) while including 15 out of 50 (30%) of most frequently mutated sites in the genome, and theoretically all mutations that change protein structure, sacrificing only the interaction of m/miRNA with three-dimensional genomic structures and *TERT* promoter events, although detection of gene rearrangements, chromothripsis, and kataegis may be limited.[141] Further dimensionality reduction may be taken by limiting genomic comparisons to amino acid substitutions/non-synonymous mutations, which bins genomic observations into one of the 170 possible amino acid substitutions (150 if mutations from a stop codon to an amino acid are excluded).

Amino acid substitution matrices. Central to the design of this chapter's experiments is biological domain knowledge, which includes three main tenets: (1) amino acids provide structural and functional information that relate genomic alterations to protein-level consequences,[142, 143] (2) physiochemical relationships between amino acids suggest substitutions that may be more or less favorable to cellular function,[144-147] and (3) SNPs are less likely to lead to amino acid substitutions detrimental to cellular function compared to SNVs.[148] Amino acid substitutions may be extracted from WES, copy number, gene expression, methylation, and proteomic investigations based on bioinformatic annotations. An investigation of amino acid substitution profiles in this context serves both to rigorously evaluate the role of amino acid substitutions in cancer genomics and to serve as a first model system that can be expanded to model genes, networks, or other genomic structures.

These features enable an evaluation of the probabilistic relationships of amino acid substitutions by comparing relationships between and among classifications. To further inform

these investigations, our work leverages CoCoPuTs, codon and codon-pair usage tables, which is a publicly available resource published by Alexaki and colleagues.[149] This knowledgebase continually extracts updated transcriptomic codon frequencies from all organisms available in GenBank (a public genomic repository) to guard against data drift. In addition to enabling subsetting based on species data (e.g., homo sapiens), CoCoPuTs modules also publish further subset species-level codon tables based on tissue of origin and cancer histology, with broad applicability to this dissertation's experiments.

Additional biological domain knowledge that informs exploration is that of known amino acid substitution probabilities. Prior to the 1990s, protein alignments were used to provide insight into gene and protein function.[150] These alignments used global alignments of pairs of proteins related by common ancestry, local alignments involving related segments of proteins, multiple alignments of protein families, and alignments made based on database searches for homology, which were evaluated using a scoring system to estimate similarity.[151-153] In seminal work by the Henikoffs in 1992, the BLOcks SUbstitution Matrix (BLOSUM) used an automated system (PROTOMAT) to pull directly measured data from the protein BLOCKS database.[154] This approach was able to offer direct observations about evolutionarily retained amino acid sequences as opposed to inferring relationships, as prior work had done. For this purpose, a logarithm of odds (lod) matrix (Eq. 3) was created based on a frequency table f_{ij}, which observed a probability of occurrence for each i, j pair represented by the following:

$$q_{ij} = f_{ij} / \sum_{i-1}^{20} \sum_{j-1}^{i} f_{ij} \tag{3.1}$$

The resultant substitution matrix scored alignments between evolutionarily divergent protein sequences with positive scores indicating the alignment was found more often than by chance, and negative scores indicating the alignment was found less often by chance.[154] The BLOSUM62 matrix, used in subsequent experiments, groups amino acids according to the chemistry of side chains and is clustered at 62%, divided by the probability that the same two amino acids might

align by chance.[155] BLOSUM62 has become the *de facto* standard for many protein alignment programs with various research applications.[156]

More recently, Kim and colleagues derived an amino acid similarity matrix for peptide-major histocompatibility complex (MHC) binding to compensate for bias in existing peptide datasets.[157] Like BLOSUM62, the matrix, named PMBEC, was derived directly from binding affinity data, capturing physicochemical properties of amino acid residues but differing markedly based on cases where residue substitution involves electrostatic charge reversal for the purpose of providing better insight into binding.[158] As human leukocyte antigen (HLA) interactions, particularly supertype B44, are a feature of interest given inconsistent ICB outcomes,[159-162] PMBEC may provide additional structural insight. A list of amino acid substitutions arranged from most negative (unfavorable) to most positive (favorable) scores based on BLOSUM62 and PMBEC arrangements can be found in Table 3.1.

Mutational signature. A key consideration in interpretation of this chapter's experiments is mutational signature. Somatic mutations characterize all cancers, and mutational signature, which varies across histology, contextualizes these mutational processes and influences amino acid substitution patterns.[119] DNA bases include purines, adenine (g.A[1]) and guanine (g.G), and pyrimidines, cytosine (g.C) and thymine (g.T), which differ in their ring structures. Classically, DNA base substitutions (mutations) are referred to by the pyrimidine of the mutated base pair given complementary binding of the DNA double helix and an inability to know on which side of the DNA helix the mutation occurred; but as the transcriptional strand is present on only one side of the helix, and either mutation can lead to an amino acid substitution, this work adopts the following notation (listed from most to least prevalent): g.C>T/G>A, g.C>A/G>T, g.G>C/C>G, g.A>T/T>A,

[1] See Preliminary section on page 28 for notation clarification

Table 3.1: Amino acid substitution matrix organization schemas.

Schema	Organization
Alphabetical	AE,AG,AP,AS,AT,AV,CF,CG,CR,CS,CW,CY,DA,DE,DG,DH,DN,DV,DY, EA,ED,EG,EK,EQ,EV,FC,FI,FL,FS,FV,FY,GA,GC,GD,GE,GR,GS,GV,GW, HD,HL,HN,HP,HQ,HR,HY,IF,IK,IL,IM,IN,IR,IS,IT,IV,KE,KI,KM,KN,KQ, KR,KT,LF,LH,LI,LM,LP,LQ,LR,LS,LV,LW,MI,MK,ML,MR,MT,MV,ND,NH,NI, NK,NS,NT,NY,PA,PH,PL,PQ,PR,PS,PT,QE,QH,QK,QL,QP,QR,RC,RG,RH, RI,RK,RL,RM,RP,RQ,RS,RT,RW,SA,SC,SF,SG,SI,SL,SN,SP,SR,ST,SW,SY, TA,TI,TK,TM,TN,TP,TR,VA,VD,VE,VF,VG,VI,VL,VM,WC,WG,WL,WR,WS, XC,XE,XG,XK,XL,XQ,XR,XS,XW,XY,YC,YD,YF,YH,YN,YS
BLOSUM62	XR,XC,XQ,XE,XG,XL,XK,XS,XW,XY,RC,RI,RW,NI,DY,DV,CR,CG,GC,GV, HL,IR,IN,IK,LH,LP,KI,PL,SW,WR,WS,YD,VD,VG,AD,RG,RL,RP,NY,DA, CF,CW,CY,QL,EG,EV,GR,GE,GW,HP,IS,LR,LQ,LS,LW,FC,FS,PR,PH,SI, SL,SF,SY,WC,WG,WL,YN,YC,YS,VE,AE,AP,RM,RS,RT,DG,DH,CS,QP,EA, GD,HD,IT,KM,KT,MR,MK,MT,FV,PA,PQ,PS,PT,SR,SC,SP,TR,TI,TK,TM,TP, VF,AG,AT,AV,RH,NK,NT,QH,GA,GS,HR,HQ,IF,LF,KN,FI,FL,SG,TA,TN,VA, AS,RQ,ND,NH,NS,DN,QR,QK,EK,HN,IM,LV,KQ,KE,MI,MV,SA,SN,ST,VL,VM, RK,DE,QE,ED,EQ,HY,IL,LI,LM,KR,ML,YH,IV,FY,YF,VI
PMBEC	IR,RI,FS,SF,LR,RL,EK,KE,HP,PH,IS,SI,IN,NI,LS,SL,SY,YS,RW,WR,LP,PL,CR,RC, MT,TM,SW,WS,RT,TR,IK,KI,PR,RP,HL,LH,KN,NK,TI,IT,KQ,QK,KT,TK,MR,RM,GW, WG,GR,RG,CS,SC,DV,VD,DY,YD,DH,HD,FV,VF,AD,AE,DA,GV,VG,MV,VM,HQ,QH, DG,GD,KM,MK,EV,VE,LW,WL,PS,SP,EG,GE,NT,TN,EA,NY,YN,LQ,QL,CY,YC, QR,RQ,CG,GC,NS,SN,DN,ND,PT,TP,RS,SR,EQ,QE,GS,SG,CF,FC,HN,NH,HY,YH, AG,GA,AS,SA,CW,WC,LV,VL,PQ,QP,FL,LF,AT,TA,IM,MI,FI,IF,DE,ED,LM,ML,AP, PA,AV,VA,ST,FY,YF,IL,LI,IV,VI,HR,RH,KR,RK

BLOSUM62[155] and PMBEC[157] organized from most negative (unfavorable) to most positive (favorable) based on weights, ties listed alphabetically. Amino acid substitutions noted with the starting (mutated) amino acid in single letter notation in the first position and substituted (variant) amino acid in the second position. X denotes a stop codon. Note stop codons are not included in PMBEC's schema.

g.A>G/T>C, g.A>C/T>G.[119]

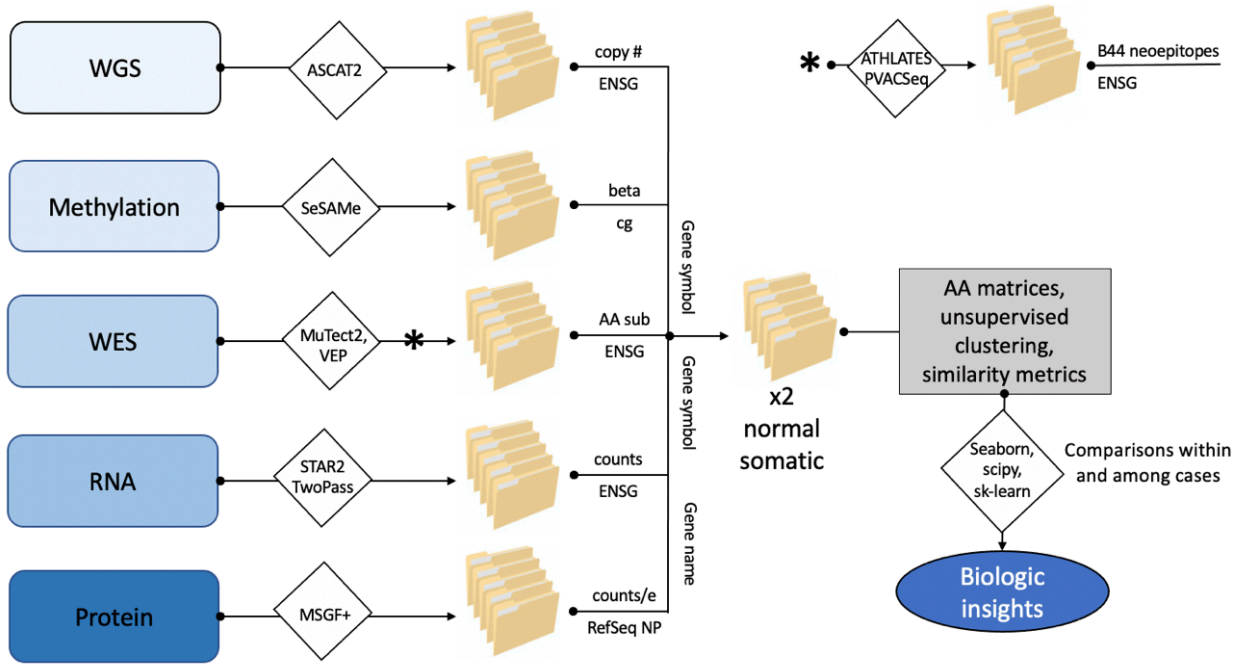
Each of these mutations can happen spontaneously, but trends emerge with distinct events and exposures. For example, tobacco smoking causes bulky adducts that lead to more g.C>A/G>T mutations,[163] while ultraviolet light leads to dimerization of thymidine (nucleoside of thymine) which causes g.C>T/G>A mutations in a second replication round.[164] g.C>T/G>A mutations are also caused by spontaneous deamination of cytosine related to aging and relate to microsatellite instability, characteristic of cancers with defective DNA mismatch repair.[165-167] g.G>C/C>G mutations suggest over activity of members of the APOBEC family of cytidine deaminases, which convert cytidine to uracil.[168, 169] g.A>C/T>G mutations associate with the combination of spontaneous deamination of cytosine with mistakes in repair mechanisms.[170] While there are no clear associations for the other two signatures,[119] as these signatures also

relate to the types of damage done by therapeutic compounds,[171] an understanding of their mapping to amino acid substitutions could provide translational opportunities.

3.3 Method

In this section, we first present an integration schema for multimodal data, which is visually summarized by Fig. 3.1. Reference creation is then discussed; histograms depicting CoCoPuTs codon and amino acid references derived from transcriptomic data and GRCh38 WES are presented as Fig. 3.2 and 3.3, respectively. We then discuss methods for data visualization, including amino acid substitution matrices (Table 3.1), histograms, boxplots, heatmaps, hierarchical clustering, and small multiples. Table 3.2 provides a reference to map amino acid substitutions to their respective DNA mutation.

Figure 3.1: Multiomic integration pipeline.



*Human leukocyte antigen and neoepitope prediction (see Chapter 5). Left-to-right: rounded rectangles – modality, diamonds – extraction/transformation software, folders – patient-specific organization, text/line mapping between files – integration approach, horizontal text – format of omic data above with omic-specific identifier below, vertical text – amino acid substitution (AA sub) mapping based on Ensembl gene identifiers (ENSG). "x2 normal, somatic" – process performed for "normal" germline (SNPs) and "somatic" tumor (SNVs). Grey rectangle – output, blue oval – purpose.

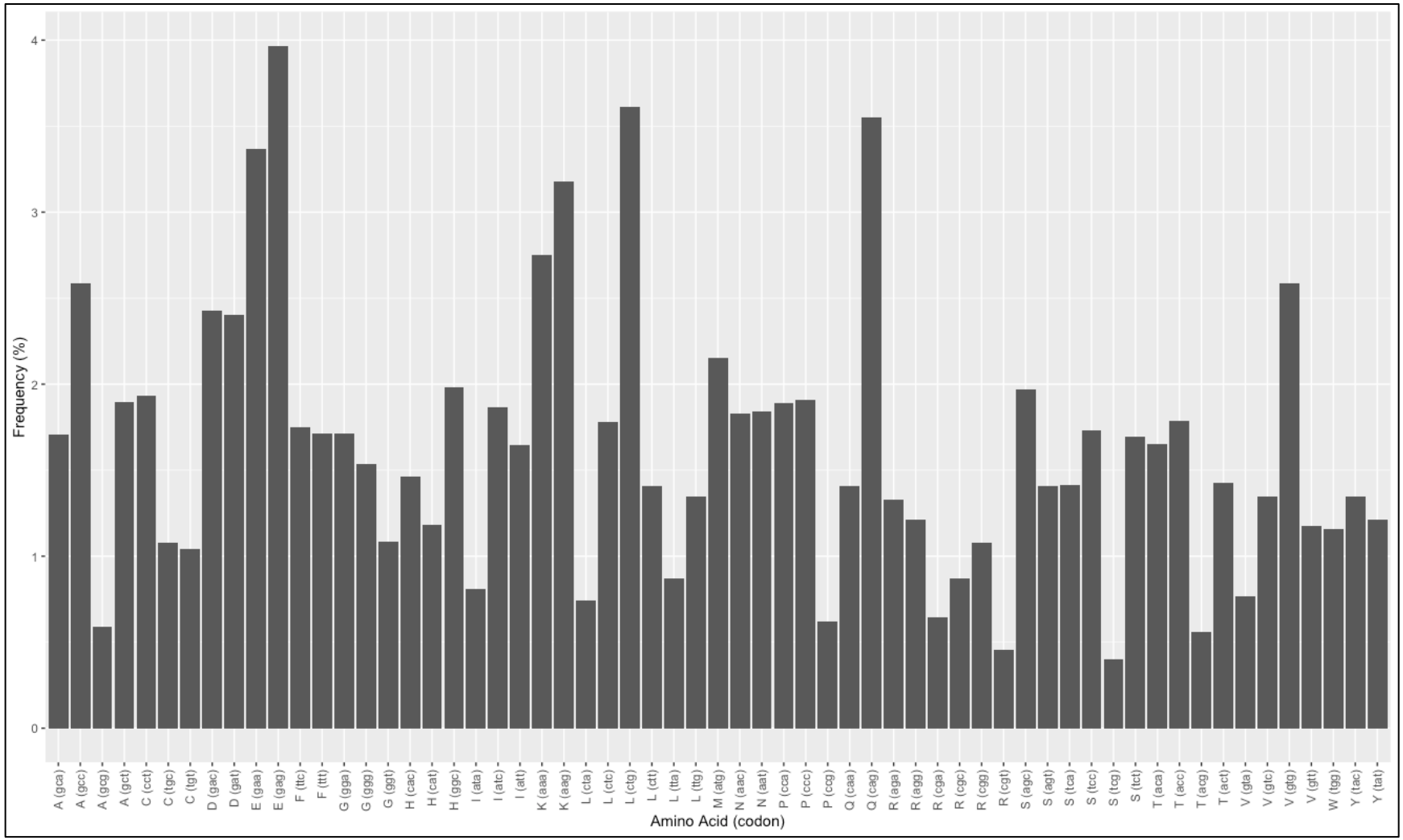
Table 3.2: Amino acid substitutions with respect to DNA mutagenesis.

Signature	DNA mutation	Amino acid substitutions (proportion)
g.C>T/G>A	g.C>T	AV (4, 1.00), HY (2, 1.00), LF (2, 0.33), PL (4, 1.00), PS (4, 1.00), RC (2, 1.00), RW (1, 0.50), SF (2, 1.00), SL (2, 1.00), TI (3, 1.00), TM (1, 1.00)
	g.G>A	AT (4, 1.00), CY (2, 1.00), DN (2, 1.00), EK (2, 1.00), GD (2, 1.00), GE (2, 1.00), GR (2, 0.33), GS (2, 1.00), MI (1, 0.33), RH (2, 1.00), RK (2, 1.00), RQ (2, 1.00), SN (2, 1.00), VI (3, 1.00), VM (1, 1.00)
g.C>A/G>T	g.C>A	AD (2, 1.00), AE (2, 1.00), DE (1, 0.25), FL (1, 0.13), HN (2, 1.00), HQ (1, 0.25), LI (3, 0.75), LM (1, 0.50), NK (1, 0.25), PH (2, 1.00), PQ (2, 1.00), PT (4, 1.00), QK (2, 1.00), RS (2, 0.33), SR (1, 0.13), SY (2, 1.00), TK (2, 1.00), TN (2, 1.00)
	g.G>T	AS (4, 1.00), CF (2, 1.00), DY (2, 1.00), ED (1, 0.25), GC (2, 1.00), GV (4, 1.00), GW (1, 1.00), KN (1, 0.25), LF (1, 0.13), MI (1, 0.33), QH (1, 0.25), RI (1, 1.00), RL (4, 1.00), RM (1, 1.00), RS (1, 0.13), SI (2, 1.00), VF (2, 1.00), VL (2, 0.33), WC (1, 0.50), WL (1, 1.00)
g.G>C/C>G	g.G>C	AP (4, 1.00), CS (2, 0.50), DH (2, 1.00), ED (1, 0.25), EQ (2, 1.00), GA (4, 1.00), GR (4, 0.67), KN (1, 0.25), LF (1, 0.13), MI (1, 0.33), QH (1, 0.25), RP (4, 1.00), RS (1, 0.13), RT (2, 1.00), ST (2, 0.33), VL (4, 0.67), WC (1, 0.50), WS (1, 1.00)
	g.C>G	AG (4, 1.00), CW (1, 0.50), DE (1, 0.25), FL (1, 0.13), HD (2, 1.00), HQ (1, 0.25), IM (1, 0.33), LV (4, 0.67), NK (1, 0.25), PA (4, 1.00), PR (4, 1.00), QE (2, 1.00), RG (4, 0.67), SC (2, 0.50), SR (1, 0.13), SW (1, 1.00), TR (2, 1.00), TS (2, 0.33)
g.A>T/T>A	g.A>T	DV (2, 1.00), ED (1, 0.25), EV (2, 1.00), HL (2, 1.00), IF (2, 1.00), IL (1, 0.25), KI (1, 1.00), KM (1, 1.00), KN (1, 0.25), LF (1, 0.13), ML (1, 0.50), NI (2, 1.00), NY (2, 1.00), QH (1, 0.25), QM (2, 1.00), RS (1, 0.13), RW (1, 0.50), SC (2, 0.50), TS (4, 0.67), YF (2, 1.00)
	g.T>A	CS (2, 0.50), DE (1, 0.25), FI (2, 1.00), FL (1, 0.13), FY (2, 1.00), HQ (1, 0.25), IK (1, 1.00), IN (2, 1.00), LH (2, 1.00), LI (1, 0.25), LM (1, 0.50), LQ (2, 1.00), MK (1, 1.00), NK (1, 0.25), SR (1, 0.13), ST (4, 0.67), VD (2, 1.00), VE (2, 1.00), WR (1, 0.50), YN (2, 1.00)
g.A>G/T>C	g.A>G	DG (2, 1.00), EG (2, 1.00), HR (2, 1.00), IM (1, 0.33), IV (3, 1.00), KE (2, 1.00), KR (2, 1.00), ND (2, 1.00), NS (2, 1.00), QR (2, 1.00), RG (2, 0.33), SG (2, 1.00), TA (4, 1.00), YD (2, 1.00)
	g.T>C	CR (2, 1.00), FL (2, 0.33), FS (2, 1.00), IT (3, 1.00), LP (4, 1.00), LS (2, 1.00), MT (1, 1.00), SP (4, 1.00), VA (4, 1.00), WR (1, 0.50), YH (2, 1.00)
g.A>C/T>G	g.A>C	DA (2, 1.00), EA (2, 1.00), ED (1, 0.25), HP (2, 1.00), IL (3, 0.75), KN (1, 0.25), KQ (2, 1.00), KT (2, 1.00), LF (1. 0.13), ML (1, 0.50), NH (2, 1.00), NT (2, 1.00), QH (1, 0.25), QP (2, 1.00), RS (1, 0.13), SR (2, 0.33), TP (4, 1.00), YS (2, 1.00)
	g.T>G	CG (2, 1.00), CW (1, 0.50), DE (1, 0.25), FC (2, 1.00), FL (1, 0.13), FV (2, 1.00), HQ (1, 0.25), IM (1, 0.33), IR (1, 1.00), IS (2, 1.00), LR (4, 1.00), LV (2, 0.33), LW (1. 100), MR (1, 1.00), NK (1, 0.25), SA (4, 1.00), SR (1, 0.13), VG (4, 1.00), WG (1, 1.00), YD (2, 1.00)

Left-to-right: DNA mutation signature[119] (see notation below), DNA base mutation, amino acid substitutions specific to each DNA base mutation. Numbers in parentheses reflect the number of mutations leading to substitution for each DNA mutation and proportion of total substitutions for which that DNA mutation is responsible (e.g., 1.00 = 100% or all). N=150 substitutions, 392 codon changes.

Preliminary. To clarify single letter notation commonly used for DNA bases and amino acids, annotations using “g.” for genomic or DNA notation and “p.” for protein or amino acid notation are used. For example, “g.A>T” refers to the DNA mutation in which an adenine becomes a thymine while “p.A>T” refers to an alanine to threonine amino acid substitution.

Figure 3.2: Amino acid codon frequencies in transcriptomic data.

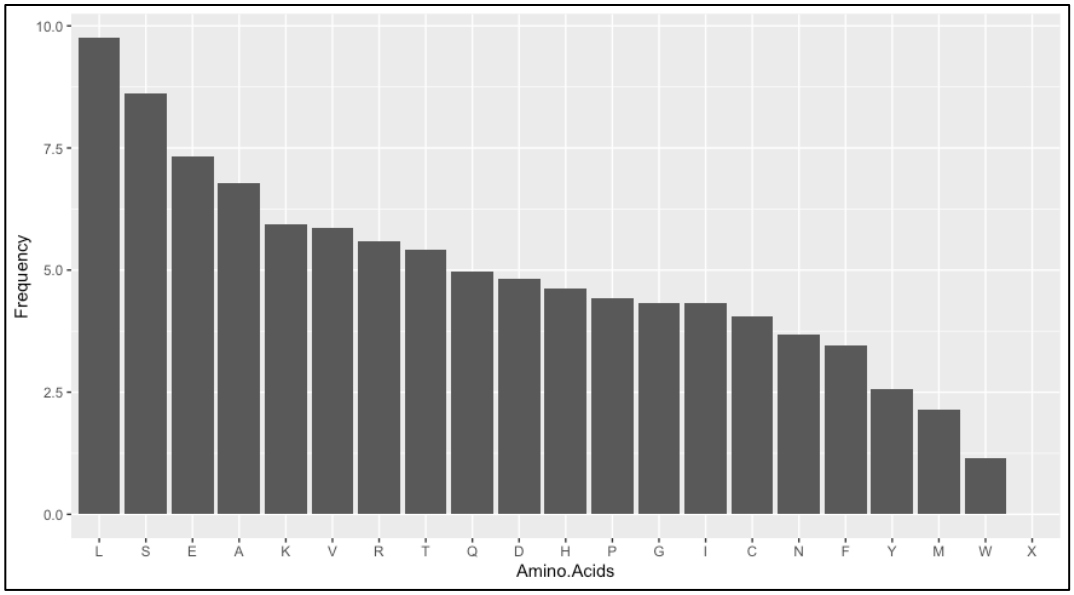


Based on data extracted from CoCoPuts 02/20/2022.[149] Listed by single letter amino acid (codon). Top three amino acid/codons (>3.5%) include p.E (g.GAA), p.L (g.CTG), and p.Q (g.CAG).

Reference creation. For comparative purposes, codon frequencies were extracted from CoCoPuTs codon usage table for homo sapiens (Fig. 3.2) and combined to represent amino acids (Fig. 3.3). The WES reference (GRCh38) was transformed to amino acids using Python v3.8 and Pandas (https://pandas.pydata.org) (Fig. 3.4). Table 3.2 provides a map of amino acid substitutions and the proportion of attributability to specific DNA base pair mutations.

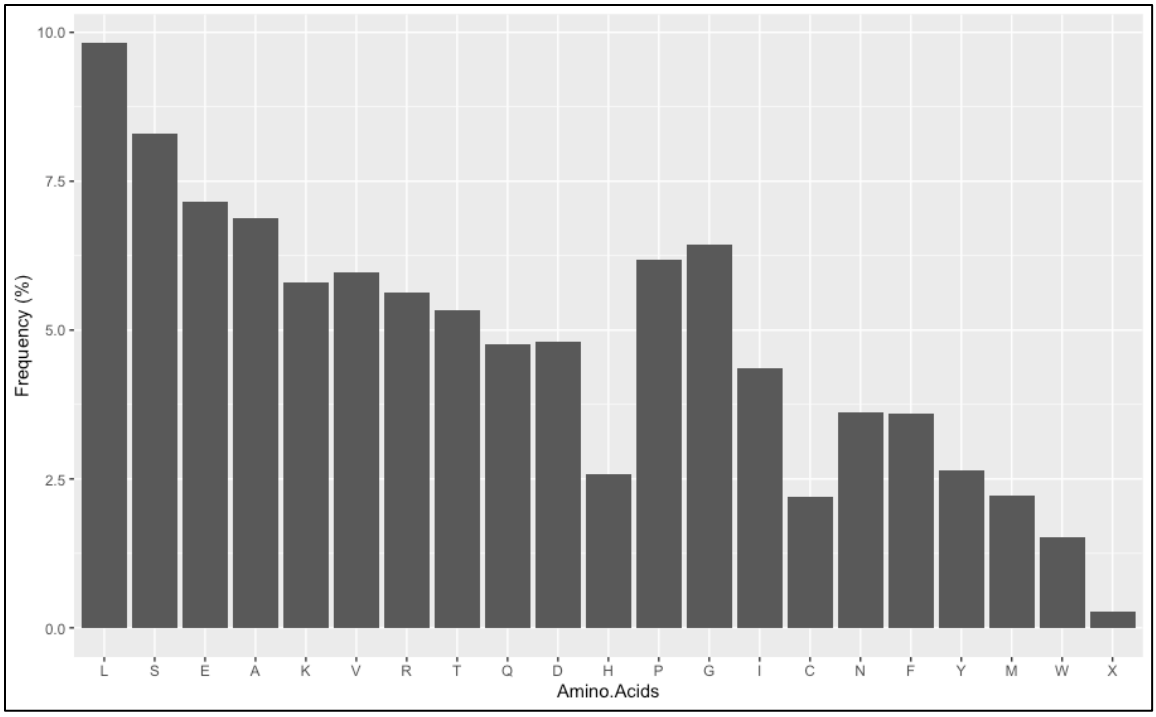
Data preprocessing. The multiomic data used for these experiments is publicly available through the Genome Data Commons (GDC) and standardized based on harmonized pipelines (https://gdc.cancer.gov/about-data/gdc-data-processing). There are three sources of genomic data for these experiments: macrodissected tumor specimens (all primary lung cancer), normal tissue (lung) specimens, and germline specimens from peripheral blood mononuclear cells (PBMC). WES data selected for these experiments included tumor and germline samples run together with MuTect2 and annotated with Variant Effect Predictor (VEP), using the 'missense' label (tumor) to identify single DNA base mutations leading to amino acid substitutions, which

Figure 3.3: Amino acid frequencies derived from transcriptomic data.



Based on data extracted from CoCoPuts 02/20/2022.[149] Amino acids listed by single letter notation, frequency by percentage.

Figure 3.4: Amino acid frequencies derived from GRCh38 (DNA reference).



Based on data extracted from CoCoPuts 02/20/2022.[149] Amino acids listed by single letter notation, frequency by percentage. X denotes stop codon.

were refined by 'PASS' to extract SNVs (tumor) and 'alt_allele_in_normal' (germline) for SNPs. Transcriptomic data (RNA) included both tumor and normal lung tissue counts without manipulation;[1] methylation data underwent SeSAMe optimization and calculation of beta-values for both tumor and normal lung tissue specimens; and multiplexed proteomics data were separated by channel into tumor and normal lung tissue files based on case IDs with incorporation of E-values and peptide fragment (ion) counts. Note RNA and proteomic data did not undergo additional normalization steps prior to incorporation as current harmonized methods have proven to be biased in the case of RNA and are not harmonized in the case of proteomics. As HLA type and neoepitopes are of interest, these were inferred from germline DNA using ATHLATES[172] and predicted with personalized variant antigens by cancer sequencing (pVAC-seq), respectively.[173]

[1] Available harmonized RNA normalization options include UQ (upper quartile) and FPKM (fragments per kilobase million), both have been proven to be biased and inferior to TMM (trimmed mean of M-values); unprocessed count data incorporated (doi: 10.1186/gb-2010-11-3-r25; 10.3389/fgene.2016.00164).

Integration schema. To build the foundational layer, SNPs and SNVs were extracted along with their Ensembl gene identifiers (ENSG), DNA base mutations, and resultant amino acid substitutions from VEP-annotated WES – while this was the dimensionality reduction step of this experiment, any gene-level information or ENSG template could be included in this layer. Separate structures were maintained for normal and tumor components at the case level. Identifiers to match gene-specific information from other omic levels were added in sequence including gene symbols, gene names, UniProt labels, NP (RefSeq) identifiers, and cg probe identifiers with the expectation there would be multiple entries for each variant to protect against data missingness. Omic data was then added to the file in sequence including RNA count numbers and copy number (CN) matched to ENSG and beta-values matched to cg probe. Proteomic sequences, E-value, and ion count/amplitude measurements matched to NP identifiers were added last. As the purpose of the structure is verify detectable protein at the gene level, entries were sorted by E-values with the entry with the lowest E-value retained; E-values less than 0.001 were used to suggest protein expression. A schema for the integration pipeline is provided in Fig. 3.1. Python v4.0 was used for pipeline construction leveraging Pandas and NumPy (https://numpy.pydata.org) for data frame manipulation.

Outcome representation. Histograms, boxplots, and scatterplots with associated summary statistics were created in R v4.1 (https://www.r-project.org). Amino acid substitutions were extracted based on count and proportions in varied organizations (alphabetical, BLOSUM62, PMBEC) as a whole and based on protein expression. Mutational signature was extracted using MuTect2 annotations ('REF' and 'ALT') and reported as complementary base pair mutations (e.g., g.C>T/G>A = g.C>T + g.G>A). These representations were visualized with heatmaps and subjected to hierarchical clustering with Seaborn (https://seaborn.pydata.org) using Euclidean, cosine, and correlational distance metrics with both x and y variables grouped to maximize

subgroup mapping to biologic correlates. PyPlot (https://pyplot.pydata.org) was used for heatmap visualization with small multiples to explore uncovered relationships.

Statistical analysis. Variance was evaluated with Pandas. Relationships identified based on various outcome representations were evaluated with two-proportional z-tests for proportions, student's T tests, and Pearson's correlations using R v4.1. Two-tailed P values of <0.05 were considered statistically significant and rounded to the nearest thousandth.

3.4 Experiments

Small multiples are used to evaluate amino acid representations, including alphabetical, BLOSUM62, and PMBEC arrangements (Fig. 3.8, also see Table 3.1) based on count and proportion. Fig. 3.9 shows a comparison of PMBEC proportion representations across

Table 3.3: CPTAC case-level data availability.

Feature	**LUAD (N, %)**		**LUSC (N, %)**	
Total cases	111	100.0	108	100.0
*Complete clinical information**	34	30.6	33	30.6
Adequate clinical information†	107	96.4	102	94.4
Age	107	96.4	107	99.1
Sex	107	96.4	107	99.1
Race/ethnicity	34	30.6	33	30.6
Stage	107	96.4	104	96.3
Smoking history	104	93.7	98	90.7
Survival >2 years	46	41.4	69	63.9
Complete genomic data‡	91	81.9	86	79.6
WES-somatic	105	94.6	105	97.2
WES-germline	92	82.9	87	80.6
RNA-somatic	111	100.0	106	98.1
RNA-normal lung	95	85.6	93	86.1
CH3-somatic	110	99.1	106	98.1
CH3-normal lung	99	89.2	90	83.3
Mass spec-somatic	111	100.0	108	100.0
Mass-spec-normal lung	102	91.9	100	92.6
WGS-somatic	111	100.0	108	100.0
WGS-normal lung	111	100.0	108	100.0
Evaluable cases§	**91**	**81.9**	**86**	**79.6**

*All clinical characteristics collected. †More than half clinical characteristics collected. ‡All files able to be processed and analyzed. §All files able to be processed and analyzed with adequate clinical information.

Table 3.4: CPTAC cohort features.

Feature	LUAD (N, %)*	LUSC (N, %)*
Age	62 (35-81)	67 (40-88)
Sex		
Female	37 (34.6)	22 (20.6)
Male	70 (65.4)	85 (79.4)
Ancestry		
Asian	53 (49.5)	23 (21.5)
Black/African	1 (0.9)	1 (0.9)
Caucasian	38 (35.5)	81 (75.7)
Hispanic	2 (1.8)	0 (0.0)
Smoking history		
Current/former	58 (54.2)	82 (76.6)
Never	46 (43.0)	16 (15.0)
Stage		
IA	24 (22.4)	17 (15.9)
IB	33 (30.8)	22 (20.6)
IIA	16 (15.0)	24 (22.4)
IIB	13 (12.1)	19 (17.8)
IIIA	20 (18.7)	20 (18.7)
IIIB-IV	1 (0.9)	2 (1.8)
Survival (days)		
PFS events	25	23
PFS	423 (7-1427)	728.5 (1-1580)
OS events	14	18
OS	444 (12-1438)	756 (1-1580)

PFS – progression-free survival, OS – overall survival. *Percentages based on N=107.

LUAD/LUSC SNVs and SNPs with protein expression. This approach is then used to compare multiomic data, including proteomic data (Fig. 3.10), transcriptomic data (Fig. 3.11), and methylation data (Fig. 3.12). Hierarchical clustering is then used to evaluate probabilistic relationships (Fig. 3.13). Finally, an ability to discern technical artifact is evaluated (Fig. 3.14).

Datasets. As a first model system, the Clinical Proteomics Tumor Analysis Consortium (CPTAC) lung cancer datasets are used.[81, 117, 118] The Cancer Genome Atlas (TCGA) lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and melanoma (SKCM) datasets are used for targeted comparisons.[104]

Results. Data availability and cohort characteristics are provided in Tables 3.3 and 3.4. Approximately 80% of CPTAC-LUAD/-LUSC cases had complete genomic data and were

Table 3.5: CPTAC summary statistics.

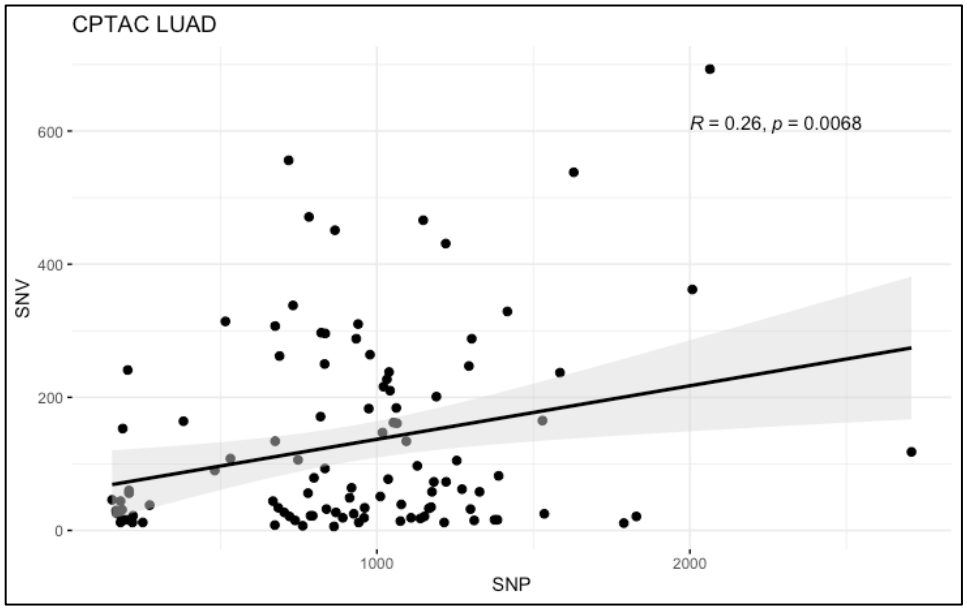
AA substitution #	LUAD	LUSC
SNP	940.0 (487.2)	264.0 (360.7)
SNP-P	87.0 (53.4)	14.0 (40.6)
SNV	588.5 (939.9)	915.0 (596.5)
SNV-P	58.0 (141.3)	109.0 (57.8)

AA – amino acid, P – protein, SNP – single nucleotide polymorphism, SNV – single nucleotide variant. Data represented as median (standard deviation) reflecting count of each subtype.

successfully processed with the pipeline. The most common reason for pipeline failure was missing normal information (LUAD N=19, LUSC N=21). A summary of SNP/SNVs is provided in Table 3.5, which exhibit large standard deviations, indicating substantial heterogeneity. Fig. 3.5 shows numbers of SNPs correlated with numbers of SNV in LUAD (R = 0.26, p = 0.007) but not in LUSC. Comparing across the 150 amino acid substitutions included in PMBEC (no stop

Figure 3.5: CPTAC comparison of SNV by SNP count by cohort.

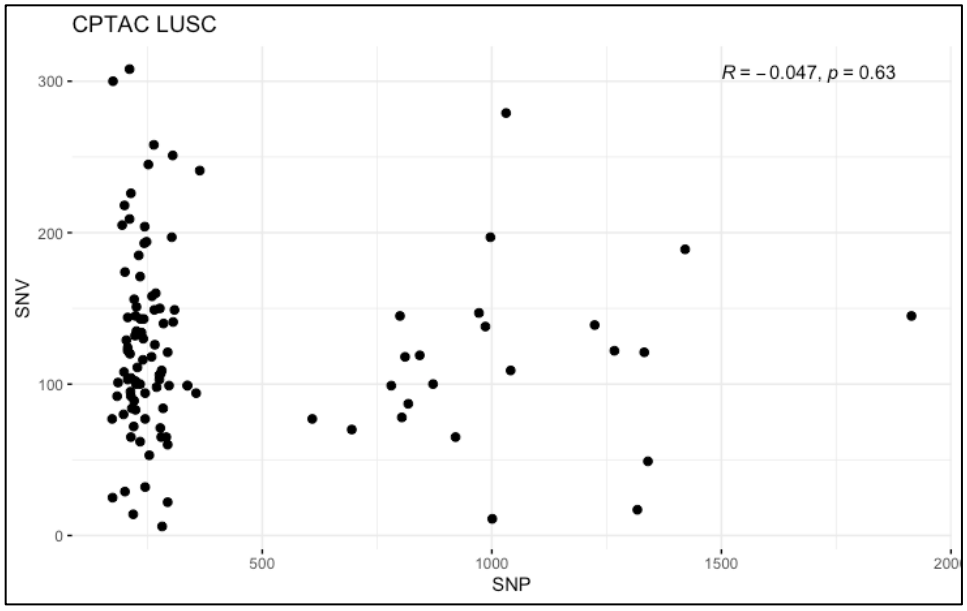
A. LUAD



Scatterplots depicting relationship of number of single nucleotide polymorphisms (SNPs) to single nucleotide variants (SNVs), correlations compared with Pearson's test. Line of best fit shown in black surrounded by gray depicting 95% confidence interval. LUAD SNP vs SNV: R=0.260, P=0.007

Figure 3.5 (con't): CPTAC LUSC comparison of SNV by SNP count by cohort.

B. LUSC



Scatterplots depicting relationship of number of single nucleotide polymorphisms (SNPs) to single nucleotide variants (SNVs), correlations compared with Pearson's test. Line of best fit shown in black surrounded by gray depicting 95% confidence interval. LUSC SNP vs SNV: R=-0.047, P=0.630.

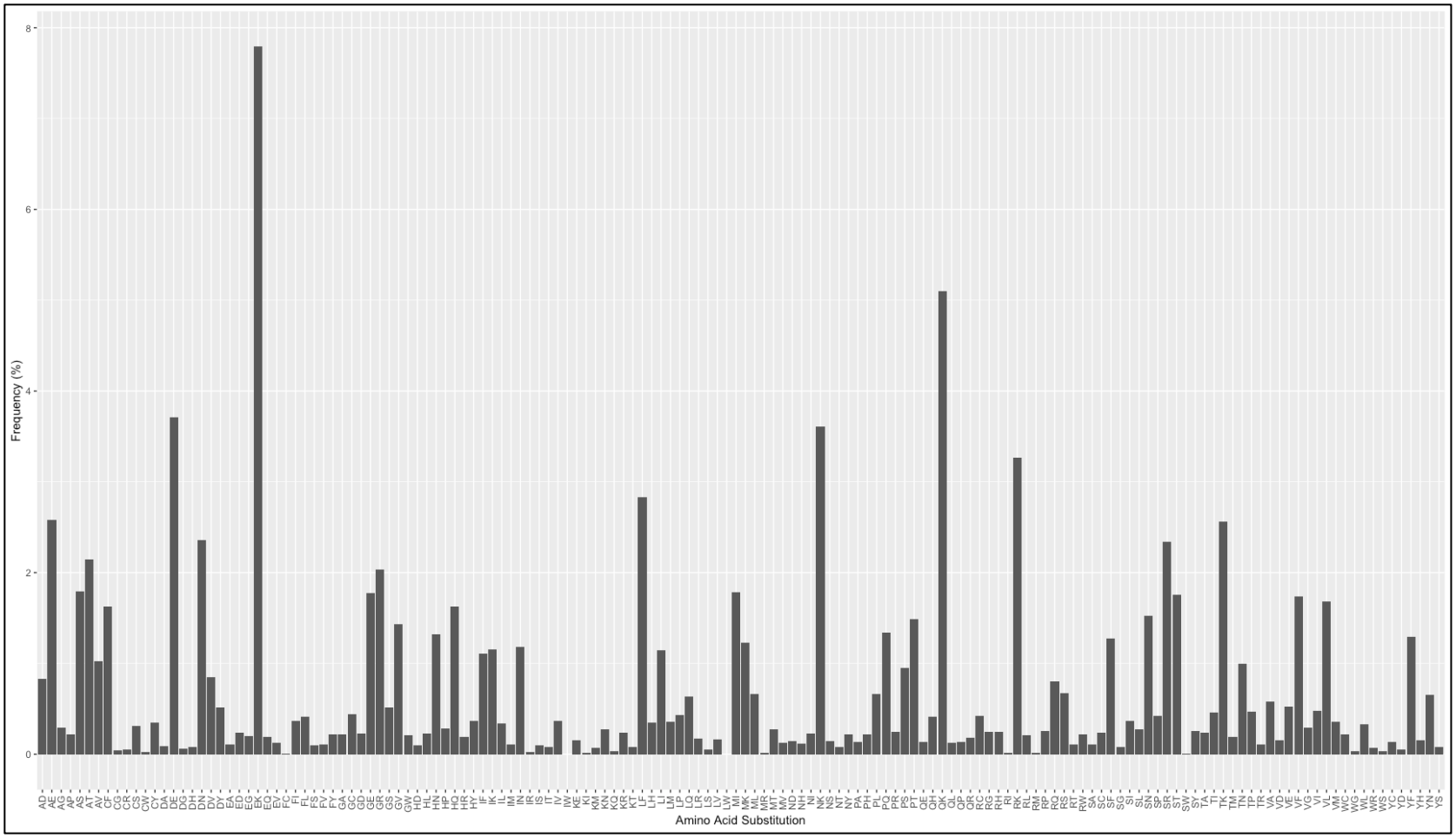
codons), the average proportion of amino acid substitutions with evidence of protein expression was higher in LUAD at 0.360 (SD 0.081) than LUSC, 0.267 (SD 0.056), p<0.001.

Fig. 3.6 displays the composition of amino acid substitutions of protein-expressed SNPs and SNVs across the dataset, demonstrating similar profiles with p.E>K substitutions representing the majority of substitutions in both SNPs and SNVs, but at a lower prevalence in the SNV dataset. Table 3.6 shows the variance of different amino acid substitutions. Variance was highest in LUSC SNV counts in DNA and lowest in LUAD SNP proportions with protein expression. The highest variance substitutions included p.E>K, p.Q>K, and p.G>V; the lowest variance substitutions included p.L>W, p.I>R, and p.W>G. Mapping to biologic correlates, the highest variance substitutions were attributable to g. C>T/G>A (aging) or g.C>A/G>T (smoking) mutations[2] from

[2] Note by convention, DNA mutations are referenced as 6 sets of complementary mutations as mutations can arise on either side of the DNA helix (see Mutational Signature).

