SPECTRA: AGNOSTIC EXPRESSION VARIABLES FOR FLEXIBLE TRANSCRIPTOME MODELING IN COMPLEX DISEASE

Rosalie G. Waller*, Heidi A. Hanson*, Michael J. Madsen, Brian Avery, Douglas Sborov, Nicola J. Camp
University of Utah School of Medicine

Abstract. Complex diseases, including cancer, are highly heterogeneous, and large molecular datasets are increasingly part of describing an individual's unique experience. Gene expression is particularly attractive because it captures both genetic and environmental consequences. Our new approach, SPECTRA, provides a framework of agnostic multi-gene linear transformations to calculate variables tuned to the needs of complex disease studies. SPECTRA variables are not supervised to an outcome and are quantitative, linearly uncorrelated variables that retain integrity to the original data and cumulatively explain the majority of the global population variance. Together these variables represent a deep dive into the transcriptome, including both large and small sources of variance. The latter is often overlooked but holds the potential for the identification of smaller groups of individuals with large effects, important for developing precision strategies. Each spectrum is a quantitative tissue phenotype that can be considered a phenotypic outcome, providing new avenues to explore disease risk. As a set, SPECTRA variables are ideal for modeling alongside other predictors for any clinical outcome of interest. We demonstrate the flexibility of SPECTRA variables for multiple endpoints using RNA sequencing from 767 myeloma patients in the CoMMpass study. SPECTRA enhances the ability to incorporating expression phenotypes in studies to advance precision screening, prevention, intervention, and survival.

Introduction

To identify risk and prognostic factors and understand complex diseases in a population, numerous data types are often collected on study participants. Transcriptomes represent the

combined effects of inherited and somatic insults as well as epigenetic and non-genetic factors and thus appeal to researchers interested in both genetic and environmental risk factors. For this reason, gene expression studies are gaining momentum in new fields, such as genomic epidemiology.[1–13] The need to incorporate transcriptomes in multivariable models alongside other risk factors brings new demands for techniques designed with this in mind.

Here we develop the SPECTRA approach to determine multi-gene transformations and calculate transcriptome variables with desirable attributes for multivariable modeling. Specifically, variables that optimize coverage of the global variance, are quantitative, uncorrelated and that retain integrity to the original data. A 'variable' has more power for modeling if it represents the variance in the population. Multiple variables may be required for broad coverage (a 'deep dive'). Knowledge of how much variance transcriptome variables represent is also important to understand the limitations of a study. Furthermore, truly quantitative variables can achieve greater power than if discretized.[14] Uncorrelated variables provide parsimony in penalized modeling and are often simpler to interpret in multivariable analyses. The integrity of variables to the original data (preservation of 'distance' between samples) is important for interpretation. Our goal for such attributes contrasts with the more common strategy to use transcriptome data to categorize samples or patients, reducing the transcriptome data to a single variable consisting of mutually exclusive categories, often called 'subtypes'.[15]

We focus on agnostic derivation. Current techniques for characterizing transcriptomes largely have a computational biology emphasis, interwoven with and constrained by biological knowledge. While these have had great success advancing our understanding of mechanism and pathway,[16–23] there remains room for complementary approaches. Sources of heterogeneity are complex and we require methods that match that complexity. Common diseases, and cancers, in particular, are multifactorial, where a wealth of other covariates may be equally important to an endpoint. New approaches that can embrace this complexity will enhance the toolset available for interrogating transcriptomes. Conceptually, the advantage of an agnostic

data-driven approach is the liberty to discover signals that may challenge conventional wisdom or defy “known” rules. Our agnostic approach is complementary and adds to current approaches for the analysis of tumor etiology, risk, treatment, and mortality.

Finally, our motivation is for universal measures, and hence our approach is unsupervised. SPECTRA produces a framework of multi-gene transformations to describe the expression space. The calculated SPECTRA variables can be used for different outcomes, providing the flexibility to explore the same variables across several different models (e.g. overall survival, progression-free survival, and time-to-treatment-failure). This can support interpretation and comparisons, improving the ability to decipher the true nature of associations and explore differences. Furthermore, the same framework of transformations can be implemented in external studies, increasing our ability to compare findings across multiple studies.

To satisfy these ideals, the core of our approach utilizes principal components analysis (PCA). PCA is an agnostic and unsupervised procedure that provides an isometry to provide a new set of orthogonal (linearly uncorrelated) variables that optimize the representation of the variance. In simple terms, PCA reveals the internal structure of the data in a way that best explains the variance in the data. Paramount in this approach is careful attention to quality control, normalization, and batch correction to ensure the variables capture meaningful variance. The results of the subsequent PCA are the rotation matrix that describes the multi-gene linear transformations; and the transformed data matrix, the quantitative variables for each individual that we refer to as a *SPECTRA variables*, or simply, *spectra*. Each measure is a spectrum that combined are spectra. The set of linear transformations provides a new reduced-dimension *framework* for the expression space. The SPECTRA variables are linearly independent, each providing additional coverage of the variance.

We previously used PCA to define a framework for the PAM50, a targeted and standardized gene-panel for breast cancer.[1,24] Using a population cohort of breast tumors, we used PCA to reduce the 50-gene space to five multi-gene expression variables. When the

framework was implemented in an external dataset of tumors from high-risk breast cancer pedigrees, these quantitative PCA variables as phenotypes proved superior to the standard PAM50 subtypes for gene mapping.[1,25] Further, when implemented in a second external clinical trial dataset, PCA variables were able to predict response to paclitaxel.[24] Here, we extend the approach to the whole transcriptome.

Improved representation of an individual's tissue (normal or diseased) will be vital to improving our ability to identify expression characteristics that are important phenotypic traits (predict disease risk) and/or important expression variables to predict patient outcomes. **Figure 1** uses a color analogy to illustrate the conceptual shift of SPECTRA, contrasting our goal of quantitative variables for direct use in outcome modeling with a more conventional categorization approach using hierarchical clustering. In our approach, each spectrum in Figure 1 (x_R, x_G, x_B) are independent variables that can be directly used to model any outcome (y_i), and other covariates/predictors can also be easily included (Figure 1d). Conversely, unsupervised hierarchical clustering uses the spectra to categorize patients into groups (Figure 1c), flattening the multiple spectra to a single categorical variable and reducing statistical power. For example, in Figure 1 analogy, x_R cannot be captured by any group ordering and associations for that spectra variable would be lost. An alternate convention is to supervise clustering to an outcome. But, while supervised clustering can improve power over unsupervised clustering for prediction of a single outcome, it also tethers the groups to the particular trained outcome and doesn't facilitate comparison to other outcomes.

We illustrate SPECTRA using the Multiple Myeloma Research Foundation (MMRF) CoMMpass Study data.[26] We derive the gene transformations (framework) for bulk whole transcriptome RNA sequencing (RNAseq) data from CD138+ myeloma cells. As a proof-of-concept, we utilize the spectra variables in various regression models to identify associations with several outcomes, including established risk scores, patient characteristics, and clinical endpoints.

Results

SPECTRA, a quantitative transcriptome approach

The motivation is the derivation of well-behaved, quantitative variables from RNAseq data to capture transcriptome variation that can be used universally as predictors for any outcome, and as novel phenotypes. The approach requires a dataset to derive the framework of transformations for the SPECTRA variables. Then multiple spectra are calculated for each individual in the dataset. An overview of the SPECTRA approach is shown in **Figure 2**. As an agnostic technique, the goal is to retain only those aspects of the RNAseq data that can represent meaningful variance. Accordingly, rigorous quality control (QC), normalization, and batch correction are performed before the derivation of the variables. Genes likely to lack precision are removed. Only coding genes with sufficient coverage across the dataset are considered. An internal normalization procedure accounts for feature-length, library size, and RNA composition. This normalization avoids the need for reference samples, real or synthetic, and provides the potential for spectra to be ported to follow-up samples and external datasets. Finally, skew and outliers are dealt with before PCA is performed. Specific details are listed in the Methods.