Figure 3.6: CPTAC cohort-wide amino acid substitutions with protein expression.

A. From single nucleotide polymorphisms (SNPs) in normal lung tissue samples.

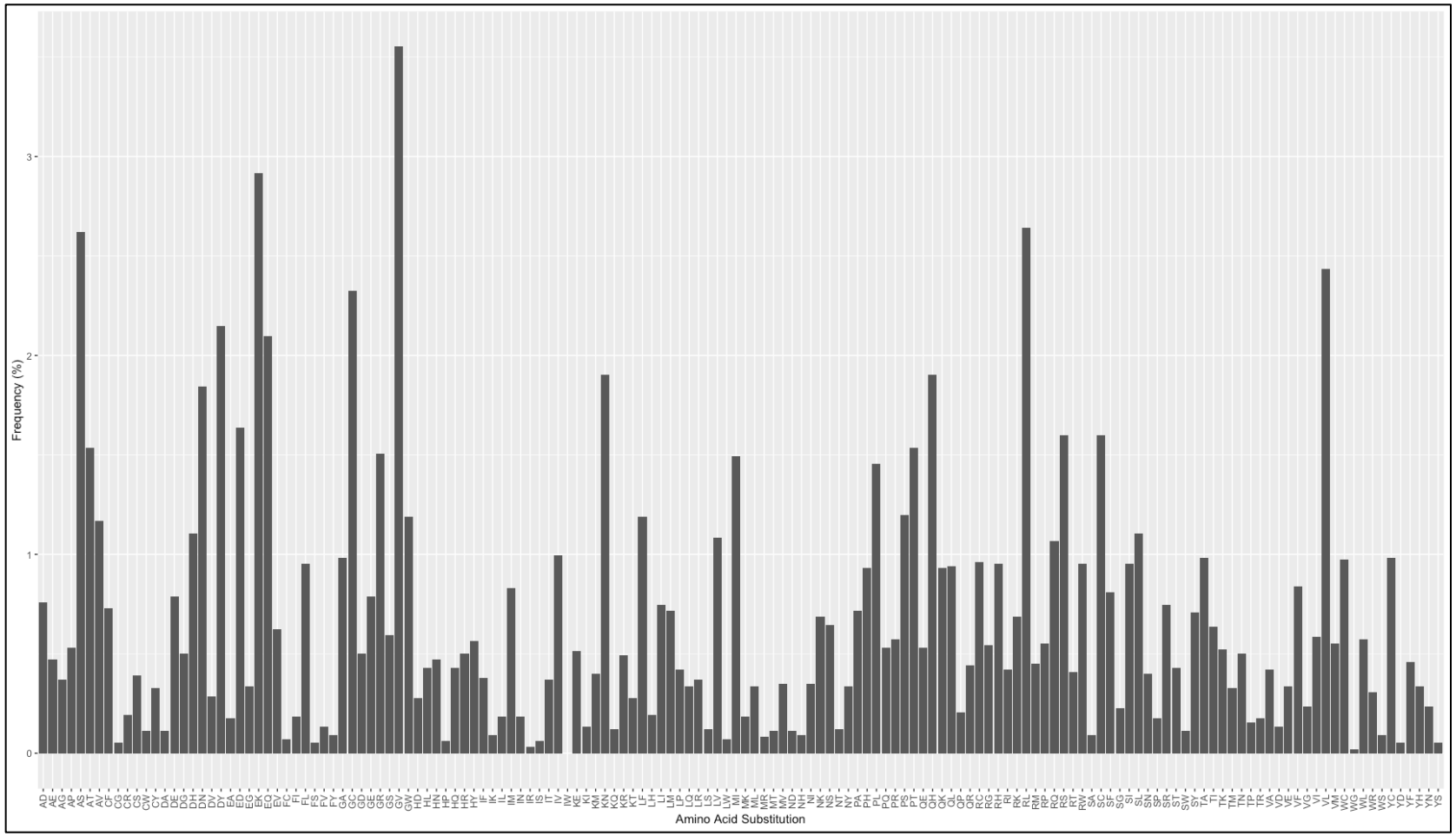


Listed by single letter notation with starting amino acid in first position and substituted amino acid in second position. Protein expression based on detection of at least one protein fragment with E value<0.001 from the gene including the SNP.

Most prevalent: p.E>K (7.791%), p.Q>K (5.099%), and p.D>E (3.711%).

Figure 3.6 (con't): CTPAC cohort-wide amino acid substitutions with protein expression.

B. From single nucleotide variants (SNVs) in non-small cell lung cancer samples.



Listed by single letter notation with starting amino acid in first position and substituted amino acid in second position. Protein expression based on detection of at least one protein fragment with E value<0.001 from the gene including the SNV.

Most prevalent: p.G>V (3.552%), p.E>K (2.918%), and p.R>L (2.641%).

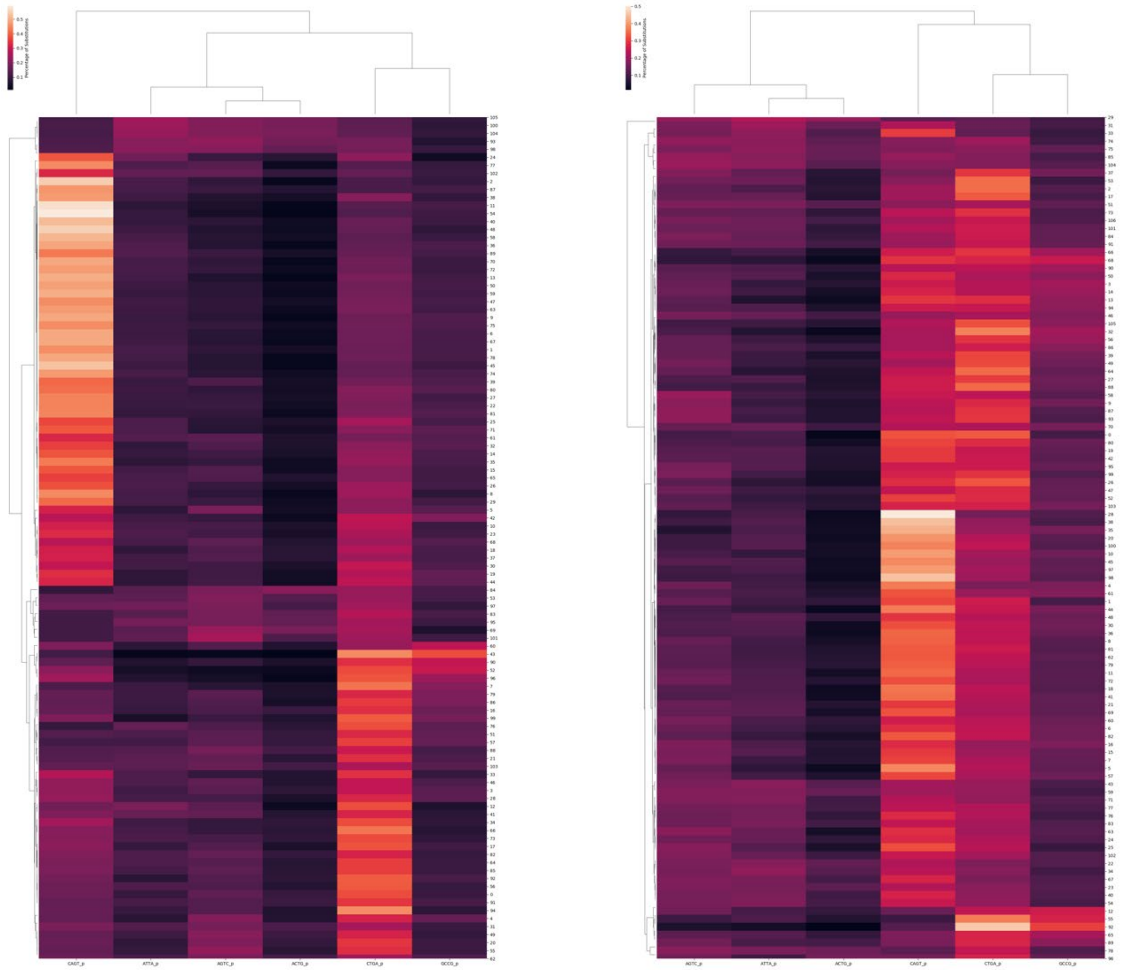
Table 3.6: CPTAC amino acid substitution variance by cohort.

	LUAD-SNP#	LUAD-SNP%	LUAD-SNPP#	LUAD-SNPP%	LUAD-SNV#	LUAD-SNV%	LUAD-SNVP#	LUAD-SNVP%	LUSC-SNP#	LUSC-SNP%	LUSC-SNPP#	LUSC-SNPP%	LUSC-SNV#	LUSC-SNV%	LUSC-SNVP#	LUSC-SNVP%
IR	0.080	0.326	0.081	0.107	0.000	0.000	0.000	0.000	0.173	0.694	0.109	0.428	0.218	1.283	0.034	0.000
RI	0.334	0.003	0.084	0.004	2.518	0.840	0.507	1.452	0.145	0.022	0.034	0.033	65.619	0.395	0.231	0.716
FS	1.163	0.101	0.300	0.253	0.227	0.566	0.032	1.210	1.419	0.323	0.293	0.539	8.269	0.135	0.023	0.165
SF	150.75	0.289	25.078	0.409	7.827	2.749	0.757	3.947	60.275	0.668	10.352	0.901	463.16	1.844	0.864	3.747
LR	2.251	0.292	0.560	0.310	1.397	4.418	0.263	10.764	1.702	0.348	0.368	0.583	30.985	0.214	0.121	0.594
RL	4.963	0.099	0.941	0.178	47.983	6.624	7.097	10.388	2.834	0.329	0.597	0.593	2377.0	2.467	2.289	5.011
EK	4633.0	9.741	869.45	10.149	113.90	20.438	6.514	34.346	1972.0	8.900	414.86	10.447	3016.9	4.326	3.806	6.955
KE	1.654	0.120	0.408	0.191	1.214	0.482	0.242	1.059	2.305	0.399	0.558	0.697	201.19	0.970	0.404	2.018
HP	5.157	0.470	1.102	0.603	0.218	1.744	0.042	2.095	3.233	0.627	0.792	0.975	1.107	0.043	0.023	0.061
PH	4.782	0.063	0.974	0.102	9.579	3.323	1.092	4.091	3.290	0.241	0.674	0.462	424.28	1.112	0.368	1.156
KQ	0.763	0.064	0.184	0.056	0.428	0.896	0.107	1.040	0.534	0.119	0.099	0.185	8.915	0.089	0.066	0.151
QK	1705.4	4.270	301.53	4.453	7.428	10.461	0.807	20.300	1026.8	5.042	199.34	5.429	568.85	1.886	0.554	5.133
KT	0.409	0.040	0.121	0.037	2.191	2.429	0.516	3.406	1.626	0.293	0.454	0.614	13.298	0.129	0.095	0.244
TK	479.09	1.076	89.827	1.347	2.525	1.977	0.307	2.979	314.37	1.209	64.726	1.579	275.11	0.828	0.421	2.175
MR	0.074	0.001	0.011	0.001	0.112	0.167	0.022	0.297	0.322	0.053	0.076	0.115	2.632	0.057	0.066	0.109
RM	0.255	0.029	0.062	0.057	3.560	0.587	0.672	1.104	0.121	0.013	0.023	0.012	68.642	0.261	0.200	0.362
GW	6.196	0.048	1.316	0.073	14.063	1.819	2.313	1.930	3.272	0.212	0.628	0.345	402.23	1.015	0.675	2.091
WG	0.547	0.046	0.160	0.080	0.000	0.000	0.000	0.000	0.400	0.050	0.095	0.069	0.739	0.026	0.023	0.073
AD	46.592	0.166	10.562	0.228	4.244	0.893	0.553	1.460	18.036	0.160	4.160	0.229	444.02	1.254	0.438	1.698
AE	438.90	1.606	82.560	1.714	2.963	0.503	0.522	0.973	240.12	1.250	44.236	1.307	156.91	0.482	0.268	0.777
DA	0.994	0.017	0.300	0.032	0.137	0.602	0.022	0.746	1.016	0.148	0.289	0.270	10.904	0.296	0.189	0.890
GV	112.69	0.375	21.139	0.529	91.962	12.636	13.919	28.863	55.987	0.749	11.879	1.474	3458.9	5.236	2.579	14.794
VG	1.079	0.266	0.512	0.739	0.447	1.118	0.124	7.298	2.323	0.405	0.788	1.219	14.052	0.757	0.127	2.036
LW	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HR	2.332	0.473	0.452	0.385	1.112	2.575	0.213	3.824	3.354	0.714	0.745	1.005	126.30	0.631	0.308	1.084
RH	2.471	0.408	0.596	0.632	2.159	4.341	0.373	6.305	3.755	0.713	0.741	0.971	650.79	2.384	0.622	3.934
KR	3.552	0.502	0.931	0.631	1.477	2.928	0.290	3.601	2.760	0.531	0.721	0.892	106.66	1.005	0.292	1.635
RK	816.20	2.089	156.89	2.293	1.835	1.065	0.246	1.771	409.84	2.109	81.649	2.388	259.22	0.962	0.722	2.274
TOTAL	15021	59	2864	73	970	362	134	588	7708	85	1561	127.16	52882	151	68	288

LUAD – adenocarcinoma, LUSC – squamous cell carcinoma, SNP – single nucleotide polymorphism, SNV – single nucleotide variant, P – protein. Amino acids substitutions with starting amino acid on left and substituted amino acid on right. Highest variance marked in orange, lowest blue.

Figure 3.7: CPTAC mutational signature by cohort.

A. LUAD-SNV **B. LUSC-SNV**



Hierarchical clustering with mutational signature proportion on the x-axis and cases on the y-axis, lighter colors depict higher proportions with white ~50%, orange ~40%, magenta ~30%, purple ~20%, dark purple ~10%. Smoking signature associated with higher g.C>A/G>T, aging with higher g.C>T/G>A, APOBEC with higher g.G>C/C>G.

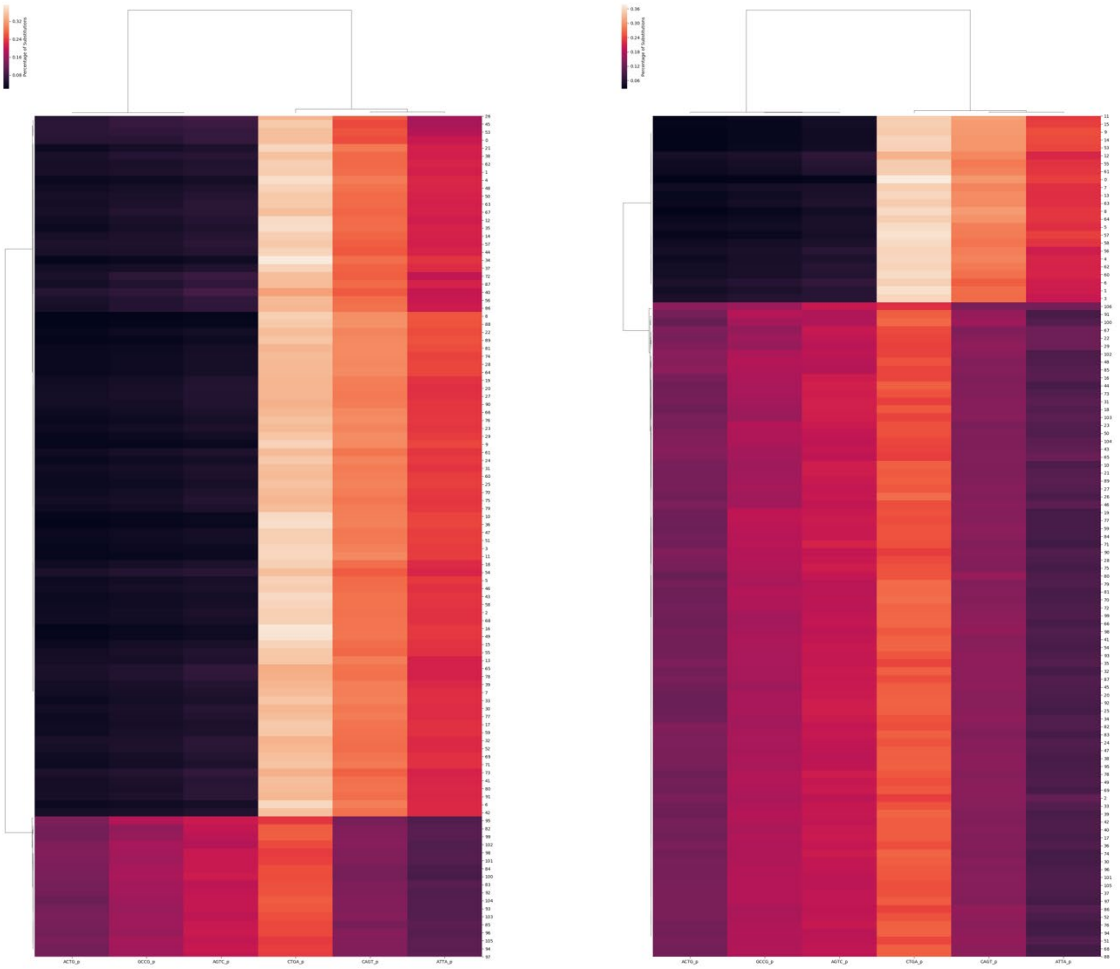
LUAD x-axis (left-to-right): **g.C>A/G>T**, g.A>T/T>A, g.A>G/T>C, g.A>C/T>G, g.C>T/G>A, g.G>C/C>G

LUSC x-axis (left-to-right): g.A>G/T>C, g.A>T/T>A, g.A>C/T>G, **g.C>A/G>T**, g.C>T/G>A, g.G>C/C>G

Figure 3.7 (con't): CPTAC mutational signature by cohort.

C. LUAD-SNP

D. LUSC-SNP



Hierarchical clustering with mutational signature proportion on the x-axis and cases on the y-axis, lighter colors depict higher proportions with white ~50%, orange ~40%, magenta ~30%, purple ~20%, dark purple ~10%.

X-axis (left-to-right: g.A>C/T>G, g.G>C/C>G, g.A>G/T>C, g.C>T/G>A, g.C>A/G>T, g.A>T/T>A

Table 3.7: CPTAC mutational signature by cohort.

	LUAD		LUSC	
Mutation	**SNP**	**SNV**	**SNP**	**SNV**
g.C>T/G>A	0.351 (0.042)	0.240 (0.084)	0.258 (0.044)	0.243 (0.055)
g.C>A/G>T	0.283 (0.056)	0.278 (0.145)	0.147 (0.060)	0.263 (0.075)
g.G>C/C>G	0.043 (0.049)	0.117 (0.050)	0.173 (0.056)	0.137 (0.040)
g.A>T/T>A	0.225 (0.050)	0.114 (0.037)	0.100 (0.053)	0.116 (0.046)
g.A>G/T>C	0.053 (0.058)	0.122 (0.050)	0.193 (0.060)	0.147 (0.031)
g.A>C/T>G	0.038 (0.038)	0.058 (0.044)	0.129 (0.040)	0.063 (0.036)

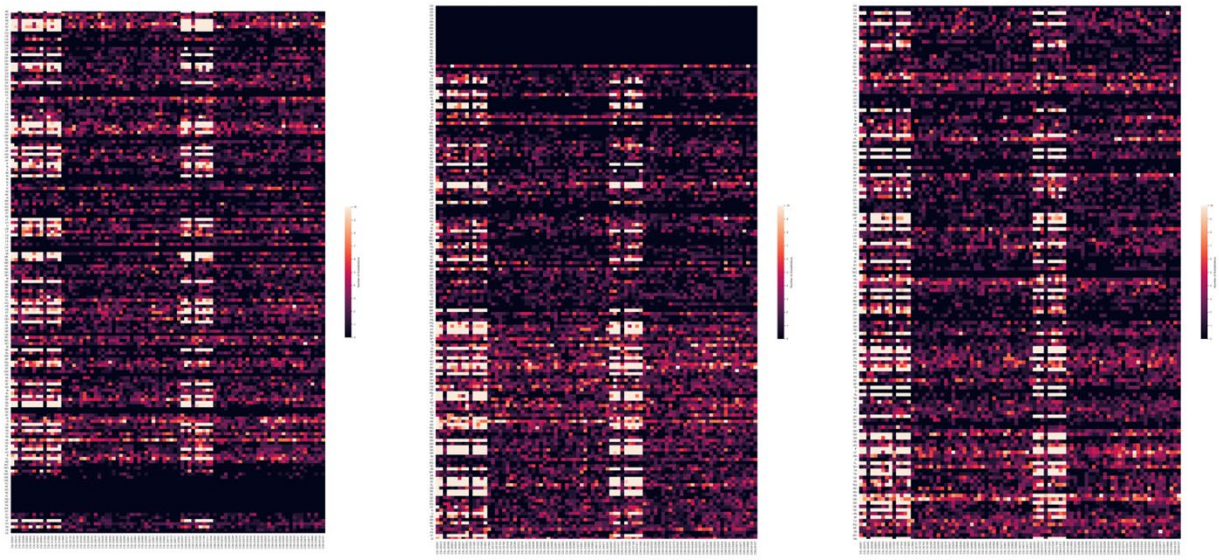
Includes DNA mutations with standard notation from most common to least common. [119] Median proportions (analyzed at the case level) are provided with standard deviations in parentheses.

amino acids codons with average or greater than average reference frequencies. Those with the lowest variance were the result of a relative absence of observations, which associated with the least common mutation, g.A>C/T>G, and small codon frequencies with only one possible codon change, suggesting a probabilistic relationship (see Table 3.2). Table 3.7 summarizes mutational signature with respect to LUAD/LUSC SNP and SNV profiles. Fig. 3.7 panels A and B depict SNV mutational signature profiles in the cohorts, which suggest a mixture of smoking and aging signatures in both LUAD and LUSC.[119] Fig. 3.7 panels C and D show SNP mutational signature profiles, which suggest that the number of events dictates the proportion of events attributable to g.C>T/G>A mutations.

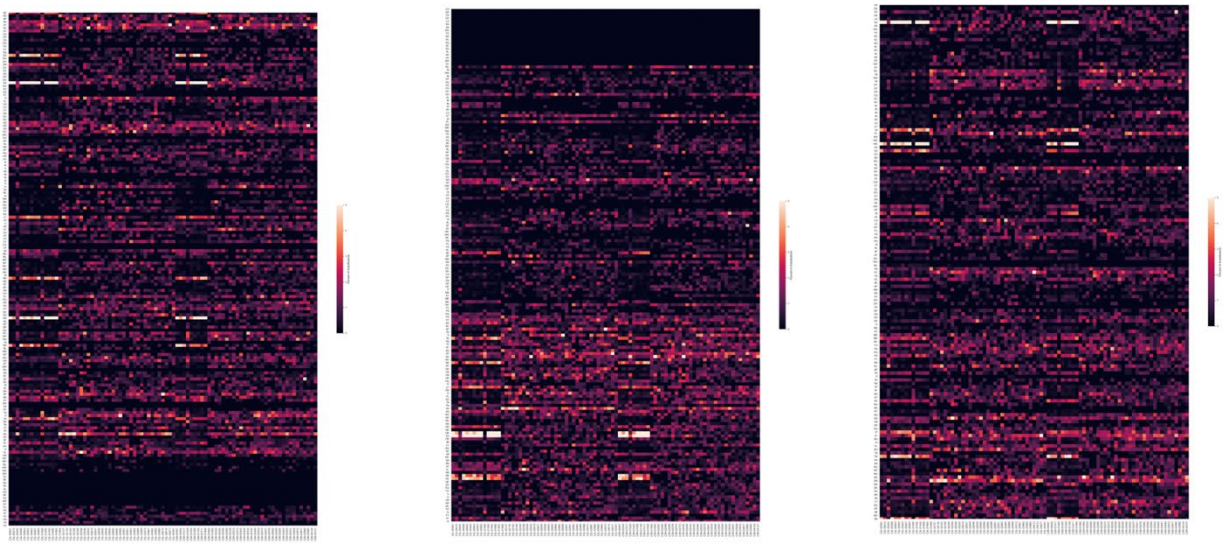
Fig 3.8 shows small multiples of alphabetical, BLOSUM62, and PMBEC visualizations. Overall, the performance of these amino acid matrix depictions was disappointing. They were able to suggest similar subgroups across count and proportion data (Fig. 3.9) but there were no clear trends suggestive of biologic correlates. Comparisons between proteomic and genomic data (Fig. 3.10), transcriptomic data (Fig. 3.11), and methylation data (Fig. 3.12), were without clear relationships. Hierarchical clustering small multiples, however, were useful. Fig. 3.13 shows that there are two clear subgroups based on SNPs with protein expression that have the same highest proportion substitution: p.E>K. These subgroups were all identified by Euclidean, correlation, and cosine distance metrics, suggesting fidelity across measures.

Figure 3.8: Example of amino acid matrices with varied representations.

A. LUSC-SNP by count (alphabetic, BLOSUM62, PMBEC)



B. LUSC-SNP by proportion (alphabetic, BLOSUM62, PMBEC)



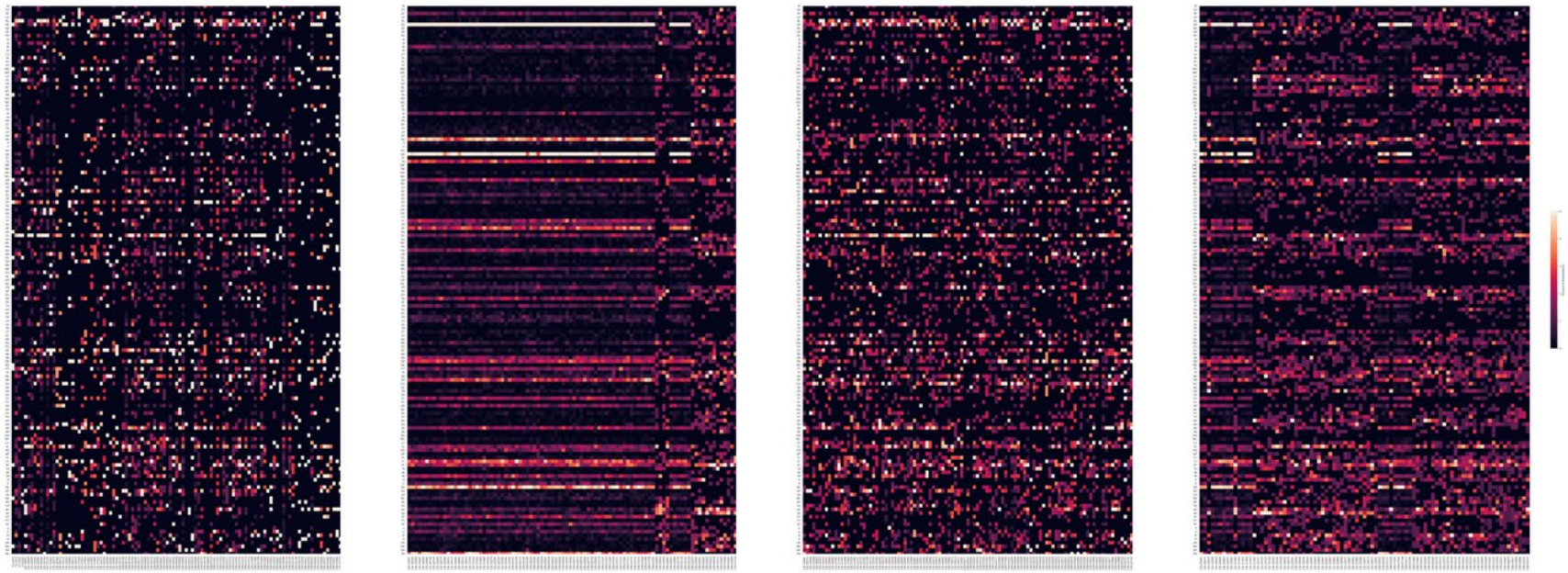
Count scale (black-to-white) 0-10, proportion scale (black-to-white) 0-5%.

Amino acids on y-axis are listed top-to-bottom according to left-to-right listing in Table 3.1.

Heatmaps depict cases on the x-axis and amino acid substitutions on the y-axis, lighter colors depict higher numbers/proportions while black represents lower numbers/proportions. Note alphabetic and BLOSUM62 representations include substitutions from a stop codons, which are represented by mostly black sections near the bottom of alphabetic representations and the top of BLOSUM62 representations. No clear patterns based on amino acid organization detected (excluding rarity of stop codon mutations).

Figure 3.9: PMBEC amino acid substitution proportions with protein expression.

A. LUAD-SNV **B. LUAD-SNP** **C. LUSC-SNV** **D. LUSC-SNP**



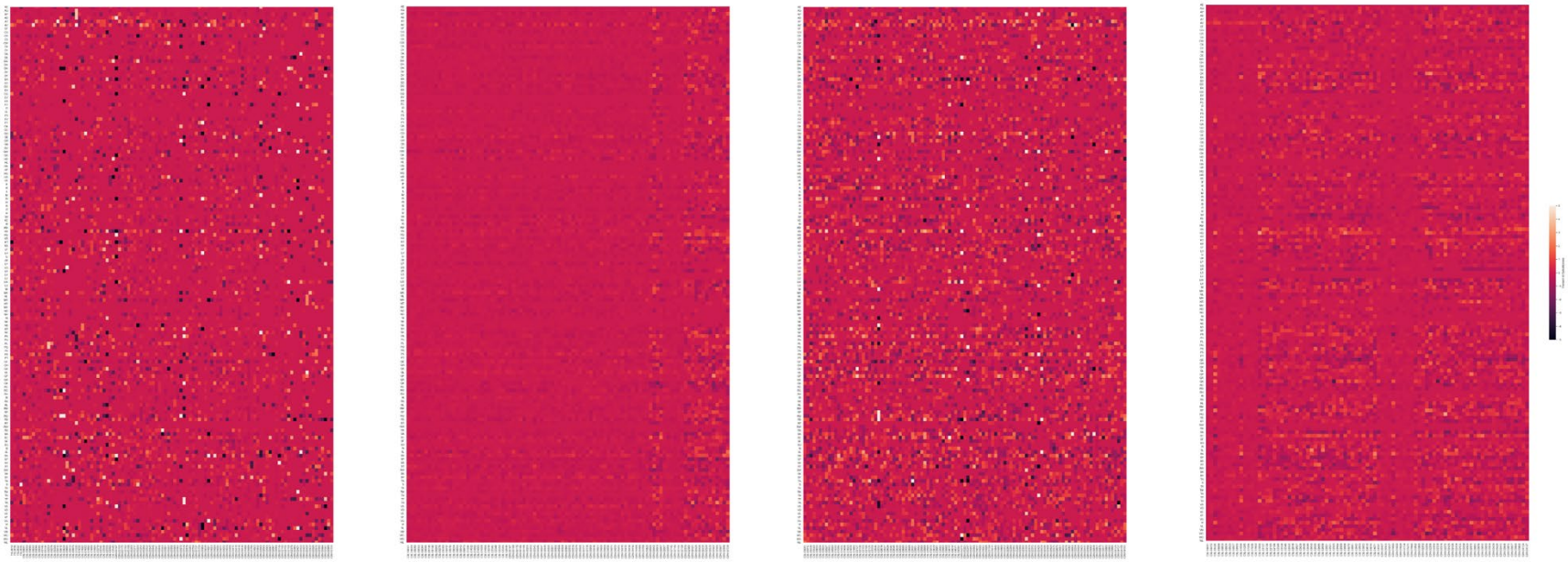
Proportion scale (black-to-white) 0-5%.

Amino acids on y-axis are listed top-to-bottom according to left-to-right PMBEC listing in Table 3.1.

Heatmaps depict cases on the x-axis and amino acid substitutions on the y-axis, lighter colors depict higher proportions while black represents lower numbers/proportions. Protein expression based on detection of at least one protein fragment with E value<0.001 with the gene including the SNP or SNV. No clear patterns detected.

Figure 3.10: Genomic vs. proteomic proportion representations of substitutions.

A. LUAD-SNV **B. LUAD-SNP** **C. LUSC-SNV** **D. LUSC-SNP**

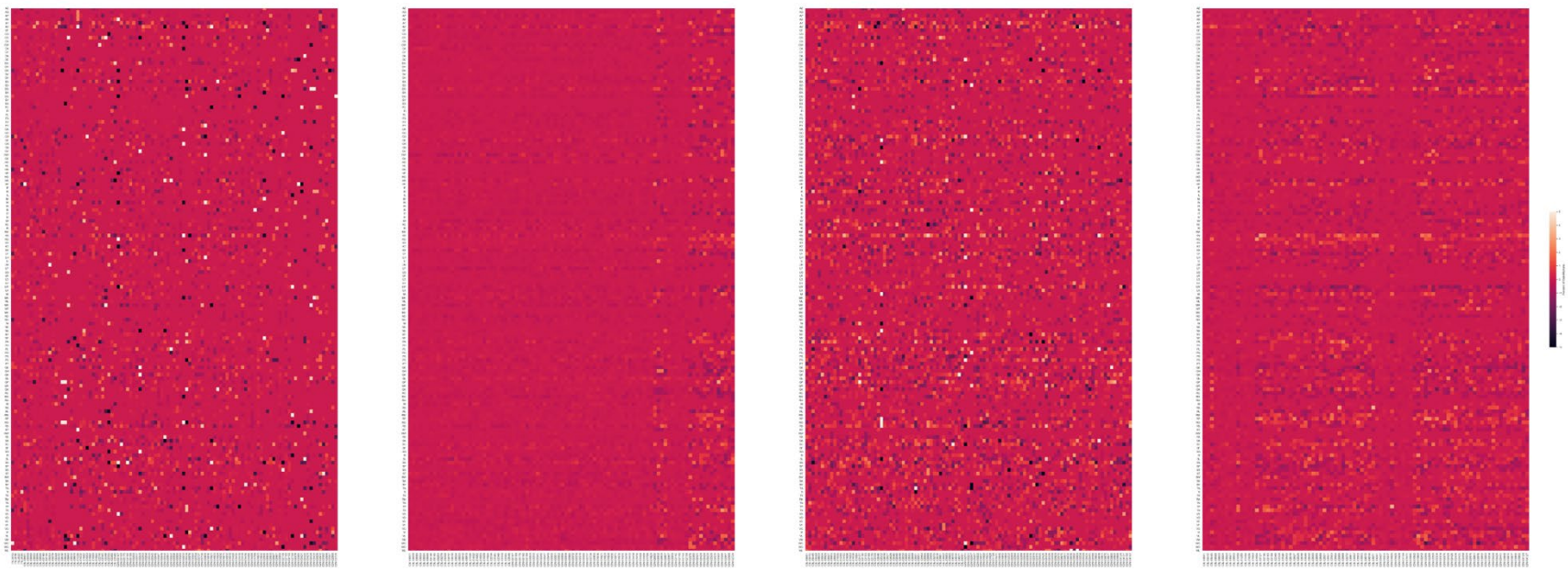


Proportion scale (black-to-white) -5% to 5% with equivalent proportions listed as magenta. Amino acids on y-axis are listed top-to-bottom according to left-to-right PMBEC listing in Table 3.1.

Heatmaps depict cases on the x-axis and amino acid substitutions on the y-axis, lighter colors depict higher proportions measured at the protein level while darker colors represent higher proportions measured at the DNA level. Protein expression based on detection of at least one protein fragment with E value<0.001 with the gene including the SNP or SNV. While there are no clear trends, SNP representations appear overall more similar than SNV representations with cases with higher numbers of SNPs suggesting the most similar comparisons.

Figure 3.11: Transcriptomic vs. proteomic proportion representations of substitutions.

A. LUAD-SNV **B. LUAD-SNP** **C. LUSC-SNV** **D. LUSC-SNP**

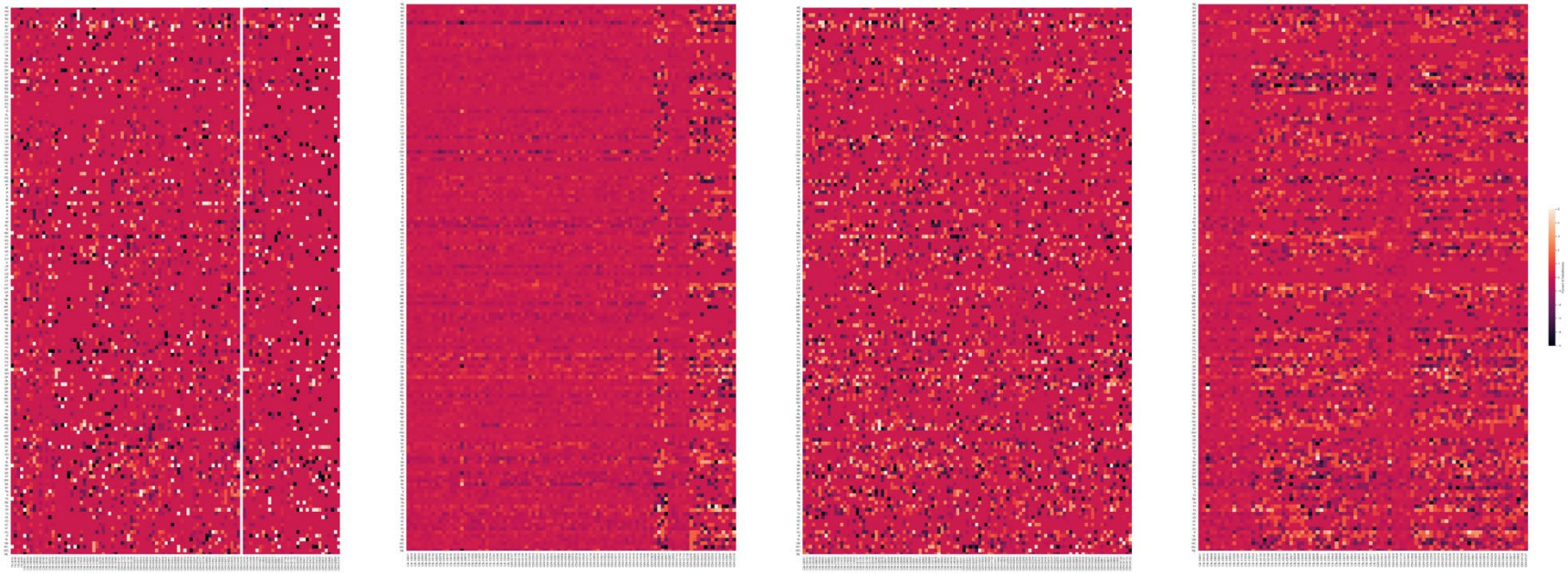


Proportion scale (black-to-white) -5% to 5% with equivalent proportions listed as magenta. Amino acids on y-axis are listed top-to-bottom according to left-to-right PMBEC listing in Table 3.1.

Heatmaps depict cases on the x-axis and amino acid substitutions on the y-axis, lighter colors depict higher proportions measured at the protein level while darker colors represent higher proportions measured based on detectable RNA (>0). Protein expression based on detection of at least one protein fragment with E value<0.001 with the gene including the SNP or SNV. While there are no clear trends, SNP representations appear overall more similar than SNV representations with cases with higher numbers of SNPs suggesting the most similar comparisons.

Figure 3.12: Methylation vs. proteomic proportion representations of substitutions.

A. LUAD-SNV **B. LUAD-SNP** **C. LUSC-SNV** **D. LUSC-SNP**

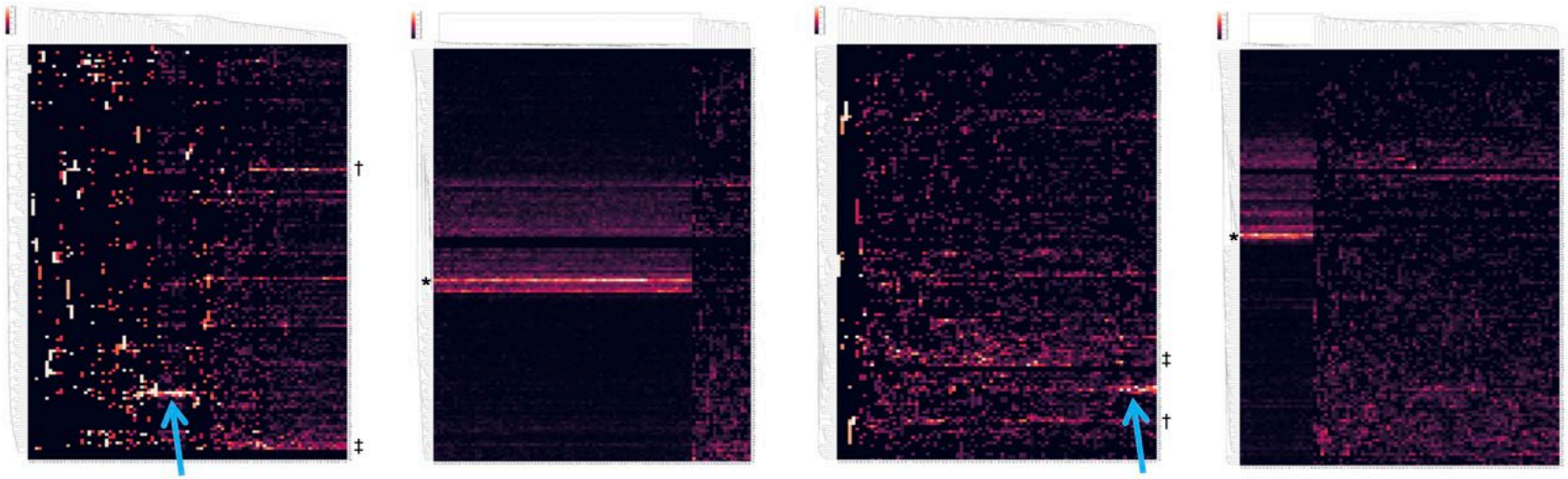


Proportion scale (black-to-white) -5% to 5% with equivalent proportions listed as magenta. Amino acids on y-axis are listed top-to-bottom according to left-to-right PMBEC listing in Table 3.1.

Heatmaps depict cases on the x-axis and amino acid substitutions on the y-axis, lighter colors depict higher proportions measured at the protein level while darker colors represent higher proportions measured based on beta values < 0.4 (note 1.0 corresponds with complete methylation and DNA inaccessibility). One patient had missing tumor methylation (represented by empty space in Panel A). Protein expression based on detection of at least one protein fragment with E value<0.001 with the gene including the SNP or SNV. While there are no clear trends, these representations seem to suggest less similarity than prior comparisons.

Figure 3.13: Hierarchical clustering of substitution proportions with protein expression.

A. LUAD-SNV **B. LUAD-SNP** **C. LUSC-SNV** **D. LUSC-SNP**

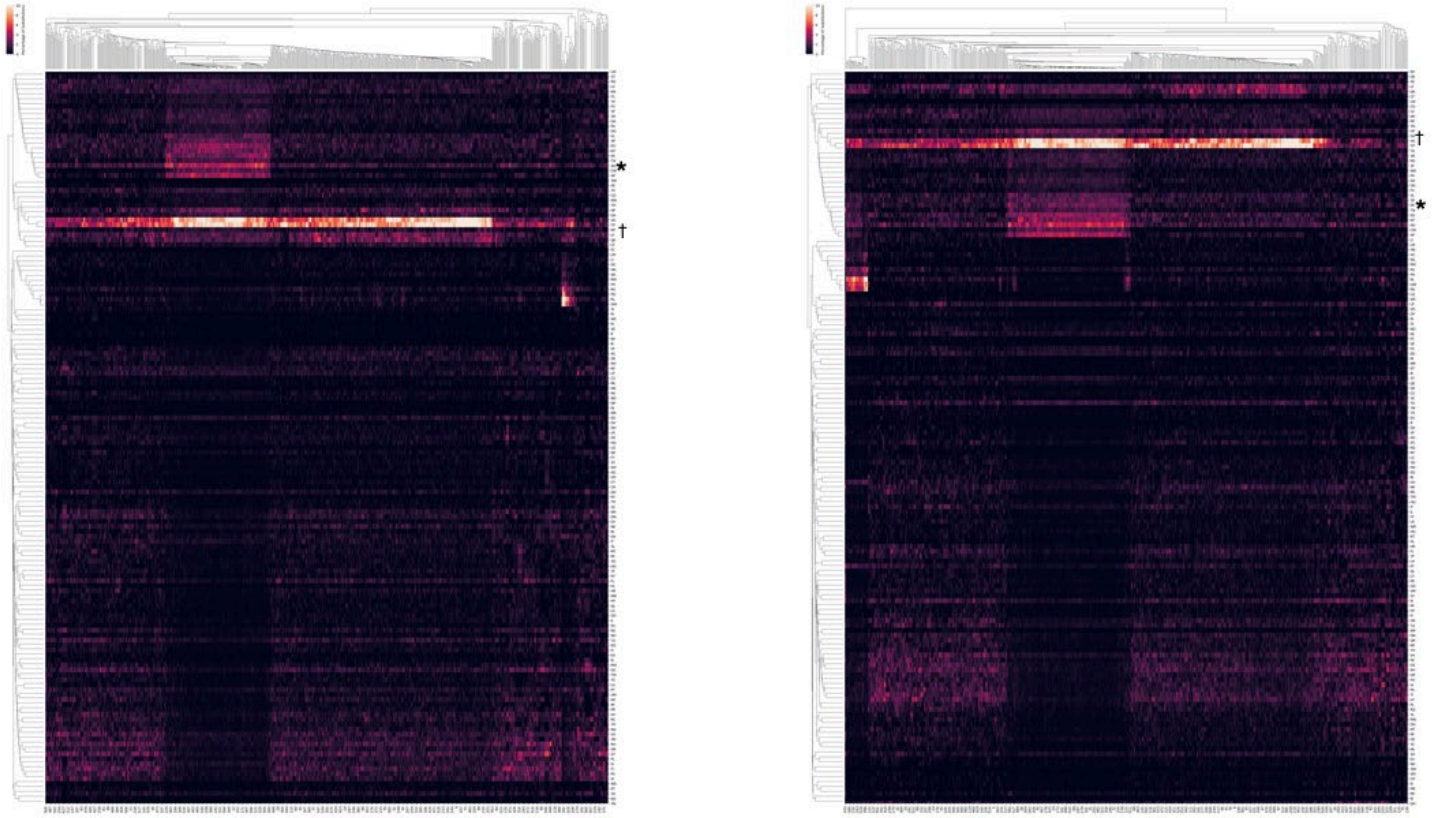


Proportion scale (black to white) 0-10%.