PCA is a well-established, data-driven method that, based on the covariance of a dataset, produces a matrix factorization which is a unique solution of linear transformations (*framework*, rotation matrix) and transformed values (*spectra*, transformed data). The linear transformations preserve the variance in the data, i.e., the transformed values preserve the distance between the sample data for individuals. Integrity to the original data provides meaningful comparisons between individuals. The resulting transformed values are quantitative variables that are orthogonal (linearly uncorrelated). For dimension reduction, components are ordered according to the amount of global variance they explain and the first k (S_1, …, S_k) selected, for which the proportion of total variance explained can be described. This reduces attention from 60,000+ expression features in a transcriptome to a handful of spectra specifically derived to represent independent components of the natural global variation across

the dataset studied; precisely the type of variables with power to identify differences and important for prediction. The procedure is unsupervised, describing only intrinsic variance in the data, hence the spectra can be incorporated into modeling any outcome in an unbiased way, and the framework of transformations can be implemented in external datasets.

When modeling, the significance and effect size for each individual spectrum as a predictor can be determined. In addition, the aggregate effect of all spectra in the model can be determined to describe the impact of the transcriptome as a whole. We define the aggregate effect of all spectra in a model as the poly-spectra liability (**PSL**) score for the outcome. This is the weighted sum of the spectra based on the model.

Illustrative case study: CD138+ spectra in multiple myeloma

The ultimate value of SPECTRA will be its use in the discovery of novel tissue phenotypes and predictors or outcomes in etiological studies. Here, as a proof-of-concept that a SPECTRA framework can capture meaningful information, we present associations between a set of spectra to several well-established outcomes or risk groups for multiple myeloma, across several different model types. These are not presented to suggest spectra could replace current clinical tests, but to illustrate the flexibility of SPECTRA to provide a universal transcriptome framework and set of variables for use in various models with disparate outcomes. We applied SPECTRA to transcriptome data for CD138+ cells from the MMRF CoMMpass study.[26] We investigated associations of CD138+ spectra with 1) existing expression-based risk scores;[27,28] 2) clinically-relevant DNA aberrations; 3) clinical prognostic outcomes, and 4) patient demographic groups with elevated myeloma risk. Also, we illustrate the potential to track transcriptome changes over time.

The CoMMpass dataset is the most extensive sequencing effort in multiple myeloma patients to date. Multiple myeloma is a malignancy of plasma cells (CD138+ cells). The publicly available transcriptome data (IA14) comprised RNAseq data for 887 CD138+ samples on 794 unique patients. Here, data for 768 patients with treatment naïve samples collected at diagnosis

were the focus. We used transcript-based expression estimates from Salmon,[29] generated by the CoMMpass study (https://research.themmrf.org). From the total 54,324 features, 7,436 genes and 767 patients' treatment-naïve CD138+ RNAseq data met quality control. The transcriptome framework and spectra were derived in a quality controlled, normalized and batch corrected data from these treatment naïve samples. The first *k* = 39 spectra (S1—S39) were selected, based on a scree test, which captured $v = 65\%$ of the global variation. No spectra showed association with batch (F-statistic, all $p > 0.8$). The details from each step of the QC process, the linear transformations necessary to calculate the 39 spectra, and the individual-level spectra variables for the patients in the IA14 CoMMpass data are provided in Supplemental Data. R markdown notebooks containing the code used to generate CD138+ spectra in the IA14 dataset, full model analyses, and results are provided in the Supplemental Materials.

As linearly uncorrelated variables, each of the 39 CD138+ spectra captures a different source of variance, and hence any spectrum has the potential to explain patient differences and provide insight. **Figure 3** shows spectra charts for 4 patients and illustrates that while patients may be similar at a high-level (overall patterning), that individual spectra may not follow that apparent high-level similarity.

Common approaches to prediction modeling include penalized or stepwise techniques to address concerns about multicollinearity and improve fit and parsimony. Here, we included all 39 spectra into each model for simplicity and to ease comparison across results. Association results for the full 39-spectra models for several different outcomes are described below. Overall model significance and the significance for individual spectra in those models are summarized in **Figure 4**.

CD138+ spectra and established expression-based risk scores. The most widely adopted and first supervised expression risk score in myeloma is the University of Arkansas UAMC 70-gene panel, developed in microarray data, and used to classify patients as low- or high-risk for

relapse.[27] Using the established classifier, we calculated each patient's risk UAMC-70 risk score and their risk status (low or high). In a multivariable linear regression with UAMC-70 risk score as a continuous outcome variable, 30 spectra were individually significant ($p<0.05$, Figure 4) and the full model predicted the UAMC 70-gene risk score with high accuracy and significance ($R^2 = 0.86$, $F_{39,727} = 118.1$, $p < 10^{-50}$). A more recent supervised expression risk score is the Shahid Bahonar University of Kerman 17-gene prognostic score (SBUK-17) published in 2020 by Zamani-Ahmadmahmudi et al.[28] We calculated each patient's SBUK-17 prognostic score. In a linear regression with SBUK-17 as a continuous outcome, 25 spectra were significant (Figure 4) and the full model predicted the SBUK 17-gene score with excellent accuracy and significance ($R^2 = 0.93$, $F_{39,272} = 252.9$, $p < 10^{-50}$). **Figure 5** illustrates the high correlation between the model PSL scores and the previously established risk scores and shows spectra can recapitulate previously established supervised expression risk scores. These results indicate that the spectra framework captures important prognostic signals.

CD138+ spectra and clinical risk. Large somatic chromosomal DNA aberrations detected by cytogenetics are used clinically to define prognostic risk groups in myeloma.[30] Clinical risk categories defined by mSMART[31] include: high risk (del(17p) and t(14;16)); intermediate risk (amp(1q) and t(4;14)); and standard risk (t(11;14)). Models for each of these five chromosomal aberrations (**Figure 4**) showed different spectra individually significant, with some spectra unique to only one aberration. Interestingly, while the models for all three translocations and amp(1q) were highly significant (all $p < 2x10^{-10}$), the full 39-spectra model for del(17p) was not. To investigate the possibility that the model was over-parameterized, we repeated the del(17p) analysis using a stepwise procedure. This produced a significant model containing only 3 spectra ($p = 0.014$, Supplemental Material). These results indicate transcriptome spectra capture signals from DNA chromosomal changes in CD138+ cells (**Figure 6a-b**).

The international staging system (**ISS**) for myeloma is also used to classify and stratify patients at diagnosis, based on somatic cytogenetics, levels of beta-2 microglobulin, albumin,

and lactate dehydrogenase in the blood.[32] In an ordinal logistic regression with the ISS stage at diagnosis as the outcome, 13 spectra were significant, providing a model that significantly differentiated the three clinical stages (**Figure 6c**). These results indicate spectra can capture signals for the disease stage.

CD138+ spectra and disease course. We used Cox proportional hazards analysis to associate spectra with overall survival (**OS**), and time to first-line treatment failure (**TTF**). Spectra significantly predicted OS (179 events, likelihood ratio test, $p = 3.1 \times 10^{-17}$, C-statistic = 0.74), with 12 spectra individually significant. A Cox proportional hazards model for TTF was also successful (369 events, likelihood ratio test, $p = 7.9 \times 10^{-10}$, C-statistic = 0.66), with 8 spectra significant. Spectrum significant in both models had effects in the same direction (**Figure 4**).

Patients were categorized into three equal tertiles based on PSL scores for OS and TTF. Kaplan Meier curves for these three equal groups for OS and TTF are shown in Figure 7. For OS, patients in OS-PLS tertile 3 had hazard ratios of 6.7 (2.9-15.3) and 8.8 (5.1-15.3) at 1 year and 3 years, respectively, compared to patients in tertile 1. For TTF, comparing TTF-PLS tertile 3 to tertile 1, hazard ratios were 4.8 (3-7.7) and 7.2 (4.3-12) at 1 year and 3 years, respectively. These results indicate spectra can capture signals and differentiate patients for disease course.

CD138+ spectra and demographic risk groups. Myeloma is an adult-onset malignancy, most frequently diagnosed at ages 65-74 years (median 69 years).[33] Incidence is higher in men (8.7 men vs. 5.6 women per 100,000) and patients self-reporting as African American (AA men 16.3, and AA women 11.9 per 100,000). Linear regression with age at diagnosis as a quantitative outcome was significant ($p = 2 \times 10^{-14}$), with 15 individually-significant spectra. Logistic regression models for gender, race (self-reported black or white; other racial categories too small to consider) and Hispanic status were all significant ($p = 4 \times 10^{-9}$, $p = 9 \times 10^{-10}$ and $p = 1 \times 10^{-3}$, respectively) (**Figure 4**). We note that associations found for demographic risk factors may be complex, as such factors involve social constructs, e.g. race and ethnicity. Transcriptomes can

harbor the effects of genetic, epigenetic, lifestyle, and environmental factors. These results indicate spectra can capture signals originating from demographic risk.

CD138+ spectra for tracking changes over time. The SPECTRA framework provides the transformations for the spectra variables such that they can be calculated in follow-up samples and tracked over time. We illustrate this potential in the eleven MM patients for whom at least three longitudinal CD138+ samples were available in the CoMMpass study. **Figure 8** shows a line graph of the PSL score for OS for these eleven patients over 80 months. In this example, the potential for tracking a patient's hazard over time is illustrated using the OS PSL score.

Discussion

The promise of personalized prevention, management, and treatment is rooted in an ability to describe an individual's unique experience and model important sources of heterogeneity.[34] In complex diseases, and cancer specifically, gene expression in diseased tissue may be an established source of heterogeneity. [35] Tools that can take a deeper dive and characterize multiple sources of expression heterogeneity will be important to advance the promise of personalized medicine. In particular, for human studies and domains such as epidemiology wishing to model multiple sources of risk in a population, transcriptome variables that can be easily incorporated with other variables are needed. The goal of this study was to provide a technique to derive an agnostic framework of variables for transcriptome data, to empower multivariable studies, and provide novel molecular phenotypes. SPECTRA identifies quantitative, orthogonal variables (non-correlated) that capture sources of transcriptome variation for use in subsequent modeling or as quantitative phenotypes. Many applications can benefit from the qualities of spectra variables, and this new framework has the potential to provide utility to numerous study designs and many outcome types.

Data quality and processing are paramount in the quest to derive informative variables. PCA itself is a simple procedure that provides linear transformations of the data to best

represent variance. If the data have technical artifacts, batch effects, unstable or non-comparable expression measures, the noise will overwhelm authentic variance. Accordingly, our technique intentionally includes strict quality control, zero-handling and normalization procedures, and batch correction (Figure 2). Without these steps, PCA can fail to provide variables with the desired qualities. An agnostic approach permits stringent data culling because the incentive to retain features based on known functional relevance is removed. The impetus is to only retain features that can contribute to meaningful variance and provide informative variables for modeling (quantitative, orthogonal, variance-representing). Of course, the limitation of an agnostic approach is reduced biological interpretation or insight into the mechanism of the variables before modeling. However, there are already many approaches that take this alternate goal of intermediate interpretation,[18] whose limitations are instead the flexibility of the variables they produce. Hence, SPECTRA offers a complementary approach to the current toolset available for all fields.

Beyond the agnosticism taken by our proposed technique, other potential advantages of SPECTRA include its unique solution within a dataset, such that the rank of the dimension reduction can be post-hoc and does not influence the definition of retained dimensions. As a statistically rigorous technique, it also provides a measured dive into the transcriptome. Each dimension (eigenvector $\boldsymbol{q}_s$) iteratively moves quantifiably deeper into the variance of the data (measurable by λ_s). Methods that iteratively find independent components (PCA and independent component analysis) have previously been shown to provide superior coverage of transcriptome data.[23] Retention of components deep in the data, representing small variances (i.e., deep dives) provide potential and power to identify small groups of individuals with large effects in outcome studies, such as a molecular phenotype that hones-in on a rare Mendelian form of cancer, or the few patients that respond to a drug. These findings could be the 'low hanging fruit' scenarios where the precision translation is more straight-forward. SPECTRA also embraces negative weights. The allowance of negative values in its matrix factorization (**MF**) is often given as a criticism of PCA,[16,36] argued as a conceptual source of its lack of biological

interpretability; a premise that components may mix biological processes due to a focus on variance. Non-negative matrix factorization (**NMF**), arguably the leading approach in the computational biology field, restricts all values in the amplitude (equivalent to $\boldsymbol{Q}^T$ in PCA) and pattern (equivalent to $\boldsymbol{T}^T$ in PCA) matrices to be non-negative. Reasoned as beneficial and a natural restriction because expression values themselves cannot be negative. Also, because NMF transformed values represent the proportions of each factor and thus provide a simple interpretation. However, non-negative values may not be a natural restriction to *systems of genes*, and over-simplicity may not adequately represent true complexity. With a non-negativity restriction, NMF limits itself to the identification of groups of over-expressed genes,[16,36] modeling only neutrality and surplus. Deficits may also be important. So, while PCA spectra may represent mixtures of different biological mechanisms, these may be important combinations, including genes acting in opposite directions, and may better reflect reality. By embracing negative values, PCA can also capture gene systems in deficit, which may be more difficult to interpret, but may equally be just as important to recognize. These differences underscore SPECTRA's value as a complementary tool to existing approaches.

Our myeloma case study illustrated derivation of a transcriptome framework and spectra variables for CD138+ cells, and the application of these in various models (linear, logistic and Cox regression, and ANOVA) with many different outcomes. We showed that the set of 39 unsupervised, agnostic spectra could significantly capture signals corresponding to published expression-based risk scores from traditional supervised approaches, known clinical DNA-based risk factors, disease stage, disease progression, survival, and demographic risk groups. We also illustrated the potential to track tissue changes using PSL scores over time. As expected for a framework of agnostically derived variables, not all spectra are relevant to every outcome. Across the 14 models presented, the number of individually significant spectra in a model ranged from 3 to 30, and only one spectra-variable (S30) showed no association with any model. Importantly, these examples show the flexibility of the framework as well as how it can support comparisons across different models and outcomes. For example, our results illustrated

how two previously established gene-expression prognostic scores could be captured using a single framework and illustrate that they are similar (**Figure 4**). Hence, the framework has the potential to provide a bridge to compare various existing categorizations (subtypes) of patients, even when no genes overlap in their signatures, or they predict different outcomes.[37] In this way, spectra provide an alternate to categorical intrinsic subtyping, a well-established practice for many cancers.[38] The ability to predict OS and TTF suggests spectra hold utility in clinical studies predicting disease course. Clinically-relevant stratification may be better represented using thresholds within a transcriptome framework.[24]

The potential for increased power using spectra variables is illustrated by the discovery of novel associations between spectra and patient demographic risk groups with known differences in incidence (age, gender, race). Prior studies, using the UAMC 70-gene panel and a Ki67 proliferation index, were not able to identify gene expression differences in CD138+ cells from self-reported AA and white patients.[39] Our multivariable results demonstrate that significant differences do exist, but also illustrate that the diseased cells in these demographic groups are not distinct entities; fewer than half the spectra variables differ significantly by these patient demographic groups. Focusing on the spectra that do show differences by demographics provides new avenues to explore why incidence varies in these groups; a key to disease prevention, intervention, and control. In particular, because transcriptomes capture both the effects of internal (inherited genetics) and external factors (lifestyle, exposures, consequences of access to care), these results could also support epidemiology and biosociology investigations into such differences. We provide the variable framework (gene transformations) and the spectra variables for the CoMMpass patients in Supplementary material to enable further study of spectra in other CoMMpass studies, as well as in other myeloma studies.