Heatmaps depict cases on the x-axis and amino acid substitutions on the y-axis, lighter colors depict higher proportions while black represents lower proportions. Protein expression based on detection of at least one protein fragment with E value<0.001 with the gene including the SNP or SNV. Hierarchical clustering performed based on cosine distance. Clear subgroups are suggested by both SNP representations with brightest horizontal line (amino acid substitution) p.E>K(*). SNV representations suggest more heterogeneity, brighter horizontal lines include p.G>V (†) and p.R>L (‡) with p.E>K bright in subset of patients (teal arrow). Higher resolution images of these heatmaps may be found in the Appendix (see 7.1).

Figure 3.14: TCGA SNP amino acid substitution proportions by cohort.

LUAD **LUSC**



Proportion scale (black to white) 0-10%. Heatmaps depict cases on the x-axis and amino acid substitutions on the y-axis, lighter colors depict higher proportions while black represents lower proportions. Hierarchical clustering performed based on cosine distance. There is a subset of cases with brighter p.E>K(*), but the brightest lines belong to p.V>G and p.T>P (†), which run across nearly all cases. Higher resolution images of these heatmaps may be found in the Appendix (see 7.2).

Subsequent evaluation revealed that these subgroups had a higher number of SNP events – in LUAD, the high prevalence p.E>K group had a median SNP count of 1014 (SD 372) while the other group had 187 (31), which was highly statistically significant ($p<0.001$). Ancestry of the low SNP group was exclusively Han, and over half of these patients had epidermal growth factor receptor (EGFR) mutations. In LUSC, there was a similar relationship with a SNP count of 921 (SD 322) vs. 241 (232) across subgroups, also highly statistically significant ($p<0.001$), although there were no clear associations based on patient characteristics. These data provide further suggestion of the probabilistic relationship of amino acid substitutions as differences in cohorts appear to be driven by the number of observations (e.g., higher total SNV or SNP counts).

Finally, an ability to discern technical artifact is evaluated (Fig. 3.14). Given the unique SNP relationships in the CPTAC dataset, TCGA LUAD and LUSC SNPs were similarly visualized. Fig. 3.14 shows an issue immediately – the same amino acid substitutions, p.T>P and p.V>G are bright across the whole dataset. The spontaneous deamination of cytosine causes both of these mutations through a g.T>C/A>G mutation, which is a known cause of NGS noise due to an artifact of thermocycling, which has since been limited by pretreatment of specimens with uracil N-glycosylase (UNG).[174] Buckley and colleagues identified similar artifacts in TCGA normal specimens in 2017, although there is no clear documentation on the GDC website about these limitations.[97]

In summary, the techniques and visualizations used in these experiments suggest current amino acid substitution matrices provide limited biologic insight, although are all able to suggest distinct cohort subgroups. Count data reveal epidemiologic patterns dependent on varying mutational burden; normalization based on proportional representations reveals variance suggestive of biologic themes. Hierarchical clustering suggests subgroups that map to measured and provided dataset features. Count data, with respect to codon references, suggest probabilistic

relationships of amino acid substitutions based on the number of observed events, which are influenced by mutational signature and explored further in Chapter 4.

3.5 Conclusion

We introduce concept-driven multiomic data architecture that uses a DNA-based foundation to map multiomic information flow from gene to protein. Our proposed method, which uses comparisons between SNP- and SNV-driven substitution profiles, reveals distinct dependencies that map to biologic correlates as well as identifies technical artifacts and associations, supporting the value of rigorous benchmarking leveraging normal samples to contextualize tumor-specific multiomic relationships. The use of normal sample comparisons may represent a tool to evaluate for unexpected similarities and differences, which could decrease both false positive and negative experiments with broad implications for the field.

CHAPTER 4

Benchmarking Similarity Metrics Using Genomic Data

Reproducibility lends credibility to scientific discovery. In this chapter, we explore triangulation, or the strategic use of multiple approaches to assess an experiment, as methodology for benchmarking and model interrogation. We first assess similarity metrics using varied structures and representations of multiomic data, demonstrating that a combination of metrics with distinct strengths enhances confidence in the evaluation of noisy datasets. We then evaluate two types of modeling: (1) personalized, using case-level features to determine the impact of each multiomic layer on prediction of an observed outcome (i.e., protein expression), and (2) general, using mathematic modeling of cohort-level labels and relationships to evaluate biological inferences suggested by personalized models. Using known and uncovered relationships from Chapter 3, probabilistic, iterated amino acid substitution model experiments reveal that starting amino acid frequency and mutational signature influence protein expression of amino acid substitutions, but are sensitive to event number (i.e., fewer substitution mutations lead to model instability), suggesting an opportunity for additional refinement.

4.1 Introduction

In 2005, the epidemiologist John Ioannidis called attention to what is now known as the "reproducibility crisis." Arguing that there is bias in study design and reporting, low statistical power in most studies, and a small number of true hypotheses, Ioannidis demonstrated that many published research studies are likely to report results that are false.[175] Great lengths since have been taken to improve reproducibility in scientific research (e.g., public availability of datasets, methods, and software, meta-research, independent verification projects, etc.), yet multiomic research remains particularly difficult to reproduce.[27, 114] While a careful assessment of statistical power and effect size, identification of cofounders, biases, and sources of variations, and cross-validation can help guard against misleading results, these techniques may not be enough to

prevent non-informative inferencing, exclusion of tangible interpretations, and overriding of true biologic signals.[114, 176] Embedding techniques in data exploration that allow for early assessment of signal robustness may enable timely identification of misguided interpretations and allow for redirection. In this light, there are two techniques this chapter explores as methodology to evaluate inference reliability in multiomic experiments: (1) triangulation, and (2) mathematical modeling.

Triangulation is built upon the premise that results are less likely to be artifactual when they agree across distinct methodologies with different sources of bias.[123, 177, 178] In stark contrast to multiple hypothesis testing that explores several ideas with the same strategy, this technique uses a number of strategies to test the same idea. Its proponents note a benefit of this technique is that, in addition to strengthening conclusions and causal inference, it allows for an exploration of inconsistencies to uncover unknown sources of bias and identify what further research may be required.[177] Here, both linear and non-linear similarity metrics are explored in context of multiomic data with convergence, or similar patterns of change, suggestive of robust findings, and divergence indicative of an opportunity to contextualize metric performance and data structure inconsistency. Through this lens, iterated models evaluated with these techniques provide a rigorous assessment of inference fidelity.

Extending experiments from Chapter 3, there are three inferences tested in this chapter: (1) biologic phenomena leading to single nucleotide polymorphisms (SNPs) are distinct from those that drive cancer mutagenesis (SNVs), (2) multiomic layers refine what is expressed at the protein level (e.g., increased methylation/decreased transcription affects protein expression), and (3) amino acid substitutions at the protein layer are probabilistic. Robust metric concordance with respect to these commonly accepted biologic themes would enable further classification and exploration, facilitating methodology to enhance supervised learning.

4.2 Related Work

Similarity metrics, repeatedly concordant evaluations, and iterated models that converge on similarity are reviewed in Chapter 2 (see also Table 2.2). Data structure and biologic correlates to data representation are detailed in Chapter 3. Other work pertinent to the experiments in this chapter include the use of models for scientific discovery; approaches to sparse matrices; and biologic domain knowledge regarding transcriptional strand bias, which leads to differences in DNA mutation representations.

Scientific discovery in a model-centric framework. Research that facilitates an exploration of model space and epistemic discovery can speed the discovery of scientific truth.[179] As such, models are loosely described as physical, conceptual, or mathematic representations of real phenomena that are difficult to observe directly.[180] Choice of methodology, complexity of truth, signal strength, and how strongly components influence a phenomenon all impact model fidelity.[179] Although some of these elements are immutable, principles that can help guide model creation and selection focus on an optimization of parsimony and minimization of complexity. To support these features, Bayesian and Akaike information criteria (BIC, AIC, respectively) are commonly referenced.[181, 182] BIC (Eq. 4.1) and AIC (Eq. 4.2) are formally defined as:

$$\mathrm{BIC} = k\ln(n) - 2\ln(\hat{L}) \tag{4.1}$$

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}) \tag{4.2}$$

where $\hat{L}$ is the maximized value of the likelihood function of the model, n is the number of observations of the data, and k is the number of parameters in the model. As BIC is an increasing function of error variance and AIC reflects information loss, minimizing BIC and/or AIC can be accomplished by decreasing k, or the number of parameters in the model. Models selected based on minimized BIC and/or AIC avoid overfitting, which is especially problematic when using smaller datasets, as most multiomic experiments do.

In more recent work, Devezer and colleagues propose that model comparison as opposed to model-driven statistical hypothesis testing is best for scientific discovery as it is generalizable and bypasses complications of hypothesis testing, such as the interpretation of p-values.[179] Arguing that model complexity falsely inflates model confirmation, they recommend linear models as an initial approach for discovery, particularly as they can be incorporated into a variety of designs and statistical analyses. Drawing from information theory, model comparisons involve the creation of first model systems that include one parameter, then systematic combination of these parameters to identify models that more closely resemble observed data.[179, 183] These principles inform the model development in this chapter's experiments, which leverage iterated parsimonious linear models compared with observed data assessed with multiple similarity metrics.

Similarity in sparse matrices. The inclusion of zeros, or "sparse" datasets, pose unique challenges in statistical assessments. Approaches to sparse datasets vary widely across academic disciplines, ranging from alternate data structures for imputation, naïve Bayes classifiers, one-hot encoding, and simple exclusion.[184-188] To identify applicable approaches to these experiments, a consideration of data structure, an understanding of the nature of the "zeros" in the dataset, and the goal of the modeling must be considered. For clarity of purpose, we define amino acid substitutions with evidence of protein expression as a compositional dataset with subsequent modeling reliant on composition classification, focusing on techniques used in mathematical geology.

Compositional datasets function as both a representation of proportions of a whole (and therefore positive and of constant sum) and a reflection of relative magnitude, in that interpretation of change is relative to the composition and not based on absolute quantification. Zeros are thus problematic, given their ambiguity – they can represent meaningfully absent data (essential zeros), data that are below a detection limit (rounded zeros), or data that are missing.[189] Metrics

that assess compositional similarity cannot account for these nuances, and many similarity metrics, such as Euclidean distance, only assess absolute changes. As an example, there are 150 different amino acid substitution categories (not counting those related to stop codons), but cases with a low tumor mutation burden (TMB) may have fewer than 50 observations, leading to $^{2}/_{3}$ or more of the composition being represented by zeros. A model that represents each of these 150 categories as a small proportion (e.g., $^{1}/_{150}$), when compared to observed proportions assessed by Euclidean distance, will have a smaller ("better") Euclidean distance than a model that correctly represents the location of half of the zeros, despite the latter being, arguably, more informative.

These concepts have been considered thoroughly in mathematical geology given the importance of compositional classification studies.[189-191] Defining sample space based on a concept of classes of equivalence, compositional observation can be represented as a ray from the origin into the positive orthant of D-dimensional real space.[189] Any point on the ray can be projected, if the projection is one to one. Experts in the field have demonstrated that replacement of zeros with masks, small values, and imputation procedures shift data structure and can grossly exaggerate similarity between observations.[184, 191, 192] Omitting uninformative components and/or dividing the sample into subsets according to patterns of zeros, however, has demonstrated model improvement in classification schemas.[189] Applications to this chapter's experiments thus include the use of PMBEC as a data structure given its systematic exclusion of stop codons (uninformative categories), treatment of zeros as essential zeros to maintain data structure (maintaining D-dimensional real space), and model stratification based on degree of sparsity (relating to genomic stability).

Transcriptional strand bias. As a refinement to the section on mutational signature in Chapter 3, prior work has shown that there are several signatures that exhibit substantial differences in mutation prevalence between transcribed and untranscribed strands, particularly in lung

cancer.[119] Representing a mutational signature in terms of six categories of complementary mutations may over- and underrepresent some amino acid substitutions. Alexandrov and colleagues suggest this is because the efficiency of DNA damage and DNA maintenance differs between transcribed and untranscribed strands of genes, particularly with respect to transcription-coupled nucleotide excision repair, which operates predominantly on the transcribed strand of genes when it encounters bulky DNA helix-distorting lesions.[119, 193] In smoking-associated lung cancers, g.C>A mutations may be overrepresented related to the propensity of tobacco carcinogens forming adducts on guanine.[194] Similarly, in melanoma, there is a higher prevalence of g.C>T mutations on the untranscribed strand due to ultraviolet light dimerization of thymine, leading to an overrepresentation of g.G>A mutations.[195] In examining mutational signatures based on transcriptional strand bias, a signal without biologic correlate also was identified for g.T>C mutations, particularly in hepatocellular cancer.[119] As these irregularities may distort models that use mutational signature to predict amino acid substitutions, evidence for these perturbations will be sought in included datasets with subsequent models that incorporate these adjustments.

4.3 Method

In this section, we first present an approach for the creation of high and low similarity datasets and similarity metric benchmarking. We then describe an analytical framework for assessing multiomic features in personalized models (Fig. 4.1) and an approach to mathematical modeling of amino acid substitutions leveraging starting amino acid prevalence (Fig. 4.2), and mutational signature with and without transcriptional strand bias (Fig. 4.3).

Preliminary. Notation using “g.” for genomic or DNA and “p.” for protein or amino acid with the starting feature listed first and resultant feature listed second is continued.

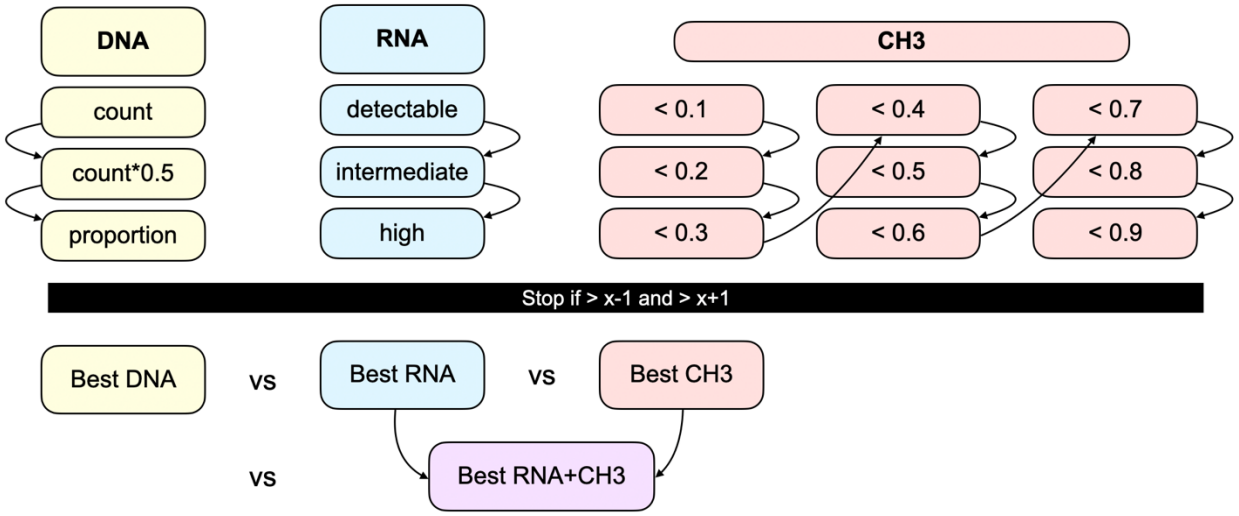
Similarity metric sensitivity to error and noise. In data frame manipulations, error and noise can contribute to flawed results. To assess the sensitivity of linear and non-linear similarity metrics

(see Table 2.2) to these features, two extreme situations are considered: (1) the comparison of highly similar datasets in which similarity metrics should converge, and (2) the comparison of dissimilar datasets in which similarity metrics should not converge. To create these two datasets, CPTAC LUAD and LUSC cohort-level SNP and SNV data structures were assessed for variance. Each dataset was then perturbed by either shifting values by one column (error) or randomly doubling one value per row (noise) and reassessed for variance. Altered datasets that increased variance were then used to compare similarity metric performance at the cohort level, with changes reported as positive values indicating less similarity (expected) and negative values indicating more similarity (artifact) (Table 4.3). Case-level comparisons were assessed manually.

To determine concordance, metric values were transformed based on absolute value and rank. For absolute value labels, Euclidean and city block distance were divided by the highest distance and reversed by subtracting the proportional value from 1, cosine distance and Jaccard distance were similarly reversed by subtracting their values from one. Values closer to zero thereby suggested less similarity and values closer to one greater similarity. Based on receiver operator curves, labels were set as low (<0.5), suggesting observed similarity no better than chance, intermediate (0.5-0.8), and high (>0.8).[196] Numerical ranking was performed using Pandas, the “qcut” function was used to create rank labels sorted into tertiles, quartiles, and quintiles. The concordance of similarity metrics in high and low similarity datasets with noise were assessed with Cohen’s kappa and Kendall’s Tau (Table 4.4). Case-level distribution was assessed manually.

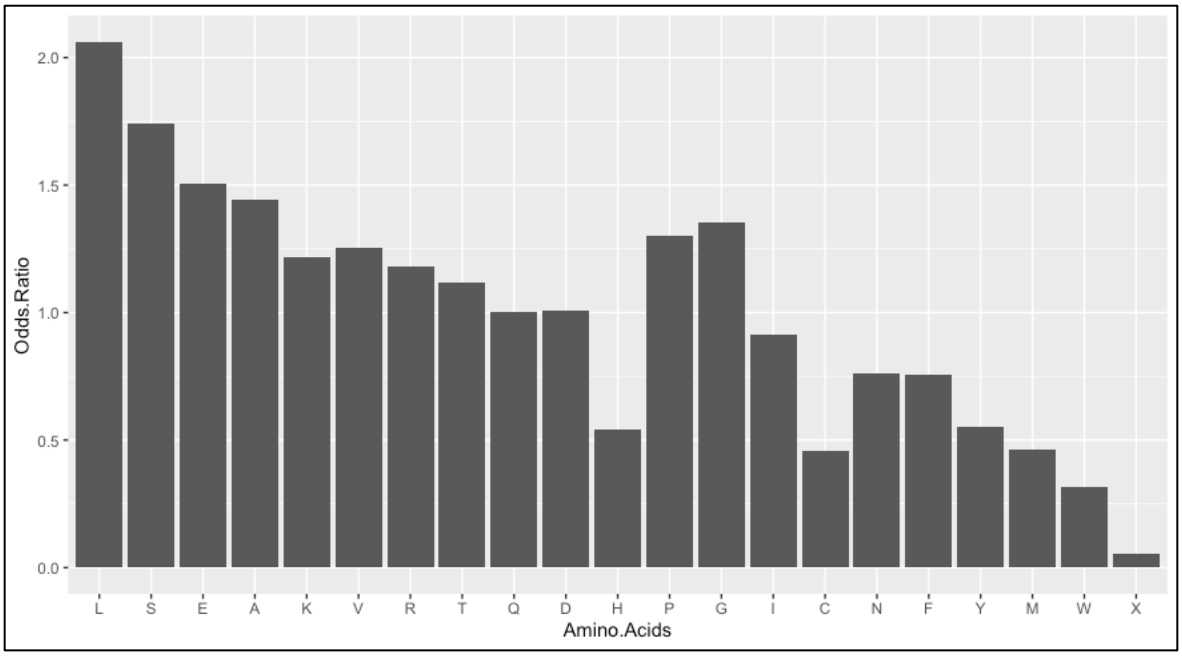
Modeling. Personalized models of amino acid substitution protein expression were created using integrated multiomic data structures including DNA, RNA, and methylation information, and compared to those including proteomic information, leveraging data structures created in Chapter 3. Entries with E-values <0.001 were used as a surrogate for protein expression. Fig. 4.1 depicts model iteration, which sequentially evaluated variables starting with DNA, RNA, and methylation

Figure 4.1: Flow diagram for personalized multiomic model iteration.



CH3 – methylation. Each layer was used to model amino acid substitutions and compared to the protein layer. If the iteration showed enhanced similarity greater than the preceding and subsequent model, iterations ceased. The best model from each layer was then compared. The best RNA and methylation model was then combined and compared to other best models.

Figure 4.2: Amino acid odds ratio based on GRCh38.



Amino acids listed by single letter notation based on frequency in transcriptomic data, odds ratio listed as proportion (compared to expected frequency of 4.76% if all represented equally). X denotes stop codon.

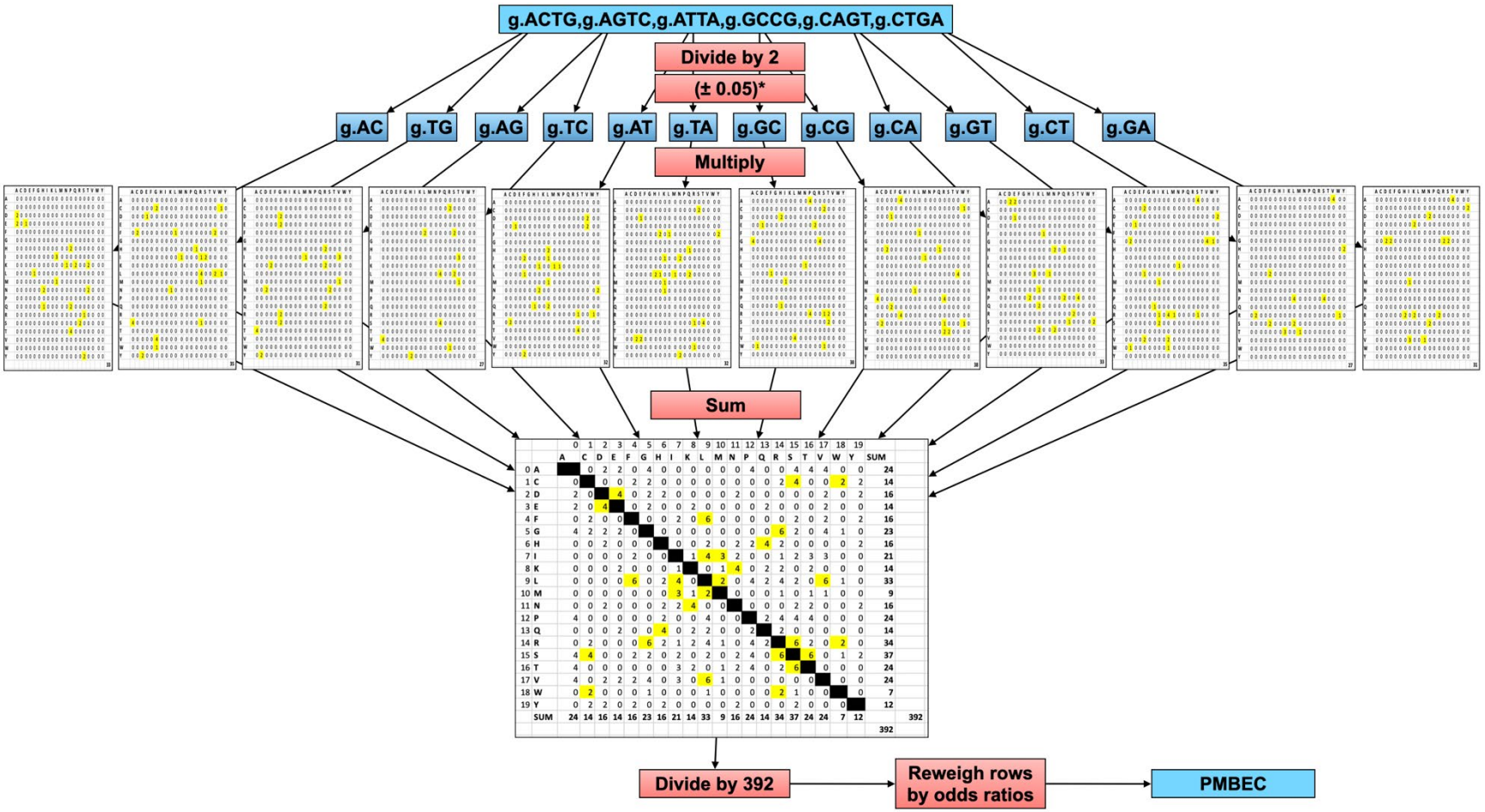
models each considered separately, then combined. RNA models were based on count data and had 3 iterations tested: detectable RNA (RNA>0), moderate expression (RNA>10), and high expression (RNA>100). Methylation models were based on beta-values and iterated with 0.1 intervals up to a level of 0.9 (each interval included methylation values less than the interval as 1.0 corresponds with complete methylation). The best RNA and methylation model were then combined. Models were compared with Pearson's correlation, cosine similarity, and Euclidean distance metrics.

General models were created using probabilistic relationships of amino acid substitutions with a schema summarized in Fig. 4.3. Three components were considered: (1) the number of ways the substitution can be made (Table 3.2), (2) the DNA mutation(s) required to cause the substitution (Table 3.2), and (3) the frequency of the starting amino acid/codon (Fig. 3.2). To create this model, each possible amino acid substitution was listed in matrix format with starting amino acids on the x-axis and resultant amino acids on the y-axis and separated into submatrices corresponding to the DNA mutation leading to the substitution. This process leads to mapping 392 possible codon changes (excluding stop codons) across a 20 x 20 amino acids matrix split across 12 different DNA mutation submatrices. To reweigh each submatrix to account for the DNA mutation proportion, the following equation (Eq. 4.3) can be used:

$$A = \left(\frac{1}{392}\right) \sum \frac{A_x \bullet R_x}{12} = \left(\frac{1}{4704}\right) \sum A_x \bullet R_x \tag{4.3}$$

Where A is a 20 x 20 matrix, x corresponds to each DNA mutation, A_x represents each of the 12 submatrices ($A_{AC} + A_{AG} + \cdots A_{TG}$), and R_x is a vector including DNA mutation proportion weights. Note when six mutational proportion categories are used (e.g., g.C>T/G>A), each is divided by 2 and duplicated to create the R_x vector. To reweigh by frequencies of starting amino acids, each row of the summed matrix can be summed, normalized, and multiplied by the odds ratio of the starting amino acid, as depicted in Fig. 4.3. A matrix map can then be used to retrieve resultant

Figure 4.3: Schema for model creation.



Mutational signature extracted and represented as six proportions as detailed in Methods. Individual proportions extracted by dividing by 2 with *optional adjustment for g.C>A/G>T, g.C>T/G>A, and g.A>G/T>C proportions based on transcriptional strand bias (i.e., g.C>A, g.G>A, g.T>C enriched). Proportions then adjust counts of substitutions that can be created for each DNA mutation (substitution pattern for each mutation highlighted). These are summed into a final matrix, that is divided by 392 and reweighed based on starting amino acid proportions (i.e., by row). Finally, a PMBEC arranged vector is extracted for further analysis.

vectors. As these models are derived from scientific inference as opposed to direct observation, cohort cross-validation techniques are not necessary and the entire cohort is retained for model evaluation.

To assess transcriptional strand bias, measured DNA base mutations were adjusted based on VEP annotation of missense mutations. For every DNA mutation, the associated amino acid substitution was checked for theoretic impossibility based on the submatrices above. If the associated DNA mutation led to a theoretically impossible substitution, the mutation was attributed to the complementary DNA base that could cause the substitution. For example, for p.G>E, the starting and resultant amino acid is comprised of codons including only guanine (g.G) and adenine (g.A) bases – if g.C>T was listed as the DNA mutation for this substitution, it was reattributed to g.G>A. For the few amino acid substitutions that could arise from either complementary base pair (e.g., p.S>R can be caused by g.A>C and g.T>G mutations), the original reported DNA mutation was maintained. For transcriptional strand bias effects that were statistically significant across cohorts, models that incorporated transcriptional strand bias conservatively adjusted impacted proportions by adding or subtracting 0.05.

The labels introduced to create general models are based on mutational signature patterns identified in Chapter 3 (Fig. 3.7) and listed in Table 4.1. The smoking label enriches for g.C>A/G>T, aging for g.C>T/G>A, and APOBEC for g.G>C/C>G. The resultant amino acid

Table 4.1: Mutational signature labels.

Mutation	AGING	APOBEC	SMOKING	UNSPEC-H	UNSPEC-L
g.C>T/G>A	0.400	0.400	0.200	0.400	0.167
g.C>A/G>T	0.200	0.050	0.400	0.300	0.167
g.G>C/C>G	0.100	0.400	0.150	0.125	0.167
g.A>T/T>A	0.150	0.050	0.150	0.125	0.167
g.A>G/T>C	0.100	0.050	0.080	0.030	0.167
g.A>C/T>G	0.050	0.050	0.020	0.020	0.167

DNA mutations with standard notation listed from most common to least common.[119]
UNSEPC-H/L – unspecified high/low.

vectors for these labels zero the 50 lowest proportions based on observations from Chapter 3. Cases that did not fit one of these labels are defined as unspecified, with unspecified-high reflective of a signature with more balanced g.C>T/G>A and g.C>A/G>T mutations, and unspecified-low with equal proportions of all mutations based on cases with few mutations (suggestive of genomic stability). Unspecified-low signatures also zero the 100 lowest proportions as a method of stratification. Label-derived models are compared at the cohort level with similarity metrics and at the case-level manually.

Outcome representation. Histograms and summary statistics were created in R v4.1 (https://www.r-project.org). Amino acid substitutions were extracted based on count and proportions in PMBEC organization using Python v4.0 Pandas and NumPy (see Methods, Chapter 3). Mutational signature was extracted using MuTect2 annotations ('REF' and 'ALT') and corrected based on transcriptional strand bias per above. Similarity metric convergence was measured based on amino acid substitution proportions represented by number, absolute number category, rank, and ranked tertiles, quartiles, and quintiles. For general model rank assessments, amino acid substitution proportions were ranked with zeros retained, the highest 10% of substitutions ranked as 2, and any other detectable substitution ranked as 1.

Statistical analysis. Variance was evaluated with Pandas. SciPy (https://scipy.org) was used to calculate distance metrics including cosine distance, Euclidean distance, city block distance, Jaccard distance, and Pearson's and Spearman's correlations. Scikit-learn (https://scikit-learn.org) was used to calculate mutual information, Cohen's kappa, and Kendall's Tau. Excellent concordance was set at 0.6, adequate at 0.2, and suggestive at 0.1.[197-200] Transcriptional strand bias was evaluated with two-proportional z-tests for proportions using R v4.1. Two-tailed P values of <0.05 were considered statistically significant and rounded to the nearest thousandth.

4.4 Experiments

Using data structures established in Chapter 3, contrived high and low similarity amino acid substitution representations are assessed for variance and then compared with linear and non-linear similarity metrics using count and proportion data (Tables 4.2 and 4.3). Similarity metric triangulation is compared based on high and low similarity comparisons with noise (Table 4.4). Multiomic models are assessed with respect to two methodologies: (1) comparisons within an individual for the purpose of showing meaningful change (Table 4.5), and (2) comparisons across a cohort for subgroup discovery (Table 4.7).

Datasets. CPTAC lung cancer datasets LUAD and LUSC were used.[81, 117, 118]

4.5 Results

Table 4.2 shows the variance in LUAD and LUSC SNP and SNV files with variance changes based on error and noise. Interestingly, both LUAD SNP and SNV files representing count data showed decreased variance with error, while LUSC SNP proportion data and SNV count data showed decreased variance with noise, suggesting the introduction of bias into the datasets. High similarity comparisons used unbiased count and proportion files with more similar variances, including LUSC-SNP count data and LUAD-SNV proportion data, comparing the measured file

Table 4.2: Variance across observed and simulated multiomic data with error and noise.

	LUAD		LUSC	
File type	**Count**	**Proportion**	**Count**	**Proportion**
SNP	**4.482**	**1.765**	**4.215**	**1.929**
Error	-0.305	0.000	+1.522	0.000
Noise	0.000	+0.029	0.000	-1.123
SNV	**3.272**	**2.559**	**2.934**	**2.168**
Error	-0.285	0.000	+1.270	+0.387
Noise	+0.109	+0.063	-2.011	0.000

Variance is represented as the log sum of each amino acid substitution variance, error and noise are reported as the change in log sum. See Methods for creation of error and noise files. Files that increased variance are highlighted in green with those that decreased variance in orange.

against its corresponding perturbed file, while low similarity comparisons used higher variance files (LUSC-SNP/SNV count and proportion data) compared across categories to result in eight comparisons per similarity metric (four high similarity and four low similarity, each assessing count-error, proportion-error, count-noise, and proportion-noise). Results are summarized in Table 4.3.

Table 4.3: Metric performance in observed and simulated multiomic data.

		Error		Noise	
Feature	**File**	**Count**	**Proportion**	**Count**	**Proportion**
Cosine similarity	high*	+0.248 (0.074)	+0.703 (0.095)	+0.003 (0.005)	+0.007 (0.012)
	low†	-0.049 (0.040)	-0.087 (0.087)	-0.004 (0.011)	-0.007 (0.018)
		0.249 (0.102)	0.461 (0.091)		
Euclidean distance	high	+79.711 (103.52)	+18.507 (6.571)	+2.806 (1.893)	+1.377 (1.789)
	low	-2.000 (2.473)	-1.132 (1.209)	-0.135 (0.524)	-0.350 (1.100)
		90.474 (110.326)	15.429 (4.218)		
City block distance	high	+609.56 (731.221)	+141.06 (17.771)	+2.953 (2.190)	+1.377 (1.789)
	low	-29.535 (26.808)	-12.346 (10.979)	-1.605 (3.115)	-1.143 (1.908)
		515.047 (502.443)	120.903 (17.448)		
Jaccard index	high	+0.910 (0.035)	+0.996 (0.020)	+0.020 (0.019)	+0.997 (0.026)
	low	-0.011 (0.032)	-0.003 (0.017)	-0.001 (0.003)	-0.003 (0.017)
		0.075 (0.034)	0.03 (0.017)		
Mutual information	high	+0.972 (0.022)	+0.984 (0.013)	+0.020 (0.010)	+0.007 (0.004)
	low	+0.039 (0.040)	+0.032 (0.041)	+0.000 (0.005)	+0.001 (0.004)
		0.050 (0.029)	0.037 (0.031)		
Pearson correlation	high	+0.315 (0.095)	+0.939 (0.045)	+0.004 (0.006)	0.010 (0.015)
	low	+0.061 (0.050)	+0.133 (0.130)	+0.005 (0.013)	+0.005 (0.014)
		0.696‡ (0.117)	0.194 (0.102)		
Spearman correlation	high	+0.857 (0.075)	+0.935 (0.045)	+0.001 (0.001)	+0.001 (0.001)
	low	+0.193 (0.127)	+0.164 (0.134)	+0.000 (0.004)	+0.001 (0.004)
		0.227 (0.089)	0.209 (0.092)		

*high = high similarity comparison, †low = low similarity comparison (see Methods). ‡Unexpectedly high.

Positive values indicate decreased similarity (expected), negative values indicate increased similarity (artifact). Average change listed with standard deviation in parentheses, negative values (orange) reflect erroneously increased similarity. For low similarity, count and proportion assessments are provided with proportions in the lower row [high similarity file had perfect similarity, i.e., 1.000 (0.000)]. **Note some signs reversed for to maintain directionality of assessments to facilitate interpretation.**

Benchmarking experiments show distance metrics incorporating absolute value and size effects (e.g., Euclidean and city block distance) are the most sensitive to error in comparisons of more similar data, but lose this sensitivity in analysis of low similarity data. Tested distance metrics (city block, cosine, Euclidean, Jaccard) are additionally more sensitive to noise in high similarity data – but alarmingly, the introduction of noise to low similarity data leads to increased artifactual similarity within these datasets. Non-distance similarity metrics (mutual information, Pearson's correlation, Spearman's correlation) show greater sensitivity to error than noise and do not inappropriately increase similarity metrics in low similarity datasets but are also less sensitive to count data. In terms of absolute metric value in low similarity data, we see that proportional data is more similar than count data with distance metrics, but not non-distance metrics. Pearson's correlation for low similarity count data is particularly high at 0.696, suggesting a potential interaction between SNP and SNV count observations.

Table 4.4 demonstrates metric concordance based on high and low similarity noise comparisons with the expectation that higher similarity files should demonstrate more concordance than low similarity comparisons. Concordance was assessed in three ways: based on measured values and absolute number categories measured by Cohen's kappa, and based on rank, which was subset as number rank, tertiles, quartiles, and quintiles, and measured by Kendall's Tau. Measured values and absolute number categories did not demonstrate any concordance (all Cohen's kappa values <0.1) and are not depicted. In the high similarity dataset, ranked quartiles demonstrated the most concordance with two exceptions (quintiles had better concordance in count data compared with Pearson's correlation and cosine distance, and tertiles performed better in count data assessed with Euclidean and city block distance).

As expected, the high similarity dataset showed more concordance across metrics than the low similarity dataset. The best high similarity concordance was noted in distance measures sensitive to size (Euclidean and city block distance), but these metrics also showed a trend

Table 4.4: Benchmarking of similarity metric robustness with noise.

		Co			E			Ci			J			M			P			S		
		r	rl_4	rl_3	r	rl_4	rl_3	r	rl_4	rl_3	r	rl_4	rl_3	r	rl_4	rl_3	r	rl_4	rl_3	r	rl_4	rl_3
Co	c				-0.04	0.02	-0.02	0.12*	0.05	0.09	0.06	-0.05	0.08	0.06	-0.01	0.11	0.01	0.09	0.14	0.03	0.09	0.08
	p				0.12*	0.02	0.09	0.09	0.07	0.11*	-0.22	-0.05	-0.11	0.00	-0.02	-0.06	0.02	0.05	-0.04	0.12	0.15	0.13*
E	c	0.12	0.12	0.11				0.07	0.16*	0.13†	0.00	-0.05	0.04	0.02	0.05	0.07	-0.01	-0.04	-0.01	0.02	0.18*	0.06
	p	0.15	0.33	0.18				0.14	0.18	0.02	-0.10	0.01	-0.06	-0.03	-0.01	-0.06	-0.01	-0.12	-0.11	0.03	0.01	-0.05
Ci	c	0.02	0.04	0.04	0.61	0.58	0.70				0.08	0.04	0.00	-0.08	0.01	-0.11	-0.00	-0.02	-0.06	0.00	0.05	0.01
	p	0.14	0.30	0.15	0.96	0.97	0.96				0.08	-0.04	0.00	0.05	0.16*	0.09	0.03	-0.02	-0.01	-0.04	-0.02	-0.13
J	c	0.05	0.12	0.20	0.19	0.10	0.09	0.15	0.18	0.00				0.07	0.01	0.11*	0.05	0.05	-0.01	0.05	0.09	0.01
	p	0.01	0.01	0.01	0.02	0.07	-0.01	0.01	0.07	-0.00				0.04	-0.01	0.08	0.09	0.09	-0.03	-0.06	-0.07	-0.10
M	c	-0.00	-0.01	-0.05	0.06	0.13	0.11	0.09	0.07	0.08	0.12	0.09	0.05				0.06	0.01	0.11	0.04	0.03	0.04
	p	0.17	0.24	0.27	-0.00	0.15	0.01	0.00	0.16	0.00	0.04	0.12	0.07				-0.02	0.02	0.04	0.04	0.10*	0.09
P	c	0.33	0.38	0.39†	0.03	0.01	-0.06	0.09	-0.02	-0.03	0.02	0.01	0.04	-0.05	0.02	-0.6				0.09	0.07	0.22*
	p	0.47	0.52	0.44	0.14	0.29	0.20	0.14	0.26	0.20	0.01	-0.02	-0.08	0.14	0.19	0.06				-0.05	0.02	-0.04
S	c	-0.04	0.02	-0.01	0.06	0.07	-0.01	0.07	0.07	0.02	-0.02	-0.05	0.08	0.09	0.12	0.15†	0.02	0.07	0.08			
	p	-0.01	0.05	-0.05	0.10	0.05	0.13	0.09	0.04	0.09	0.01	-0.04	-0.10	-0.00	0.01	-0.08	0.06	0.10	0.02			

C – count, Co – cosine distance, E – Euclidean distance, Ci – city block distance, J – Jaccard distance, M – mutual information, P – Pearson correlation, p – proportion, r – rank, r-l – rank-label quartiles, S – Spearman correlation. *Concordance improved w/ noise, †rank-quintile with best performance. Bottom/left high similarity, top/right less. Rank assessed by Kendell's Tau, rank-label assessed by Cohen's kappa. Excellent concordance 0.6 (blue), adequate 0.2 (green), suggestive 0.1 (orange). Absolute value and labels not pictured given poor concordance.

towards concordance based on artifactual similarity in the low similarity dataset. Measures with the best concordance in the high similarity dataset and lowest concordance in the low similarity dataset were cosine distance and Pearson's correlation, particularly in proportional data. The worst performance was noted in Pearson's and Spearman's correlations, which showed little concordance in the high similarity data and artifactual concordance in ranked tertiles. There was adequate performance (more true positives than false positives) with cosine and Euclidean distance, mutual information and cosine distance in proportional data, Pearson's correlation and Euclidean distance in proportional data, and Pearson's correlation and city block distance in proportional data. Based on these experiments, model performance was assessed based on proportional assessments with cosine distance, Pearson's correlation, and Euclidean distance.

To assess the value of multiomic data in amino acid substitution protein expression prediction, personalized models of amino acid substitution proportions were created based on different layers/features of the multiomic data structure and assessed based on similarity to proportions of amino acid substitutions detected at the protein layer. Results are depicted in Table 4.5 and separated based on LUAD and LUSC cohorts and SNP and SNV measurements. Overall, a high degree of similarity (correlation/cosine similarity >0.900 across all groups) was noted based on DNA-level extraction, suggesting a probabilistic nature of amino acid substitution protein expression (i.e., amino acid substitutions proportions at the protein level are driven by proportions at the DNA level). Multiomic effects were also demonstrated with comparable performance in LUAD based on detectable RNA, although moderate RNA expression (>10 count) had better performance in LUSC cohorts. The impact of methylation was suggested with improved performance up to beta values of 0.9, suggesting protein impact only with high levels of gene methylation. The best model performance included the combination of the best RNA model with the best methylation model, suggesting subtle and distinct effects of each, except for LUSC SNV, which had the best Euclidean distance in the combined model but a higher correlation and cosine

Table 4.5: Relation of varied multiomic features to protein expression.

	LUAD						LUSC					
	SNP			SNV			SNP			SNV		
	Corr	Cosine	Euc	Corr	Cosine	Euc	Corr	Cosine	Euc	Corr	Cosine	Euc
DNA-only (count)	0.997 (0.019)	0.997 (0.016)	163.713 (95.412)	0.960 (0.041)	0.965 (0.039)	12.042 (24.416)	0.962 (0.027)	0.969 (0.021)	29.762 (67.987)	0.914 (0.049)	0.926 (0.042)	32.187 (17.068)
DNA*0.5 (count)	0.997 (0.019)	0.997 (0.016)	23.775 (12.519)	0.960 (0.041)	0.965 (0.039)	3.000 (5.259)	0.962 (0.027)	0.969 (0.021)	7.591 (7.331)	**0.914 (0.049)**	**0.926 (0.042)**	**11.225 (5.514)**
DNA-only (prop)	0.993 (0.038)	0.995 (0.028)	1.676 (1.770)	0.931 (0.044)	0.941 (0.042)	8.033 (6.165)	0.905 (0.057)	0.933 (0.039)	5.296 (2.047)	0.806 (0.064)	0.853 (0.054)	9.196 (4.638)
RNA > 0	0.994 (0.032)	0.995 (0.023)	1.582 (1.578)	0.935 (0.041)	0.946 (0.039)	7.339 (6.094)	0.821 (0.063)	0.864 (0.053)	8.921 (4.227)	0.820* (0.063)	0.864* (0.052)	8.920* (4.211)
RNA >10	0.994 (0.044)	0.995 (0.032)	1.676 (1.902)	0.934 (0.045)	0.946 (0.043)	7.523 (6.047)	0.898 (0.064)	0.931 (0.046)	5.528 (2.075)	0.830* (0.059)	0.874* (0.051)	8.567* (3.849)
RNA > 100	0.992 (0.057)	0.994 (0.043)	1.886 (2.354)	0.918* (0.061)	0.926* (0.057)	8.416* (6.486)	0.881 (0.070)	0.918 (0.048)	5.931 (2.184)	0.828* (0.073)	0.866* (0.066)	8.992* (4.641)
B < 0.1	0.635 (0.237)	0.671 (0.223)	20.835 (19.009)	0.224 (0.191)	0.270 (0.179)	61.871 (24.328)	0.240 (0.232)	0.309 (0.219)	37.384 (19.245)	0.160 (0.134)	0.201 (0.133)	62.758 (23.868)
B < 0.2	0.856 (0.253)	0.880 (0.233)	9.617 (12.781)	0.430 (0.194)	0.489 (0.195)	38.333 (27.181)	0.386 (0.235)	0.468 (0.211)	24.881 (10.153)	0.371 (0.150)	0.437 (0.149)	30.671 (19.608)
B < 0.3	0.930 (0.195)	0.945 (0.173)	6.240 (7.962)	0.570 (0.194)	0.599 (0.197)	27.441 (20.777)	0.554 (0.195)	0.642 (0.164)	16.580 (6.234)	0.504 (0.126)	0.578 (0.124)	20.330 (14.025)
B < 0.4	0.962 (0.168)	0.969 (0.138)	4.495 (5.236)	0.680 (0.167)	0.711 (0.167)	16.583 (14.071)	0.620 (0.178)	0.711 (0.140)	12.731 (4.755)	0.617 (0.120)	0.694 (0.113)	15.446 (10.485)
B < 0.5	0.974 (0.140)	0.979 (0.173)	3.607 (4.264)	0.772 (0.113)	0.802 (0.110)	14.756 (9.600)	0.668 (0.153)	0.760 (0.114)	11.272 (3.836)	0.687 (0.108)	0.759 (0.098)	12.235 (6.808)
B < 0.6	0.982 (0.127)	0.986 (0.098)	2.907 (3.855)	0.837 (0.097)	0.868 (0.093)	11.227 (8.110)	0.702 (0.149)	0.800 (0.104)	9.538 (2.971)	0.751 (0.095)	0.812 (0.082)	10.548 (5.013)
B < 0.7	0.988 (0.098)	0.991 (0.073)	2.385 (3.056)	0.893 (0.066)	0.913 (0.061)	9.700 (6.261)	0.808 (0.098)	0.867 (0.068)	7.650 (2.609)	0.777 (0.080)	0.830 (0.069)	9.758 (4.723)
B < 0.8	0.990 (0.070)	0.993 (0.051)	2.099 (2.438)	0.916 (0.052)	0.931 (0.048)	8.664 (6.191)	0.862 (0.075)	0.906 (0.051)	6.403 (2.193)	0.802 (0.067)	0.850 (0.057)	9.336 (4.298)
B < 0.9	0.993 (0.029)	0.995 (0.029)	1.743 (1.753)	0.928 (0.044)	0.939 (0.042)	8.028 (6.107)	0.895 (0.056)	0.929 (0.038)	5.343 (1.955)	0.812 (0.066)	0.859 (0.056)	9.129 (4.556)
Max-RNA+B	**0.999 (0.008)**	**0.999 (0.006)**	**0.763 (0.853)**	**0.995 (0.015)**	**0.996 (0.014)**	**1.339 (2.847)**	**0.989 (0.013)**	**0.992 (0.010)**	**1.829 (0.988)**	**0.838* (0.059)**	**0.875* (0.051)**	**8.547* (3.685)**

B – beta value, Corr – Pearson's correlation, Euc – Euclidean distance. *One case removed from these analyses given lack of RNA data. Median score with standard deviation in parentheses. Note Cosine refers to cosine similarity (1 – cosine distance).