There are numerous potential applications beyond those undertaken here that could benefit from a statistically rigorous transcriptome framework of expression variables. As shown previously for the PAM50 panel in breast tumors, differences can be observed between familial and sporadic tissues, suggesting familial components,[1] and defining powerful new phenotypes

for genetic, exposure, and gene-environment studies. Future avenues for spectra as quantitative phenotypes could include expression-quantitative trait locus analyses, Mendelian randomization, seeding machine-learning applications,[40] tissue measures for pre-clinical models, and corollary studies in clinical trials.

As for any approach, there are limitations. A key question of one of representation. For epidemiology studies, for example, spectra should ideally be representative of the entire disease population. This requires that the derivation dataset is a random sample from that population, or based on a known selective sampling scheme. While there are many publicly available transcriptome datasets,[41,42] most fall short of this ideal. Thus, the spectra variables derived from these will have inherent limitations in representation. An investigator should consider if a derivation dataset is adequate to represent their study goals. We note that the goal of the MMRF CoMMpass study was intentionally designed to represent myeloma patients from diagnosis through treatment, and is the largest existing cohort of treatment-naïve CD138+ transcriptomes, with sampling continuing over time. However, the demographic representation of patients was not achieved, and this remains a limitation of that study. Another limitation is that, as a simple variance-based procedure, PCA models all sources of variance in the dataset. If artifacts remain in the data, the resulting spectra will also represent these. To minimize this issue, we employed a strict data quality and batch correction process in our workflow, concentrating only on a subset of genes for which PCA is likely to be meaningful: well-mapped, stable, with sufficient depth, and with batch correction. We also removed genes known to be unstable across different RNAseq pipelines.[43] A third limitation is the ability to use the framework of spectra in external studies. As a data-driven technique, the complete PCA decomposition is overfitted to the derivation dataset. To limit this, we use dimension reduction and focus on the first k spectra (largest k components of variation), selected using a scree test[44] to be those before decreasing marginal returns. Last, SPECTRA is intentionally agnostic, designed for modeling, and dimensions are not pre-interpreted for functional relevance. Hence,

post-hoc analyses will be required to uncover the mechanism/s that underlie the associations identified.

In conclusion, we present a new technique, SPECTRA, to derive an agnostic transcriptome framework of quantitative, orthogonal variables for a dataset. These multi-gene expression variables are designed specifically to capture transcriptome variation, providing new transcriptome phenotypes and variables for flexible modeling, along with other covariates, to better differentiate individuals for any outcome. Applied to CD138+ transcriptomes for myeloma patients, we defined CD138+ spectra and implemented these in many different outcome models. We illustrated an ability to predict prognosis, survival, clinical risk, and provide new insight into potential differences between patients from demographic groups. Fundamentally, the technique shifts from categorization to multiple quantitative measures. SPECTRA variables provide a new paradigm and toolset for exploring transcriptomes that hold promise for discoveries to advance precision screening, prevention, intervention, and survival studies.

Methods

Conceptual construction

Here we establish the matrix factorization (**MF**) natural for individual-based outcome modeling. Data matrices, $\boldsymbol{X}$ and $\boldsymbol{T}$, are <u>*oriented with individuals as subjects (n rows) and genes as variables (g columns)*</u>. Given a $n \times g$ design matrix, $\boldsymbol{X}$ (mean-centered expression values for n individuals on g genes), PCA is the MF

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{Q}^T \qquad \textit{Equation 1}$$

where $\boldsymbol{T}$ contains the transformed values (the dimension variables), and $\boldsymbol{Q}$ is the PCA 'rotation' matrix. Each row in $\boldsymbol{Q}^T = \left(\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_g\right)^T$ is an orthogonal eigenvector (or component) which holds the coefficients for the linear model to transform the observed gene values into the

spectra variables. The set of linear transformations are the transcriptome framework. The rotation matrix can be derived from the eigen decomposition of the covariance matrix, $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T \qquad \textit{Equation 2}$$

where $\boldsymbol{\Sigma}$ is proportional to $\boldsymbol{X}^T\boldsymbol{X}$, and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. Each eigenvalue, λ_s, is a scalar indicating the proportion of the global variance represented by the transformed value defined by the sth eigenvector, $\boldsymbol{q}_s$, in $\boldsymbol{Q}$. Eigenvalues are ranked, such that the first PC, defined by $\boldsymbol{q}_1$ captures the most variance, $\boldsymbol{q}_2$ the next highest, and so on. We note that there can only be $\min(n, g)$ non-zero eigenvalues, because by definition, beyond this no variance remains. In most, if not all, existing RNAseq studies, there are more genes than individuals and hence n is the limiting rank.

Dimensionality can be reduced to k dimensions by utilizing $\boldsymbol{Q}_k$; only the first k columns (PCs) of $\boldsymbol{Q}$. After selection of k PCs, transformed values are represented as:

$$\boldsymbol{T}_k = \boldsymbol{X}\boldsymbol{Q}_k \qquad \textit{Equation 3}$$

We note that PCA is deterministic and therefore the selection of k is a post-procedure decision that does not influence the MF. The proportion of variance explained by the retained dimensions ($\sum_{s=1}^{k} \lambda_s / \sum_{\forall s} \lambda_s$) can be used as a measure of coverage.

SPECTRA workflow

Careful attention to quality control, normalization, and batch correction are used to ensure the spectra capture meaningful variation. Gene expression counts from bulk RNAseq are the input data. The four steps in the workflow are (1) quality control; (2) internal normalization; (3) correction for batch effects; (4) PCA and dimension reduction (**Figure 2**).

Quality control. QC is essential to ensure the transcriptome dimensions capture meaningful variation across the individuals. Features in the transcriptome likely to be unduly influenced by poor alignment or lacking precision due to sequencing depth were removed as potentials for introducing spurious and unstable variation. Accordingly, we removed all non-autosomal and non-protein-coding genes as well as features with low counts. A feature was considered to have inadequate data for precision if more than 5% of samples had fewer than 100 read counts. After the removal of features, individuals were removed from consideration if more than 10% of the remaining features had fewer than 100 read counts.

Normalization. This is required for comparisons across genes and individuals and includes adjustment for gene length, sequencing depth (library size), and RNA composition. Zero-handling is also necessary to appropriately incorporate counts of zero for a feature or transcript). We chose to use a robust internal (single sample) normalization to obviate the need for a 'reference' sample and to provide the possibility for portability across datasets. While our technique is gene-focused, our processing is designed to handle transcript-based alignment and quantification because these have been suggested to be more accurate.[45] Normalized gene expression estimates, e_g, were calculated according to the following procedure:

$$\mathrm{e}_g = \log_2\left(\frac{\sum_{t=1}^{m}\frac{\mathrm{c}_t+1/m}{\mathrm{l}_t}}{\mathrm{median}\left(\sum_{t=1}^{m}\frac{\mathrm{c}_t+1/m}{\mathrm{l}_t}\right)}\right) \qquad \textbf{\textit{Equation 4}}$$

where c_t is the read count for transcript t, l_t is the transcript length in kilobases (extracted from the GTF used to align and quantify the RNAseq data), and m is the number of transcripts for the gene. Zero-handling is achieved by adding $1/m$ to the transcript counts: $\mathrm{c}_t + 1/m$. Division by l_t corrects for transcript length. Summing the length-corrected transcript counts results in a gene-level count per kilobase (CPK) measure. **Equation 4** may also be used for gene-level read counts (equivalent to m =1). Adjustments for sequencing depth and RNA composition (often

referred to as the $size\ factor$) is achieved via division of each gene-based CPK measure by the median of CPK-values for retained features. We note that the more usual upper-quartile adjustment also provides robust internal normalization;[46] however, since our implementation is post-QC after numerous features have been removed for low counts, the median is more suitable. Normalized data are log_2 transformed to account for skew. We also truncate outliers beyond the five standard deviation thresholds from the mean of the normalized gene counts to the relevant threshold value.

Batch correction. Sequencing is often generated in batches, and it is necessary to correct for the potential of technical artifacts and any spurious variation introduced. We adjust for sequence batch using ComBat[47] as implemented in the sva R package,[48] with patient characteristics that are unbalanced by batch included as covariates.

PCA. We implement PCA with the covariance matrix. For functions that use singular value decomposition to perform PCA, it is necessary to center the expression values first to ensure the MF is performed for the covariance. Expression values (e_g) are centered on the mean across individuals for gene g. These centered data represent the design matrix, $\boldsymbol{X}$ $(n \times g)$ (**Equation 1**) for the PCA. The R core function $prcomp(x = \boldsymbol{X}, center = TRUE, scale = FALSE, retx = TRUE)$ was used to perform PCA. We use a scree test[44] (the inflection point of the rank-ordered plot of λ_s, or elbow method) to select the k spectra to retain. The proportion of variance explained by this k-dimensional space ($\sum_{s=1}^{k} \lambda_s / \sum_{\forall s} \lambda_s$) indicates the depth of the dive into the transcriptome data.

CD138+ spectra in myeloma

Data were generated as part of the MMRF CoMMpass Study (release IA14)[26] and downloaded from the MMRF web portal (https://research.themmrf.org/). Clinical data and CD138+ RNAseq were available for 781 patients at baseline (newly diagnosed bone marrow

samples) and 123 follow-up bone marrow samples. Transcript-based expression estimates processed by Salmon (version 0.7.2) were used. The 768 baseline samples were used in the PCA to derive the CD138+ transcriptome framework and SPECTRA variables. Covariates included in batch correction were age, gender, overall survival, progression-free survival, and time to first-line treatment failure. The first 39 components were selected based on the scree test.[44] All 39 spectra were forced variables in all regression models.

To illustrate the flexibility of the transcriptome framework, linear, logistic, and Cox regression were performed for several different clinical outcomes and demographic risk groups. In each analysis, all 39 CD138+ spectra were entered into the model as independent, predictor variables. No model fitting was performed. An individual spectrum was considered significant in a model if its model coefficient was significantly different from 1.0 ($p < 0.05$). A likelihood ratio test comparing the full 39-spectra model to the null model was used to determine the significance of the overall model fit. To illustrate the aggregate effect of all spectra in the model we used a poly-spectra liability (**PSL**) score. This score is the weighted sum of the spectra values based on the spectra coefficients in the model. In our illustrations here, the PSL scores contain all 39 spectra. In other applications, such as penalized modeling, a PSL score may include only those spectra retained in the model.

To illustrate the potential to track longitudinal changes, spectra and PSL scores were calculated for follow-up longitudinal samples. To enable this, batch corrected gene-level measures for the follow-up samples were centered on the mean of the baseline data ($\bar{\boldsymbol{e}}$), and then multiplied with the rotation matrix ($\boldsymbol{Q}_k$) which holds the linear transformations for the spectra framework.

Data availability. Processed RNAseq data from the CoMMpass Study can be downloaded from https://research.themmrf.org/. Dimension variables for the IA14 CoMMpass data are provided in the Supplement. We also provide the details of the QC process and the transcriptome

framework (linear equations for the gene transformations) necessary to calculate the 39-spectra variables in other studies in the Supplement.

Code availability. R markdown notebooks used to derive the CD138+ transcriptome spectra and generate the myeloma results are included in the Supplement.

Funding. The research reported in this publication was supported by the National Cancer Institute (Award Numbers F99CA234943, K00CA234943, K07CA230150, and P30CA042014-29S9), the National Center for Advancing Translational Sciences (Award Number UL1TR002538), the National Library of Medicine (Award Number T15LM007124) of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Madsen, M. J. *et al.* Reparameterization of PAM50 Expression Identifies Novel Breast Tumor Dimensions and Leads to Discovery of a Genome-Wide Significant Breast Cancer Locus at *12q15*. *Cancer Epidemiol Biomarkers Prev* **27**, 644–652 (2018).
2. Sweeney, C. *et al.* Intrinsic subtypes from PAM50 gene expression assay in a population-based breast cancer cohort: differences by age, race, and tumor characteristics. *Cancer Epidemiol. Biomarkers Prev.* **23**, 714–724 (2014).
3. Stopsack, K. H. *et al.* Regular aspirin use and gene expression profiles in prostate cancer patients. *Cancer Causes Control* **29**, 775–784 (2018).
4. Wang, S. *et al.* Gene expression in triple-negative breast cancer in relation to survival. *Breast Cancer Res. Treat.* **171**, 199–207 (2018).
5. Bhattacharya, A. *et al.* A framework for transcriptome-wide association studies in breast cancer in diverse study populations. *Genome Biol.* **21**, 42 (2020).
6. Caan, B. J. *et al.* Intrinsic subtypes from the PAM50 gene expression assay in a population-based breast cancer survivor cohort: prognostication of short- and long-term outcomes. *Cancer Epidemiol. Biomarkers Prev.* **23**, 725–734 (2014).
7. Allott, E. H. *et al.* Bimodal age distribution at diagnosis in breast cancer persists across molecular and genomic classifications. *Breast Cancer Res. Treat.* **179**, 185–195 (2020).
8. Troester, M. A. *et al.* Racial Differences in PAM50 Subtypes in the Carolina Breast Cancer Study. *J. Natl. Cancer Inst.* **110**, (2018).
9. Huo, D. *et al.* Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol* **3**, 1654–1662 (2017).

10. Millstein, J. *et al.* Prognostic gene expression signature for high-grade serous ovarian cancer. *Ann. Oncol.* (2020) doi:10.1016/j.annonc.2020.05.019.

11. López, C. *et al.* Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. *Nat Commun* **10**, 1459 (2019).

12. Zhu, B. *et al.* Immune gene expression profiling reveals heterogeneity in luminal breast tumors. *Breast Cancer Res.* **21**, 147 (2019).

13. Zhang, M. *et al.* Characterising cis-regulatory variation in the transcriptome of histologically normal and tumour-derived pancreatic tissues. *Gut* **67**, 521–533 (2018).