Table 4.6: CPTAC transcriptional strand bias by cohort.

	LUAD				LUSC			
Mutation	**SNP**	**P**	**SNV**	**P**	**SNP**	**P**	**SNV**	**P**
g.C>T/G>A	-0.096 (0.040)	<0.001	-0.067 (0.099)	0.004	-0.119 (0.035)	<0.001	-0.087 (0.067)	<0.001
g.C>A/G>T	0.089 (0.049)	<0.001	-0.104 (0.104)	<0.001	-0.019 (0.053)	0.443	-0.089 (0.035)	<0.001
g.G>C/C>G	-0.004 (0.012)	0.852	0.000 (0.056)	1.000	-0.020 (0.053)	0.441	-0.016 (0.046)	0.449
g.A>T/T>A	-0.061 (0.030)	0.009	0.016 (0.048)	0.485	-0.014 (0.026)	0.454	0.021 (0.035)	0.441
g.A>G/T>C	0.015 (0.034)	0.514	0.046 (0.048)	0.050*	0.059 (0.034)	0.014	0.067 (0.046)	0.005
g.A>C/T>G	0.007 (0.011)	0.780	0.000 (0.056)	1.000	0.020 (0.019)	0.441	0.000 (0.031)	1.000

DNA mutations with standard notation listed from most common to least common.[119] Median changes are listed with reference to the first mutation (e.g., for g.C>T/G>A, negative values indicate that g.G>A leads to more amino acid substitutions than g.C>T), standard deviations are provided in parentheses. P values are provided with those <0.05 highlighted in green.

*Based on the statistical methods, this value was not considered significant although is suggestive.

similarity in the DNA count model*0.5, although the inability to include a sample may have impacted model performance.

In preparation for general models, evidence for transcriptional strand bias was sought with results summarized in Table 4.6. Strong evidence was found for transcriptional strand bias in g.C>T/G>A mutations with 6.7-11.9% of g.C>T mutations reattributed to g.G>A mutations, which was highly statistically significant across all groups. g.C>A/G>T mutations also suggested bias, although with an inconsistent signal: g.G>T mutations were favored in SNV assessments and g.C>A mutations favored in LUAD SNP (LUSC SNP assessments were not statistically significant). Evidence for g.T>C bias was also found with g.A>G mutations statistically significant in LUSC SNP and SNV groups and close to significance in LUAD SNV with 4.6-6.7% of mutations reattributed. These findings recapitulate what was found by Alexandrov and colleagues, increasing confidence in their results and techniques used in these experiments.

Results from general model assessments are shown in Table 4.7. Starting with personalized models, mutational signature was extracted from each case and then used to create a model of amino acid substitution proportions. Pearson's correlation and cosine similarity improved as starting amino acid proportions and transcriptional bias were included across all groupings except LUSC SNP, although the absolute metric values were overall suboptimal with cosine similarity ranging from 0.480-0.566 in SNV groups and 0.630-0.726 in SNP groups (SNV results were comparable to the range of cosine similarity in the low similarity proportional dataset). As general models progressed through iterations starting with equal probability of DNA mutations, then corrected for starting amino acids, with refinements to mutational signature and then the addition of transcriptional strand bias, there is a trend toward improvement across groupings based on cosine similarity, although Euclidean distance does not show consistent directionality. Interestingly, labels for mutational signature (see Table 4.1) show higher cosine similarity than personalized models in LUAD SNV and LUSC SNP groupings, suggesting improvement based

Table 4.7: Comparison of amino acid substitution models.

	LUAD						LUSC					
	SNP			SNV			SNP			SNV		
	Corr	Cosine	Euc	Corr	Cosine	Euc	Corr	Cosine	Euc	Corr	Cosine	Euc
A	0.491 (0.091)	0.653 (0.045)	12.431 (1.023)	0.251 (0.168)	0.420 (0.167)	20.931 (0.168)	0.350 (0.099)	0.630 (0.050)	11.489 (1.270)	0.253 (0.111)	0.511 (0.103)	15.079 (6.342)
B	0.581 (0.133)	0.703 (0.064)	16.243 (0.899)	0.273 (0.160)	0.441 (0.158)	22.810 (10.181)	0.319 (0.141)	0.619 (0.065)	14.792 (1.156)	0.286 (0.114)	0.523 (0.106)	17.342 (5.867)
C	0.619 (0.133)	0.726 (0.067)	16.237 (0.898)	0.326 (0.166)	0.480 (0.173)	22.808 (10.182)	0.344 (0.140)	0.627 (0.068)	14.788 (1.155)	0.355 (0.105)	0.566 (0.104)	17.336 (5.866)
1	*	0.515 (0.043)	13.974 (1.162)	*	0.378 (0.139)	21.134 (10.786)	*	0.595 (0.061)	11.891 (1.573)	*	0.495 (0.090)	15.062 (6.234)
2	*	0.520 (0.037)	13.979 (1.094)	*	0.385 (0.150)	21.157 (10.891)	*	0.583 (0.055)	12.149 (1.499)	*	0.515 (0.097)	15.067 (6.294)
3	*	0.614 (0.056)	13.028 (1.043)	*	0.388 (0.178)	21.256 (11.148)	*	0.601 (0.050)	11.780 (1.381)	*	0.536 (0.106)	14.697 (6.375)
4	*	0.651 (0.074)	16.249 (0.901)	*	0.401 (0.168)	22.820 (10.181)	*	0.610 (0.054)	14.795 (1.159)	*	0.508 (0.097)	17.341 (5.866)
5	*	0.608 (0.034)	13.129 (1.027)	*	0.567 (0.182)	22.009 (10.131)	*	0.634 (0.047)	11.538 (1.440)	*	0.539 (0.115)	17.344 (5.994)
6	0.470 (0.109)	0.781 (0.120)	7.874 (0.967)	0.214 (0.155)	0.430 (0.191)	10.536 (1.262)	0.361 (0.102)	0.572 (0.119)	9.434 (0.990)	0.316 (0.115)	0.552 (0.110)	8.916 (1.418)

Reported as median (standard deviation). Models A-C are personalized and created with case-specific information. Models 1-6 are generalized. Best score is highlighted in green, the lowest score in orange.*Correlations not included given ambiguous arrays (unstable models).

Model A uses each patient's measured mutational signature (unmanipulated counts of amino acids).
Model B uses each patient's measured mutational signature and starting proportions of amino acids.
Model C uses each patient's measured mutational signature corrected for transcriptional strand bias and starting proportions of amino acids.

Model 1 includes counts of amino acids with every DNA mutation equally likely.
Model 2 includes counts of amino acids with every DNA mutation equally likely and less probable events set to zero.
Model 3 uses starting proportions of amino acids and corrects for average mutational signature (see Methods).
Model 4 uses starting proportions of amino acids and corrects for average mutational signature and transcriptional strand bias (see Methods).
Model 5 uses starting proportions of amino acids and labels to estimate mutational signatures (see Methods).
Model 6 uses starting proportions of amino acids and labels to estimate mutational signatures assessed by rank (see Methods).

on stratification of zeros. The use of rank labels shows the best performance out of any general model in all groupings except LUSC SNP with cosine similarity ranging from 0.552-0.781, although offer the possibility for additional refinement. Manual review suggests that the worst model performance involved cases with few mutations, which were typically labeled as unspecified-low (cosine similarity 0.200-0.300), and best model performance with the smoking label (cosine similarity 0.700-0.800). As LUSC SNP had the most cases labeled with unspecified-low, this may explain this group's poorer performance, demonstrating model sensitivity to number of mutation/substitution events and a condition in which interference failed.

4.6 Conclusion

In these experiments, we suggest the value of triangulation as methodology for benchmarking and model interrogation as a strategy to assess inference. Incorporating multiple statistical assessments with distinct biases can increase confidence in the identification of meaningful change and signal strength, particularly in noisy datasets. As distance measures alone may be misleading, we recommend the inclusion of two distance similarity metrics, one sensitive to scale and one not, with a non-distance metric for enhanced signal discrimination. We additionally demonstrate that iterative modeling can reveal flawed inference and offer the use of labels and stratification based on patterns of zeros as techniques to improve general multiomic models. We conclude that iterative modeling may be considered as a discovery technique in multiomic assessments, particularly when power calculations cannot be performed and/or signal strength is unknown *a priori*.

CHAPTER 5

Features of Cancer-Immune System Interactions

This chapter is adapted from the paper, "Mutational landscape influences immunotherapy outcomes among patients with non-small-cell lung cancer with human leukocyte antigen supertype B44" published in Nature Cancer in 2020.[89]

Identifying biomarkers of response to immune checkpoint blockade (ICB) remains a priority in clinical oncology. Human leukocyte antigen (HLA)-B has been recognized as a major determinant of discrepancies in disease outcomes yet exhibits inconsistent associations with ICB treatment. Here we use approaches developed in Chapters 3 and 4 to show that the B44 supertype, which features an electropositive binding pocket that preferentially displays peptides with negatively charged amino acid anchors, performs similarly when potential neoantigens with enhanced binding ("motif neoepitopes") are identified. We then demonstrate that the likelihood of motif neoepitopes depends on mutational landscape and show evidence of immunoediting and immune escape based on motif features.

5.1 Introduction

ICB is arguably the most important therapeutic advance in cancer in the past decade, yet prediction of clinical benefit remains challenging.[201] For ICB to be effective, CD8 T-cells must be able to engage and activate to kill cancer cells, requiring both the presence of a stimulatory signal (engagement of receptors with an antigen bound to the cancer cell's major histocompatibility complex-1 or HLA) and absence of an inhibitory signal (lack of engagement of co-inhibitory receptors, such as PD-L1, which is a target of ICB).[202] Current ICB biomarkers act as surrogates that enhance the probability of a favorable T-cell-cancer interaction, including high tumor programmed death ligand 1 (PD-L1) expression (a sign the cancer may have established itself by inhibiting T-cells),[203] high tumor mutation burden (more options for stimulatory binding),[204] DNA mismatch repair-deficiency (more mistakes to lead to more options for stimulatory binding),[205] and

gene signatures associated with enhanced T-cell activity.[206-208] Of these, PD-L1 is the only Food & Drug Administration (FDA) non-conditionally approved biomarker, despite its inconsistent performance.[209-211] PD-L1, like other ICB biomarkers, is able to enrich the likelihood of ICB response in identified subgroups but does not preclude ICB treatment response in those without this marker, and similarly, does not assure response in those with these markers. The experiments in this chapter suggest heterogeneous HLA-specific interactions provide a rationale for inconsistent ICB biomarker performance and suggest an approach to improve assessment of HLA-related outcomes.

5.2 Related Work

Related work pertinent to these experiments includes biologic domain knowledge regarding cancer-immune interactions, human leukocyte antigen (HLA) and HLA supertypes, and B44-associations to ICB response.

Cancer-immune interactions. It is generally accepted that the host immune system shapes tumor fate in three phases through the activation of innate (general) and adaptive (specific) immune mechanisms.[212] In the first “elimination” phase, abnormal cells are destroyed by a competent immune system. Sporadic tumor cells that manage to survive immune destruction then enter an “equilibrium” phase where editing occurs, referring to the selective pruning of cells with features that elicit immune response. This, in turn, leads to the immunoselection of tumor cells more capable of surviving in an immunocompetent host, which characterizes the “escape” phase, in which immunologically sculpted tumors begin to grow progressively and become clinically apparent.[213] This process is of great interest to translational work in oncology as identification and optimization of features that shift escaped tumors back to equilibrium/elimination phases can inform drug development and ideally, novel therapeutic strategies for cancer control.

Neoantigens, or abnormal peptide fragments that are encountered by T-cell receptors, provide the media through which T-cells recognize tumor cells as foreign. Initial evidence for the role of neoantigens emerged when immunocompetent and immunodeficient genetic mouse models of sarcoma were exposed to lentivirus encoding strong epitopes p.SIINKFEKL and p.SIYRYYGL – these epitopes subsequently were found only in the sarcomas of immunodeficient mice and not in those of immunocompetent mice, suggesting the immunocompetent mice were able to effectively kill tumor cells that included these epitopes.[214] A suggestive analysis in a pan-cancer study using The Cancer Genome Atlas (TCGA) additionally identified a relatively depleted neoantigen burden compared to what would be expected theoretically.[206]

The issue with neoantigens, however, is predicting which are clinically relevant. Neoantigens are highly specific to individual tumors, rarely shared across people, and out of hundreds to thousands of predicted candidates, only 8% are truly oncogenic, and only one or two of these will lead to clonal T-cell expansion.[215, 216] This problem has led to poor agreement among numerous computational approaches to neoantigen prediction, which are unable to provide insight into the impact of functional neoantigens and their correlation with clinical endpoints.[217, 218] In an evaluation of pancreatic cancer, for instance, tumors with the highest predicted neoantigen number in silico and the most abundant CD8+ T-cell infiltrate in biopsy specimens together associated with longer survival, but neither was statistically significant on its own.[219]

HLA supertypes & B44. As a method to refine identification of clinically relevant neoantigens, HLA and HLA supertypes can be considered. HLA class I moieties (HLA-A, HLA-B, HLA-C) are found in all nucleated cells and are the scaffolds that present intracellular peptides (antigens) to CD8+ T-cells.[220] While HLA has been implicated in immune responses to cancer for decades,[221, 222] the large number of class I alleles (over 6500 to date)[223] made HLA-based evaluations exceedingly difficult until the identification of HLA supertypes.[158] HLA supertypes leverage structural features of binding pocket residues to group HLA alleles.[158, 224, 225] These designations

rely in particular on residues that interact with peptide anchors, which generate most HLA-peptide binding energy, intimately relating supertype to antigenic peptide motifs (conserved amino acids in specific positions of presented peptides).

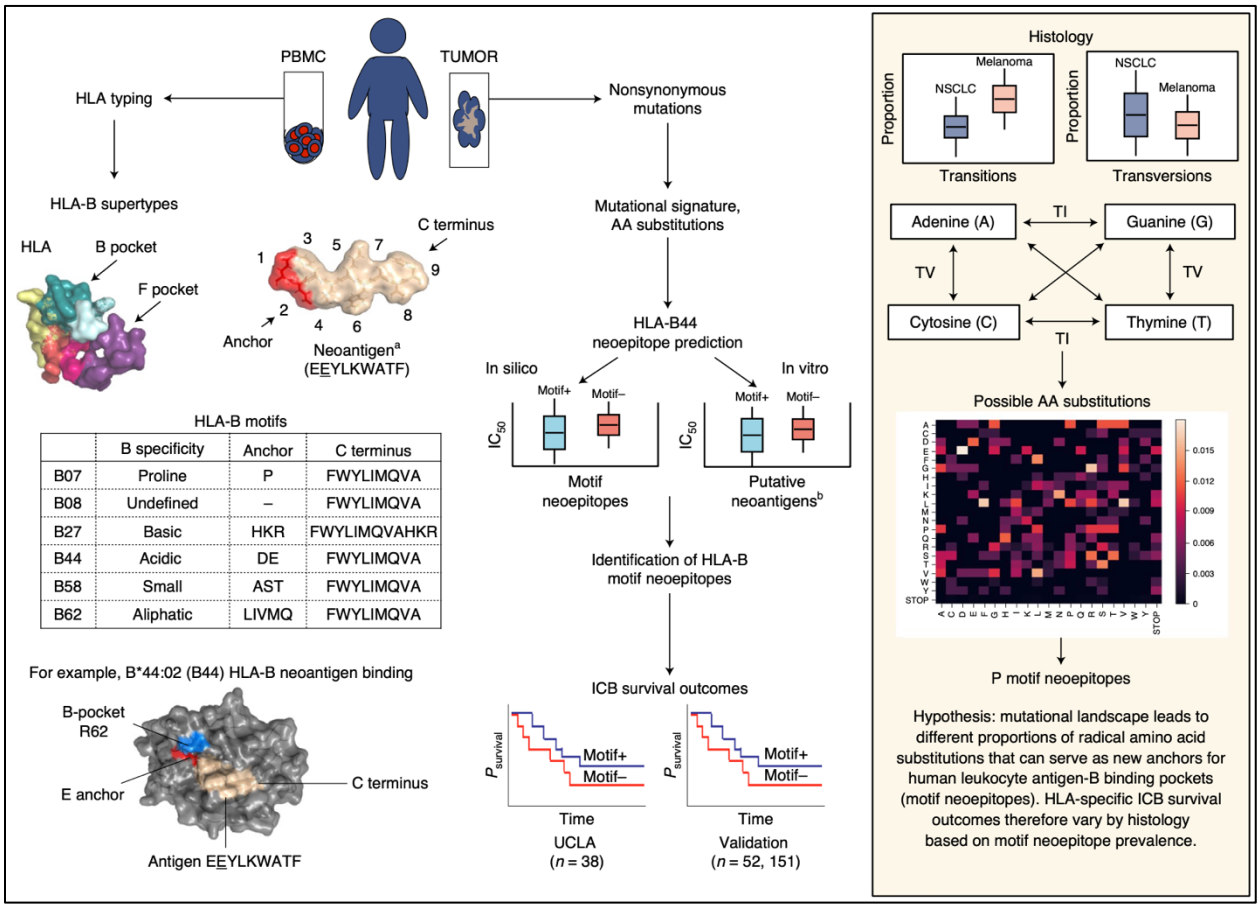
In ICB-treated melanoma, the B44 supertype (B44), which is found in approximately half of people irrespective of race, is associated with greater overall survival, especially in those with increased glycine to glutamic acid anchor substitutions.[159] As melanoma and NSCLC have relatively similar prevalence of somatic mutations and response to ICB,[119, 226] it seemed likely that B44 would associate with clinical benefit across histology. Yet in one study of ICB-treated NSCLC featuring targeted sequencing panels, a protective effect was not found.[161] In this chapter, we evaluate NSCLC and melanoma cohorts from our institution along with other publicly available cohorts to examine the role of B44 in ICB-treatment and cancer-immune interactions through the lens of mutational landscape and favorable neoantigens.

5.3 Method

In this section, we detail our clinical sample processing and approach to HLA identification, tumor mutation burden (TMB), neoepitope prediction, mutational landscape, and clinical outcomes (Fig. 5.1). We then describe methods to evaluate antigen presentation machinery (APM), immune cell infiltrates, cytokine levels, and immune checkpoint expression. In vitro competition assays that evaluate HLA binding are then discussed. Multiomic data structures, modeling approaches, and adjustment for transcriptional strand bias previously were described in Chapters 3 and 4.

Preliminary. The same annotations using “g.” for genomic or DNA notation and “p.” for protein or amino acid notation with single letter symbols are continued. We introduce the concept of radical substitutions, which refer to changes in standard amino acid physiochemical properties including charge, size, hydrophobicity, and polarity based on side chain properties.[227]

Figure 5.1: Experiment schema.



	B specificity	Anchor	C terminus
B07	Proline	P	FWYLIMQVA
B08	Undefined	–	FWYLIMQVA
B27	Basic	HKR	FWYLIMQVAHKR
B44	Acidic	DE	FWYLIMQVA
B58	Small	AST	FWYLIMQVA
B62	Aliphatic	LIVMQ	FWYLIMQVA

+, present; -,absent; AA, amino acid; PBMC, peripheral blood mononuclear cells; TI, transition; TV, transversion. [a]Not drawn to scale. [b]Refers to peptide form of neoepitopes.

Standard single letter notation used for amino acids, substitutions underlined. An example of B44 HLA and neoantigen binding is shown (3KPM).[228] We hypothesized that mutational signature dictates radical amino acid substitutions, which in turn influences the likelihood of HLA-B motif neoepitopes (amino acid nonamers inferred from nonsynonymous mutations) and response to ICB.

Definitions. Amino acid charge is determined by average side chain pK_a in protein conformation.[229] Aspartic acid (D) and glutamic acid (E) are considered negatively charged, histidine (H), lysine (K), and arginine (R) positively charged. Radical charged substitutions include mutations from oppositely charged or uncharged wildtype (starting) amino acids to p.D/E (negatively charged), or p.H/K/R (positively charged); stop codons are considered uncharged. Transition mutations are defined as DNA nucleotide mutations between purines (adenine/guanine) or pyrimidines (cytosine/thymine); transversion mutations include all other permutations.[163, 164] Motif neoepitopes are defined by 2008 criteria, requiring the presence of a

predicted nonamer with a radical substitution in the anchor (second) position and C-terminus that matches supertype motif.[224, 225] B27-motif neoepitopes require a radial mutation substituting a positively charged amino acid in the second position of a predicted nonamer; B44-motif neoepitopes require a radical glutamic acid in the second position of a nonamer (Fig. 5.1).

Sample processing. For patients treated at the University of California, Los Angeles (UCLA), peripheral blood mononuclear cells (PBMC) and tissue samples underwent multiplexed paired-end whole exome sequencing (WES) to a target depth of 100-150X on HiSeq 2000/3000 (Ilumina) performed by the UCLA Technology Center for Genomics and Bioinformatics. Macrodissection was not performed. DNA isolation was performed with DNeasy Blood and Tissue Kit (Qiagen); exon capture and library preparation used the KAPA HyperPrep Kit and Nimblegen SeqCap EZ Human Exome Library v3.0 (Roche). Publicly available multiomic data used for these experiments were procured through the Genome Data Commons (GDC) and standardized based on harmonized pipelines (https://gdc.cancer.gov/about-data/gdc-data-processing), see Chapter 3 "Data preprocessing" for more information.

Datasets. Of 67 advanced patients with NSCLC treated on clinical trials with single-agent pembrolizumab at UCLA who consented to our tissue-banking protocol approved by the UCLA Institutional Review Board, 65 had adequate germline samples and were included in our retrospective survival analysis. Thirty-eight of these patients had matched PBMC and tissues samples and were included in the UCLA NSCLC cohort. SRP067938 comprised patients from the UCLA melanoma cohort (N=14) and was used for genomic analysis only. The DF-NSCLC and DF-melanoma cohort (N=52, N=151) included patients from phs000452, phs000980, phs00694, phs001041, phs001565, and SRP067938, excluding five patients included in the UCLA NSCLC cohort.[163, 164, 203, 238, 247, 248] Unique patients from TCGA-LUAD (lung adenocarcinoma), LUSC (lung squamous cell carcinoma), and SKCM (melanoma) with clinical information available (N=518, 496, 470) were used for genomic analyses. Only one TCGA cancer sample was allowed per

patient; for patients with more than one cancer sample, the sample including the highest number of VEP annotations that passed all filters was included. CPTAC lung cancer datasets LUAD and LUSC were used for protein-based analyses.[81, 117, 118]

HLA, tumor mutation burden (TMB), and neoepitope prediction. For UCLA samples, HLA type was obtained by aligning germline WES to the UCSC hg38 primary assembly reference via Burrows-Wheeler Aligner v0.7 (http://bio-bwa.sourceforge.net), which was filtered by Sequence Alignment/Map tools (SAMtools) v1.7 (http://www.htslib.org).[230] HLA calling was performed with ATHLATES software.[172] Supertype was determined by 2008 criteria and included alleles with an experimentally established motif or B and F pocket exact match with the exception of B*44:05, which does not have a B and F pocket exact match but previously has been included in B44 analyses (see Fig. 5.1).[159, 225] Somatic mutations were identified using Falcon Computing optimized pipelines leveraging Genomic Analysis Toolkit software v.3.8, including MuTect2 requiring four calls to confirm a variant.[231]

Annotations were performed with snpEff v.4.1, Ensemble VEP v.94/v.99, and VCF Annotation Tools.[82, 232] TMB was calculated by summing the number of protein coding mutations, allowing only one mutation to be counted per reference base position, and dividing this number by 38 to create a mutations per megabase statistic. High TMB was considered the top 40% of observed values per cohort.[233] Personalized variant antigens by cancer sequencing (pVAC-seq) software v.1.5 was used to predict nonamer HLA-B neoepitopes and half-maximal inhibitor concentrations (IC_{50}) from nonsynonymous mutations, reporting only those with a predicted IC_{50} less than 500 nM.[173, 234] The NetMHC 4.0 algorithm was used for well characterized alleles, NetMHCpan 4.0 for alleles with limited information and otherwise not eligible for NetMHC.[235-237] HLA and neoepitope prediction for publicly available ICB-treated cohorts was performed by Miao and colleagues.[238] The code used to produce these data is openly available

(https://www.broadinstitute.org, https://www.ensembl.org, https://www.pvac-seq.readthedocs.io, http://vatools.org).

Antigen presentation machinery (APM), immune cell infiltrate, cytokine levels, and immune checkpoint gene data. TCGAbiolinks was used for analysis of antigen presentation machinery and immune checkpoint related gene expression.[239, 240] Masked somatic Mutation Annotation Format (MAF) files were first preprocessed to remove low quality and potential germline variants based on VEP annotation. Genes involved in HLA-B antigen processing and presentation (*HLA-B, B2M, TAP1, TAP2, TAPBP, PSMB5-10, ERAP1, ERAP2, PDIA3, CANX, CALR*) were selected for further analysis.[241] Samples that contained annotations including missense, nonsense, nonstop, translation-start site, in-frame deletions, in-frame insertions, frameshift deletions, frameshift insertions, splice-site, and splice-region mutations were considered as having mutations in APM. To account for high polymorphism, HLA somatic mutations were obtained using Polysolver software[242] as MAF files with mutations calls against a reference genome are inaccurate. Current biomarkers including checkpoint inhibitory co-receptors *PD-L1 (CD274), CTLA-4, LAG-3*, and *CD8A*;[243] cytokines/chemokines *CXCL9/CXCL10/CXCL13* associated with the recruitment of T-cells;[27] and pan-cancer gene signatures including the cytolytic signature (*GZMA,* PRF1)[206] and T-cell inflammation signature (Tinflam, including *CD2, CD3D/E, CCL5, CIITA, CXCL13, GZMK, HLA-DRA, HLA-E, IDO1, IL2RG, LAG3, NKG7, STAT1,* and *TAGAP*),[208] also were evaluated. Cell enrichment score was determined by gene set enrichment analysis in xCell software (https://xcell.ucsf.edu) for tumor infiltrating immune cells including memory-/B-cells, memory-/CD4+T-cells, memory-/CD8+T-cells, regulatory T-cells, natural killer cells, dendritic cells, monocytes, and M1/M2 macrophages.

In vitro competition assays. The DUCAF cell line (HLA-B*18:01, B44 supertype) and Sweig007 cell line (HLA-B*40:02, B44 supertype) were cultured in RPMI 1640, supplemented with 2 mM L-glutamine, 1% penicillin-streptomycin and 10% fetal bovine serum. Cultures were maintained

between 3 x 10^5 - 2 x 10^6 cells per mL. Wildtype and mutant peptides synthesized for HLA-B*18:01 and HLA-B*40:02 competition assays were based on neoepitope prediction per methods above from two patients with B*18:01 and the only patient with B*40:02. Neoepitopes were selected based on lowest percentile rank overall, motif, and those featuring radical substitutions with no more than two neoepitopes coming from the same patient in any one category. B*40:02 neoepitopes were used to synthesize wildtype, mutant, and artificial nonamers (Bachem). Leucine and methionine were selected for uncharged amino acid comparisons given similarity in molecular weight and shape to glutamic acid. Differences in IC_{50} were assessed based on a competition-based cellular peptide binding assay.[244] For any category assessed with only one representative neoepitope, assays were run twice with both values included in the analysis. Data were acquired on an LSRFortessa Cell Analyzer DIVA v8.0 (BD Biosciences) and analyses performed with FlowJo v.10.5 (Tree Star). IC_{50} were calculated with Prism (GraphPad).

Outcome representation. Histograms, boxplots, scatterplots, and survival curves with associated summary statistics were created in R v4.1 (www.r-project.org). Multiomic data structures were created based on techniques used in Chapter 3. Mutational signature was extracted using MuTect2 annotations ('REF' and 'ALT') and reported based on corrections for transcriptional cell bias detailed in Chapter 4. Amino acid substitution representations were created as described in Chapter 4 and visualized with heatmaps with Seaborn (https://seaborn.pydata.org). The code used to support the findings of this study is publicly available at https://github.com/garon-lab/hlab44.

Statistical analysis. Overall survival (OS) and progression-free survival (PFS) were estimated using the Kaplan-Meier method and compared between groups using nonparametric log-rank tests. Hazard ratios were estimated using proportional hazards; a hypothesis test based on standard errors compared B44 hazard ratios. Univariable and multivariable analyses were conducted using proportional hazard ratios assessed with chi-square likelihood ratio tests. DNA

mutation and amino acid substitution comparisons used standard t-tests for count data and the generalized estimating equations method for proportions; Tukey's correction for multiple comparisons was used. Correlations were assessed by Pearson's method. IC_{50} were compared with Wilcoxon tests. The difference between motif and non-motif neoepitope gene expression was based on logarithmically-transformed counts[245] and assessed with a clustered Wilcoxon rank sum test using the Rosner-Glynn-Lee method to account for repeated measures.[246] Two-tailed P values of <0.05 were considered statistically significant and rounded to the nearest thousandth. Statistical analyses were performed in R v4.0/4.1 and SAS v9.4 with the exception of the generalized estimating equations method, which was performed in SPSS v.24.

5.4 Experiments

Standard evaluation techniques are first employed. ICB survival outcomes are assessed based on B44 (Fig. 5.2). Univariable and multivariable logistic regression models assess macroscopic features such as age, sex, ancestry, and smoking history as well as microscopic features including histology and expression of immune checkpoints (e.g., PD-L1) (Table 5.2). Leveraging modeling techniques of the prior chapters, radical glutamic acid substitution proportions are compared between NSCLC and melanoma samples (Fig. 5.3). Mutational signature is then used to compare average proportions of DNA mutations leading to radical substitutions (Table 5.3) and correlate radical glutamic acid substitutions (Fig. 5.4) to glutamic acid ("motif") neoepitopes (Fig. 5.5). In silico motif neoepitope features are characterized (Fig. 5.5), and competition assays evaluate predicted neoantigens based on motif (Fig. 5.6 and 5.7, see Appendix 7.3). ICB-survival based on the presence of motif neoepitopes (see Appendix 7.4) is then examined in multiple cohorts (Fig. 5.8 and 5.9). ICB-biomarkers are then evaluated in B44 based on motif (Fig. 5.10) and compared to TMB (Table 5.4), as are antigen presentation machinery and immune cell enrichment (Fig. 5.11). TCGA datasets compare expression and methylation of motif vs. non-motif neoepitopes in B44 (Table 5.5). Based on experiments in Chapter 4, CPTAC integrated data

structures are then used to compare motif neoepitope protein expression and protein expression of amino acid substitutions controlled by the total number of amino acid substitutions and stratified by HLA supertype (Table 5.6, Fig. 5.12 and 5.13).

5.5 Results

B44 associates with poor outcomes in NSCLC

The median age of the UCLA NSCLC cohort was 68 (range 32-91), 63.1% were smokers and 9% (25/65) were women. Cohort features are summarized in Table 5.1. Approximately half (35 out of 65) of the UCLA NSCLC cohort had at least one B44 allele, with similar prevalence in non-Hispanic white (58.0%, 29 out of 50) and Hispanic white/non-white (40.0%, 6 out of 15) patients. Fig. 5.2 shows the median OS of patients with the B44 supertype was 9.3 months (95% CI 3.9-18.7) versus 18.8 months (95% CI 9.2-38.8, P=0.024); median progression-free survival (PFS) was 2.1 months (95% CI 1.9-4.9) versus 10.2 months (95% CI 5.5-14.5, P=0.040). The B44 hazard ratio for death was 2.02 for NSCLC compared to 0.61 for melanoma (P<0.001),[159] suggesting clearly different B44 associations for these cancer types.

In univariable and multivariable analyses, B44 had a hazard ratio of 2.13-2.88 and associated with worse OS (hazard ratio 2.45, P=0.005; univariable P=0.007, multivariable P=0.075) and PFS (hazard ratio 4.18, P=0.002, univariable P=0.003, multivariable P=0.096). High PD-L1 and adenocarcinoma histology, previously associated with improved ICB outcomes,[203] were the only protective features (P=0.059 and 0.051 in OS, 0.022 and 0.030 in PFS). Additional features are in Table 5.2.

Radical glutamic acid substitutions vary based on mutational signature

Based on the previously described enrichment of p.G>E anchors in melanoma responders,[159] we hypothesized that negatively charged glutamic acid substitutions in the anchor position were beneficial for immune presentation due to enhanced binding to B44's positively

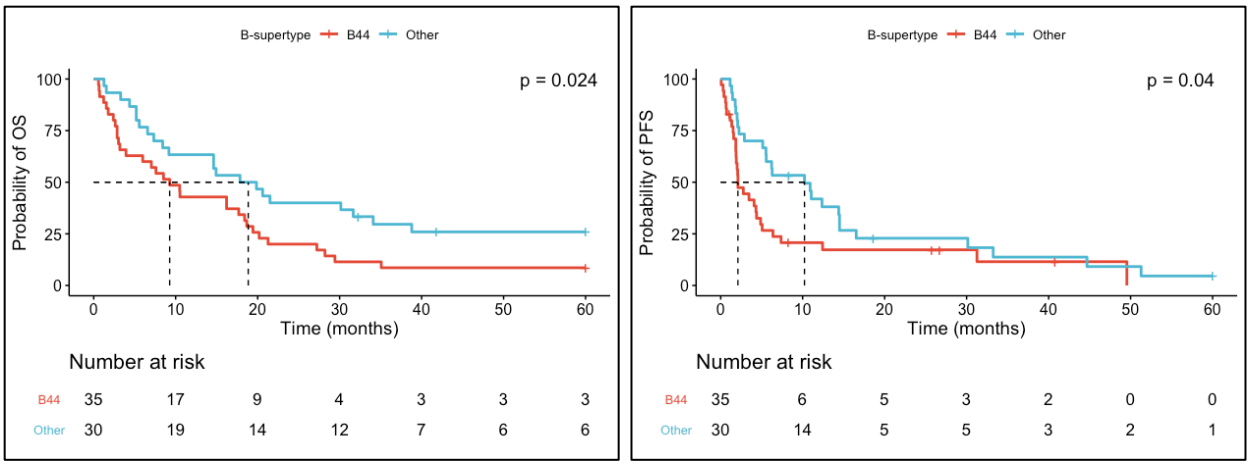
Table 5.1: Cohort features.

	UCLA	DF		TCGA			CPTAC	
	NSCLC	NSCLC	MEL	LUAD	LUSC	SKCM	LUAD	LUSC
Total cases	65	52	151	511	482	466	111	108
Age	68 (32-91)	66 (42-87)	62 (18-86)†	66 (33-88)	68 (39-90)	58 (15-90)	62 (35-81)	67 (40-88)
Sex								
Female	25 (38.5)	26 (50.0)	57 (37.7)	275 (53.8)	125 (25.9)	178 (38.2)	37 (34.6)	22 (20.6)
Male	40 (61.5)	26 (50.0)	94 (42.3)	236 (46.2)	357 (74.1)	288 (61.8)	70 (65.4)	85 (79.4)
Ancestry								
Asian	6 (9.2)	-	-	7 (1.4)	8 (1.7)	12 (2.6)	53 (49.5)	23 (21.5)
Black/African	2 (3.1)	-	-	52 (10.2)	29 (6.0)	1 (0.2)	1 (0.9)	1 (0.9)
Caucasian	51 (78.5)	-	-	385 (75.3)	338 (70.1)	443 (95.6)	38 (35.5)	81 (75.7)
Hispanic	6 (9.2)	-	-	-	-	-	2 (1.8)	0 (0.0)
Smoking history								
Current/former	40 (63.1)	38 (73.1)	-	334 (65.4)	399 (82.8)	-	58 (54.2)	82 (76.6)
Never	25 (36.9)	15 (26.9)	-	177 (34.6)	83 (17.2)	-	46 (43.0)	16 (15.0)
Stage								
I	0	-	-	273 (53.4)	233 (48.3)	76 (16.3)	57 (53.2)	39 (36.5)
II	0	-	-	120 (23.5)	157 (32.6)	140 (30.0)	29 (27.1)	43 (40.2)
III-IV	65 (100.0)	-	-	110 (21.5)	88 (18.4)	191 (41.0)	21 (19.6)	25 (18.5)
B44	35 (53.8)	20 (38.5)	65 (43.0)	239 (46.8)	228 (47.3)	227 (48.7)	35 (31.5)	37 (34.2)
+motif*	14 (40.0)	13 (65.0)	48 (73.8)	100 (41.8)	95 (41.7)	156 (68.7)	7 (6.3)	16 (14.8)

*percentage reflects out of total B44 patients; †missing on 52 patients (34.4%), - data not available. Features represented as total number followed by percentage in parenetheses. Note not every category includes full number of cases.

Figure 5.2: Survival based on B44 supertype in UCLA NSCLC cohort.

A. Overall Survival (OS) **B. Progression-Free Survival (PFS)**



Survival estimated using Kaplan-Meier method and compared with a nonparametric log-rank test. Dashed lines represent the median. B44 supertype includes all patients with at least one B44 allele. Patients with B44 N=35, other N=30.

A. Patients with B44 had a median OS 9.3 months (95% CI 3.9 – 18.7) versus 18.8 months (95% CI 9.2 – 38.8) in others, P=0.024.

B. Patients with B44 had a median PFS 2.1 months (I5% CI 1.9 – 4.9) versus 10.2 months (95% CI 5.5 – 14.5) in others, P=0.040.

charged binding pocket.[142, 249, 250] We further proposed that radical glutamic acid substitutions, or those that substitute glutamic acid for an uncharged or positively charged wildtype amino acid, may be distributed unevenly in melanoma and NSCLC, decreasing the likelihood of substitutions that could act as new B44 anchors in NSCLC tumors.

Fig. 5.3 shows the comparison of UCLA NSCLC and melanoma cohort proportions of radical substitutions. An average of 2.8% of somatic substitutions were to glutamic acid in NSCLC compared to 5.9% in melanoma. We anticipated the mutational landscape was responsible for this difference as melanoma preferentially exhibits transition mutations, particularly g.C>T due to ultraviolet light dimerization, while NSCLC has relatively more transversion mutations, particularly g.C>A caused by bulky adducts from smoking.[163, 164] While transition and transversion mutations are equally likely to substitute negatively charged amino acids (4.2 versus 4.2%, P=1.0), transversion mutations are more likely to substitute positively charged amino acids (12.5 versus

Table 5.2: Univariable and multivariable analyses of B44 subset survival.

A.

B44 OS Univariable analysis

Characteristic	HR (95% CI)	P
Age	0.98 (0.96 – 1.01)	0.228
Male	1.64 (0.94 – 2.83)	0.082
White	1.02 (0.56 – 2.03)	0.943
Histology-AD	1.45 (0.76 – 3.05)	0.278
PDL1-high*	0.71 (0.38 – 1.28)	0.263
Smoking+	0.61 (0.36 – 1.08)	0.088
TMB-high†	0.85 (0.43 – 1.62)	0.623
B44	2.13 (1.23 – 3.69)	0.007

B.

B44 OS Multivariable analysis

Characteristic	HR (95% CI)	P
Age	1.02 (0.98 – 1.07)	0.291
Male	1.26 (0.58 – 2.81)	0.564
White	0.62 (0.23 – 2.00)	0.394
Histology-AD	0.32 (0.11 – 1.00)	0.051
PDL1-high*	0.42 (0.16 – 1.03)	0.059
Smoking+	0.61 (0.25 – 1.49)	0.278
TMB-high†	0.86 (0.36 – 2.05)	0.740
B44[1]	2.88 (0.89 – 8.71)	0.075

C.

B44 PFS Univariable analysis

Characteristic	HR (95% CI)	P
Age	0.98 (0.96 – 1.01)	0.196
Male	1.84 (1.05 – 3.17)	0.033
White	0.98 (0.55 – 1.83)	0.946
Histology-AD	1.35 (0.73 – 2.70)	0.349
PDL1-high*	0.70 (0.37 – 1.27)	0.248
Smoking+	0.59 (0.34 – 1.04)	0.070
TMB-high†	0.80 (0.41 – 1.52)	0.504
B44	2.40 (1.36 – 4.20)	0.003

D.

B44 PFS Multivariable analysis

Characteristic	HR (95% CI)	P
Age	1.01 (0.97 – 1.06)	0.562
Male	0.81 (0.34 – 1.97)	0.635
White	0.86 (0.3 – 2.73)	0.785
Histology-AD	0.28 (0.10 – 0.88)	0.030
PDL1-high*	0.33 (0.12 – 0.86)	0.022
Smoking+	0.66 (0.27 – 1.63)	0.364
TMB-high†	1.05 (0.44 – 2.45)	0.903
B44[2]	2.61 (0.83 – 7.27)	0.096

AD – adenocarcinoma, PDL1 – programmed death ligand 1 (high considered ≥ 50% by tumor proportion score);[203] PFS – progression-free survival, OS – overall survival, Smoking+ – ≥100 lifetime cigarettes;[251] TMB – tumor mutation burden (high – top 40%, corresponding to ≥ 38 mutations/megabase).[233]

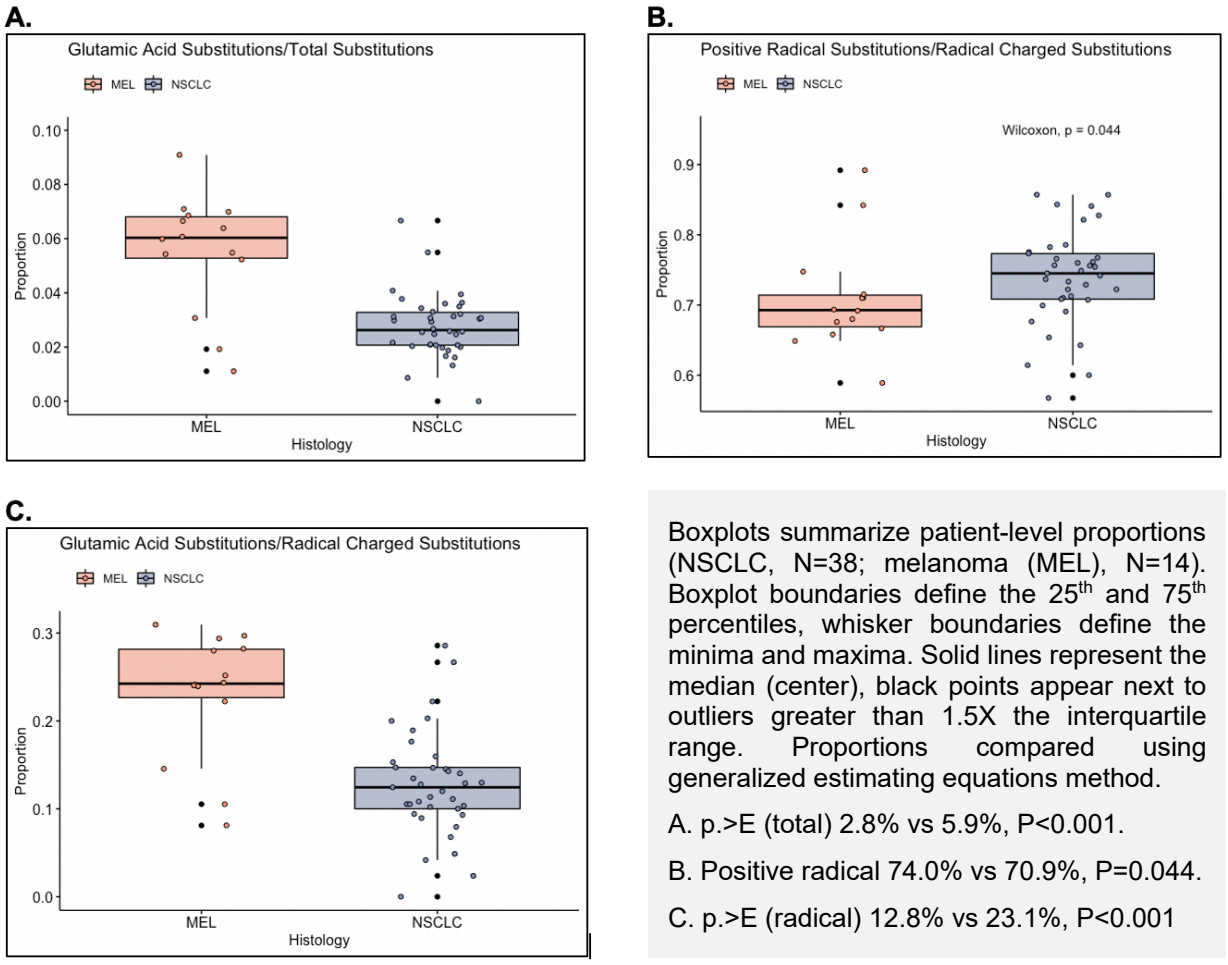
*Missing on 8 patients (N=57). †Reported for patients with WES (N=38). [1]Including a B44 interaction term for race, which was not significant (P=0.233). [2]Race interaction term (P=0.498).

Cox proportional hazard ratios assessed by chi-square tests and compared with likelihood ratio tests (95% confidence interval).

6.3%, P=0.021), particularly g.C>A, which decreases radical negative substitutions by 0.2% for every 10% of enrichment (see Table 3.2).

To model this phenomenon, we determined that if all nonsynonymous mutations occurred equally, based on amino acids present in the GRCh38 reference, we would expect 14.0% to result in a radical positive substitution and 5.5% to result in a radical negative substitution (P<0.001), indicating a natural predisposition for radical positive substitutions (see Table 3.2). Using corrected average proportions of DNA base mutations for our NSCLC and melanoma cohorts

Figure 5.3: Proportion of glutamic acid and radical charged substitutions by histology.



Boxplots summarize patient-level proportions (NSCLC, N=38; melanoma (MEL), N=14). Boxplot boundaries define the 25th and 75th percentiles, whisker boundaries define the minima and maxima. Solid lines represent the median (center), black points appear next to outliers greater than 1.5X the interquartile range. Proportions compared using generalized estimating equations method.