14. Altman, D. G. & Royston, P. The cost of dichotomising continuous variables. *BMJ* **332**, 1080 (2006).

15. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).

16. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 4164–4169 (2004).

17. Reich, M. *et al.* The GenePattern Notebook Environment. *Cell Syst* **5**, 149-151.e1 (2017).

18. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).

19. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 2907–2912 (1999).

20. Zhang, S., Li, X., Lin, Q., Lin, J. & Wong, K.-C. Uncovering the key dimensions of high-throughput biomolecular data using deep learning. *Nucleic Acids Res.* **48**, e56 (2020).

21. Chen, K. M. *et al.* PathCORE-T: identifying and visualizing globally co-occurring pathways in large transcriptomic compendia. *BioData Min* **11**, 14 (2018).

22. Sompairac, N. *et al.* Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. *Int J Mol Sci* **20**, (2019).

23. Way, G. P., Zietz, M., Rubinetti, V., Himmelstein, D. S. & Greene, C. S. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome Biol.* **21**, 109 (2020).

24. Camp, N. J. *et al.* Re-interpretation of PAM50 gene expression as quantitative tumor dimensions shows utility for clinical trials: application to prognosis and response to paclitaxel in breast cancer. *Breast Cancer Res Treat* **175**, 129–139 (2019).

25. Hanson, H. A. *et al.* Family Study Designs Informed by Tumor Heterogeneity and Multi-Cancer Pleiotropies: The Power of the Utah Population Database. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* **29**, 807–815 (2020).

26. Keats, J. J. *et al.* Interim Analysis Of The Mmrf Commpass Trial, a Longitudinal Study In Multiple Myeloma Relating Clinical Outcomes To Genomic and Immunophenotypic Profiles. *Blood* **122**, 532–532 (2013).

27. Shaughnessy, J. D. *et al.* A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* **109**, 2276–2284 (2007).

28. Zamani-Ahmadmahmudi, M., Nassiri, S. M. & Soltaninezhad, F. Development of a RNA sequencing-based prognostic gene signature in multiple myeloma. *Br J Haematol* bjh.16744 (2020) doi:10.1111/bjh.16744.

29. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417–419 (2017).

30. Paner, A., Patel, P. & Dhakal, B. The evolving role of translocation t(11;14) in the biology, prognosis, and management of multiple myeloma. *Blood Rev.* 100643 (2019) doi:10.1016/j.blre.2019.100643.

31. Mikhael, J. R. *et al.* Management of Newly Diagnosed Symptomatic Multiple Myeloma: Updated Mayo Stratification of Myeloma and Risk-Adapted Therapy (mSMART) Consensus Guidelines 2013. *Mayo Clinic Proceedings* **88**, 360–376 (2013).

32. Palumbo, A. *et al.* Revised International Staging System for Multiple Myeloma: A Report From International Myeloma Working Group. *JCO* **33**, 2863–2869 (2015).

33. Myeloma - Cancer Stat Facts. https://seer.cancer.gov/statfacts/html/mulmy.html.

34. Ramón y Cajal, S. *et al.* Clinical implications of intratumor heterogeneity: challenges and opportunities. *J Mol Med* **98**, 161–177 (2020).

35. Kwa, M., Makris, A. & Esteva, F. J. Clinical utility of gene-expression signatures in early stage breast cancer. *Nat Rev Clin Oncol* **14**, 595–610 (2017).

36. Stein-O'Brien, G. L. *et al.* Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet.* **34**, 790–805 (2018).

37. Szalat, R., Avet-Loiseau, H. & Munshi, N. C. Gene Expression Profiles in Myeloma: Ready for the Real World? *Clin Cancer Res* **22**, 5434–5442 (2016).

38. Dai, X. *et al.* Breast cancer intrinsic subtype classification, clinical use and future trends. *Am J Cancer Res* **5**, 2929–2943 (2015).

39. Manojlovic, Z. *et al.* Comprehensive molecular profiling of 718 Multiple Myelomas reveals significant differences in mutation frequencies between African and European descent cases. *PLoS Genet* **13**, e1007087 (2017).

40. Ma, S. & Dai, Y. Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics* **12**, 714–722 (2011).

41. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

42. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

43. Arora, S., Pattwell, S. S., Holland, E. C. & Bolouri, H. Variability in estimated gene expression among commonly used RNA-seq pipelines. *Sci Rep* **10**, 2734 (2020).

44. Cattell, R. B. The Scree Test For The Number Of Factors. *Multivariate Behav Res* **1**, 245–276 (1966).

45. Zhao, S., Xi, L. & Zhang, B. Union Exon Based Approach for RNA-Seq Gene Quantification: To Be or Not to Be? *PLoS ONE* **10**, e0141910 (2015).

46. Shahriyari, L. Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. *Briefings in Bioinformatics* **20**, 985–994 (2019).

47. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

48. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

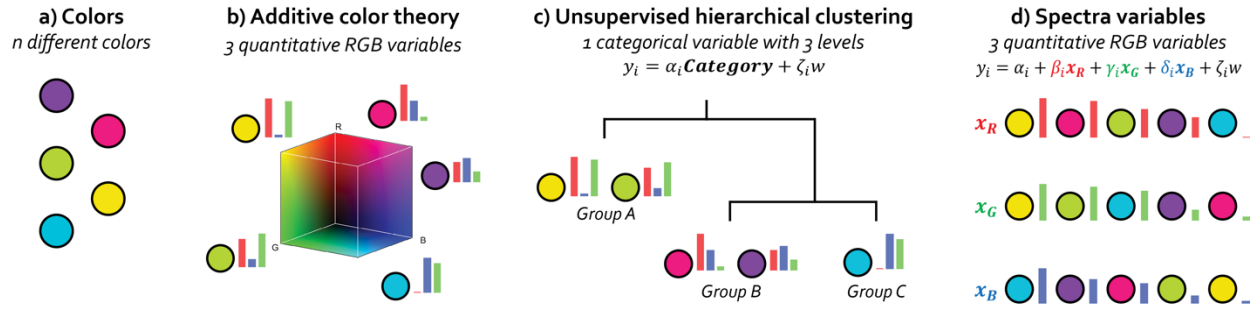


Figure 1. **A color analogy to illustrate the advantages of spectra variables for modeling**. a) Individual observations of color. b) Dimension Reduction (additive color theory), all colors can be represented using 3 quantitative RGB variables. c) Standard-use, modeling on the 3 RGB variables used to identify structure across samples using hierarchical clustering. This derives groups based on the complete 3-variable RGB profile to derive one polychotomous meta-variable (different groups are non-ordinal levels). d) Multivariable modeling implementation of spectra variables, multiple separate spectrum integrated directly into a multivariable analysis. Each uncorrelated variable can be assessed separately for its predictive value for an outcome. This implementation retains the full resolution of the initial data because the variables are quantitative and retain integrity to the initial data. Note, lower-resolution versions of x_B and x_G can be achieved using hierarchical groups but the loss of quantification will likely also lose power. x_R cannot be captured by any group ordering and associations for this spectrum would be lost using hierarchical groups.

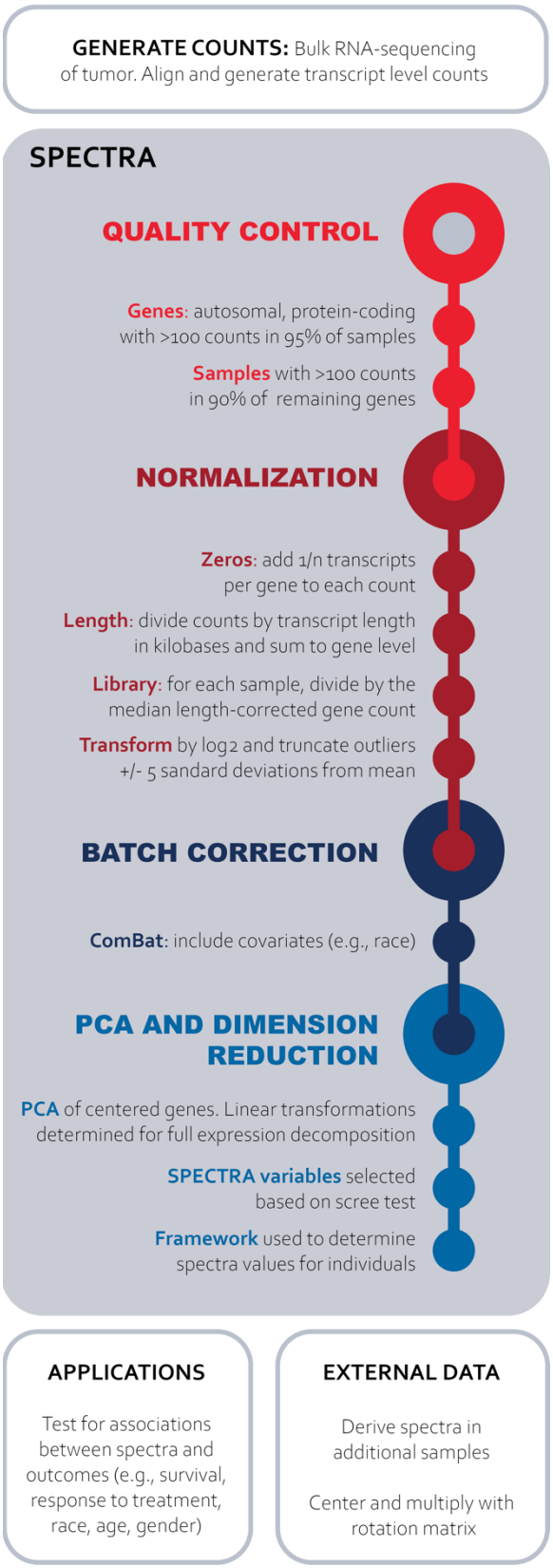


Figure 2. Overview of SPECTRA workflow.

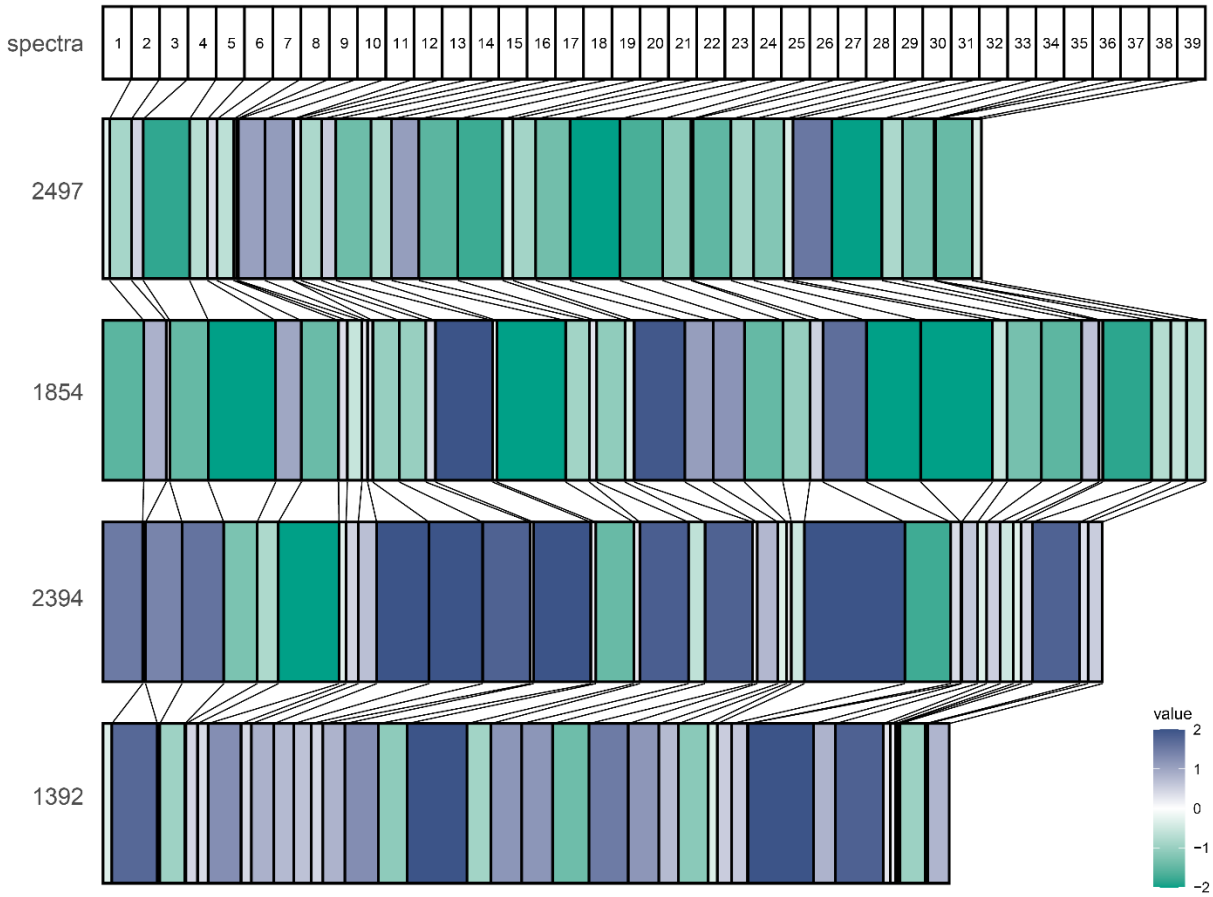


Figure 3. Spectra charts in four CoMMpass patients. For each patient, all 39 spectra are illustrated with the value represented by the bar width and intensity. The color indicates if the patient's spectra value is positive or negative. Each patient has a unique profile across the 39 spectra. At a high-level patients 2497 & 1854 may appear most similar (mostly green) and 2394 & 1392 (mostly blue). However, at a finer resolution, similarities vary. For example, for spectrum S1 and S2, patients 2497 and 1854 are quite different. In fact, for spectrum S1, 2497 is more similar to 1392, for spectrum S15, 1854 is more similar to 2394.