A. p.>E (total) 2.8% vs 5.9%, P<0.001.

B. Positive radical 74.0% vs 70.9%, P=0.044.

C. p.>E (radical) 12.8% vs 23.1%, P<0.001

(Table 5.3), we predicted, on average, NSCLC would have 12.5% radical positive substitutions and 6.4% radical negative substitutions while melanoma would have 11.1% radical positive substitutions and 6.6% radical negative substitutions. Reframing as proportions of radical substitutions, this corresponds to 73.8% of radical charged substitutions being positive in NSCLC (74.0% were observed, Fig. 5.3B) versus 71.8% in melanoma (70.9% were observed, Fig. 5.3B). In this model, radical negatively charged amino acid substitutions would be distributed equally between aspartic and glutamic acid with glutamic acid representing 11.0% radical substitutions in NSCLC (12.8% were observed, Fig. 5.3C) and 14.1% in melanoma (23.1% were observed, Fig. 5.3C). Table 5.3 shows the average proportions of DNA mutations leading to radical substitutions in NSCLC and melanoma cohorts. Fig. 5.4 shows enrichment of p.>E with g.C>T/G>A.

Table 5.3: Average proportions of DNA mutations leading to radical substitutions.

	NSCLC ($n=38$)	Melanoma ($n=14$)	Difference	*P* value
g.A>C	0.020 (0.02)	0.010 (0.01)	0.010	0.18
g.A>G	0.086 (0.04)	0.058 (0.03)	0.028	<0.001
g.A>T	0.007 (0.02)	0.001 (0.01)	0.001	0.058
g.C>A	0.207 (0.04)	0.076 (0.02)	0.131	<0.001
g.C>G	0.085 (0.04)	0.045 (0.02)	0.040	<0.001
g.C>T[a]	-	-	-	-
g.G>A	0.384 (0.10)	0.713 (0.05)	−0.329	<0.001
g.G>C	0.053 (0.02)	0.026 (0.01)	0.027	<0.001
g.G>T	0.037 (0.05)	0.007 (0.03)	0.030	<0.001
g.T>A	0.048 (0.02)	0.015 (0.00)	0.033	<0.001
g.T>C	0.028 (0.02)	0.028 (0.02)	0.000	0.88
g.T>G	0.037 (0.02)	0.015 (0.01)	0.022	0.065

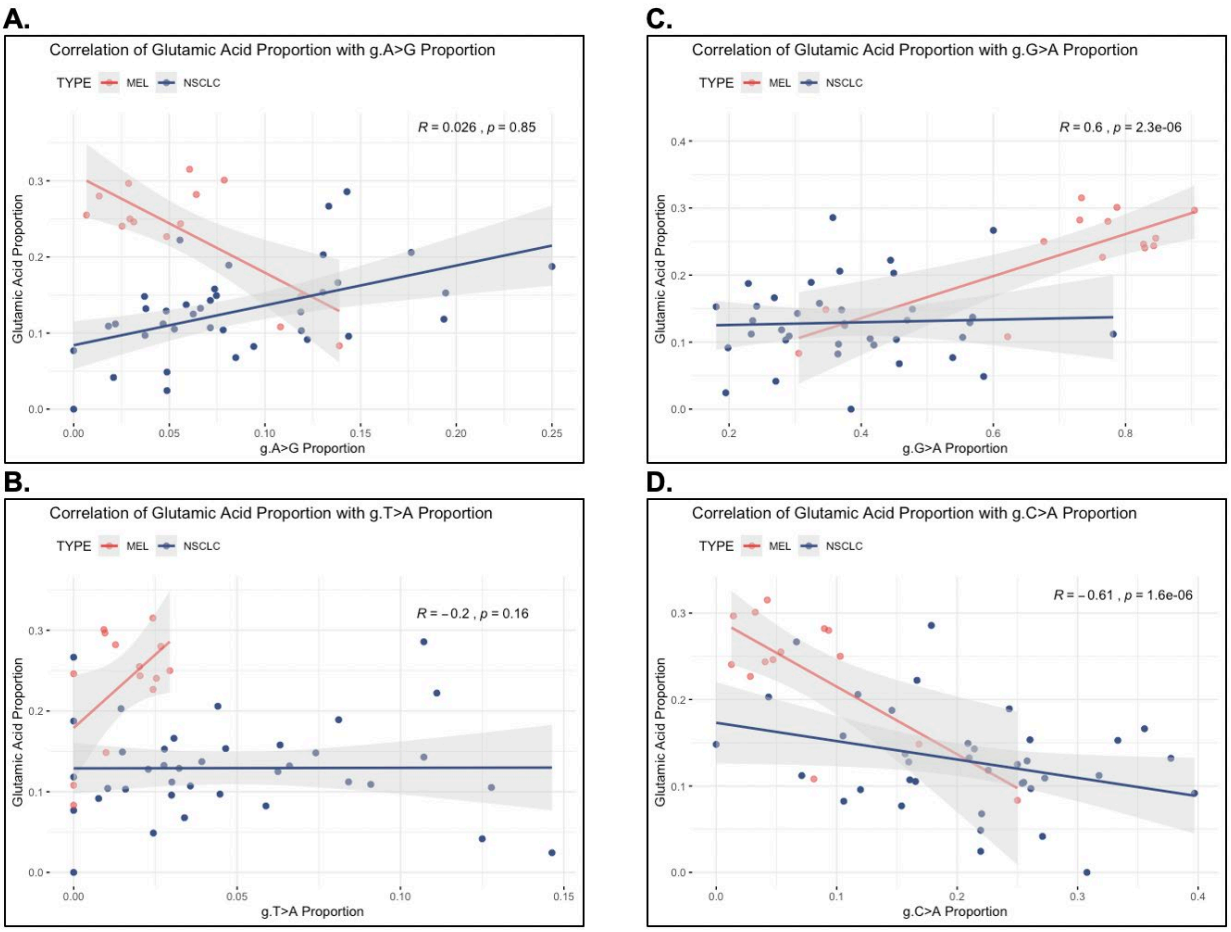
DNA mutations: g.A>C, adenine to cytosine; g.A>G, adenine to guanine; g.A>T, adenine to thymine; g.C>A, cytosine to adenine; g.C>G, cytosine to guanine; g.C>T, cytosine to guanine; g.G>A, guanine to adenine; g.G>C, guanine to cytosine; g.G>T, guanine to thymine; g.T>A, thymine to adenine; g.T>C, thymine to cytosine; g.T>G, thymine to guanine. [a]No radical substitutions are caused by g.C>T (cytosine to guanine) mutations. Proportions represent the number of DNA mutations leading to radical substitutions out of all mutations leading to radical substitutions. Mean proportion is listed with s.d. in parentheses. *P* values assessed using generalized estimating equations method for proportions.

for radical glutamic acid substitutions correlated with g.G>A mutation proportions and anti-correlated with g.C>A mutation proportions, suggesting that g.G>A was the most influential in enriched proportions of radical glutamic acid mutations in melanoma.

Motif neoepitopes suggest improved B44 binding in silico and in vitro

To understand the impact of radical glutamic acid substitutions on B44 binding and motif, we compared all NSCLC B44 neoepitope predictions with $IC_{50} \leq 500$ nM. Fig. 5.5 shows that among NSCLC B44 cases (N=21), the number of radical glutamic acid substitutions correlated with the number of predicted neoepitopes featuring glutamic acid substitutions (R=0.56, P=0.009), and the number of predicted neoepitopes featuring glutamic acid substitutions correlated with the number of B44 motif neoepitopes (R=0.89, P<0.001). A trend toward an association between B44 motif neoepitopes and radical glutamic acid substitutions also was seen (R=0.42, P=0.06). Fig. 5.6 characterizes predicted B44 neoepitopes based on motif, demonstrating that B44 motif neoepitopes have similar IC_{50} and percentile rank, but their wildtype epitopes have significantly

Figure 5.4: Correlation of radical glutamic acid and DNA mutation proportions.



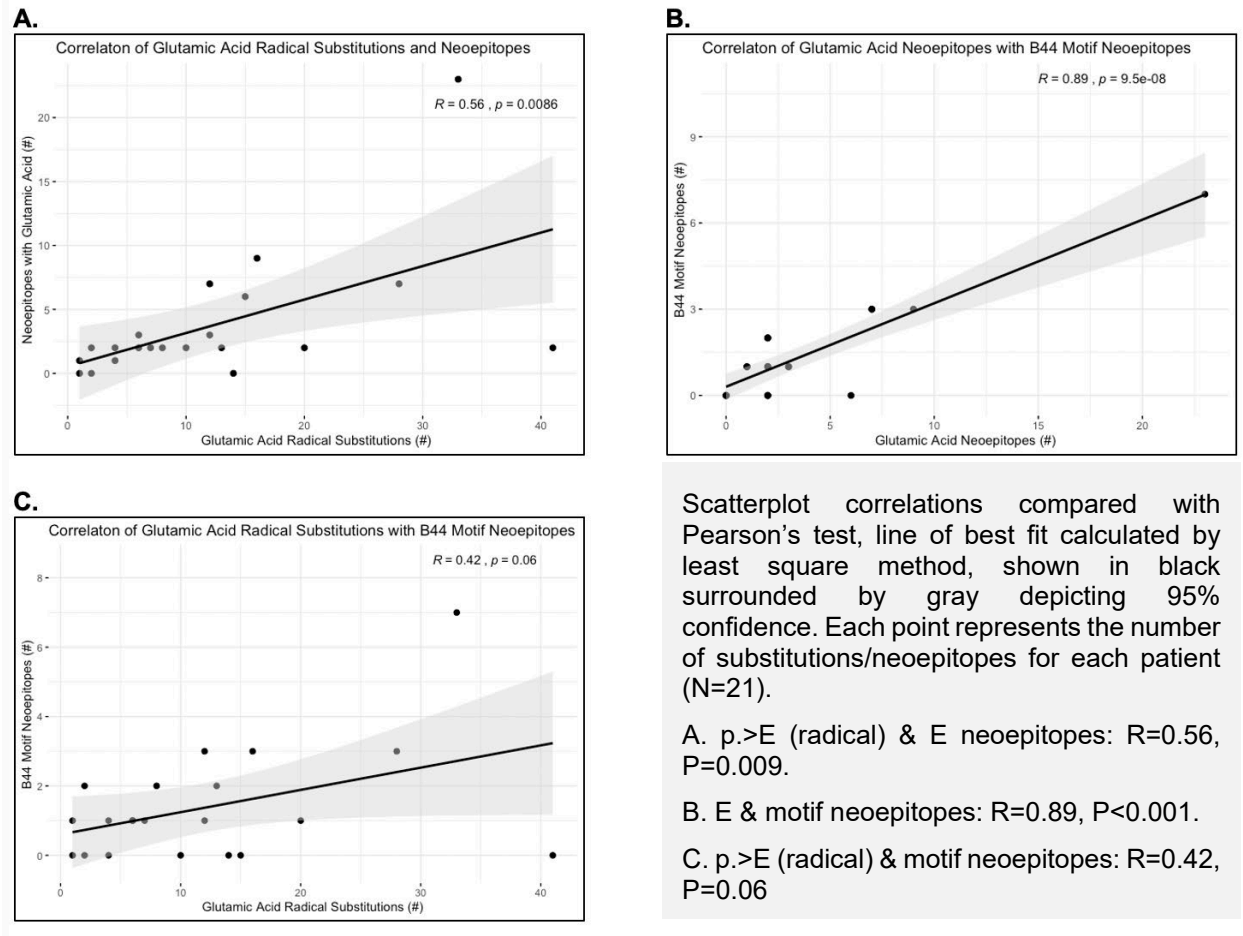
MEL – melanoma, NSCLC – non-small cell lung cancer, g. – genomic, A – adenine, C – cytosine, G – guanine, T – thymine. Scatterplot correlations compared with Pearson's test, line of best fit calculated by least square method, shown in black surrounded by gray depicting 95% confidence. Mutations that can result in radical glutamic acid substitutions (g.A>G, g.T>A, g.G>A, and g.C>A) are pictured. Each point represents a patient (NSCLC N=38, MEL N=14). Glutamic acid and mutation proportions calculated as proportion of all radical mutations.

A. g.A>G to p.>E (radical) R = 0.03, P=0.85. B. g.T>A to p.>E (radical) R = -0.20, P=0.16.

C. g.G>A to p.>E (radical) R = 0.60, P<0.001. D. g.C>A to p.>E (radical) R = -0.61, P<0.001.

higher predicted IC_{50} (P<0.001), suggesting improved mutant binding compared to wildtype. We then assessed mutant and wildtype epitopes in vitro with HLA-B*18:01 and B*40:02 (B44) cell models. Fig. 5.7A-B shows there were no significant differences in motif compared to other neoepitopes in terms of mutant IC_{50} (P=0.21) but that the difference between mutant and wildtype

Figure 5.5: Correlation of glutamic acid substitutions to (B44 motif) neoepitopes in NSCLC.



Scatterplot correlations compared with Pearson's test, line of best fit calculated by least square method, shown in black surrounded by gray depicting 95% confidence. Each point represents the number of substitutions/neoepitopes for each patient (N=21).

A. p.>E (radical) & E neoepitopes: R=0.56, P=0.009.

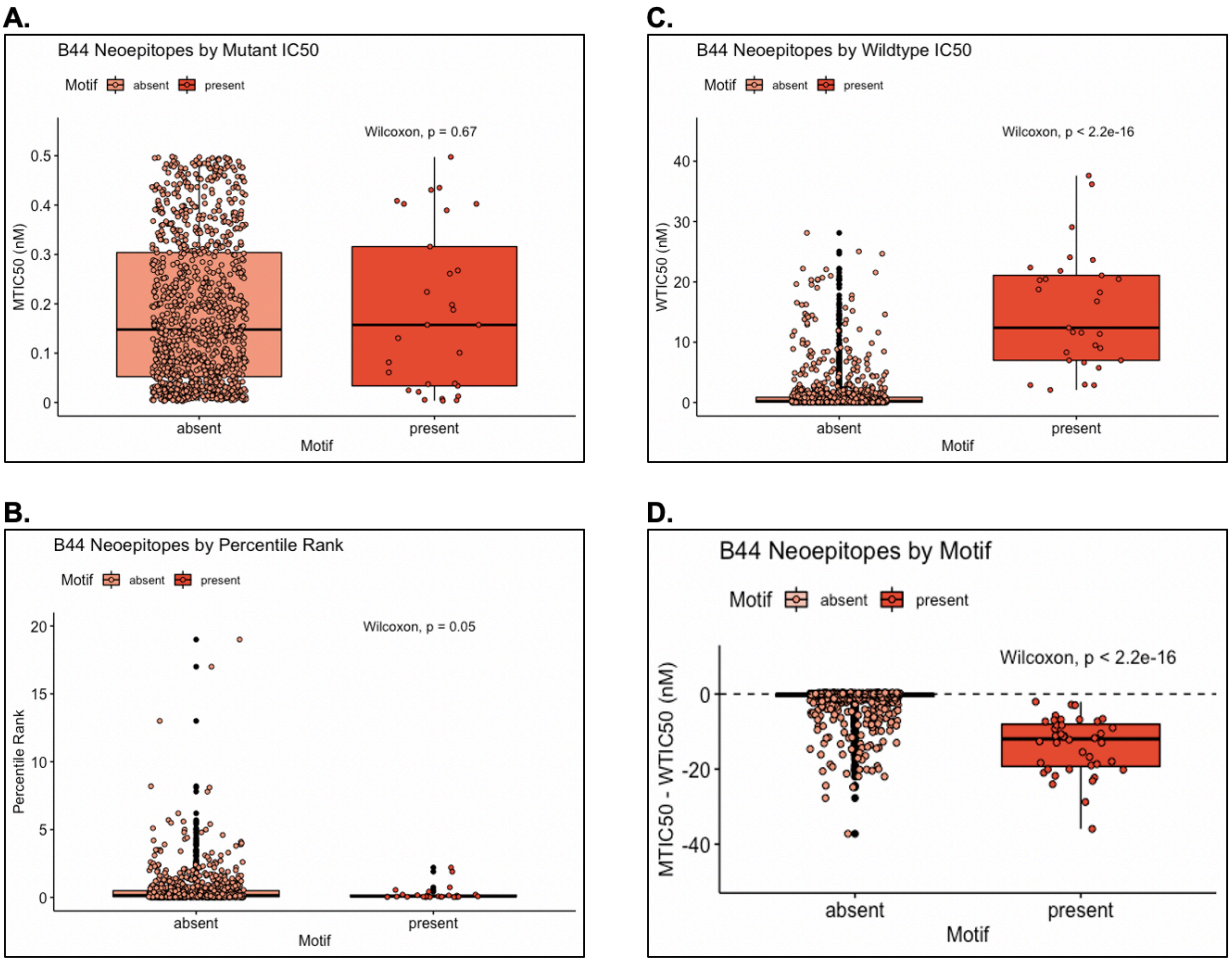
B. E & motif neoepitopes: R=0.89, P<0.001.

C. p.>E (radical) & motif neoepitopes: R=0.42, P=0.06

IC_{50} demonstrated greater comparative binding of motif neoepitopes (P=0.026, 0.016), recapitulating findings of our in silico experiments.

With the B*40:02 cell model, we explored the impact of amino acid charge on IC_{50} by using all predicted B*40:02 neoepitopes from our cohort that exhibited a radical charged substitution, synthesizing wildtype, mutant, and artificial peptides substituting other charged amino acids in the same position (see Appendix 7.3). Assessing IC_{50} differences among oppositely charged artificial and mutant peptides in relation to wildtype peptides, we demonstrated B*40:02 binding affinity was commensurate with an electrostatic gradient: peptides with negatively charged amino acid anchors exhibited significantly improved binding compared to those with positively charged anchors (P=0.031). Evaluating radical charged substitutions in positions 1 and 3-9, artificial and

Figure 5.6: In silico B44 neoepitope differences in half-maximal inhibitory concentrations.



Boxplots summarize distribution of points with solid lines representing the median proportion, black points appear next to outliers greater than 1.5 times the interquartile range. Each point represents a neoepitope prediction. B44 motif present refers to neoepitopes featuring a radical substitution to glutamic acid in the anchor position with a known C-terminus (FWYLIMQVA). Assessed with Wilcoxon test of difference.

A. Predicted B44 neoepitope IC_{50} P=0.67. B. Predicted B44 neoepitope percentile rank P=0.05.

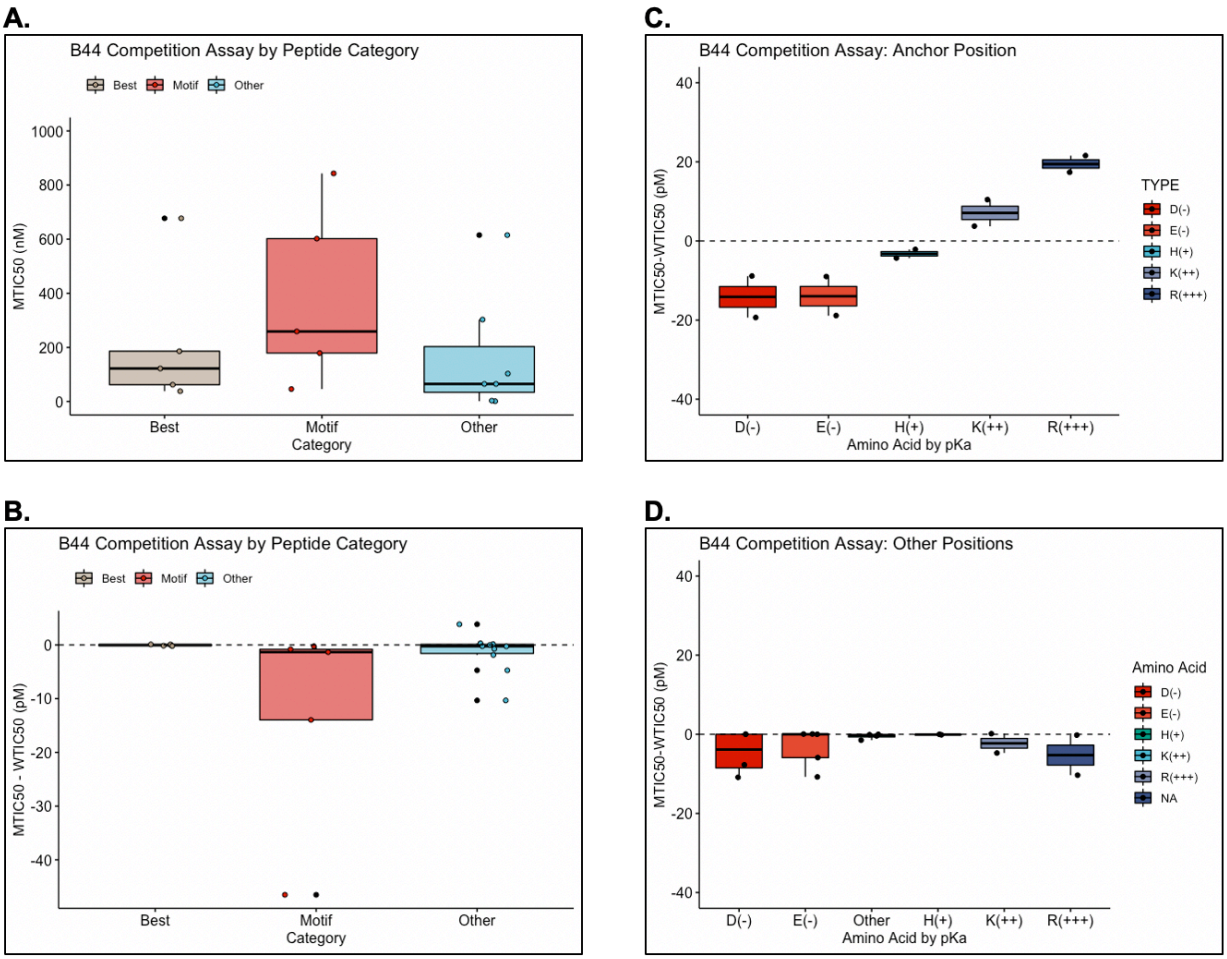
C. Predicted B44 wildtype epitope IC_{50} P<0.001. D. Predicted B44 IC_{50} mutant-wildtype P<0.001.

mutant B*40:02 peptides did not reveal significant IC_{50} differences based on charge (P=0.90), limiting electrostatic influence on B44 peptide binding to the anchor position.

Motif neoepitopes associate with ICB survival in B44 supertypes

Examining B44 survival based on the presence or absence of B44 motif neoepitopes, we found that the presence of motif neoepitopes identified subpopulations with distinct ICB survival.

Figure 5.7: In vitro B44 neoepitope differences in half-maximal inhibitory concentrations.



Boxplots summarize distribution of points with solid lines representing the median proportion, black points appear next to outliers greater than 1.5 times the interquartile range. Each point represents a peptide. IC50 – half-maximal inhibitory concentration (IC_{50}). MT – amino acid substitution (mutant), WT – wildtype. Best reflects mutant peptides with lowest IC_{50} rank. Motif peptides are defined by a radical charged substitution to glutamic acid in the second position and a known C-terminus (FWYLIMQVA). Additional details are available in Methods and Table 5.4.

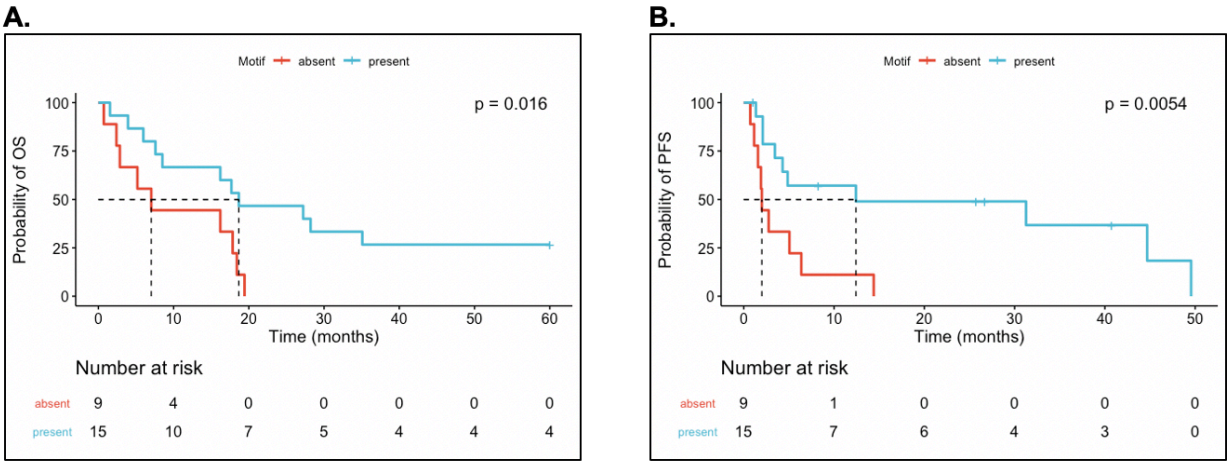
A. B44 competition assay (B*18:01, B*40:02) comparing predicted peptides based on mutant IC_{50}. Wilcoxon test of difference between motif and non-motif peptides P=0.21, best and motif peptides P=0.19.

B. B44 competition assay (B*18:01, B*40:02) comparing predicted peptides based on difference between mutant and wildtype binding (MT IC_{50} – WT IC_{50}). Wilcoxon test of difference between motif and non-motif peptides and P=0.026, best and motif peptides P=0.016.

C. B44 competition assay comparing predicted and artificial peptides featuring radical substitutions in the anchor position. Wilcoxon test of difference between negatively and positively charged amino acids P=0.031.

D. B44 competition assay comparing predicted and artificial peptides featuring radical substitutions in non-anchor positions. Wilcoxon test of difference between mutant and wildtype binding not significant between negatively and positively charged amino acids, P=0.90.

Figure 5.8: Survival based on motif neoepitopes in UCLA NSCLC.



Survival estimated using the Kaplan-Meier method and compared between groups with a non-parametric log-rank test. Dashed lines represent medians. Motif neoepitopes are listed in Appendix 7.4.

A. Motif neoepitopes median OS 18.7 (95% CI 6.0-NR) vs. 7.0 (95% CI 0.7-18.4) months, P=0.016.

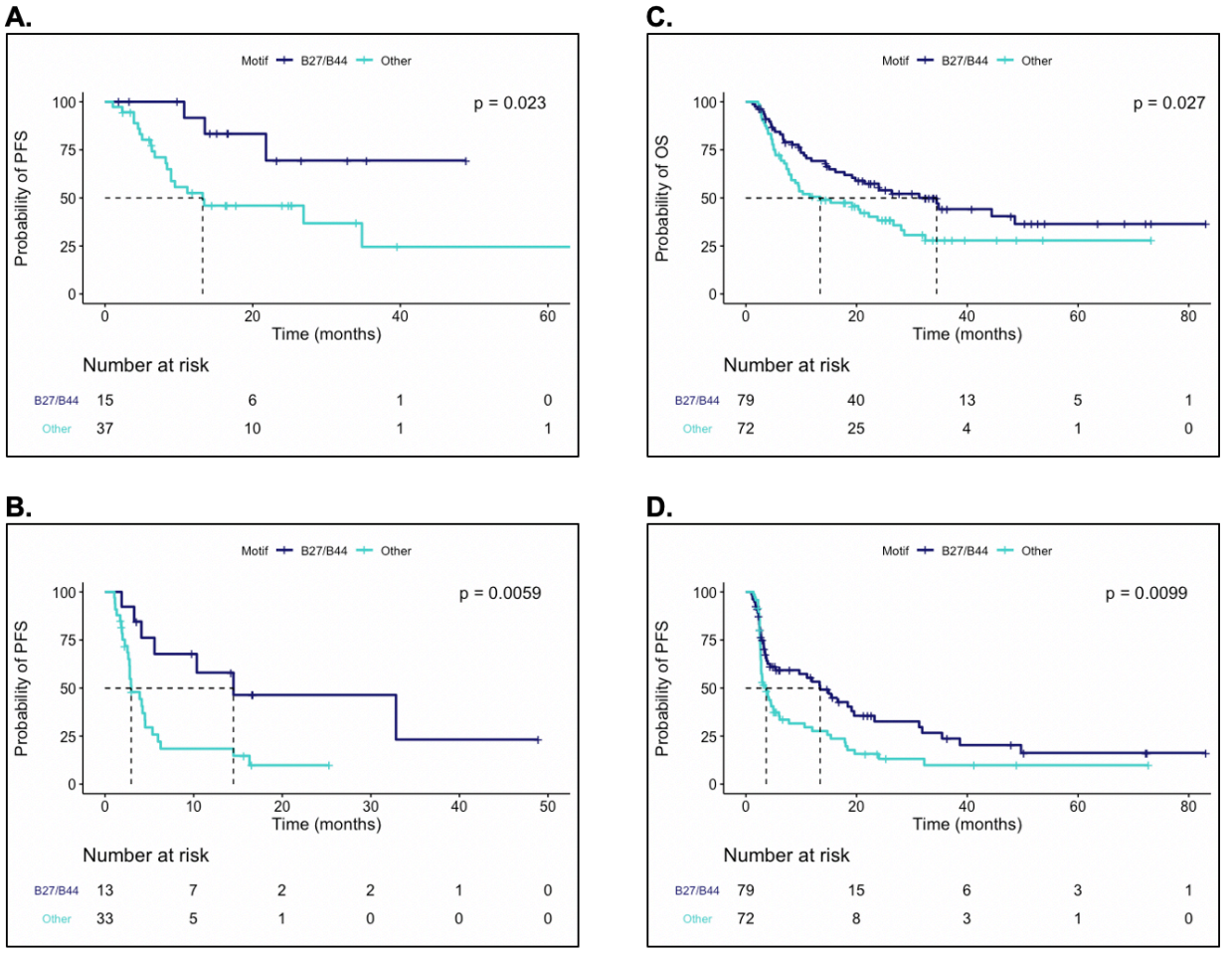
B. Motif neoepitopes median PFS12.4 (95% CI 2.1-NR) vs. 2.8 (95% CI 1.2-12.4) months, P=0.005.

Median OS in those with motif neoepitopes was 18.2 versus 7.0 months in those without (HR 0.38, P=0.038); median PFS was 8.7 versus 3.5 months (HR 0.40, P=0.072). We examined other B-supertypes (B07: N=19, motif N=15; B27: N=5, motif N=3; B58 N=4, motif=3; B62: N=8, motif=7) for similar survival curve separation based on the presence of favorable neoepitopes, but only B27, the only other supertype with a charged binding pocket, demonstrated a trend towards significance with median OS with motif neoepitopes not reached versus 11.5 months in those without (HR could not be calculated, P=0.039) and median PFS 45.3 vs. 8.2 months (HR 0.18, P=0.062). Fig. 5.8 shows composite survival outcomes in patients with B44 and/or B27. The presence of motif neoepitopes associated with a median OS of 18.7 months (95% CI 6.0-not reached) versus 7.0 months (95% CI 0.7-18.4, P=0.016) and median PFS of 12.4 months (95% CI 1.2-not reached) versus 2.8 months (95% CI 1.2-12.4, P=0.005). Appendix 7.4 lists observed B44 and B27 motif neoepitopes.

To determine the robustness of the association between motif neoepitopes in charged HLA supertypes with ICB survival outcomes, we validated our findings using publicly available

data generated at Dana Farber (DF-NSCLC, DF-melanoma).[238] Motif neoepitopes were present in 47.1% (16/34) of patients from the DF-NSCLC B44/B27 cohort and 87.9% (80/91) patients from DF-melanoma B44/B27 cohort, suggesting motif neoepitopes are more common in melanoma than NSCLC, as we originally hypothesized. Fig. 5.9 shows the presence of B44 and B27 motif neoepitopes was protective in all cohorts (DF-NSCLC: median OS not reached vs. 13.2 months

Figure 5.9: Survival based on motif neoepitopes in DF NSCLC and melanoma cohorts.



Survival estimated using the Kaplan-Meier method and compared between groups with a non-parametric log-rank test. Dashed lines represent medians.

A. NSCLC motif median OS NR (95% CI 21.8.-NR) vs. 13.2 (95% CI 8.9-NR) months, P=0.023.

B. NSCLC motif median PFS14.4 (95% CI 5.6-NR) vs. 3.0 (95% CI 2.8-5.4) months, P=0.006.

C. Melanoma motif median OS 34.5 (95% CI 19.8-NR) vs. 13.4 (95% CI 8.2-28.0) months, P=0.027.

D. Melanoma motif median PFS13.4 (95% CI 5.4-23.2) vs. 3.7 (95% CI 2.8-6.0) months, P=0.010.

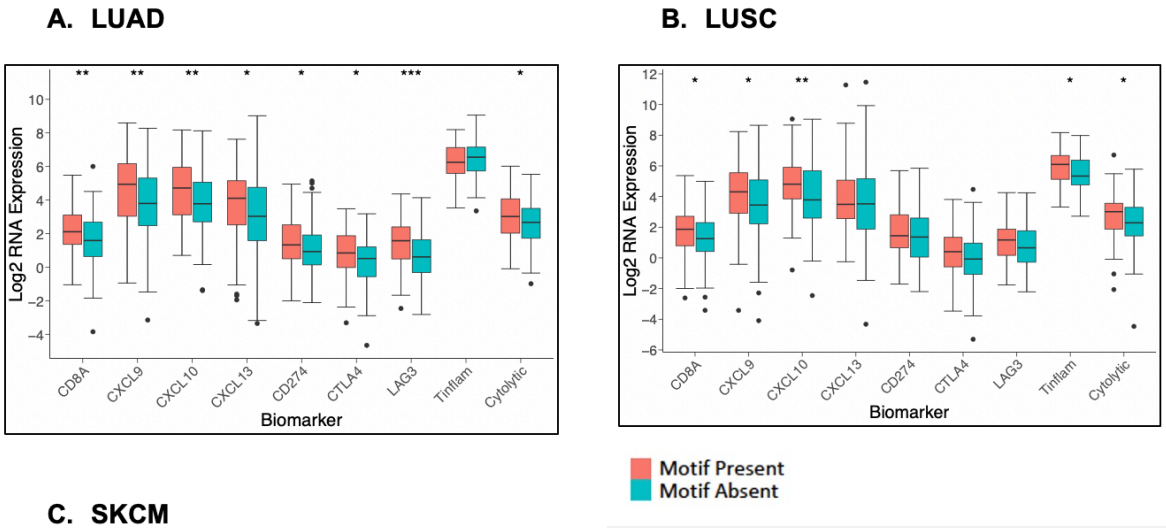
P=0.023, median PFS 14.4 vs. 3.0 months P=0.006; DF-melanoma: median OS 34.5 vs. 13.4 months P=0.027, median PFS 13.4 vs. 3.7 months P=0.010).

Immunoediting and immunosuppression suggest motif neoepitope functionality

While these findings are suggestive, it is unclear whether motif neoepitopes are functional entities and immune "targets" or simply biomarkers of response. We hypothesized that evidence of immune escape or immunoediting based on the presence of motif neoepitopes would be supportive of immunogenicity/functionality and evaluated TCGA and CPTAC datasets. While there was no enrichment for HLA B44 or B27 alleles in NSCLC or melanoma cohorts, Fig. 5.10 depicts an evaluation of checkpoint and other biomarkers of ICB response based on RNA expression, which shows most biomarkers are significantly upregulated in cases with motif neoepitopes, although aside from *CXCL9*, which has been suggested to be the most robust marker,[27] none are significantly upregulated in all cohorts. Table 5.4 compares motif neoepitopes against TMB, showing enhanced performance across gene expression biomarkers. Fig. 5.11 evaluates antigen presentation machinery and immune cell composition. Cases with motif neoepitopes show proportionately higher loss of APM in all TCGA cohorts, which is statistically significant in LUSC (P=0.048) and SKCM cohorts (P=0.010), but does not reach statistical significance in LUAD (P=0.130), and demonstrates varied enrichment of subcomponents. Immune cell composition does not demonstrate any clear signals across NSCLC cohorts.

Evaluating B44 neoepitopes, Table 5.5 examines TCGA gene expression and methylation based on findings from Chapter 4, using detectable gene expression (RNA>0) and beta values less than 0.9 as cut points. The proportion of motif neoepitopes with gene expression is less than that of non-motif neoepitopes (LUAD 0.664 vs. 0.830, LUSC 0.852 vs. 0.893, SKCM 0.796 vs. 0.884) although at the case level, is only statistically significant in SKCM (0.2 vs. 1.2, P=0.001), although directionality is retained in LUAD (0.7 vs. 1.5, P=0.083) and LUSC (1.3 vs. 1.5, P=0.215). The proportion of motif neoepitopes with highly methylated genes is not consistent in LUAD (0.953

Figure 5.10: Relation of ICB biomarkers to motif neoepitopes in TCGA B44 patients.



LUAD – lung adenocarcinoma (B44 N=239), LUSC – lung squamous cell carcinoma (B44 N=228), SKCM – melanoma (B44 N=227).

*P<0.050, **P<0.01, ***P<0.001.

Orange – patients with at least one B44 motif neoepitope, teal – patients w/o B44 motif neoepitopes.

Differences assessed based on a clustered Wilcoxon rank sum test using the Rosner-Glynn-Lee method to account for repeated measures.

A. LUAD B44 patients with motif neoepitopes (N=88) have significantly higher *CD8A, CXCL9, CXCL10, CXCL13, CD274, CTLA4, LAG3,* and the cytolytic signature (8/9 markers).

B. LUSC B44 patients with motif neoepitopes (N=85) have significantly higher *CD8A, CXCL9, CXCL10, Tinflam* and cytolytic signature (5/9 markers).

C. SKCM B44 patients with motif neoepitopes (N=148) have significantly higher *CXCL9, CXCL10, CD274, CTLA4, LAG3,* and the cytolytic signature (6/9 markers).

vs. 0.942), although suggestive in LUSC (0.878 vs. 0.894) and SKCM (0.831 vs. 0.849), although absolute differences are small and only statistically significant in SKCM (0.3 vs. 0.1, P=0.006). Table 5.6 summarizes CPTAC neoepitopes with evidence of gene protein expression, only possible with the CPTAC LUAD and LUSC B44 cohorts given few protein-level events in B27 patients. Overall, the proportion of neoepitopes with protein expression at the case level was 0.235 for LUAD and 0.364 for LUSC; 0.375 and 0.636 had detectable RNA, respectively; 0.975

Table 5.4: Immunomodulatory gene expression in TCGA B44 patients based on motif neoepitopes vs. high tumor mutation burden.

Histology	Marker	P-value	Motif	Non-motif	P-value	TMB-H	TMB-L
Cases			N = 88	N = 151		N = 92	N = 147
LUAD	*CD8A*	0.011	2.110 (1.412)	1.587 (1.472)	0.030	2.025 (1.386)	1.585 (1.545)
LUAD	*CXCL9*	0.003	4.940 (2.094)	3.796 (2.122)	0.001	5.023 (2.106)	3.786 (2.173)
LUAD	*CXCL10*	0.007	4.717 (1.816)	3.778 (1.817)	0.002	4.789 (1.704)	3.843 (1.923)
LUAD	*CXCL13*	0.040	4.105 (2.118)	3.030 (2.416)	0.212	3.799 (2.106)	3.631 (5.838)
LUAD	*CD274*	0.052	1.327 (1.553)	0.929 (1.374)	0.055	1.389 (1.649)	0.980 (1.352)
LUAD	*CTLA4*	0.025	0.845 (1.365)	0.514 (1.334))	0.430	0.649 (1.389)	0.554 (1.376)
LUAD	*LAG3*	<0.001	1.575 (1.479)	0.611 (1.425)	0.003	1.342 (1.460)	0.636 (1.826)
LUAD	*Tinflam*	0.261	6.248 (1.131)	6.554 (1.038)	0.103	6.321 (1.190)	6.679 (1.017)
LUAD	*Cytolytic*	0.056	3.019 (1.402)	2.669 (1.295)	0.156	2.863 (1.338)	2.778 (1.386)
Cases			N = 85	N = 143		N = 87	N = 141
LUSC	*CD8A*	0.033	1.866 (1.561)	1.255 (1.546)	0.509	1.646 (1.676)	1.511 (1.535)
LUSC	*CXCL9*	0.022	4.305 (2.135)	3.444 (2.343)	0.065	4.556 (2.231)	3.867 (2.366)
LUSC	*CXCL10*	0.007	4.807 (1.830)	3.782 (2.180)	0.044	4.904 (2.042)	4.053 (2.143)
LUSC	*CXCL13*	0.271	3.492 (1.957)	3.522 (2.409)	0.032	3.853 (2.356)	3.333 (2.227)
LUSC	*CD274*	0.298	1.444 (1.482)	1.361 (1.696)	0.862	1.362 (1.744)	1.480 (1.595)
LUSC	*CTLA4*	0.045	0.405 (1.419)	-0.076 (1.635)	0.149	0.243 (1.616)	-0.112 (1.575)
LUSC	*LAG3*	0.083	1.169 (1.336)	0.658 (1.373)	0.211	1.088 (1.396)	0.578 (1.376)
LUSC	*Tinflam*	0.012	6.094 (1.059)	5.335 (1.168)	0.121	6.023 (1.159)	5.547 (1.156)
LUSC	*Cytolytic*	0.027	3.013 (1.524)	2.285 (1.554)	0.067	2.876 (1.494)	2.390 (1.619)
Cases			N = 148	N = 79		N = 150	N = 77
SKCM	*CD8A*	0.060	1.930 (2.521)	0.833 (2.630)	0.094	2.109 (2.515)	0.842 (2.768)
SKCM	*CXCL9*	0.010	4.296 (2.792)	2.897 (2.941)	0.037	4.418 (2.812)	2.734 (3.079)
SKCM	*CXCL10*	0.012	4.103 (2.676)	3.162 (2.700)	0.044	4.124 (2.707)	2.805 (2.812)
SKCM	*CXCL13*	0.120	2.257 (2.874)	1.657 (3.591)	0.029	2.325 (2.961)	1.004 (3.573)
SKCM	*CD274*	0.034	0.098 (1.828)	-0.374 (1.728)	0.062	0.176 (1.856)	-0.416 (1.771)
SKCM	*CTLA4*	0.001	0.541 (2.029)	-0.473 (2.453)	0.046	0.358 (2.129)	-0.115 (2.417)
SKCM	*LAG3*	0.014	1.415 (2.335)	0.055 (2.475)	0.144	1.325 (3.329)	0.267 (2.650)
SKCM	*Tinflam*	0.133	5.800 (1.356)	5.369 (1.500)	0.311	5.939 (1.379)	5.374 (1.538)
SKCM	*Cytolytic*	0.264	2.178 (2.186)	1.237 (2.248)	0.193	2.332 (2.146)	1.257 (2.454)

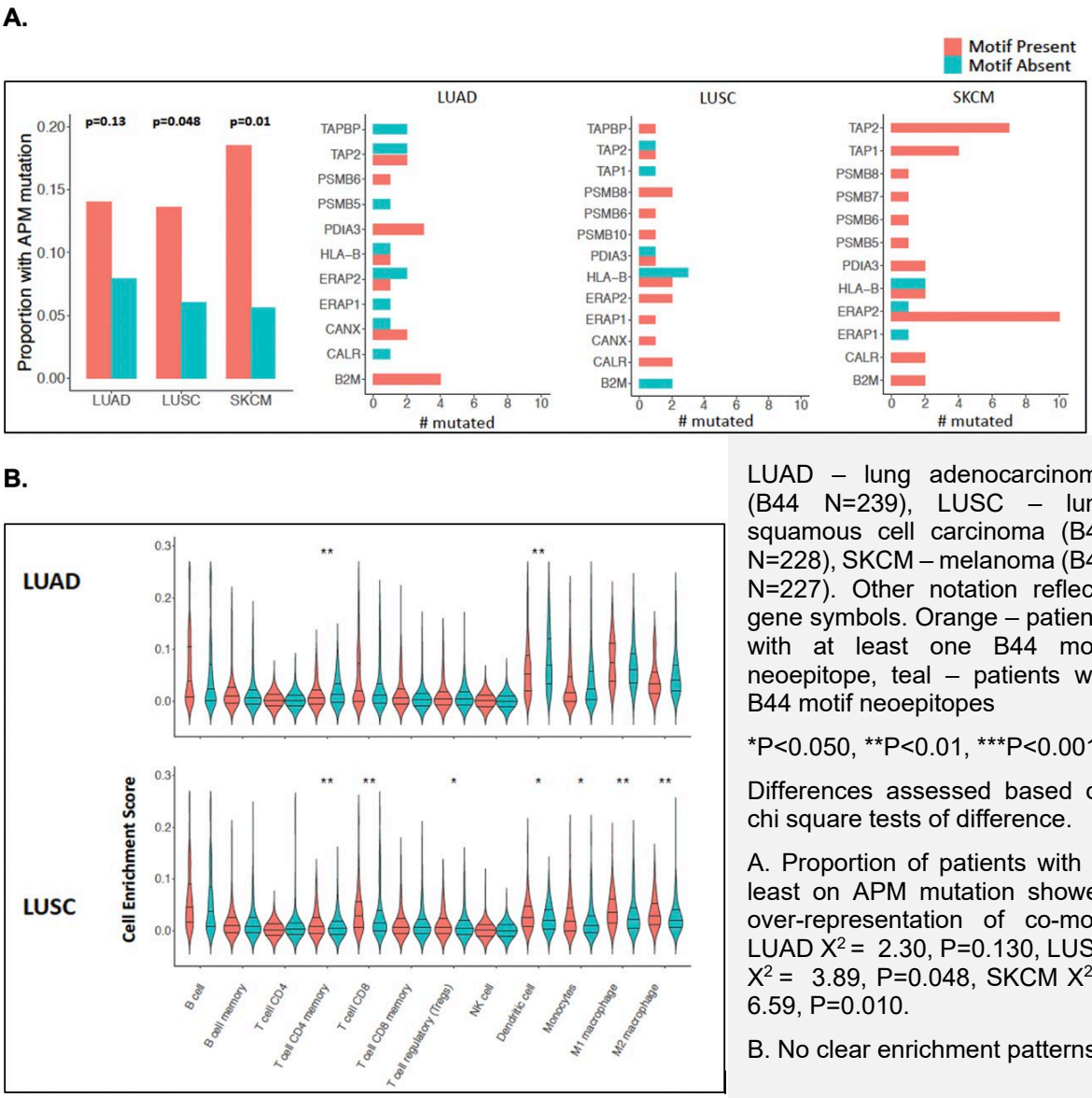
CD274 – PD-L1 gene expression, LUAD – lung adenocarcinoma, LUSC – lung squamous cell carcinoma, SKCM – melanoma, Tinflam – T-inflamed signature TMB-H – tumor mutation burden high (≥5 mutations/megabase), TMB-L – tumor mutation burden low (<5 mutations/megabase). Measured by log count of gene expression displayed as median (standard deviation).

Those with P-values <0.01 are highlighted in green, those <0.05 are highlighted in orange.

and 0.855 had methylation beta values less than 0.9. There were very few cases with protein expression of B44 motif neoepitopes (7 in LUAD and 16 in LUSC with only 1-3 per case), so B44 neoepitopes were considered in total. The total proportion of B44 motif vs. non-motif neoepitopes with protein expression in LUAD was 0.478 vs. 0.610 (P=0.226) and in LUSC was 0.171 vs. 0.512 (P<0.001). RNA and methylation findings were similar to TCGA results.

Findings at the cohort level did not show relative suppression of radical glutamic acid or any other particular amino acid substitution proportion in DNA and RNA layers based on HLA

Figure 5.11: Relation of multiomic analyses to motif neoepitopes in TCGA B44 patients.



LUAD – lung adenocarcinoma (B44 N=239), LUSC – lung squamous cell carcinoma (B44 N=228), SKCM – melanoma (B44 N=227). Other notation reflects gene symbols. Orange – patients with at least one B44 motif neoepitope, teal – patients w/o B44 motif neoepitopes

*P<0.050, **P<0.01, ***P<0.001.

Differences assessed based on chi square tests of difference.

A. Proportion of patients with at least on APM mutation showed over-representation of co-motif LUAD X^2 = 2.30, P=0.130, LUSC X^2 = 3.89, P=0.048, SKCM X^2 = 6.59, P=0.010.

B. No clear enrichment patterns.

status in TCGA or CPTAC cohorts. Given the probabilistic relationships depicted by protein expression of amino acid substitutions relative to mutation number in prior experiments, we explored HLA-specific effects by examining the number of SNVs with protein expression compared to the number of amino acid substitutions with protein expression, compared to SNPs with protein expression as a control. Based on Chapter 3 experiments, we selected amino acid substitutions with similar frequencies for comparison (radical glutamic acid substitutions have a

Table 5.5: TCGA neoepitope gene expression and methylation by cohort.

A. Patient-level characteristics.

	TCGA-LUAD (N=518)	TCGA-LUSC (N=496)	TCGA-SKCM (N=470)
*Patients with HLA-B predictions**	513 (0.990)	484 (0.976)	468 (0.996)
Patients with charged HLA-B	331 (0.645)	296 (0.612)	305 (0.652)
+ motif neoepitopes (DNA)†	76 (0.230)	99 (0.334)	137 (0.449)
+ motif neoepitopes with gene expression†	59 (0.178)	85 (0.287)	126 (0.413)

B. Proportion of charged HLA-B neoepitopes with detectable gene expression by motif.

	TCGA-LUAD (N=7,999)	TCGA-LUSC (N=7,863)	TCGA-SKCM (N=17,283)
Proportion of motif neoepitopes with gene expression*	0.664	0.852	0.796
Proportion of non-motif neoepitopes with gene expression	0.830	0.893	0.884

C. Charged HLA-B neoepitope gene expression by motif.

	TCGA-LUAD (N=321)	TCGA-LUSC (N=291)	TCGA-SKCM (N=291)
Motif neoepitopes	0.7 (0.1 – 4.4)	1.3 (0.1 – 5.5)	0.2 (0.0 – 1.7)
Non-motif neoepitopes	1.5 (0.1 – 6.3)	1.5 (0.2 – 6.2)	1.2 (0.1 – 6.3)
P-value	0.083	0.215	0.001

Charged HLA-B include all alleles belonging to B44 and B27 supertypes. *Patients without HLA-B allelic pair inference possible with ATHLATES excluded. †Represented as proportion of patients out of all patients with at least one charged HLA-B allele.

A. Patients depicted as raw number (proportion). HLA alleles were predicted for nearly all patients. Charged HLA alleles were present in over half. Approximately a quarter of LUAD patients had motif neoepitopes with a higher proportion in LUSC and particular enrichment in SKCM.

B. Proportion of neoepitopes with at least 0.001 fragments per kilobase million (FPKM) gene expression out of all predicted charged HLA-B neoepitopes with $IC_{50} \leq 500$ nM. Motif neoepitopes (LUAD N=125, LUSC N=162, SKCM N=338). Motif neoepitopes were less likely to be expressed than others.

C. All predicted charged HLA-B neoepitopes with at least 0.001 FPKM gene expression and $IC_{50} \leq 500$ nM included. Depicted as median FPKM (interquartile range). Difference between motif and non-motif neoepitopes assessed with clustered Wilcoxon rank sum test using Rosner-Glynn-Lee method to account for repeated measures and highly significant in SKCM.

Table 5.5 (con't): Neoepitope gene expression and methylation in TCGA cohorts.

D. Proportion of charged HLA-B neoepitopes with beta-values < 0.9 by motif.

	TCGA-LUAD (N=7,999)	TCGA-LUSC (N=7,863)	TCGA-SKCM (N=17,283)
Proportion of motif neoepitopes with beta-values < 0.9*	0.953	0.878	0.831
Proportion of non-motif neoepitopes with beta-values < 0.9	0.942	0.894	0.849
Motif neoepitopes beta values	0.5 (0.1 – 0.7)	0.3 (0.0 – 0.7)	0.3 (0.0 – 0.8)
Non-motif neoepitopes beta values	0.4 (0.1 – 0.8)	0.3 (0.0 – 0.7)	0.1 (0.1 – 0.7)
Motif vs. non-motif beta value P-value	0.939	0.850	0.006

*Patients without HLA-B allelic pair inference possible with ATHLATES excluded. All predicted charged HLA-B neoepitopes with at least 0.001 FPKM gene expression and $IC_{50} \leq 500$ nM included and evaluated based on their gene beta value. Difference between motif and non-motif neoepitopes assessed with clustered Wilcoxon rank sum test using Rosner-Glynn-Lee method to account for repeated measures.

There was not a significant difference in the proportion of genes with beta-values greater than 0.9, which would be suggestive of immunoediting. Only SKCM showed a statistically significant increase in methylation based on motif neoepitope classification.

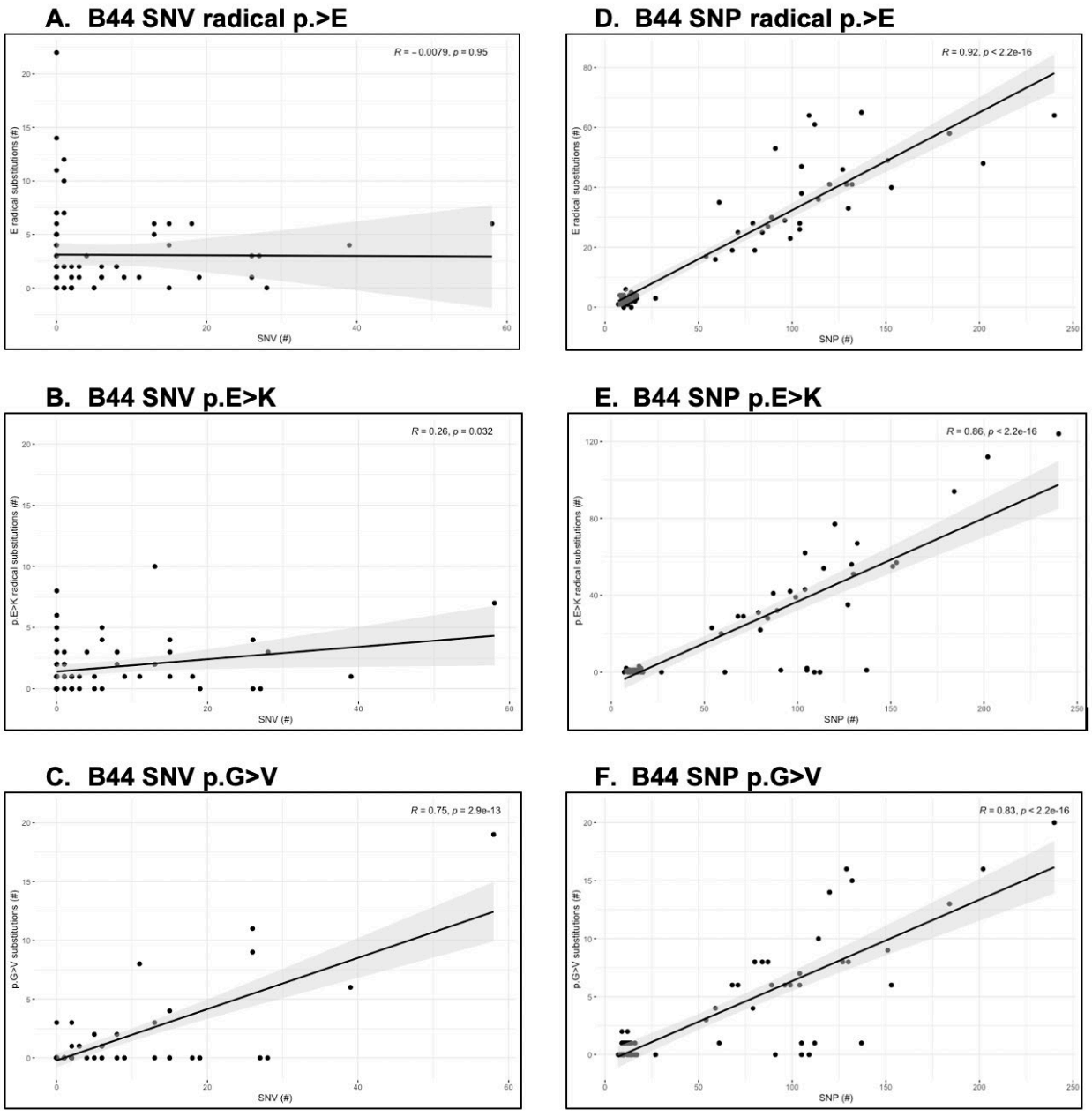
Table 5.6: CPTAC B44 neoepitope multiomic characteristics.

	LUAD (N=35)		LUSC (N=37)	
	Count	Proportion	Count	Proportion
B44-neoep	6 (0-62)		11 (1-103)	
+CH3	2 (0-25)	0.975 (0.249)	5 (0-31)	0.855 (0.165)
+RNA	2 (0-25)	0.375 (0.249)	7 (0-43)	0.636 (0.193)
+protein	1 (0-18)	0.235 (0.173)	4 (0-24)	0.364 (0.127)
*B44-motif**	23		35	
+CH3	15	0.952	10	0.816
+RNA	15	0.652	16	0.457
+protein	11	0.478	6	0.171
B44-non-motif	187		373	
+CH3	187	1.000	291	0.880
+RNA	187	1.000	373	1.000
+protein	114	0.610	191	0.512

CH3 – methylation, LUAD – lung adenocarcinoma, LUSC – lung squamous cell carcinoma.*LUAD had 7 cases with motif neoepitopes with protein expression, LUSC 16 cases (1-3 per case).

Represented as median number/proportion (range for count or standard deviation for proportion). B44-neoep refers to assessments at the case level of all B44 patients; B44-motif and B44-non-motif categories are considered in total. LUSC showed a greater proportional difference in protein expression compared to LUSC based on B44 motif.

Figure 5.12: HLA B44 supertype reveals immunoediting based on probabilistic relationships.



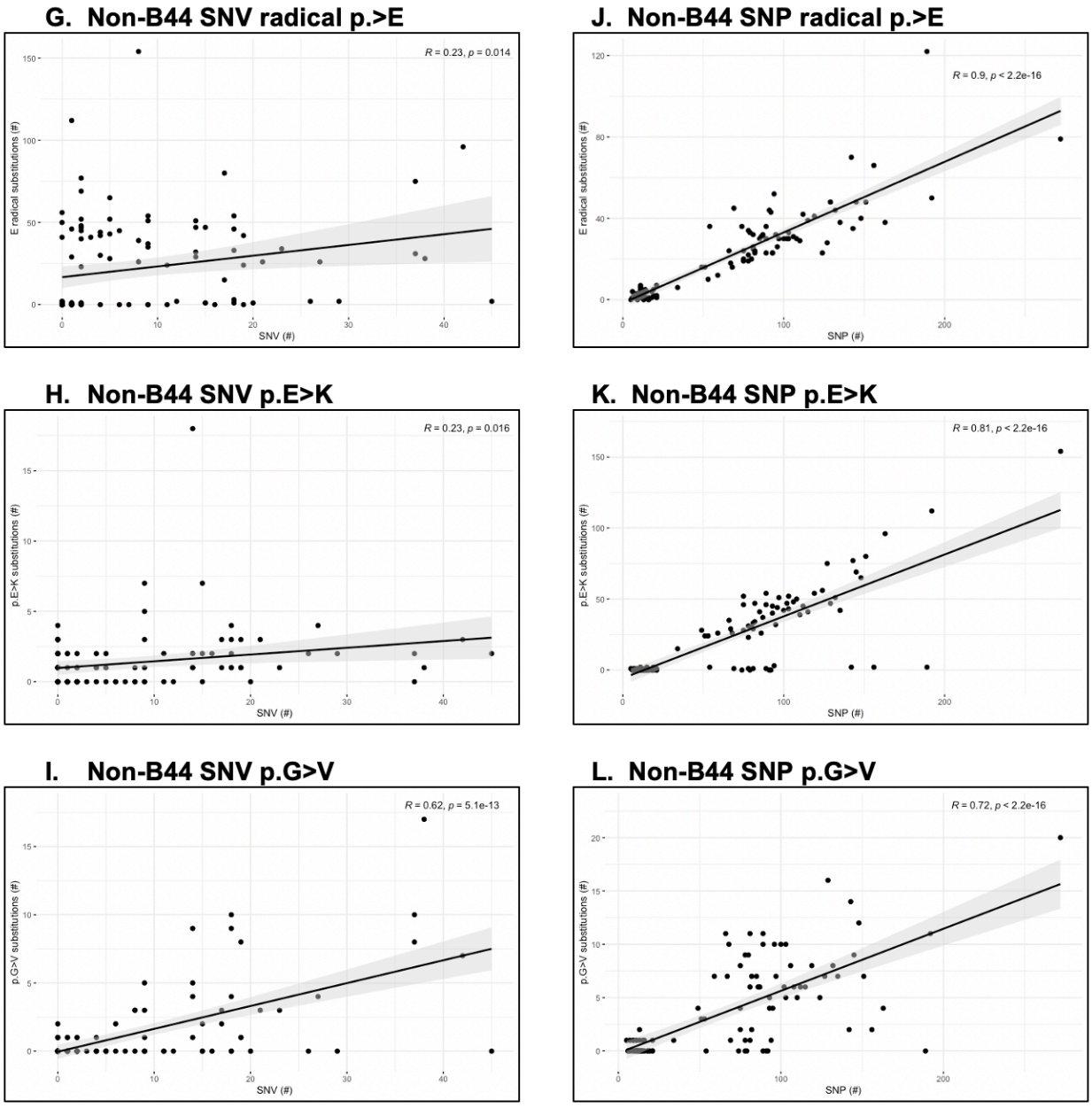
CPTAC-LUAD and LUSC grouped together and subset on B44 (N=67), correlations compared with Pearson's test, line of best fit shown in black surrounded by gray depicting 95% confidence interval.

A. SNV count vs. p.>E (radical), R= -0.008, P=0.950. D. SNP count vs. p.>E (radical), R=0.920, P<0.001.

B. SNV count vs. p.E>K, R=0.260, P=0.032. E. SNP count vs. p.E>K, R=0.860, P<0.001.

C. SNV count vs. p.G>V, R=0.750, P<0.001. F. SNP count vs. p.G>V, R=0.830, P<0.001.

Figure 5.12 (con't): HLA B44 supertype reveals immunoediting based on probabilistic relationships.



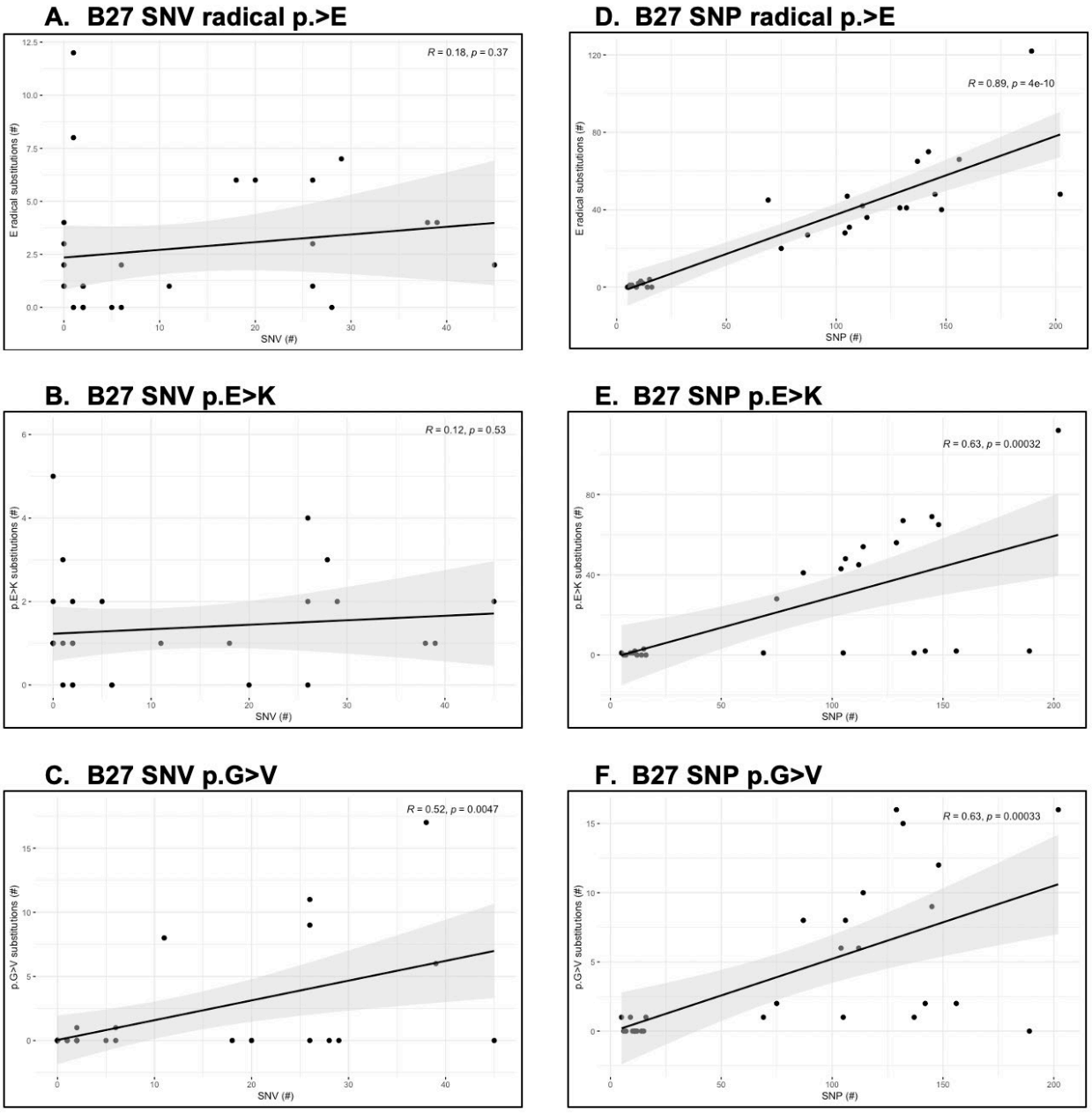
CPTAC-LUAD and LUSC grouped together and subset on non-B44 (N=110), correlations compared with Pearson's test, line of best fit shown in black surrounded by gray depicting 95% confidence interval.

G. SNV count vs. p.>E (radical), R= 0.230, P=0.014. J. SNP count vs. p.>E (radical), R=0.900, P<0.001.

H. SNV count vs. p.E>K, R=0.230, P=0.016. K. SNP count vs. p.E>K, R=0.810, P<0.001.

I. SNV count vs. p.G>V, R=0.620, P<0.001. L. SNP count vs. p.G>V, R=0.720, P<0.001.

Figure 5.13: HLA B27 supertype reveals immunoediting based on probabilistic relationships.



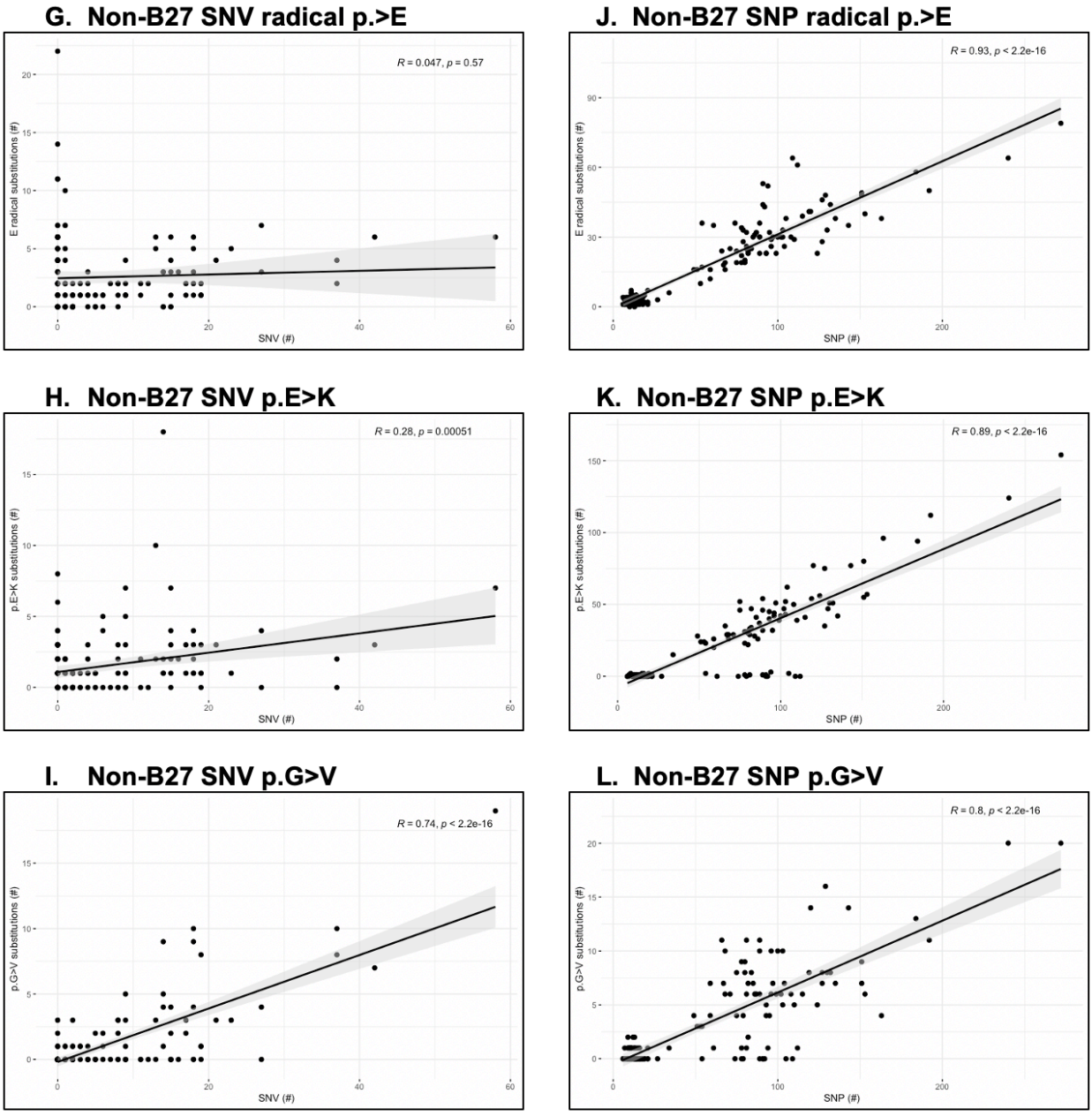
CPTAC-LUAD and LUSC grouped together and subset on B27 (N=28), correlations compared with Pearson's test, line of best fit shown in black surrounded by gray depicting 95% confidence interval.

A. SNV count vs. p.>E (radical), R= 0.180, P=0.370. D. SNP count vs. p.>E (radical), R=0.890, P<0.001.

B. SNV count vs. p.E>K, R=0.120, P=0.530. E. SNP count vs. p.E>K, R=0.630, P<0.001.

C. SNV count vs. p.G>V, R=0.520, P=0.005. F. SNP count vs. p.G>V, R=0.630, P<0.001.

Figure 5.13 (con't): HLA B27 supertype reveals immunoediting based on probabilistic relationships.



CPTAC-LUAD and LUSC grouped together and subset on non-B27 (N=149), correlations compared with Pearson's test, line of best fit shown in black surrounded by gray depicting 95% confidence interval.

G. SNV count vs. p.>E (radical), R= 0.047, P=0.570. J. SNP count vs. p.>E (radical), R=0.930, P<0.001.

H. SNV count vs. p.E>K, R=0.280, P=0.016. K. SNP count vs. p.E>K, R=0.890, P=0.010.

I. SNV count vs. p.G>V, R=0.740, P<0.001. L. SNP count vs. p.G>V, R=0.800, P<0.001.

frequency of ~2.5%, p.E>K is an opposite radical substitution with similar frequency at ~2.8%, p.G>V represents an uncharged substitution with the highest frequency at ~3.5%, see Fig. 3.6). Fig. 5.12 and 5.13 summarize results, which show number of events correlate significantly in every scenario except in radical glutamic acid substitutions caused by SNVs in B44 cases, and radical glutamic acid substitutions and p.E>K substitutions caused by SNVs in B27 cases. While radical mutations may be unfavorable overall based on SNV comparisons (SNV p.>E correlations range 0.05-0.2 excluding B44, SNV p.E>K correlations range 0.2-0.3 excluding B27, p.G>V correlations in SNV range 0.5-0.7, SNP correlations range 0.7-0.9 with $P<0.001$), glutamic acid substitutions are less than expected based on SNV protein expression in B44 patients with a correlation of -0.01 (distinctly not statistically significant and the only negative correlation in the dataset), and p.E>K substitutions are present in proportions less than expected based on SNV expression in B27 patients with a correlation of 0.12 (distinctly not statistically significant). This is a novel finding that provides the first evidence for immunoediting based on HLA-B supertypes and supports the functionality of motif neoepitopes in B44 and B27.

5.6 Conclusion

The application of techniques and known and uncovered relationships from Chapter 3 and 4 resolve the discrepancy in ICB survival based on HLA B44 in NSCLC and melanoma by considering the presence of favorable mutations that can be targeted by the adaptive immune system. Isolated feature assessment demonstrates how multiomic analyses may lead to conflicting results, compounded by increasing feature complexity and lack of power. The application of individual (personalized) models accounting for mutational signature resolve these discrepancies and suggest translational relevance. Refinement of currently available neoantigen prediction models based on “motif neoepitopes” suggests the first evidence for clinically relevant neoantigen prediction. Comparisons of protein-validated amino acid substitutions suggest subtle immunoediting effects based on HLA, which is a novel finding.

CHAPTER 6

Conclusion

This chapter summarizes the results and contributions from this dissertation. Based on the findings presented, we suggest research directions to further improve multiomic discovery and translation, particularly with respect to neoepitope prediction and HLA-based therapies.

6.1 Summary of Research

This dissertation advances our understanding of practical applications of multiomic-based model development and establishes techniques to improve reproducibility. In this dissertation, we provide the following research developments that address bias and type two errors in multiomic discovery:

1. **Inclusion of normal tissue in multiomic data structure.** This dissertation supports an expanded role for germline and normal tissue in cancer multiomic analyses. Multiomic data from normal tissue can provide a comparative model that does not require *a priori* knowledge of a biologic system. This can identify not only features that are statistically different, but also those that are expectedly different and unexpectedly similar. Using this framework, these experiments reveal technical artifacts and associations that otherwise could lead to spurious results, as well as demonstrate the relationship of probabilistic distributions to numbers of observed events, thereby enabling discernment of nuanced phenomena. To this point, models that correct for transcriptional bias and depict amino acid substitutions based on mutational signature created as a part of this work will be made openly available.
2. **Inclusion of similarity, iteration, and multiple metrics in model development.** Experiments in this dissertation suggest that distance measures and statistically significant differences may be misleading in complex analyses. We offer techniques that evaluate data similarities as opposed to differences and recommend iterative modeling to recapitulate observed data to assess scientific inference. The inclusion of two distance similarity metrics

(one sensitive to scale and one not) and a non-distance metric can be used to better understand model performance, particularly when using sparse or noisy datasets. These techniques are particularly effective when signal strength is unknown and/or when power calculations are not possible. Notably, the most important finding in this dissertation stemmed from a loss of statistical significance, which would have been overlooked by an analysis employing differential expression.

Chapters 3 and 4 presented key issues related to multiomic modeling with approaches that overcome these challenges. Demonstrations of this research were conducted in Chapter 5 using public datasets (e.g., TGCA, CPTAC) and specific institutional datasets obtained through UCLA Health/the Garon Lab to illustrate the problems and advances of these techniques over conventional approaches.

6.2 Future Directions

We identify limitations in works of this dissertation and subsequently suggest several directions for extending this work to improve the reliability of multiomic analyses and translational innovation.

Application of approach. Amino acid substitutions were a first model system and initially used for dimensionality reduction. While undoubtedly, models of amino acid substitutions could be refined, particularly with respect to prediction involving cases with low tumor mutation burden and more sparse datasets (and the unexpectedly high proportion of radical glutamic acid mutations in melanoma, which was unusual given how closely our model predicted other values), the approaches developed in this dissertation could be applied to other expanded or nuanced datasets and/or clinical problems. Evaluating targeted gene panels and/or microRNA, which to date have been notoriously difficult to incorporate in multiomic data structure, could enable additional insight and potentially lead to novel translational opportunities.

Additionally, as these experiments evaluate SNPs and SNVs that account for most cancer genomic changes – but do not account for other types of mutations, such as insertions, deletions, frameshifts, or splice site changes – an expanded view of cancer mutagenesis could be evaluated to determine compositional impact. For example, efforts of Kames and colleagues have created compendia for codon pair usage bias for multiple human tissues that cannot be predicted from codon frequencies alone, which could provide additional insight with respect to probable mutations.[252] Mutational signature also could be applied to amino acid substitution representations and/or codon pair usage tables to provide novel ways to identify likely neoantigens and/or likely resistance mechanisms.

Similarly, further defining the use of multiple similarity metrics could enhance multimodal analytic benchmarking. While in these experiments, the combination of metrics was useful for model insight, it is not clear that a combination would be useful for all multiomic experiments or applicable outside of multiomic experiments. One potential direction is to further investigate the utility of multiple metrics as a summary statistic that could be used to define the robustness of the information gained by a particular experiment or methodology. To support the implementation and adoption of this approach, well known datasets and established findings, such as those used by the Swanton group,[27] could be explored with respect to number of observations and power to further characterize metric performance. Alternatively, these metrics could be evaluated as an initial biomarker discovery tactic to suggest more robust signals.

Identification and translation of HLA-specific effects. The most biologically important discovery of this dissertation is the approach to identifying functionally relevant neoepitopes based on HLA supertypes. The focus on B44, in this regard, is not arbitrary. Approximately half of patients have B44, enabling statistically significant conclusions in modestly sized cohorts. Other HLA-B supertypes, such as B27, had significant survival benefits but could not be further defined due to smaller numbers. It is unclear why B27 motif neoepitopes were not more common in

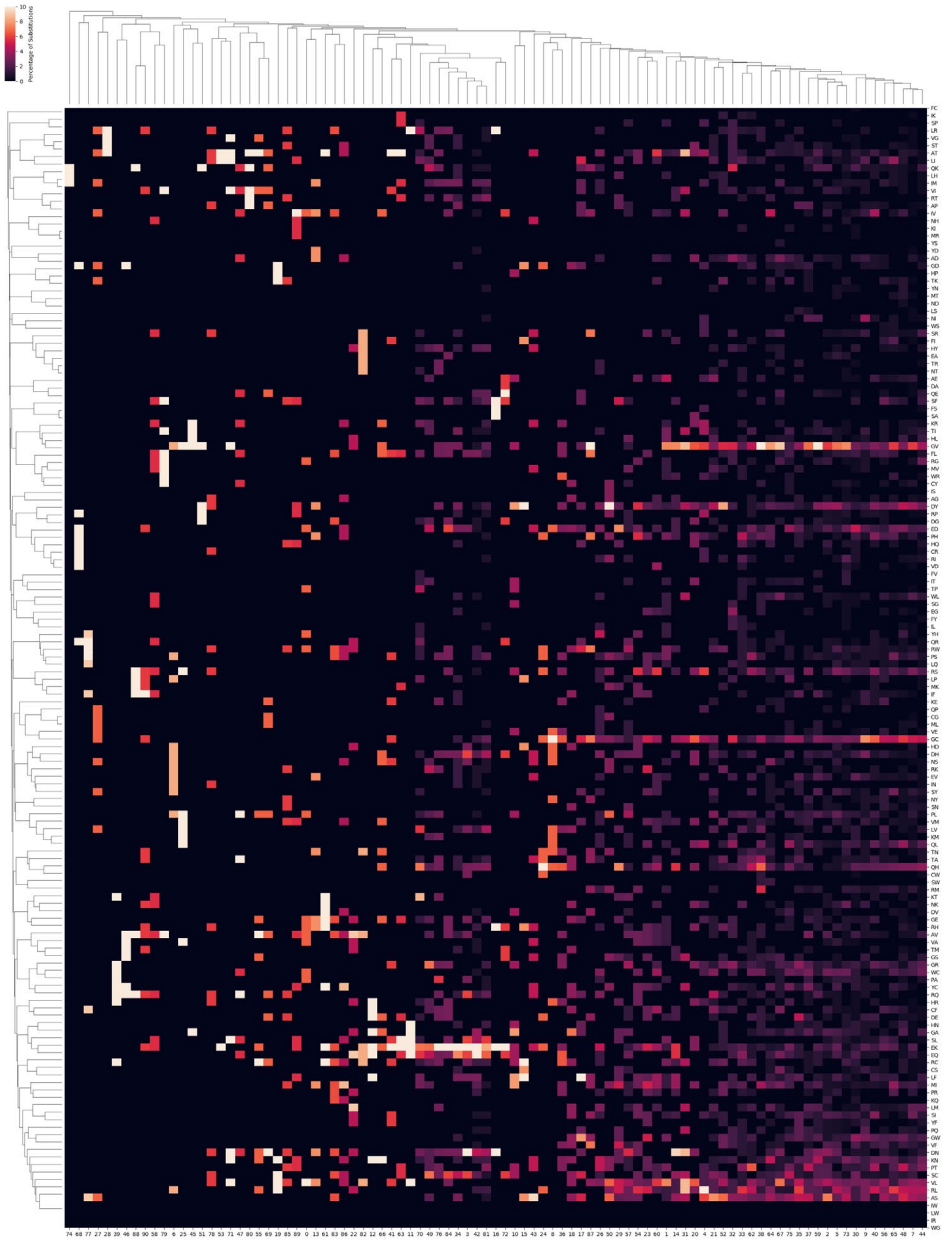
NSCLC, especially given the enrichment of positive radical substitutions, and currently available proteomic data is unable to provide additional information. NSCLC-specific purifying selection against these mutations is an exciting prospect, but additional evaluation is required. Additionally, given the strong autoimmune associations of B27, the possibility of SNPs or other somatic non-oncogenic mutations functioning similarly to motif neoepitopes is intriguing, as is whether B27 is enriched in those with significant ICB toxicities.

Incorporating recent work by Pataskar and colleagues that revealed tryptophan depletion results in tryptophan-to-phenylalanine substitutants, additional modifications can be made to neoepitope prediction approaches as phenylalanine is a common C-terminus for antigens.[253] Exposing organoid models to interferon-gamma also could potentially identify peptides that may be shared across patients that could be targeted by cancer vaccines or provided to dendritic cells to enhance immune recognition in other cellular therapeutic strategies. Other translational approaches that could be considered include the targeted use of chemotherapeutic agents that induce damage with particular signatures of DNA damage, such as temozolomide, which increases g.G>A mutations particularly in those with O6-methylguanine-DNA methyltransferase (MGMT) or other DNA repair deficiency,[170, 254] which could enhance the likelihood of substitutions leading to new neoepitope anchors. Further work in immunopeptidomics and proteomics with respect to HLA and motif neoepitopes is of great interest.

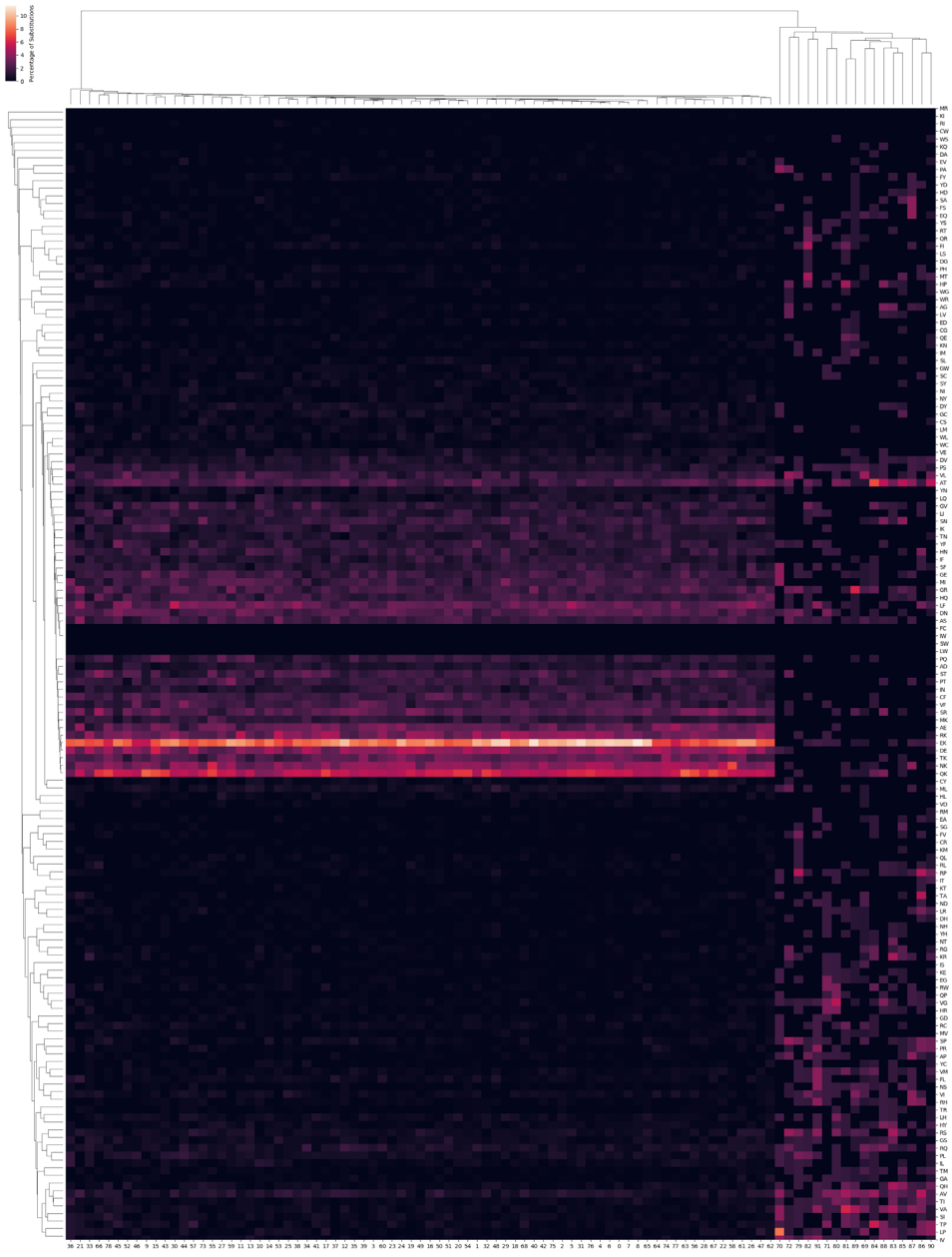
APPENDIX

7.1 Fig 3.12: Hierarchical clustering, CPTAC, high resolution.

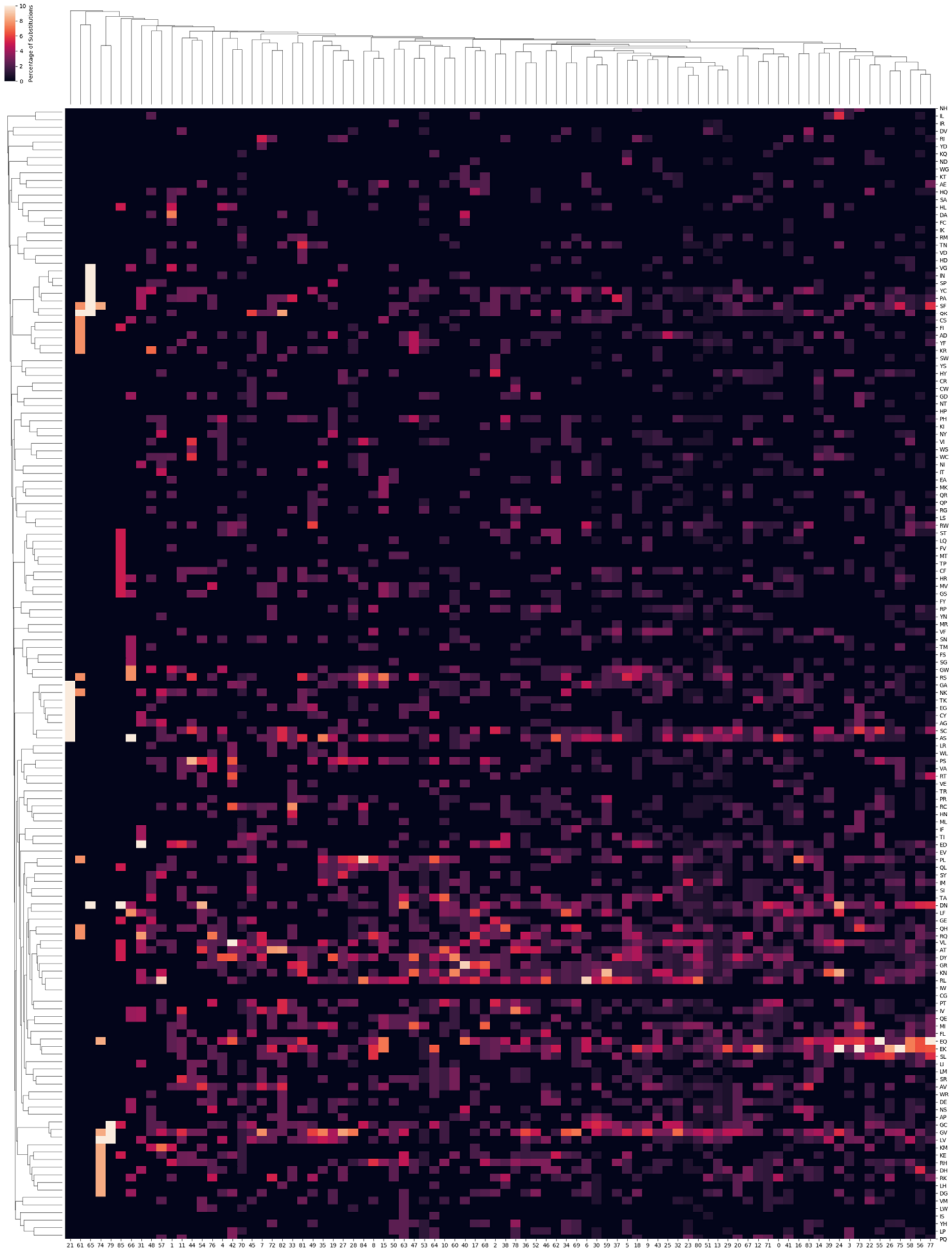
LUAD-SNV



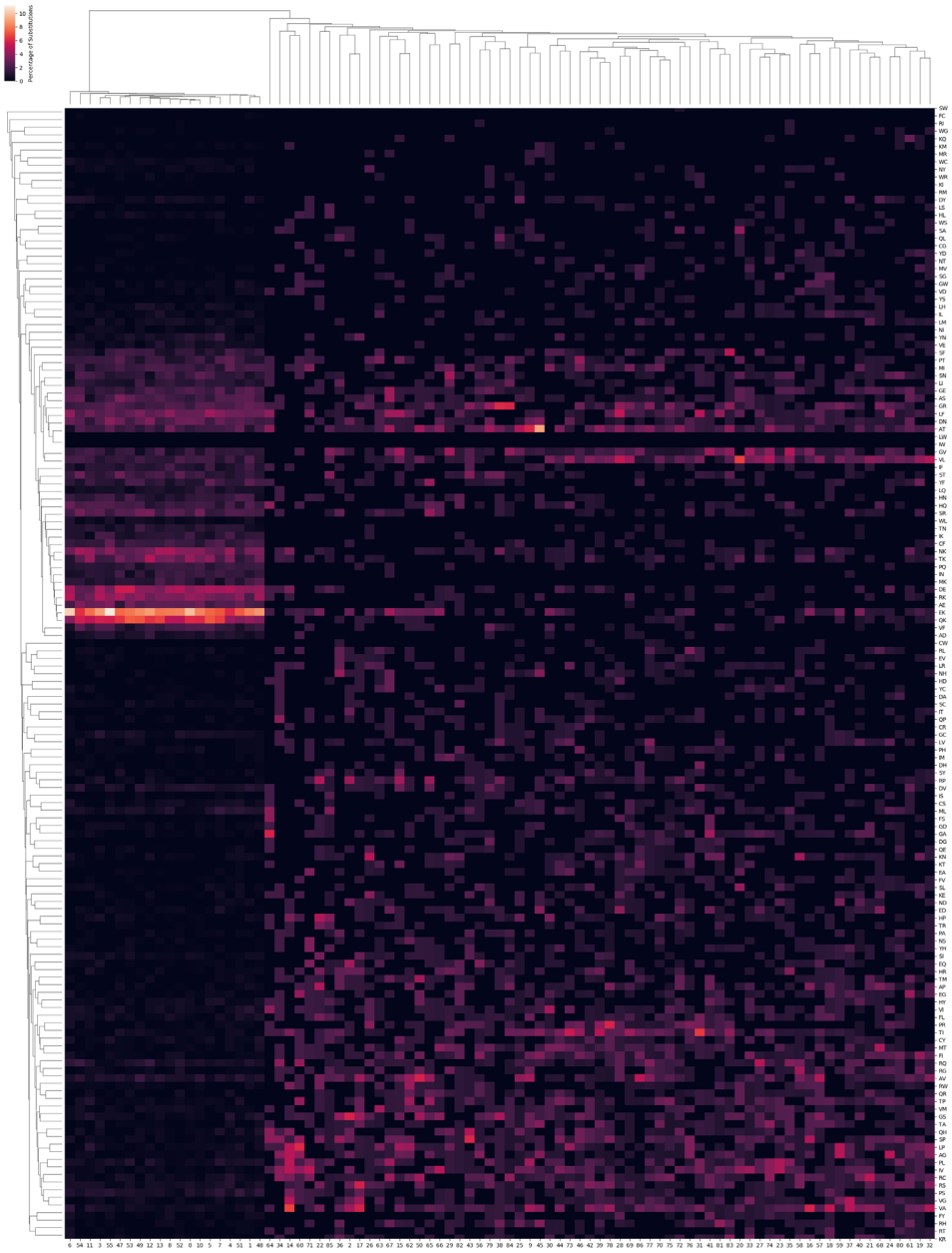
LUAD-SNP



LUSC-SNV

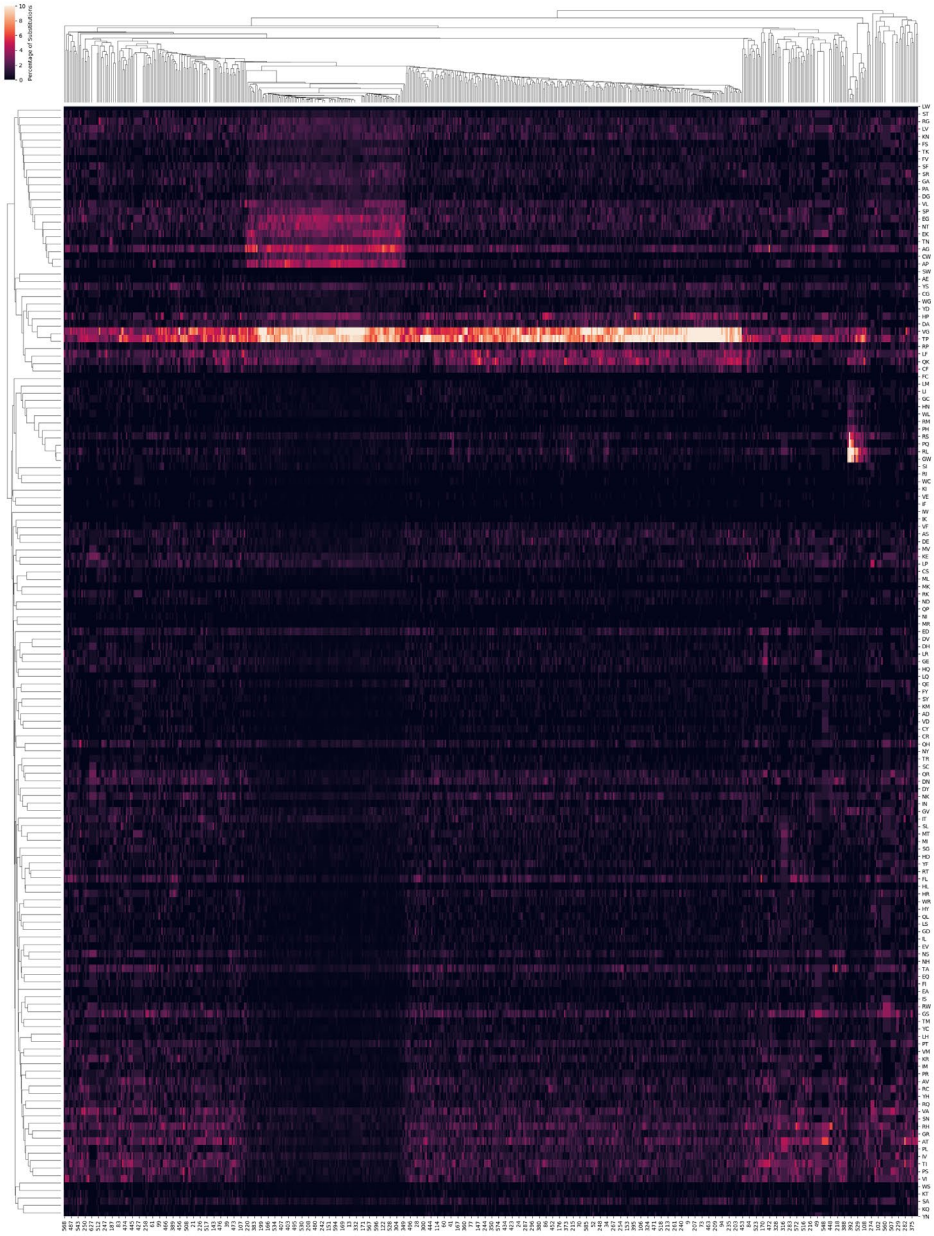


LUSC-SNP

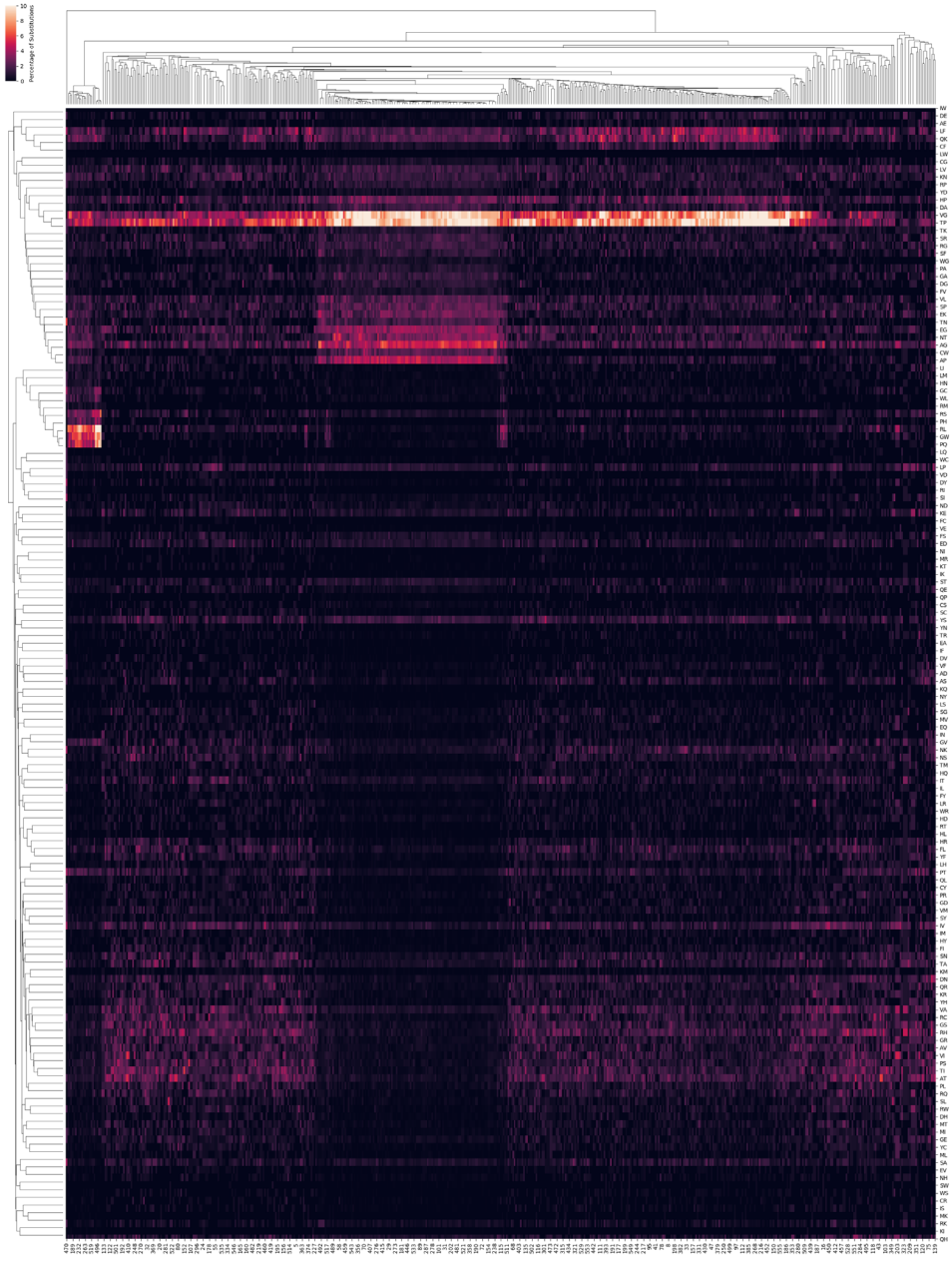


7.2 Fig 3.14: Hierarchical clustering, TCGA, high resolution

LUAD-SNP



LUSC-SNP



7.3 Experimental peptides (Chapter 5)

WT epitope	MT epitope	HLA Allele	MT	MT pos	MT IC_{50}	WT IC_{50}	Category
DAQKLLEKM	DEQKLLEKM	B*18:01	p.A>E	2	0.843	2.718	motif
DEELEQMLD	DEELEQMLY	B*18:01	p.D>Y	9	0.103	0.370	other
DQSLIYTLL	DESLIYTLL	B*18:01	p.Q>E	2	0.259	1.071	motif
EELEADTEY	EELEAHTEY	B*18:01	p.D>H	6	0.062	0.206	best
GEVAPSMFL	GEVAPRMFL	B*40:02	p.S>R	6	0.677	0.662	predicted/best
GEVAPSMFL	GEVAPDMFL	B*40:02	p.S>D	6	0.116	0.079	artificial
GEVAPSMFL	GEVAPEMFL	B*40:02	p.S>E	6	0.142	0.079	artificial
HEKVLNEAV	HEKVLNKAV	B*40:02	p.E>K	7	2.218	2.062	predicted/other
HEKVLNEAV	HEKVLNLAV	B*40:02	p.E>L	7	0.568	2.062	artificial
HEKVLNEAV	HEKVLNMAV	B*40:02	p.E>M	7	0.025	0.435	artificial
IAERYGFQY	IEERYGFQY	B*18:01	p.A>E	2	0.046	46.560	motif
KGPSDLLTV	KDPSDLLTV	B*40:02	p.G>D	2	0.430*	14.558*	artificial
KGPSDLLTV	KEPSDLLTV	B*40:02	p.G>E	2	0.602*	14.558*	predicted/motif
KGPSDLLTV	KHPSDLLTV	B*40:02	p.G>H	2	3.580*	6.821*	artificial
KGPSDLLTV	KKPSDLLTV	B*40:02	p.G>K	2	13.907*	6.821*	artificial
KGPSDLLTV	KRPSDLLTV	B*40:02	p.G>R	2	26.270*	6.821*	artificial
LALTAPRPY	LELTAPRPY	B*18:01	p.A>E	2	0.179	0.417	motif
LEIDNRLCL	LEIDHRLCL	B*40:02	p.N>H	5	0.001	1.870	predicted/other
MEMFLFFTA	MEDFLFFTA	B*40:02	p.M>D	3	0.080	10.949	artificial
MEMFLFFTA	MEEFLFFTA	B*40:02	p.M>E	3	0.182	10.949	artificial
MEMFLFFTA	MERFLFFTA	B*40:02	p.M>R	3	0.615	10.949	predicted/other
MEPGNNPIF	MESGNNPIF	B*18:01	p.P>S	3	0.038	0.062	best
NGLEIIWAE	NELEIIWAE	B*18:01	p.G>E	2	0.303	†	other
QEFWISQAS	QEFWISDAS	B*40:02	p.Q>D	7	0.006	0.002	artificial
QEFWISQAS	QEFWISEAS	B*40:02	p.Q>E	7	0.009	0.002	artificial
QEFWISQAS	QEFWISHAS	B*40:02	p.Q>H	7	0.003	0.002	predicted/other
QENEVLFTM	LENEVLFTM	B*18:01	p.Q>L	1	0.122	0.027	best
RERTMVSTR	RERDMVSTR	B*40:02	p.T>D	4	14.545	22.253	artificial
RERTMVSTR	REREMVSTR	B*40:02	p.T>E	4	9.448	15.344	artificial
RERTMVSTR	RERKMVSTR	B*40:02	p.T>K	4	10.617	15.344	predicted/other
RQFYLWTCL	RHFYLWTCL	B*40:02	p.Q>H	2	4.760	0.909	other
SELLKEFPY	SELLEEFPY	B*18:01	p.K>E	5	0.186	0.111	best
VENSFFLNV	VENSFFLEV	B*40:02	p.N>E	8	0.649	0.798	other
YDGAYAPVL	YEGAYAPVL	B*18:01	p.D>E	2	0.065	0.289	other

Standard single letter notation used for amino acids. All predicted neoepitopes had an $IC_{50} \leq 500$ nM. HLA – human leukocyte antigen, MT – mutant, p. – protein (amino acid), pos – position in peptide (1-9), WT – wildtype. IC_{50} displayed in picomolars (pM).

*Average of two experimental values included for calculations in which only one neoepitope met criteria for inclusion (see Methods). †No binding observed, not included in wildtype comparison – also note this peptide does not feature a known C-terminus and does not meet motif criteria.

7.4 Observed motif neoepitopes (Chapter 5)

ID	Chr:position	Gene	Motif	HLA	Sub*	P	MT epitope	WT epitope	$WTIC_{50}$
L01	19:21294033	ZNF708	B44	B*45:01	G>E	2	CEECGKAFA	CGECGKAFA	37.6
L02	10:128108409	MKI67	B44	B*44:02	K>E	2	TEWPKRSL	TKQWPKRSL	18.3
L05	17:18264369	MIEF2	B44	B*40:01	G>E	2	EELAGNLWL	EGLAGNLWL	24.1
L05	15:24678591	NPAP1	B44	B*40:01	K>E	2	QESDSSFIL	QKSDSSFIL	5.8
L05	2:162423340	KCNH7	B44	B*44:02	N>D	2	MDMVCMSVF	MNMVCMSVF	2.9
L08	5: 141211193	PCDHB12	B44	B*44:02	K>E	2	NEFKFLKPI	NKFKFLKPI	18.8
L13	11:105905218	GRIA4	B44	B*44:02	G>E	2	TENVQFDHY	TGNVQFDHY	22.4
L16*	2: 129980461	RAB6C	B44	B*37:01	G>E	2	EELAGNLWL	RGSDVIITL	12.4
L16*	2: 69840988	GMCL1	B44	B*37:01	K>E	2	QESDSSFIL	IKPSRVVAI	11.7
L17	1:77297600	AK5	B44	B*18:01	A>E	2	TENVQFDHY	IAERYGFQY	11.4
L17	11:49183163	FOLH1	B44	B*18:01	A>E	2	DEQKLLEKM	DAQKLLEKM	9.0
L17	17:75831123	UNC13D	B44	B*18:01	K>E	2	IEPSRVVAI	PKALHTATF	20.5
L17	20:9516068	LAMP5	B44	B*18:01	Q>E	2	EEFPYGIEA	SQSELQVFW	11.6
L17	5: 96783147	ERAP1	B44	B*18:01	Q>E	2	EELMELLAA	SQLLLLACV	9.5
L17	5:141414675	PCDHGA10	B44	B*18:01	Q>E	2	IEERYGFQY	IQGVPLSSY	7.0
L17	5: 181005746	BTNL3	B44	B*18:01	Q>E	2	DESLIYTLL	DQSLIYTLL	2.1
L21	11:122850063	BSX	B27	B*38:01	L>H	2	QHSGLEKRF	QLSGLEKRF	35.1
L22	1:209801358	IRF6	B27	B*39:01	D>H	2	VHSGLYPGL	VDSGLYPGL	20.8
L22	2:54844094	EML6	B44	B*44:02	G>E	2	AETQDSEIF	AGTQDSEIF	6.6
L24	1:74641334	ERICH3	B44	B*18:01	A>E	2	LELTAPRPY	LALTAPRPY	20.5
L24	17:75831123	UNC13D	B44	B*18:01	K>E	2	PEALHTATF	PKALHTATF	7.0
L24	5:141414675	PCDHGA10	B44	B*18:01	Q>E	2	IEGVPLSSY	IQGVPLSSY	21.8
L33	X:3343810	MXRA5	B44	B*40:01	V>E	2	WEALSVVLI	WGALSVVLI	2.9
L33	2:68513170	APLF	B44	B*40:01	G>E	2	SELEGSTEI	SQLEGSTEI	8.3
L33	19:51746899	FPR1	B44	B*40:01	G>E	2	LEFAVTFVL	LVFAVTFVL	29.1
L35	X:110173392	TMEM164	B44	B*40:01	G>E	2	AEPLCKYLL	AGPLCKYLL	23.6
L35	X:110173392	TMEM164	B44	B*44:02	G>E	2	AEPLCKYLL	AGPLCKYLL	20.3
L39	16:30967034	SETD1A	B44	B*44:02	G>E	2	REALRLPSF	RGALRLPSF	8.5
L47	8:69621089	SULF1	B44	B*40:02	G>E	2	KEPSDLLTV	KGPSDLLTV	16.8
L53	11:55994561	OR5F1	B27	B*27:05	E>K	2	LKLQIILFL	LELQIILFL	9.8
L53	X:19355426	PDHA1	B27	B*27:05	G>R	2	YRMGTSVER	YGMGTSVER	15.8
L53	9:110375436	SVEP1	B27	B*27:05	G>R	2	NRGRCVAPY	NGGRCVAPY	17.4
L53	15:42178275	VPS39	B44	B*40:01	A>E	2	METQIQQLL	MATQIQQLL	16.8
L53	3:172334967	FNDC3B	B44	B*40:01	K>E	2	GESPCPSEVL	GKSCPSEVL	3.0

Chr – chromosome, HLA – human leukocyte antigen, MT – mutant, P – peptide position, Sub – amino acid substitution, $WTIC_{50}$ – wildtype IC_{50} (in mM). Standard single letter notation used for amino acids.

*All substitutions are p. – protein (amino acid). All predicted neoepitopes had an $IC_{50} \leq 500$ nM.

REFERENCES

1. Linzer DI, Levine AJ. Characterization of a 54K dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected embryonal carcinoma cells. Cell. 1979;17(1):43-52. Epub 1979/05/01. doi: 10.1016/0092-8674(79)90293-9. PubMed PMID: 222475.

2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531-7. Epub 1999/10/16. doi: 10.1126/science.286.5439.531. PubMed PMID: 10521349.

3. Sgroi DC, Teng S, Robinson G, LeVangie R, Hudson JR, Jr., Elkahloun AG. In vivo gene expression profile analysis of human breast cancer progression. Cancer Res. 1999;59(22):5656-61. Epub 1999/12/03. PubMed PMID: 10582678.

4. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52. Epub 2000/08/30. doi: 10.1038/35021093. PubMed PMID: 10963602.

5. Wiestner A, Rosenwald A, Barry TS, Wright G, Davis RE, Henrickson SE, Zhao H, Ibbotson RE, Orchard JA, Davis Z, Stetler-Stevenson M, Raffeld M, Arthur DC, Marti GE, Wilson WH, Hamblin TJ, Oscier DG, Staudt LM. ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. Blood. 2003;101(12):4944-51. Epub 2003/02/22. doi: 10.1182/blood-2002-10-3306. PubMed PMID: 12595313.

6. Jarzab B, Wiench M, Fujarewicz K, Simek K, Jarzab M, Oczko-Wojciechowska M, Wloch J, Czarniecka A, Chmielik E, Lange D, Pawlaczek A, Szpak S, Gubala E, Swierniak A. Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications. Cancer Res. 2005;65(4):1587-97. Epub 2005/03/01. doi: 10.1158/0008-5472.CAN-04-3078. PubMed PMID: 15735049.

7. Lanza G, Ferracin M, Gafa R, Veronese A, Spizzo R, Pichiorri F, Liu CG, Calin GA, Croce CM, Negrini M. mRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer. Mol Cancer. 2007;6:54. Epub 2007/08/25. doi: 10.1186/1476-4598-6-54. PubMed PMID: 17716371; PMCID: PMC2048978.

8. de Souto MC, Costa IG, de Araujo DS, Ludermir TB, Schliep A. Clustering cancer gene expression data: a comparative study. BMC Bioinformatics. 2008;9:497. Epub 2008/11/29. doi: 10.1186/1471-2105-9-497. PubMed PMID: 19038021; PMCID: PMC2632677.

9. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science. 1987;235(4785):177-82. Epub 1987/01/09. doi: 10.1126/science.3798106. PubMed PMID: 3798106.

10. Carroll M, Ohno-Jones S, Tamura S, Buchdunger E, Zimmermann J, Lydon NB, Gilliland DG, Druker BJ. CGP 57148, a tyrosine kinase inhibitor, inhibits the growth of cells expressing BCR-ABL, TEL-ABL, and TEL-PDGFR fusion proteins. Blood. 1997;90(12):4947-52. Epub 1998/01/07. PubMed PMID: 9389713.

11. Rikova K, Guo A, Zeng Q, Possemato A, Yu J, Haack H, Nardone J, Lee K, Reeves C, Li Y, Hu Y, Tan Z, Stokes M, Sullivan L, Mitchell J, Wetzel R, Macneill J, Ren JM, Yuan J, Bakalarski CE, Villen J, Kornhauser JM, Smith B, Li D, Zhou X, Gygi SP, Gu TL, Polakiewicz RD, Rush J, Comb MJ. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell. 2007;131(6):1190-203. Epub 2007/12/18. doi: 10.1016/j.cell.2007.11.025. PubMed PMID: 18083107.

12. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, Baselga J, Norton L. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001;344(11):783-92. doi: 10.1056/NEJM200103153441101. PubMed PMID: 11248153.

13. Druker BJ, Sawyers CL, Kantarjian H, Resta DJ, Reese SF, Ford JM, Capdeville R, Talpaz M. Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. N Engl J Med. 2001;344(14):1038-42. doi: 10.1056/NEJM200104053441402. PubMed PMID: 11287973.

14. Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304(5676):1497-500. doi: 10.1126/science.1099314. PubMed PMID: 15118125.

15. Martinez-Jimenez F, Muinos F, Sentis I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H, Gonzalez-Perez A, Lopez-Bigas N. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20(10):555-72. Epub 20200810. doi: 10.1038/s41568-020-0290-x. PubMed PMID: 32778778.

16. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci U S A. 2003;100(15):8817-22. Epub 20030711. doi: 10.1073/pnas.1133470100. PubMed PMID: 12857956; PMCID: PMC166396.

17. Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, Li X, Marma MS, Shi S, Wu J, Edwards JR, Romu A, Turro NJ. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci U S A.

2006;103(52):19635-40. Epub 20061214. doi: 10.1073/pnas.0609513103. PubMed PMID: 17170132; PMCID: PMC1702316.

18. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007;39(12):1522-7. Epub 20071104. doi: 10.1038/ng.2007.42. PubMed PMID: 17982454.

19. Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH, Sano Marma M, Meng Q, Cao H, Li X, Shi S, Yu L, Kalachikov S, Russo JJ, Turro NJ, Ju J. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc Natl Acad Sci U S A. 2008;105(27):9145-50. Epub 20080630. doi: 10.1073/pnas.0804023105. PubMed PMID: 18591653; PMCID: PMC2442126.

20. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10(12):1213-8. Epub 20131006. doi: 10.1038/nmeth.2688. PubMed PMID: 24097267; PMCID: PMC3959825.

21. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, Eichler EE. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015;517(7536):608-11. Epub 20141110. doi: 10.1038/nature13907. PubMed PMID: 25383537; PMCID: PMC4317254.

22. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med. 2016;375(12):1109-12. doi: 10.1056/NEJMp1607591. PubMed PMID: 27653561; PMCID: PMC6309165.

23. Mertins P, Tang LC, Krug K, Clark DJ, Gritsenko MA, Chen L, Clauser KR, Clauss TR, Shah P, Gillette MA, Petyuk VA, Thomas SN, Mani DR, Mundt F, Moore RJ, Hu Y, Zhao R, Schnaubelt M, Keshishian H, Monroe ME, Zhang Z, Udeshi ND, Mani D, Davies SR, Townsend RR, Chan DW, Smith RD, Zhang H, Liu T, Carr SA. Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry. Nat Protoc. 2018;13(7):1632-61. doi: 10.1038/s41596-018-0006-9. PubMed PMID: 29988108; PMCID: PMC6211289.

24. Zhang Z, Hernandez K, Savage J, Li S, Miller D, Agrawal S, Ortuno F, Staudt LM, Heath A, Grossman RL. Uniform genomic data analysis in the NCI Genomic Data Commons. Nat Commun. 2021;12(1):1226. Epub 20210222. doi: 10.1038/s41467-021-21254-9. PubMed PMID: 33619257; PMCID: PMC7900240.

25. Salama AKS, Li S, Macrae ER, Park JI, Mitchell EP, Zwiebel JA, Chen HX, Gray RJ, McShane LM, Rubinstein LV, Patton D, Williams PM, Hamilton SR, Armstrong DK, Conley BA, Arteaga CL, Harris LN, O'Dwyer PJ, Chen AP, Flaherty KT. Dabrafenib and Trametinib in Patients With Tumors With BRAF(V600E) Mutations: Results of the NCI-MATCH Trial Subprotocol H. J Clin Oncol. 2020;38(33):3895-904. Epub 20200806. doi: 10.1200/JCO.20.00762. PubMed PMID: 32758030; PMCID: PMC7676884.

26. Middleton G, Robbins H, Andre F, Swanton C. A state-of-the-art review of stratified medicine in cancer: towards a future precision medicine strategy in cancer. Ann Oncol. 2022;33(2):143-57. Epub 20211119. doi: 10.1016/j.annonc.2021.11.004. PubMed PMID: 34808340.

27. Litchfield K, Reading JL, Puttick C, Thakkar K, Abbosh C, Bentham R, Watkins TBK, Rosenthal R, Biswas D, Rowan A, Lim E, Al Bakir M, Turati V, Guerra-Assuncao JA, Conde L, Furness AJS, Saini SK, Hadrup SR, Herrero J, Lee SH, Van Loo P, Enver T, Larkin J, Hellmann MD, Turajlic S, Quezada SA, McGranahan N, Swanton C. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. Cell. 2021;184(3):596-614 e14. Epub 20210127. doi: 10.1016/j.cell.2021.01.002. PubMed PMID: 33508232; PMCID: PMC7933824.

28. Petricoin EF, 3rd, Hackett JL, Lesko LJ, Puri RK, Gutman SI, Chumakov K, Woodcock J, Feigal DW, Jr., Zoon KC, Sistare FD. Medical applications of microarray technologies: a regulatory science perspective. Nat Genet. 2002;32 Suppl:474-9. Epub 2002/11/28. doi: 10.1038/ng1029. PubMed PMID: 12454641.

29. Ring BZ, Ross DT. Microarrays and molecular markers for tumor classification. Genome Biol. 2002;3(5):comment2005. Epub 2002/06/07. doi: 10.1186/gb-2002-3-5-comment2005. PubMed PMID: 12049658; PMCID: PMC139355.

30. Coughlin CR, 2nd, Scharer GH, Shaikh TH. Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. Genome Med. 2012;4(10):80. Epub 2012/11/02. doi: 10.1186/gm381. PubMed PMID: 23114084; PMCID: PMC3580449.

31. McShane LM. Statistical challenges in the development and evaluation of marker-based clinical tests. BMC Med. 2012;10:52. Epub 2012/05/31. doi: 10.1186/1741-7015-10-52. PubMed PMID: 22642713; PMCID: PMC3379945.

32. Tarca AL, Lauria M, Unger M, Bilal E, Boue S, Kumar Dey K, Hoeng J, Koeppl H, Martin F, Meyer P, Nandy P, Norel R, Peitsch M, Rice JJ, Romero R, Stolovitzky G, Talikka M, Xiang Y, Zechner C, Collaborators ID. Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge. Bioinformatics. 2013;29(22):2892-9. Epub 2013/08/24. doi: 10.1093/bioinformatics/btt492. PubMed PMID: 23966112; PMCID: PMC3810846.

33. Shi W, Ng CKY, Lim RS, Jiang T, Kumar S, Li X, Wali VB, Piscuoglio S, Gerstein MB, Chagpar AB, Weigelt B, Pusztai L, Reis-Filho JS, Hatzis C. Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity. Cell Rep. 2018;25(6):1446-57. Epub 2018/11/08. doi: 10.1016/j.celrep.2018.10.046. PubMed PMID: 30404001; PMCID: PMC6261536.

34. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science.

2006;313(5795):1929-35. Epub 2006/09/30. doi: 10.1126/science.1132939. PubMed PMID: 17008526.

35. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, Lasorella A, Aldape K, Califano A, Iavarone A. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 2010;463(7279):318-25. Epub 2009/12/25. doi: 10.1038/nature08712. PubMed PMID: 20032975; PMCID: PMC4011561.

36. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26(12):i237-45. Epub 2010/06/10. doi: 10.1093/bioinformatics/btq182. PubMed PMID: 20529912; PMCID: PMC2881367.

37. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001;98(24):13790-5. Epub 2001/11/15. doi: 10.1073/pnas.191502998. PubMed PMID: 11707567; PMCID: PMC61120.

38. Catto JW, Linkens DA, Abbod MF, Chen M, Burton JL, Feeley KM, Hamdy FC. Artificial intelligence in predicting bladder cancer outcome: a comparison of neuro-fuzzy modeling and artificial neural networks. Clin Cancer Res. 2003;9(11):4172-7. Epub 2003/10/02. PubMed PMID: 14519642.

39. Yuan Y, Savage RS, Markowetz F. Patient-specific data fusion defines prognostic cancer subtypes. PLoS Comput Biol. 2011;7(10):e1002227. Epub 2011/10/27. doi: 10.1371/journal.pcbi.1002227. PubMed PMID: 22028636; PMCID: PMC3197649.

40. Ren S, Shao Y, Zhao X, Hong CS, Wang F, Lu X, Li J, Ye G, Yan M, Zhuang Z, Xu C, Xu G, Sun Y. Integration of Metabolomics and Transcriptomics Reveals Major Metabolic Pathways and Potential Biomarker Involved in Prostate Cancer. Mol Cell Proteomics. 2016;15(1):154-63. Epub 2015/11/08. doi: 10.1074/mcp.M115.052381. PubMed PMID: 26545398; PMCID: PMC4762514.

41. Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. Biomed Res Int. 2018;2018:9836256. Epub 2018/11/08. doi: 10.1155/2018/9836256. PubMed PMID: 30402498; PMCID: PMC6192166.

42. Jiang X, Wells A, Brufsky A, Neapolitan R. A clinical decision support system learned from data to personalize treatment recommendations towards preventing breast cancer metastasis. PLoS One. 2019;14(3):e0213292. Epub 2019/03/09. doi: 10.1371/journal.pone.0213292. PubMed PMID: 30849111; PMCID: PMC6407919.

43. Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. Front Oncol.

2020;10:1030. Epub 20200630. doi: 10.3389/fonc.2020.01030. PubMed PMID: 32695678; PMCID: PMC7338582.

44. Cai L, Yuan W, Zhang Z, He L, Chou KC. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Sci Rep. 2016;6:36540. Epub 2016/11/23. doi: 10.1038/srep36540. PubMed PMID: 27874022; PMCID: PMC5118795.

45. Research AM. Immune checkpoint inhibitors market by type and application: global opportunity analysis and industry forecase, 2021-2030 2021 [cited 2022 May 25]. Available from: https://www.alliedmarketresearch.com/immune-check-point-inhibitors-market.

46. Wang DY, Salem JE, Cohen JV, Chandra S, Menzer C, Ye F, Zhao S, Das S, Beckermann KE, Ha L, Rathmell WK, Ancell KK, Balko JM, Bowman C, Davis EJ, Chism DD, Horn L, Long GV, Carlino MS, Lebrun-Vignes B, Eroglu Z, Hassel JC, Menzies AM, Sosman JA, Sullivan RJ, Moslehi JJ, Johnson DB. Fatal Toxic Effects Associated With Immune Checkpoint Inhibitors: A Systematic Review and Meta-analysis. JAMA Oncol. 2018;4(12):1721-8. doi: 10.1001/jamaoncol.2018.3923. PubMed PMID: 30242316; PMCID: PMC6440712.

47. Haslam A, Prasad V. Estimation of the Percentage of US Patients With Cancer Who Are Eligible for and Respond to Checkpoint Inhibitor Immunotherapy Drugs. JAMA Netw Open. 2019;2(5):e192535. Epub 20190503. doi: 10.1001/jamanetworkopen.2019.2535. PubMed PMID: 31050774; PMCID: PMC6503493.

48. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74. doi: 10.1016/j.cell.2011.02.013. PubMed PMID: 21376230.

49. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F,

Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowki J, International Human Genome Sequencing C. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921. Epub 2001/03/10. doi: 10.1038/35057062. PubMed PMID: 11237011.

50. Waddington CH. Preliminary Notes on the Development of the Wings in Normal and Mutant Strains of Drosophila. Proc Natl Acad Sci U S A. 1939;25(7):299-307. Epub 1939/07/01. doi: 10.1073/pnas.25.7.299. PubMed PMID: 16577903; PMCID: PMC1077909.

51. Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature. 1983;301(5895):89-92. Epub 1983/01/06. doi: 10.1038/301089a0. PubMed PMID: 6185846.

52. Holliday R. The inheritance of epigenetic defects. Science. 1987;238(4824):163-70. Epub 1987/10/09. doi: 10.1126/science.3310230. PubMed PMID: 3310230.

53. Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell. 1993;75(5):855-62. Epub 1993/12/03. doi: 10.1016/0092-8674(93)90530-4. PubMed PMID: 8252622.

54. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75(5):843-54. Epub 1993/12/03. doi: 10.1016/0092-8674(93)90529-y. PubMed PMID: 8252621.

55. Merlo A, Herman JG, Mao L, Lee DJ, Gabrielson E, Burger PC, Baylin SB, Sidransky D. 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. Nat Med. 1995;1(7):686-92. Epub 1995/07/01. doi: 10.1038/nm0795-686. PubMed PMID: 7585152.

56. Sager R. Expression genetics in cancer: shifting the focus from DNA to RNA. Proc Natl Acad Sci U S A. 1997;94(3):952-5. Epub 1997/02/04. doi: 10.1073/pnas.94.3.952. PubMed PMID: 9023363; PMCID: PMC19620.

57. Seligson DB, Horvath S, Shi T, Yu H, Tze S, Grunstein M, Kurdistani SK. Global histone modification patterns predict risk of prostate cancer recurrence. Nature. 2005;435(7046):1262-6. Epub 2005/07/01. doi: 10.1038/nature03672. PubMed PMID: 15988529.

58. Halford S, Rowan A, Sawyer E, Talbot I, Tomlinson I. O(6)-methylguanine methyltransferase in colorectal cancers: detection of mutations, loss of expression, and weak association with G:C>A:T transitions. Gut. 2005;54(6):797-802. Epub 2005/05/13. doi: 10.1136/gut.2004.059535. PubMed PMID: 15888787; PMCID: PMC1774551.

59. Saito Y, Liang G, Egger G, Friedman JM, Chuang JC, Coetzee GA, Jones PA. Specific activation of microRNA-127 with downregulation of the proto-oncogene BCL6 by chromatin-modifying drugs in human cancer cells. Cancer Cell. 2006;9(6):435-43. Epub 2006/06/13. doi: 10.1016/j.ccr.2006.04.020. PubMed PMID: 16766263.

60. Joshi R, Passner JM, Rohs R, Jain R, Sosinsky A, Crickmore MA, Jacob V, Aggarwal AK, Honig B, Mann RS. Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell. 2007;131(3):530-43. Epub 2007/11/06. doi: 10.1016/j.cell.2007.09.024. PubMed PMID: 17981120; PMCID: PMC2709780.

61. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. The role of DNA shape in protein-DNA recognition. Nature. 2009;461(7268):1248-53. Epub 2009/10/30. doi: 10.1038/nature08473. PubMed PMID: 19865164; PMCID: PMC2793086.

62. Lange SS, Takata K, Wood RD. DNA polymerases and cancer. Nat Rev Cancer. 2011;11(2):96-110. Epub 2011/01/25. doi: 10.1038/nrc2998. PubMed PMID: 21258395; PMCID: PMC3739438.

63. Lazarovici A, Zhou T, Shafer A, Dantas Machado AC, Riley TR, Sandstrom R, Sabo PJ, Lu Y, Rohs R, Stamatoyannopoulos JA, Bussemaker HJ. Probing DNA shape and methylation state on a genomic scale with DNase I. Proc Natl Acad Sci U S A. 2013;110(16):6376-81. Epub 2013/04/12. doi: 10.1073/pnas.1216822110. PubMed PMID: 23576721; PMCID: PMC3631675.

64. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016;17 Suppl 2:15. Epub 20160120. doi: 10.1186/s12859-015-0857-9. PubMed PMID: 26821531; PMCID: PMC4959355.

65. Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol. 2006;7(3):198-210. doi: 10.1038/nrm1857. PubMed PMID: 16496022.

66. Silverbush D, Cristea S, Yanovich-Arad G, Geiger T, Beerenwinkel N, Sharan R. Simultaneous Integration of Multi-omics Data Improves the Identification of Cancer Driver Modules. Cell Syst. 2019;8(5):456-66 e5. Epub 20190515. doi: 10.1016/j.cels.2019.04.005. PubMed PMID: 31103572.

67. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56-68. doi: 10.1038/nrg2918. PubMed PMID: 21164525; PMCID: PMC3140052.

68. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol. 2000;7(3-4):601-20. doi: 10.1089/106652700750050961. PubMed PMID: 11108481.

69. Kristensen VN, Lingjaerde OC, Russnes HG, Vollan HK, Frigessi A, Borresen-Dale AL. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer. 2014;14(5):299-313. doi: 10.1038/nrc3721. PubMed PMID: 24759209.

70. Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics. 2011;27(6):887-8. Epub 2011/01/14. doi: 10.1093/bioinformatics/btr019. PubMed PMID: 21228048.

71. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, Ladanyi M, Sander C. Integrative subtype discovery in glioblastoma using iCluster. PLoS One. 2012;7(4):e35236. Epub 20120423. doi: 10.1371/journal.pone.0035236. PubMed PMID: 22539962; PMCID: PMC3335101.

72. Nowak G, Hastie T, Pollack JR, Tibshirani R. A fused lasso latent feature model for analyzing multi-sample aCGH data. Biostatistics. 2011;12(4):776-91. Epub 20110603. doi: 10.1093/biostatistics/kxr012. PubMed PMID: 21642389; PMCID: PMC3169672.

73. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. Nat Genet. 2004;36(10):1090-8. Epub 20040926. doi: 10.1038/ng1434. PubMed PMID: 15448693.

74. Mayer CD, Lorent J, Horgan GW. Exploratory analysis of multiple omics datasets using the adjusted RV coefficient. Stat Appl Genet Mol Biol. 2011;10:Article 14. doi: 10.2202/1544-6115.1540. PubMed PMID: 21381439.

75. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc. 1995;57:289-300.

76. Tarazona S, Balzano-Nogueira L, Gomez-Cabrero D, Schmidt A, Imhof A, Hankemeier T, Tegner J, Westerhuis JA, Conesa A. Harmonization of quality metrics and power calculation in multi-omic studies. Nat Commun. 2020;11(1):3092. Epub 20200618. doi: 10.1038/s41467-020-16937-8. PubMed PMID: 32555183; PMCID: PMC7303201.

77. Rainer J, Gatto L, Weichenberger CX. ensembldb: an R package to create and use Ensembl-based annotation resources. Bioinformatics. 2019;35(17):3151-3. doi: 10.1093/bioinformatics/btz031. PubMed PMID: 30689724; PMCID: PMC6736197.

78. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733-45. Epub 20151108. doi: 10.1093/nar/gkv1189. PubMed PMID: 26553804; PMCID: PMC4702849.

79. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T. A uniform system for microRNA annotation. RNA. 2003;9(3):277-9. doi: 10.1261/rna.2183803. PubMed PMID: 12592000; PMCID: PMC1370393.

80. Sindhu KJ, Venkatesan N, Karunagaran D. MicroRNA Interactome Multiomics Characterization for Cancer Research and Personalized Medicine: An Expert Review. OMICS. 2021;25(9):545-66. Epub 20210826. doi: 10.1089/omi.2021.0087. PubMed PMID: 34448651.

81. Rudnick PA, Markey SP, Roth J, Mirokhin Y, Yan X, Tchekhovskoi DV, Edwards NJ, Thangudu RR, Ketchum KA, Kinsinger CR, Mesri M, Rodriguez H, Stein SE. A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline. J Proteome Res. 2016;15(3):1023-32. Epub 20160225. doi: 10.1021/acs.jproteome.5b01091. PubMed PMID: 26860878; PMCID: PMC5117628.

82. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. Ensembl 2018. Nucleic Acids Res. 2018;46(D1):D754-D61. Epub 2017/11/21. doi: 10.1093/nar/gkx1098. PubMed PMID: 29155950; PMCID: PMC5753206.

83. Technology NIoSa. Software for peptide mass spectral libraries 2022. Available from: https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:pepsoftware.

84. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun. 2014;5:5277. Epub 20141031. doi: 10.1038/ncomms6277. PubMed PMID: 25358478; PMCID: PMC5036525.

85. Zhou W, Triche TJ, Jr., Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res. 2018;46(20):e123. doi: 10.1093/nar/gky691. PubMed PMID: 30085201; PMCID: PMC6237738.

86. Veeneman BA, Shukla S, Dhanasekaran SM, Chinnaiyan AM, Nesvizhskii AI. Two-pass alignment improves novel splice junction quantification. Bioinformatics. 2016;32(1):43-9. Epub 20151030. doi: 10.1093/bioinformatics/btv642. PubMed PMID: 26519505; PMCID: PMC5006238.

87. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20. Epub 2013/09/28. doi: 10.1038/ng.2764. PubMed PMID: 24071849; PMCID: PMC3919969.

88. Hess JF, Kohl TA, Kotrova M, Ronsch K, Paprotka T, Mohr V, Hutzenlaub T, Bruggemann M, Zengerle R, Niemann S, Paust N. Library preparation for next generation sequencing: A review of automation strategies. Biotechnol Adv. 2020;41:107537. Epub 20200319. doi: 10.1016/j.biotechadv.2020.107537. PubMed PMID: 32199980.

89. Cummings AL, Gukasyan J, Lu HY, Grogan T, Sunga G, Fares CM, Hornstein N, Zaretsky J, Carroll J, Bachrach B, Akingbemi WO, Li D, Noor Z, Lisberg A, Goldman JW, Elashoff D, Bui AAT, Ribas A, Dubinett SM, Rossetti M, Garon EB. Mutational landscape influences immunotherapy outcomes among patients with non-small-cell lung cancer with human leukocyte antigen supertype B44. Nat Cancer. 2020;1(12):1167-75. Epub 20201116. doi: 10.1038/s43018-020-00140-1. PubMed PMID: 35121931.

90. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-

Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9. Epub 2008/11/07. doi: 10.1038/nature07517. PubMed PMID: 18987734; PMCID: PMC2581791.

91. Webb-Robertson BJ, Wiberg HK, Matzke MM, Brown JN, Wang J, McDermott JE, Smith RD, Rodland KD, Metz TO, Pounds JG, Waters KM. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. J Proteome Res. 2015;14(5):1993-2001. Epub 20150422. doi: 10.1021/pr501138h. PubMed PMID: 25855118; PMCID: PMC4776766.

92. McAlister GC, Nusinow DP, Jedrychowski MP, Wuhr M, Huttlin EL, Erickson BK, Rad R, Haas W, Gygi SP. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Anal Chem. 2014;86(14):7150-8. Epub 20140703. doi: 10.1021/ac502040v. PubMed PMID: 24927332; PMCID: PMC4215866.

93. Huang T, Choi M, Tzouros M, Golling S, Pandya NJ, Banfai B, Dunkley T, Vitek O. MSstatsTMT: Statistical Detection of Differentially Abundant Proteins in Experiments with Isobaric Labeling and Multiple Mixtures. Mol Cell Proteomics. 2020;19(10):1706-23. Epub 20200717. doi: 10.1074/mcp.RA120.002105. PubMed PMID: 32680918; PMCID: PMC8015007.

94. O'Connell JD, Paulo JA, O'Brien JJ, Gygi SP. Proteome-Wide Evaluation of Two Common Protein Quantification Methods. J Proteome Res. 2018;17(5):1934-42. Epub 20180419. doi: 10.1021/acs.jproteome.8b00016. PubMed PMID: 29635916; PMCID: PMC5984592.

95. Brenes A, Hukelmann J, Bensaddek D, Lamond AI. Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. Mol Cell Proteomics. 2019;18(10):1967-80. Epub 20190722. doi: 10.1074/mcp.RA119.001472. PubMed PMID: 31332098; PMCID: PMC6773557.

96. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007;4(3):207-14. doi: 10.1038/nmeth1019. PubMed PMID: 17327847.

97. Buckley AR, Standish KA, Bhutani K, Ideker T, Lasken RS, Carter H, Harismendy O, Schork NJ. Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls. BMC Genomics. 2017;18(1):458. Epub 20170612. doi: 10.1186/s12864-017-3770-y. PubMed PMID: 28606096; PMCID: PMC5467262.

98. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, Gunderson KL. Genome-wide DNA methylation profiling using Infinium(R) assay. Epigenomics. 2009;1(1):177-200. doi: 10.2217/epi.09.14. PubMed PMID: 22122642.

99. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587. Epub 2010/12/02. doi: 10.1186/1471-2105-11-587. PubMed PMID: 21118553; PMCID: PMC3012676.

100. Triche TJ, Jr., Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 2013;41(7):e90. Epub 20130309. doi: 10.1093/nar/gkt090. PubMed PMID: 23476028; PMCID: PMC3627582.

101. Poplawski A, Binder H. Feasibility of sample size calculation for RNA-seq studies. Brief Bioinform. 2018;19(4):713-20. doi: 10.1093/bib/bbw144. PubMed PMID: 28100468.

102. Boutros PC, Margolin AA, Stuart JM, Califano A, Stolovitzky G. Toward better benchmarking: challenge-based methods assessment in cancer genomics. Genome Biol. 2014;15(9):462. Epub 20140917. doi: 10.1186/s13059-014-0462-7. PubMed PMID: 25314947; PMCID: PMC4318527.

103. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018;46(20):10546-62. doi: 10.1093/nar/gky889. PubMed PMID: 30295871; PMCID: PMC6237755.

104. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061-8. Epub 2008/09/06. doi: 10.1038/nature07385. PubMed PMID: 18772890; PMCID: PMC2671642.

105. Huang FW, Mosquera JM, Garofalo A, Oh C, Baco M, Amin-Mansour A, Rabasha B, Bahl S, Mullane SA, Robinson BD, Aldubayan S, Khani F, Karir B, Kim E, Chimene-Weiss J, Hofree M, Romanel A, Osborne JR, Kim JW, Azabdaftari G, Woloszynska-Read A, Sfanos K, De Marzo AM, Demichelis F, Gabriel S, Van Allen EM, Mesirov J, Tamayo P, Rubin MA, Powell IJ, Garraway LA. Exome Sequencing of African-American Prostate Cancer Reveals Loss-of-Function ERF Mutations. Cancer Discov. 2017;7(9):973-83. Epub 20170517. doi: 10.1158/2159-8290.CD-16-0960. PubMed PMID: 28515055; PMCID: PMC5836784.

106. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief

Bioinform. 2016;17(4):628-41. Epub 20160311. doi: 10.1093/bib/bbv108. PubMed PMID: 26969681; PMCID: PMC4945831.

107. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Graf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Group M, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Borresen-Dale AL, Brenton JD, Tavare S, Caldas C, Aparicio S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346-52. Epub 20120418. doi: 10.1038/nature10983. PubMed PMID: 22522925; PMCID: PMC3440846.

108. Liu Q, Li C, Wanga V, Shepherd BE. Covariate-adjusted Spearman's rank correlation with probability-scale residuals. Biometrics. 2018;74(2):595-605. Epub 20171113. doi: 10.1111/biom.12812. PubMed PMID: 29131931; PMCID: PMC5949238.

109. Quigley DA, To MD, Perez-Losada J, Pelorosso FG, Mao JH, Nagase H, Ginzinger DG, Balmain A. Genetic architecture of mouse skin inflammation and tumour susceptibility. Nature. 2009;458(7237):505-8. Epub 20090111. doi: 10.1038/nature07683. PubMed PMID: 19136944; PMCID: PMC4460995.

110. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012;40(Database issue):D1301-7. Epub 20111116. doi: 10.1093/nar/gkr1074. PubMed PMID: 22096230; PMCID: PMC3245032.

111. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50. Epub 20050930. doi: 10.1073/pnas.0506580102. PubMed PMID: 16199517; PMCID: PMC1239896.

112. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics. 2008;24(13):i223-31. doi: 10.1093/bioinformatics/btn161. PubMed PMID: 18586718; PMCID: PMC2718639.

113. Omberg L, Ellrott K, Yuan Y, Kandoth C, Wong C, Kellen MR, Friend SH, Stuart J, Liang H, Margolin AA. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. Nat Genet. 2013;45(10):1121-6. doi: 10.1038/ng.2761. PubMed PMID: 24071850; PMCID: PMC3950337.

114. Krassowski M, Das V, Sahu SK, Misra BB. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Front Genet. 2020;11:610798. Epub 20201210. doi: 10.3389/fgene.2020.610798. PubMed PMID: 33362867; PMCID: PMC7758509.

115. Fagan A, Culhane AC, Higgins DG. A multivariate analysis approach to the integration of proteomic and gene expression data. Proteomics. 2007;7(13):2162-71. doi: 10.1002/pmic.200600898. PubMed PMID: 17549791.

116. Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. Molecular mechanisms of disease-causing missense mutations. J Mol Biol. 2013;425(21):3919-36. Epub 20130716. doi: 10.1016/j.jmb.2013.07.014. PubMed PMID: 23871686; PMCID: PMC3796015.

117. Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K, Petralia F, Li Y, Liang WW, Reva B, Krek A, Ji J, Song X, Liu W, Hong R, Yao L, Blumenberg L, Savage SR, Wendl MC, Wen B, Li K, Tang LC, MacMullan MA, Avanessian SC, Kane MH, Newton CJ, Cornwell M, Kothadia RB, Ma W, Yoo S, Mannan R, Vats P, Kumar-Sinha C, Kawaler EA, Omelchenko T, Colaprico A, Geffen Y, Maruvka YE, da Veiga Leprevost F, Wiznerowicz M, Gumus ZH, Veluswamy RR, Hostetter G, Heiman DI, Wyczalkowski MA, Hiltke T, Mesri M, Kinsinger CR, Boja ES, Omenn GS, Chinnaiyan AM, Rodriguez H, Li QK, Jewell SD, Thiagarajan M, Getz G, Zhang B, Fenyo D, Ruggles KV, Cieslik MP, Robles AI, Clauser KR, Govindan R, Wang P, Nesvizhskii AI, Ding L, Mani DR, Carr SA, Clinical Proteomic Tumor Analysis C. Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. Cell. 2020;182(1):200-25 e35. doi: 10.1016/j.cell.2020.06.013. PubMed PMID: 32649874; PMCID: PMC7373300.

118. Satpathy S, Jaehnig EJ, Krug K, Kim BJ, Saltzman AB, Chan DW, Holloway KR, Anurag M, Huang C, Singh P, Gao A, Namai N, Dou Y, Wen B, Vasaikar SV, Mutch D, Watson MA, Ma C, Ademuyiwa FO, Rimawi MF, Schiff R, Hoog J, Jacobs S, Malovannaya A, Hyslop T, Clauser KR, Mani DR, Perou CM, Miles G, Zhang B, Gillette MA, Carr SA, Ellis MJ. Microscaled proteogenomic methods for precision oncology. Nat Commun. 2020;11(1):532. Epub 20200127. doi: 10.1038/s41467-020-14381-2. PubMed PMID: 31988290; PMCID: PMC6985126.

119. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Australian Pancreatic Cancer Genome I, Consortium IBC, Consortium IM-S, PedBrain I, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-21. doi: 10.1038/nature12477. PubMed PMID: 23945592; PMCID: PMC3776390.

120. Li B, Brady SW, Ma X, Shen S, Zhang Y, Li Y, Szlachta K, Dong L, Liu Y, Yang F, Wang N, Flasch DA, Myers MA, Mulder HL, Ding L, Liu Y, Tian L, Hagiwara K, Xu K,

Zhou X, Sioson E, Wang T, Yang L, Zhao J, Zhang H, Shao Y, Sun H, Sun L, Cai J, Sun HY, Lin TN, Du L, Li H, Rusch M, Edmonson MN, Easton J, Zhu X, Zhang J, Cheng C, Raphael BJ, Tang J, Downing JR, Alexandrov LB, Zhou BS, Pui CH, Yang JJ, Zhang J. Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. Blood. 2020;135(1):41-55. Epub 2019/11/08. doi: 10.1182/blood.2019002220. PubMed PMID: 31697823; PMCID: PMC6940198.

121. Degasperi A, Amarante TD, Czarnecki J, Shooter S, Zou X, Glodzik D, Morganella S, Nanda AS, Badja C, Koh G, Momen SE, Georgakopoulos-Soares I, Dias JML, Young J, Memari Y, Davies H, Nik-Zainal S. A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. Nat Cancer. 2020;1(2):249-63. Epub 2020/03/03. doi: 10.1038/s43018-020-0027-5. PubMed PMID: 32118208; PMCID: PMC7048622.

122. Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S. Machine learning and its applications to biology. PLoS Comput Biol. 2007;3(6):e116. doi: 10.1371/journal.pcbi.0030116. PubMed PMID: 17604446; PMCID: PMC1904382.

123. Yu B. Stability. Bernoulli. 2013;19(4):1484-500.

124. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333-7. Epub 2014/01/28. doi: 10.1038/nmeth.2810. PubMed PMID: 24464287.

125. Yona G, Dirks W, Rahman S, Lin DM. Effective similarity measures for expression profiles. Bioinformatics. 2006;22(13):1616-22. Epub 2006/04/06. doi: 10.1093/bioinformatics/btl127. PubMed PMID: 16595558.

126. Kraskov A, Stogbauer H, Grassberger P. Estimating mutual information. Phys Rev E Stat Nonlin Soft Matter Phys. 2004;69(6 Pt 2):066138. Epub 2004/07/13. doi: 10.1103/PhysRevE.69.066138. PubMed PMID: 15244698.

127. Stogbauer H, Kraskov A, Astakhov SA, Grassberger P. Least-dependent-component analysis based on mutual information. Phys Rev E Stat Nonlin Soft Matter Phys. 2004;70(6 Pt 2):066123. Epub 2005/02/09. doi: 10.1103/PhysRevE.70.066123. PubMed PMID: 15697450.

128. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet. 2005;37(11):1243-6. Epub 2005/10/18. doi: 10.1038/ng1653. PubMed PMID: 16228001.

129. McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM, Mesirov JP, Polley MY, Kim KY, Tricoli JV, Taylor JM, Shuman DJ, Simon RM, Doroshow JH, Conley BA. Criteria for the use of omics-based predictors in clinical trials. Nature.

2013;502(7471):317-20. Epub 2013/10/18. doi: 10.1038/nature12564. PubMed PMID: 24132288; PMCID: PMC4180668.

130. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference: Morgan Kaufmann; 1988.

131. Nguyen H, Shrestha S, Draghici S, Nguyen T. PINSPlus: a tool for tumor subtype discovery in integrated genomic data. Bioinformatics. 2019;35(16):2843-6. Epub 2018/12/28. doi: 10.1093/bioinformatics/bty1049. PubMed PMID: 30590381.

132. Planell N, Lagani V, Sebastian-Leon P, van der Kloet F, Ewing E, Karathanasis N, Urdangarin A, Arozarena I, Jagodic M, Tsamardinos I, Tarazona S, Conesa A, Tegner J, Gomez-Cabrero D. STATegra: Multi-Omics Data Integration - A Conceptual Scheme With a Bioinformatics Pipeline. Front Genet. 2021;12:620453. Epub 20210304. doi: 10.3389/fgene.2021.620453. PubMed PMID: 33747045; PMCID: PMC7970106.

133. Ramstead MJD, Badcock PB, Friston KJ. Answering Schrodinger's question: A free-energy formulation. Phys Life Rev. 2018;24:1-16. Epub 2017/10/17. doi: 10.1016/j.plrev.2017.09.001. PubMed PMID: 29029962; PMCID: PMC5857288.

134. Sudarikov K, Tyakht A, Alexeev D. Methods for The Metagenomic Data Visualization and Analysis. Curr Issues Mol Biol. 2017;24:37-58. Epub 20170706. doi: 10.21775/cimb.024.037. PubMed PMID: 28686567.

135. Midway SR. Principles of Effective Data Visualization. Patterns (N Y). 2020;1(9):100141. Epub 20201111. doi: 10.1016/j.patter.2020.100141. PubMed PMID: 33336199; PMCID: PMC7733875.

136. Kotu V, Deshpande B. Data Science: Concepts and Practice. Second ed. Cambridge, MA: Morgan Kaufmann; 2019.

137. Krzywinski M, Altman N. Points of significance: error bars. Nat Methods. 2013;10(10):921-2. doi: 10.1038/nmeth.2659. PubMed PMID: 24161969.

138. Tufte ER. The visual display of quantitative information. Cheshire, Conn: Graphics Press; 2001.

139. Ryan L. The Visual Imperative. Cambridge, MA: Morhan Kaufmann; 2016.

140. Lane DM, Sandor A. Designing better graphs by including distributional information and integrating words, numbers, and images. Psychol Methods. 2009;14(3):239-57. doi: 10.1037/a0016620. PubMed PMID: 19719360.

141. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, Hornshoj H, Hess JM, Juul RI, Lin Z, Feuerbach L, Sabarinathan R, Madsen T, Kim J, Mularoni L, Shuai S, Lanzos A, Herrmann C, Maruvka YE, Shen C, Amin SB, Bandopadhayay P, Bertl J, Boroevich KA, Busanovich J, Carlevaro-Fita J, Chakravarty D, Chan CWY, Craft D, Dhingra P, Diamanti K, Fonseca NA, Gonzalez-Perez A, Guo Q, Hamilton MP,

Haradhvala NJ, Hong C, Isaev K, Johnson TA, Juul M, Kahles A, Kahraman A, Kim Y, Komorowski J, Kumar K, Kumar S, Lee D, Lehmann KV, Li Y, Liu EM, Lochovsky L, Park K, Pich O, Roberts ND, Saksena G, Schumacher SE, Sidiropoulos N, Sieverling L, Sinnott-Armstrong N, Stewart C, Tamborero D, Tubio JMC, Umer HM, Uuskula-Reimand L, Wadelius C, Wadi L, Yao X, Zhang CZ, Zhang J, Haber JE, Hobolth A, Imielinski M, Kellis M, Lawrence MS, von Mering C, Nakagawa H, Raphael BJ, Rubin MA, Sander C, Stein LD, Stuart JM, Tsunoda T, Wheeler DA, Johnson R, Reimand J, Gerstein M, Khurana E, Campbell PJ, Lopez-Bigas N, Drivers P, Functional Interpretation Working G, Group PSVW, Weischenfeldt J, Beroukhim R, Martincorena I, Pedersen JS, Getz G, Consortium P. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature. 2020;578(7793):102-11. Epub 2020/02/07. doi: 10.1038/s41586-020-1965-x. PubMed PMID: 32025015.

142. Sinha N, Smith-Gill SJ. Electrostatics in protein binding and function. Curr Protein Pept Sci. 2002;3(6):601-14. Epub 2002/12/10. PubMed PMID: 12470214.

143. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118. Epub 2011/07/06. doi: 10.1093/nar/gkr407. PubMed PMID: 21727090; PMCID: PMC3177186.

144. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863-74. Epub 2001/05/05. doi: 10.1101/gr.176601. PubMed PMID: 11337480; PMCID: PMC311071.

145. Vitkup D, Sander C, Church GM. The amino-acid mutational spectrum of human genetic disease. Genome Biol. 2003;4(11):R72. Epub 2003/11/13. doi: 10.1186/gb-2003-4-11-r72. PubMed PMID: 14611658; PMCID: PMC329120.

146. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005;15(7):978-86. Epub 2005/06/21. doi: 10.1101/gr.3804205. PubMed PMID: 15965030; PMCID: PMC1172042.

147. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7(10):e46688. Epub 2012/10/12. doi: 10.1371/journal.pone.0046688. PubMed PMID: 23056405; PMCID: PMC3466303.

148. Akashi H. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics. 1995;139(2):1067-76. doi: 10.1093/genetics/139.2.1067. PubMed PMID: 7713409; PMCID: PMC1206357.

149. Alexaki A, Kames J, Holcomb DD, Athey J, Santana-Quintero LV, Lam PVN, Hamasaki-Katagiri N, Osipova E, Simonyan V, Bar H, Komar AA, Kimchi-Sarfaty C. Codon and Codon-Pair Usage Tables (CoCoPUTs): Facilitating Genetic Variation Analyses and Recombinant Gene Design. J Mol Biol. 2019;431(13):2434-41. Epub 20190426. doi: 10.1016/j.jmb.2019.04.021. PubMed PMID: 31029701.

150. Dayhoff MO, Eck RV. Atlas of Protein Sequence and Structure. Hersh RT, editor. Silver Spring, MD: National Biomedical Research Foundation; 1968.

151. McLachlan AD, Boswell DR. Confidence limits for homology in protein or gene sequences. The c-myc oncogene and adenovirus E1a protein. J Mol Biol. 1985;185(1):39-49. doi: 10.1016/0022-2836(85)90181-0. PubMed PMID: 4046040.

152. Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol. 1987;25(4):351-60. doi: 10.1007/BF02603120. PubMed PMID: 3118049.

153. Rao JKM. New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. Int J Pept Protein Res. 1987;29:276-81.

154. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89(22):10915-9. doi: 10.1073/pnas.89.22.10915. PubMed PMID: 1438297; PMCID: PMC50453.

155. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S. Increased coverage of protein families with the blocks database servers. Nucleic Acids Res. 2000;28(1):228-30. doi: 10.1093/nar/28.1.228. PubMed PMID: 10592233; PMCID: PMC102407.

156. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol. 2004;22(8):1035-6. doi: 10.1038/nbt0804-1035. PubMed PMID: 15286655.

157. Kim Y, Sidney J, Pinilla C, Sette A, Peters B. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. BMC Bioinformatics. 2009;10:394. Epub 20091130. doi: 10.1186/1471-2105-10-394. PubMed PMID: 19948066; PMCID: PMC2790471.

158. Sidney J, del Guercio MF, Southwood S, Engelhard VH, Appella E, Rammensee HG, Falk K, Rotzschke O, Takiguchi M, Kubo RT, et al. Several HLA alleles share overlapping peptide specificities. J Immunol. 1995;154(1):247-59. Epub 1995/01/01. PubMed PMID: 7527812.

159. Chowell D, Morris LGT, Grigg CM, Weber JK, Samstein RM, Makarov V, Kuo F, Kendall SM, Requena D, Riaz N, Greenbaum B, Carroll J, Garon E, Hyman DM, Zehir A, Solit D, Berger M, Zhou R, Rizvi NA, Chan TA. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. Science. 2017. doi: 10.1126/science.aao4572. PubMed PMID: 29217585.

160. Chowell D, Krishna C, Pierini F, Makarov V, Rizvi NA, Kuo F, Morris LGT, Riaz N, Lenz TL, Chan TA. Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. Nat Med. 2019;25(11):1715-20. Epub 2019/11/09. doi: 10.1038/s41591-019-0639-4. PubMed PMID: 31700181.

161. Negrao MV, Lam VK, Reuben A, Rubin ML, Landry LL, Roarty EB, Rinsurongkawong W, Lewis J, Roth JA, Swisher SG, Gibbons DL, Wistuba, II,

Papadimitrakopoulou V, Glisson BS, Blumenschein GR, Jr., Lee JJ, Heymach JV, Zhang J. PD-L1 expression, tumor mutational burden and cancer gene mutations are stronger predictors of benefit from immune checkpoint blockade than HLA class I genotype in non-small cell lung cancer. J Thorac Oncol. 2019. Epub 2019/02/20. doi: 10.1016/j.jtho.2019.02.008. PubMed PMID: 30780001.

162. Cummings AL, Gukasyan J, Lu HY, Grogan TR, Sunga G, Fares CM, Hornstein N, Zaretsky J, Carroll J, Bachrach B, Akingbemi WO, Li D, Noor Z, Lisberg A, Goldman JW, Elashoff D, Bui AAT, Ribas A, Dubinett SM, Rossetti M, Garon EB. Mutational landscape influences immunotherapy outcomes among non-small cell lung cancer patients with human leukocyte antigen supertype B44. Nature Cancer. 2020:in press.

163. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, Miller ML, Rekhtman N, Moreira AL, Ibrahim F, Bruggeman C, Gasmi B, Zappasodi R, Maeda Y, Sander C, Garon EB, Merghoub T, Wolchok JD, Schumacher TN, Chan TA. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science. 2015;348(6230):124-8. doi: 10.1126/science.aaa1348. PubMed PMID: 25765070; PMCID: PMC4993154.

164. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C, Dicara D, Ramos AH, Lawrence MS, Cibulskis K, Sivachenko A, Voet D, Saksena G, Stransky N, Onofrio RC, Winckler W, Ardlie K, Wagle N, Wargo J, Chong K, Morton DL, Stemke-Hale K, Chen G, Noble M, Meyerson M, Ladbury JE, Davies MA, Gershenwald JE, Wagner SN, Hoon DS, Schadendorf D, Lander ES, Gabriel SB, Getz G, Garraway LA, Chin L. A landscape of driver mutations in melanoma. Cell. 2012;150(2):251-63. Epub 2012/07/24. doi: 10.1016/j.cell.2012.06.024. PubMed PMID: 22817889; PMCID: PMC3600117.

165. Pfeifer GP. Mutagenesis at methylated CpG sequences. Curr Top Microbiol Immunol. 2006;301:259-81. doi: 10.1007/3-540-31390-7_10. PubMed PMID: 16570852.

166. Boland CR, Goel A. Microsatellite instability in colorectal cancer. Gastroenterology. 2010;138(6):2073-87 e3. doi: 10.1053/j.gastro.2009.12.064. PubMed PMID: 20420947; PMCID: PMC3037515.

167. Thompson LH. Recognition, signaling, and repair of DNA double-strand breaks produced by ionizing radiation in mammalian cells: the molecular choreography. Mutat Res. 2012;751(2):158-246. Epub 20120626. doi: 10.1016/j.mrrev.2012.06.002. PubMed PMID: 22743550.

168. Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic hypermutation. Annu Rev Biochem. 2007;76:1-22. doi: 10.1146/annurev.biochem.76.061705.090740. PubMed PMID: 17328676.

169. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR,

Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jonsson G, Garber JE, Silver D, Miron P, Fatima A, Boyault S, Langerod A, Tutt A, Martens JW, Aparicio SA, Borg A, Salomon AV, Thomas G, Borresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR, Breast Cancer Working Group of the International Cancer Genome C. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149(5):979-93. Epub 20120517. doi: 10.1016/j.cell.2012.04.024. PubMed PMID: 22608084; PMCID: PMC3414841.

170. Hunter C, Smith R, Cahill DP, Stephens P, Stevens C, Teague J, Greenman C, Edkins S, Bignell G, Davies H, O'Meara S, Parker A, Avis T, Barthorpe S, Brackenbury L, Buck G, Butler A, Clements J, Cole J, Dicks E, Forbes S, Gorton M, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Kosmidou V, Laman R, Lugg R, Menzies A, Perry J, Petty R, Raine K, Richardson D, Shepherd R, Small A, Solomon H, Tofts C, Varian J, West S, Widaa S, Yates A, Easton DF, Riggins G, Roy JE, Levine KK, Mueller W, Batchelor TT, Louis DN, Stratton MR, Futreal PA, Wooster R. A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. Cancer Res. 2006;66(8):3987-91. doi: 10.1158/0008-5472.CAN-06-0127. PubMed PMID: 16618716; PMCID: PMC7212022.

171. Pich O, Muinos F, Lolkema MP, Steeghs N, Gonzalez-Perez A, Lopez-Bigas N. The mutational footprints of cancer therapies. Nat Genet. 2019;51(12):1732-40. Epub 2019/11/20. doi: 10.1038/s41588-019-0525-5. PubMed PMID: 31740835; PMCID: PMC6887544.

172. Liu C, Yang X, Duffy B, Mohanakumar T, Mitra RD, Zody MC, Pfeifer JD. ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. Nucleic Acids Res. 2013;41(14):e142. doi: 10.1093/nar/gkt481. PubMed PMID: 23748956; PMCID: PMC3737559.

173. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, Griffith M. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Med. 2016;8(1):11. Epub 2016/01/31. doi: 10.1186/s13073-016-0264-5. PubMed PMID: 26825632; PMCID: PMC4733280.

174. Chen G, Mosier S, Gocke CD, Lin MT, Eshleman JR. Cytosine deamination is a major cause of baseline noise in next-generation sequencing. Mol Diagn Ther. 2014;18(5):587-93. doi: 10.1007/s40291-014-0115-2. PubMed PMID: 25091469; PMCID: PMC4175022.

175. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124. Epub 20050830. doi: 10.1371/journal.pmed.0020124. PubMed PMID: 16060722; PMCID: PMC1182327.

176. McShane LM, Polley MY. Development of omics-based clinical tests for prognosis and therapy selection: the challenge of achieving statistical robustness and clinical utility. Clin Trials. 2013;10(5):653-65. Epub 2013/09/04. doi: 10.1177/1740774513499458. PubMed PMID: 24000377; PMCID: PMC4410005.

177. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. Int J Epidemiol. 2016;45(6):1866-86. doi: 10.1093/ije/dyw314. PubMed PMID: 28108528; PMCID: PMC5841843.

178. Munafo MR, Davey Smith G. Robust research needs many lines of evidence. Nature. 2018;553(7689):399-401. doi: 10.1038/d41586-018-01023-3. PubMed PMID: 29368721.

179. Devezer B, Nardin LG, Baumgaertner B, Buzbas EO. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. PLoS One. 2019;14(5):e0216125. Epub 20190515. doi: 10.1371/journal.pone.0216125. PubMed PMID: 31091251; PMCID: PMC6519896.

180. Cartwright N. How the Laws of Physics Lie. Oxford, New York: Oxford Univesrity Press; 1983.

181. Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978;6:461-4.

182. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974;19:716-23.

183. Nearing GS, Ruddell BL, Bennett AR. Does information theory provide a new paradigm for earth science? Hypothesis testing. Water Resources Research. 2020;56(2).

184. Martin-Fernandez JA, Barcelo-Vidal C, Pawlowsky-Glahn V, editors. Zero replacement in compositional data sets. Conference of the International Federation of Classification Societies; 2000; Berlin: Springer-Verlag.

185. Delucchi KL, Bostrom A. Methods for analysis of skewed data distributions in psychiatric clinical studies: working with many zero values. Am J Psychiatry. 2004;161(7):1159-68. doi: 10.1176/appi.ajp.161.7.1159. PubMed PMID: 15229044.

186. Gomez-Rubio V, Lopez-Quilez A. Statistical methods for the geographical analysis of rare diseases. Adv Exp Med Biol. 2010;686:151-71. doi: 10.1007/978-90-481-9485-8_10. PubMed PMID: 20824445.

187. Liu L, Shih YT, Strawderman RL, Zhang D, Johnson BA, Chai H. Statistical analysis of zero-inflated nonnegative continuous data: a review. Statistical Science. 2019;34(2):253-79.

188. Xia Y. Correlation and association analyses in microbiome study integrating multiomics in health and disease. Prog Mol Biol Transl Sci. 2020;171:309-491. Epub 20200523. doi: 10.1016/bs.pmbts.2020.04.003. PubMed PMID: 32475527.

189. Martin-Fernandez JA, Barcelo-Vidal C, Pawlowsky-Glahn V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Mathematical Geology. 2003;35(3):253-78.

190. Aitchison J. The statistical analysis of compositional data. J Roy Statist Soc. 1982;44(2):139-60.

191. Tauber F. Spurious clusters in granulometric data caused by logratio transformation. Mathematical Geology. 1999;31(5):491-504.

192. Krzanowski WJ. Principles of multivariate analysis. Oxford: Clarendon Press; 1988.

193. Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. Nat Rev Mol Cell Biol. 2008;9(12):958-70. doi: 10.1038/nrm2549. PubMed PMID: 19023283.

194. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene. 2002;21(48):7435-51. doi: 10.1038/sj.onc.1205803. PubMed PMID: 12379884.

195. Pfeifer GP, You YH, Besaratinia A. Mutations induced by ultraviolet light. Mutat Res. 2005;571(1-2):19-31. Epub 20050120. doi: 10.1016/j.mrfmmm.2004.06.057. PubMed PMID: 15748635.

196. Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol. 1979;14(2):109-21. doi: 10.1097/00004424-197903000-00002. PubMed PMID: 478799.

197. Daniel WW. Kendall’s tau. Second ed. Boston, Massachusetts: PWS-Kent; 1990.

198. Garner JB. The standard error of Cohen's Kappa. Stat Med. 1991;10(5):767-75. Epub 1991/05/01. doi: 10.1002/sim.4780100512. PubMed PMID: 2068430.

199. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82. Epub 2012/10/25. PubMed PMID: 23092060; PMCID: PMC3900052.

200. Chang CH. Cohen's kappa for capturing discrimination. Int Health. 2014;6(2):125-9. Epub 2014/04/03. doi: 10.1093/inthealth/ihu010. PubMed PMID: 24691677.

201. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70(1):7-30. Epub 2020/01/09. doi: 10.3322/caac.21590. PubMed PMID: 31912902.

202. Leach DR, Krummel MF, Allison JP. Enhancement of antitumor immunity by CTLA-4 blockade. Science. 1996;271(5256):1734-6. doi: 10.1126/science.271.5256.1734. PubMed PMID: 8596936.

203. Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, Patnaik A, Aggarwal C, Gubens M, Horn L, Carcereny E, Ahn MJ, Felip E, Lee JS, Hellmann MD, Hamid O, Goldman JW, Soria JC, Dolled-Filhart M, Rutledge RZ, Zhang J, Lunceford JK, Rangwala R, Lubiniecki GM, Roach C, Emancipator K, Gandhi L, Investigators K-.

Pembrolizumab for the treatment of non-small-cell lung cancer. N Engl J Med. 2015;372(21):2018-28. doi: 10.1056/NEJMoa1501824. PubMed PMID: 25891174.

204. Hellmann MD, Rizvi NA, Goldman JW, Gettinger SN, Borghaei H, Brahmer JR, Ready NE, Gerber DE, Chow LQ, Juergens RA, Shepherd FA, Laurie SA, Geese WJ, Agrawal S, Young TC, Li X, Antonia SJ. Nivolumab plus ipilimumab as first-line treatment for advanced non-small-cell lung cancer (CheckMate 012): results of an open-label, phase 1, multicohort study. Lancet Oncol. 2017;18(1):31-41. doi: 10.1016/S1470-2045(16)30624-6. PubMed PMID: 27932067.

205. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, Skora AD, Luber BS, Azad NS, Laheru D, Biedrzycki B, Donehower RC, Zaheer A, Fisher GA, Crocenzi TS, Lee JJ, Duffy SM, Goldberg RM, de la Chapelle A, Koshiji M, Bhaijee F, Huebner T, Hruban RH, Wood LD, Cuka N, Pardoll DM, Papadopoulos N, Kinzler KW, Zhou S, Cornish TC, Taube JM, Anders RA, Eshleman JR, Vogelstein B, Diaz LA, Jr. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med. 2015;372(26):2509-20. doi: 10.1056/NEJMoa1500596. PubMed PMID: 26028255; PMCID: PMC4481136.

206. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015;160(1-2):48-61. doi: 10.1016/j.cell.2014.12.033. PubMed PMID: 25594174; PMCID: PMC4856474.

207. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, Berent-Maoz B, Pang J, Chmielowski B, Cherry G, Seja E, Lomeli S, Kong X, Kelley MC, Sosman JA, Johnson DB, Ribas A, Lo RS. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell. 2016;165(1):35-44. doi: 10.1016/j.cell.2016.02.065. PubMed PMID: 26997480; PMCID: PMC4808437.

208. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, Sher X, Liu XQ, Lu H, Nebozhyn M, Zhang C, Lunceford JK, Joe A, Cheng J, Webber AL, Ibrahim N, Plimack ER, Ott PA, Seiwert TY, Ribas A, McClanahan TK, Tomassini JE, Loboda A, Kaufman D. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science. 2018;362(6411). doi: 10.1126/science.aar3593. PubMed PMID: 30309915; PMCID: PMC6718162.

209. Cummings AL, Garon EB. The ascent of immune checkpoint inhibitors: is the understudy ready for a leading role? Cancer Biol Med. 2017;14(4):341-7. doi: 10.20892/j.issn.2095-3941.2017.0090. PubMed PMID: 29372100; PMCID: PMC5785166.

210. Cummings AL, Garon EB. KEYNOTE-042 rolls back programmed cell death ligand 1 threshold for non-small cell lung cancer pembrolizumab monotherapy without new insight into those deriving benefit. Transl Lung Cancer Res. 2019;8(Suppl 4):S403-S6. doi: 10.21037/tlcr.2019.07.06. PubMed PMID: 32038925; PMCID: PMC6987348.

211. Cummings AL, Santoso KM, Goldman JW. KEYNOTE-021 cohorts D and H suggest modest benefit in combining ipilimumab with pembrolizumab in second-line or

later advanced non-small cell lung cancer treatment. Transl Lung Cancer Res. 2019;8(5):706-9. doi: 10.21037/tlcr.2019.08.11. PubMed PMID: 31737507; PMCID: PMC6835110.

212. Shankaran V, Ikeda H, Bruce AT, White JM, Swanson PE, Old LJ, Schreiber RD. IFNgamma and lymphocytes prevent primary tumour development and shape tumour immunogenicity. Nature. 2001;410(6832):1107-11. doi: 10.1038/35074122. PubMed PMID: 11323675.

213. Mittal D, Gubin MM, Schreiber RD, Smyth MJ. New insights into cancer immunoediting and its three component phases--elimination, equilibrium and escape. Curr Opin Immunol. 2014;27:16-25. Epub 20140214. doi: 10.1016/j.coi.2014.01.004. PubMed PMID: 24531241; PMCID: PMC4388310.

214. DuPage M, Mazumdar C, Schmidt LM, Cheung AF, Jacks T. Expression of tumour-specific antigens underlies cancer immunoediting. Nature. 2012;482(7385):405-9. Epub 20120208. doi: 10.1038/nature10803. PubMed PMID: 22318517; PMCID: PMC3288744.

215. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science. 2015;348(6230):69-74. doi: 10.1126/science.aaa4971. PubMed PMID: 25838375.

216. McGranahan N, Furness AJ, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, Jamal-Hanjani M, Wilson GA, Birkbak NJ, Hiley CT, Watkins TB, Shafi S, Murugaesu N, Mitter R, Akarca AU, Linares J, Marafioti T, Henry JY, Van Allen EM, Miao D, Schilling B, Schadendorf D, Garraway LA, Makarov V, Rizvi NA, Snyder A, Hellmann MD, Merghoub T, Wolchok JD, Shukla SA, Wu CJ, Peggs KS, Chan TA, Hadrup SR, Quezada SA, Swanton C. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science. 2016;351(6280):1463-9. doi: 10.1126/science.aaf1490. PubMed PMID: 26940869; PMCID: PMC4984254.

217. Wells DK, van Buuren MM, Dang KK, Hubbard-Lucey VM, Sheehan KCF, Campbell KM, Lamb A, Ward JP, Sidney J, Blazquez AB, Rech AJ, Zaretsky JM, Comin-Anduix B, Ng AHC, Chour W, Yu TV, Rizvi H, Chen JM, Manning P, Steiner GM, Doan XC, Tumor Neoantigen Selection A, Merghoub T, Guinney J, Kolom A, Selinsky C, Ribas A, Hellmann MD, Hacohen N, Sette A, Heath JR, Bhardwaj N, Ramsdell F, Schreiber RD, Schumacher TN, Kvistborg P, Defranoux NA. Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. Cell. 2020;183(3):818-34 e13. Epub 20201009. doi: 10.1016/j.cell.2020.09.015. PubMed PMID: 33038342; PMCID: PMC7652061.

218. Gubin MM, Zhang X, Schuster H, Caron E, Ward JP, Noguchi T, Ivanova Y, Hundal J, Arthur CD, Krebber WJ, Mulder GE, Toebes M, Vesely MD, Lam SS, Korman AJ, Allison JP, Freeman GJ, Sharpe AH, Pearce EL, Schumacher TN, Aebersold R, Rammensee HG, Melief CJ, Mardis ER, Gillanders WE, Artyomov MN, Schreiber RD. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature. 2014;515(7528):577-81. Epub 2014/11/28. doi: 10.1038/nature13988. PubMed PMID: 25428507; PMCID: PMC4279952.

219. Balachandran VP, Luksza M, Zhao JN, Makarov V, Moral JA, Remark R, Herbst B, Askan G, Bhanot U, Senbabaoglu Y, Wells DK, Cary CIO, Grbovic-Huezo O, Attiyeh M, Medina B, Zhang J, Loo J, Saglimbeni J, Abu-Akeel M, Zappasodi R, Riaz N, Smoragiewicz M, Kelley ZL, Basturk O, Australian Pancreatic Cancer Genome I, Garvan Institute of Medical R, Prince of Wales H, Royal North Shore H, University of G, St Vincent's H, Institute QBMR, University of Melbourne CfCR, University of Queensland IfMB, Bankstown H, Liverpool H, Royal Prince Alfred Hospital COBL, Westmead H, Fremantle H, St John of God H, Royal Adelaide H, Flinders Medical C, Envoi P, Princess Alexandria H, Austin H, Johns Hopkins Medical I, Cancer AR-NCfARo, Gonen M, Levine AJ, Allen PJ, Fearon DT, Merad M, Gnjatic S, Iacobuzio-Donahue CA, Wolchok JD, DeMatteo RP, Chan TA, Greenbaum BD, Merghoub T, Leach SD. Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. Nature. 2017;551(7681):512-6. Epub 20171108. doi: 10.1038/nature24462. PubMed PMID: 29132146; PMCID: PMC6145146.

220. Dausset J. The major histocompatibility complex in man. Science. 1981;213(4515):1469-74. Epub 1981/09/25. doi: 10.1126/science.6792704. PubMed PMID: 6792704.

221. Algarra I, Collado A, Garrido F. Altered MHC class I antigens in tumors. Int J Clin Lab Res. 1997;27(2):95-102. Epub 1997/01/01. PubMed PMID: 9266279.

222. Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. Nature. 2017;541(7637):321-30. doi: 10.1038/nature21349. PubMed PMID: 28102259.

223. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2015;43(Database issue):D423-31. Epub 20141120. doi: 10.1093/nar/gku1161. PubMed PMID: 25414341; PMCID: PMC4383959.

224. Sette A, Sidney J. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. Immunogenetics. 1999;50(3-4):201-12. PubMed PMID: 10602880.

225. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and updated classification. BMC Immunol. 2008;9:1. doi: 10.1186/1471-2172-9-1. PubMed PMID: 18211710; PMCID: PMC2245908.

226. Brahmer JR, Tykodi SS, Chow LQ, Hwu WJ, Topalian SL, Hwu P, Drake CG, Camacho LH, Kauh J, Odunsi K, Pitot HC, Hamid O, Bhatia S, Martins R, Eaton K, Chen S, Salay TM, Alaparthy S, Grosso JF, Korman AJ, Parker SM, Agrawal S, Goldberg SM, Pardoll DM, Gupta A, Wigginton JM. Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. N Engl J Med. 2012;366(26):2455-65. doi: 10.1056/NEJMoa1200694. PubMed PMID: 22658128; PMCID: 3563263.

227. Jencks WP RJ. Handbook of Biochemistry and Molecular Biology. 4th ed. Boca Raton, FL: CRC Press; 2010.

228. Macdonald WA, Purcell AW, Mifsud NA, Ely LK, Williams DS, Chang L, Gorman JJ, Clements CS, Kjer-Nielsen L, Koelle DM, Burrows SR, Tait BD, Holdsworth R, Brooks AG, Lovrecz GO, Lu L, Rossjohn J, McCluskey J. A naturally selected dimorphism within the HLA-B44 supertype alters class I structure, peptide repertoire, and T cell recognition. J Exp Med. 2003;198(5):679-91. Epub 2003/08/27. doi: 10.1084/jem.20030066. PubMed PMID: 12939341; PMCID: PMC2194191.

229. Grimsley GR, Scholtz JM, Pace CN. A summary of the measured pK values of the ionizable groups in folded proteins. Protein Sci. 2009;18(1):247-51. Epub 2009/01/30. doi: 10.1002/pro.19. PubMed PMID: 19177368; PMCID: PMC2708032.

230. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. Epub 2009/06/10. doi: 10.1093/bioinformatics/btp352. PubMed PMID: 19505943; PMCID: PMC2723002.

231. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8. Epub 2011/04/12. doi: 10.1038/ng.806. PubMed PMID: 21478889; PMCID: PMC3083463.

232. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80-92. Epub 2012/06/26. doi: 10.4161/fly.19695. PubMed PMID: 22728672; PMCID: PMC3679285.

233. Hellmann MD, Ciuleanu TE, Pluzanski A, Lee JS, Otterson GA, Audigier-Valette C, Minenza E, Linardou H, Burgers S, Salman P, Borghaei H, Ramalingam SS, Brahmer J, Reck M, O'Byrne KJ, Geese WJ, Green G, Chang H, Szustakowski J, Bhagavatheeswaran P, Healey D, Fu Y, Nathan F, Paz-Ares L. Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. N Engl J Med. 2018;378(22):2093-104. Epub 2018/04/17. doi: 10.1056/NEJMoa1801946. PubMed PMID: 29658845.

234. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2018. Epub 2018/10/26. doi: 10.1093/nar/gky1006. PubMed PMID: 30357391.

235. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Roder G, Peters B, Sette A, Lund O, Buus S. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS One. 2007;2(8):e796. Epub 2007/08/30. doi: 10.1371/journal.pone.0000796. PubMed PMID: 17726526; PMCID: PMC1949492.

236. Lundegaard C, Lund O, Nielsen M. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. Bioinformatics. 2008;24(11):1397-8. Epub 2008/04/17. doi: 10.1093/bioinformatics/btn128. PubMed PMID: 18413329.

237. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol. 2017;199(9):3360-8. Epub 2017/10/06. doi: 10.4049/jimmunol.1700893. PubMed PMID: 28978689; PMCID: PMC5679736.

238. Miao D, Margolis CA, Vokes NI, Liu D, Taylor-Weiner A, Wankowicz SM, Adeegbe D, Keliher D, Schilling B, Tracy A, Manos M, Chau NG, Hanna GJ, Polak P, Rodig SJ, Signoretti S, Sholl LM, Engelman JA, Getz G, Janne PA, Haddad RI, Choueiri TK, Barbie DA, Haq R, Awad MM, Schadendorf D, Hodi FS, Bellmunt J, Wong KK, Hammerman P, Van Allen EM. Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. Nat Genet. 2018;50(9):1271-81. Epub 2018/08/29. doi: 10.1038/s41588-018-0200-2. PubMed PMID: 30150660; PMCID: PMC6119118.

239. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44(8):e71. Epub 20151223. doi: 10.1093/nar/gkv1507. PubMed PMID: 26704973; PMCID: PMC4856967.

240. Mounir M, Lucchetta M, Silva TC, Olsen C, Bontempi G, Chen X, Noushmehr H, Colaprico A, Papaleo E. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. PLoS Comput Biol. 2019;15(3):e1006701. Epub 20190305. doi: 10.1371/journal.pcbi.1006701. PubMed PMID: 30835723; PMCID: PMC6420023.

241. Leone P, Shin EC, Perosa F, Vacca A, Dammacco F, Racanelli V. MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells. J Natl Cancer Inst. 2013;105(16):1172-87. Epub 20130712. doi: 10.1093/jnci/djt184. PubMed PMID: 23852952.

242. Castro A, Ozturk K, Pyke RM, Xian S, Zanetti M, Carter H. Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes. BMC Med Genomics. 2019;12(Suppl 6):107. Epub 2019/07/28. doi: 10.1186/s12920-019-0544-1. PubMed PMID: 31345234; PMCID: PMC6657029.

243. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer. 2012;12(4):252-64. doi: 10.1038/nrc3239. PubMed PMID: 22437870; PMCID: PMC4856023.

244. Kessler JH, Benckhuijsen WE, Mutis T, Melief CJ, van der Burg SH, Drijfhout JW. Competition-based cellular peptide binding assay for HLA class I. Curr Protoc Immunol. 2004;Chapter 18:Unit 18 2. Epub 2008/04/25. doi: 10.1002/0471142735.im1812s61. PubMed PMID: 18432926.

245. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-8. Epub 2008/06/03. doi: 10.1038/nmeth.1226. PubMed PMID: 18516045.

246. Rosner B, Glynn RJ, Lee ML. Extension of the rank sum test for clustered data: two-group comparisons with group membership defined at the subunit level. Biometrics. 2006;62(4):1251-9. Epub 2006/12/13. doi: 10.1111/j.1541-0420.2006.00582.x. PubMed PMID: 17156300.

247. Herbst RS, Baas P, Kim DW, Felip E, Perez-Gracia JL, Han JY, Molina J, Kim JH, Arvis CD, Ahn MJ, Majem M, Fidler MJ, de Castro G, Jr., Garrido M, Lubiniecki GM, Shentu Y, Im E, Dolled-Filhart M, Garon EB. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. Lancet. 2016;387(10027):1540-50. doi: 10.1016/S0140-6736(15)01281-7. PubMed PMID: 26712084.

248. Ribas A, Shin DS, Zaretsky J, Frederiksen J, Cornish A, Avramis E, Seja E, Kivork C, Siebert J, Kaplan-Lefko P, Wang X, Chmielowski B, Glaspy JA, Tumeh PC, Chodon T, Pe'er D, Comin-Anduix B. PD-1 Blockade Expands Intratumoral Memory T Cells. Cancer Immunol Res. 2016;4(3):194-203. doi: 10.1158/2326-6066.CIR-15-0210. PubMed PMID: 26787823; PMCID: PMC4775381.

249. DiBrino M, Parker KC, Margulies DH, Shiloach J, Turner RV, Biddison WE, Coligan JE. Identification of the peptide binding motif for HLA-B44, one of the most common HLA-B alleles in the Caucasian population. Biochemistry. 1995;34(32):10130-8. Epub 1995/08/15. doi: 10.1021/bi00032a005. PubMed PMID: 7543776.

250. Rammensee HG, Friede T, Stevanoviic S. MHC ligands and peptide motifs: first listing. Immunogenetics. 1995;41(4):178-228. Epub 1995/01/01. PubMed PMID: 7890324.

251. Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, Sunpaweravong P, Han B, Margono B, Ichinose Y, Nishiwaki Y, Ohe Y, Yang JJ, Chewaskulyong B, Jiang H, Duffield EL, Watkins CL, Armour AA, Fukuoka M. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med. 2009;361(10):947-57. doi: 10.1056/NEJMoa0810699. PubMed PMID: 19692680.

252. Kames J, Alexaki A, Holcomb DD, Santana-Quintero LV, Athey JC, Hamasaki-Katagiri N, Katneni U, Golikov A, Ibla JC, Bar H, Kimchi-Sarfaty C. TissueCoCoPUTs: Novel Human Tissue-Specific Codon and Codon-Pair Usage Tables Based on Differential Tissue Gene Expression. J Mol Biol. 2020;432(11):3369-78. Epub 20200123. doi: 10.1016/j.jmb.2020.01.011. PubMed PMID: 31982380.

253. Pataskar A, Champagne J, Nagel R, Kenski J, Laos M, Michaux J, Pak HS, Bleijerveld OB, Mordente K, Navarro JM, Blommaert N, Nielsen MM, Lovecchio D, Stone E, Georgiou G, de Gooijer MC, van Tellingen O, Altelaar M, Joosten RP, Perrakis A, Olweus J, Bassani-Sternberg M, Peeper DS, Agami R. Tryptophan depletion results in tryptophan-to-phenylalanine substitutants. Nature. 2022;603(7902):721-7. Epub

20220309. doi: 10.1038/s41586-022-04499-2. PubMed PMID: 35264796; PMCID: PMC8942854.

254. Fan CH, Liu WL, Cao H, Wen C, Chen L, Jiang G. O6-methylguanine DNA methyltransferase as a promising target for the treatment of temozolomide-resistant gliomas. Cell Death Dis. 2013;4:e876. Epub 20131024. doi: 10.1038/cddis.2013.388. PubMed PMID: 24157870; PMCID: PMC4648381.