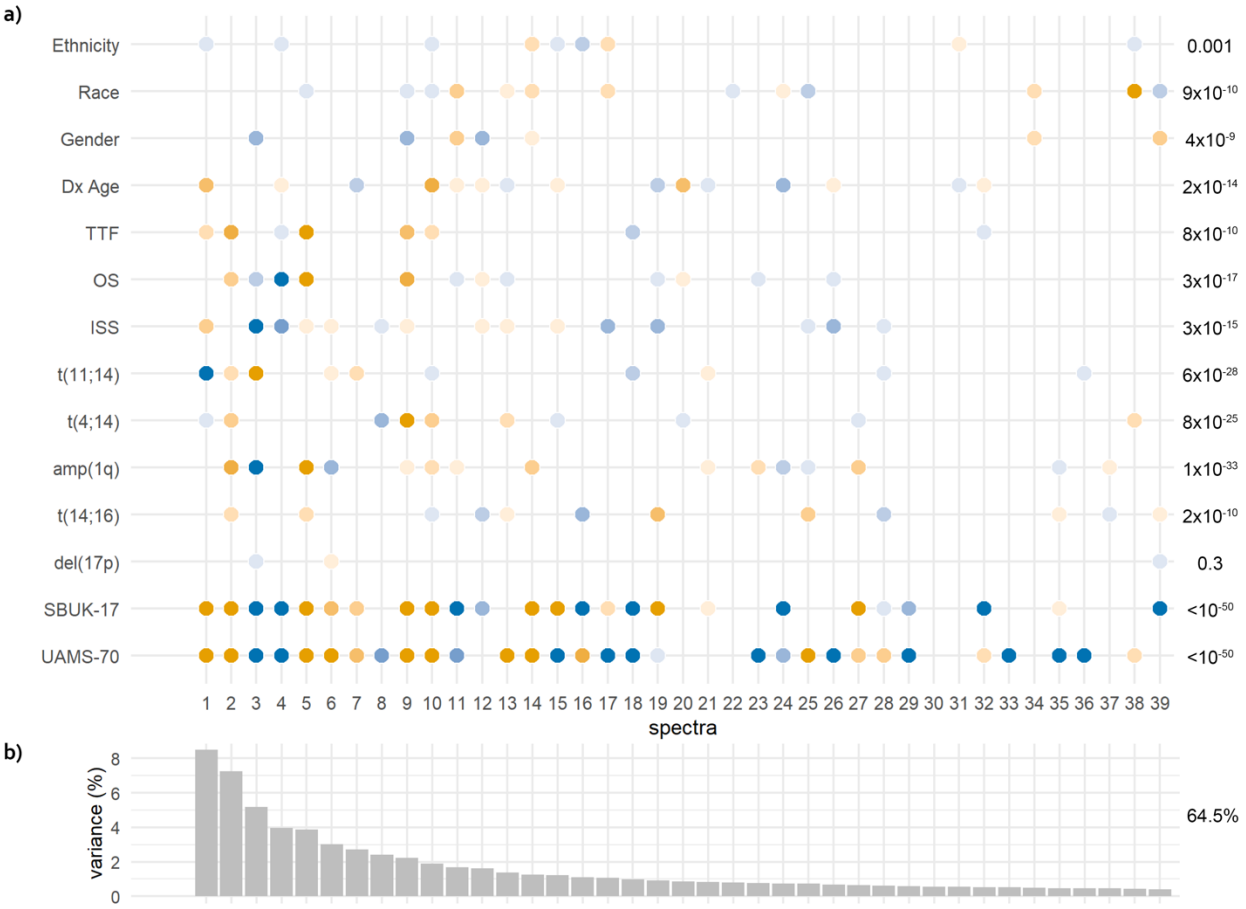


Figure 4. Overview of CD138+ spectra outcome modeling. a) Multivariable modeling results. The outcome of interest (left y-axis) and the significance of the full 39-spectra model (right y-axis). Spectra variables are illustrated on the x-axis. The significance and direction of each spectrum are indicated: blue negative beta coefficient, orange positive beta coefficient. No dot is shown if a spectrum was not significantly associated at $p<0.05$ level. **b)** Percent of the global variance captured by each spectra variable. The total variance captured by all 39 spectra is shown at the right.

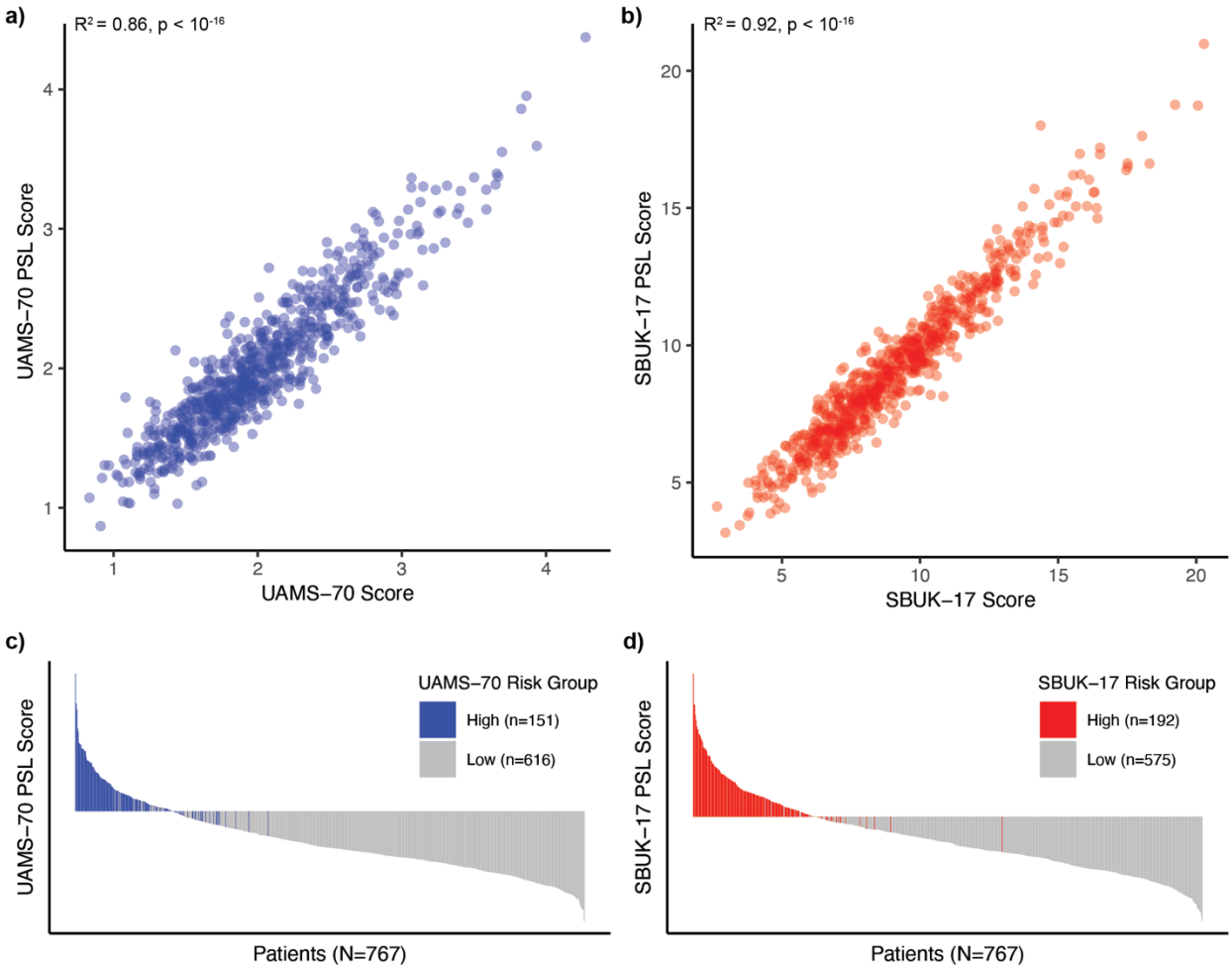


Figure 5. CD138+ spectra and two established expression scores. Correlation of the 39-spectra PSL score and established gene expression profiles from **a)** University of Alabama School of Medicine 70 gene risk score (UAMS-70) and **b)** Shahid Bahonar University of Kerman 17-gene prognostic score (SBUK-17). Waterfall plots for **c)** UAMS-70 and **d)** SBUK-17 low- and high-risk scores. Patients were ordered by their PSL score and colored by high/low risk as predicted by the UAMS-70 score. (UAMS-70 high-risk cutoff determined by clustering).

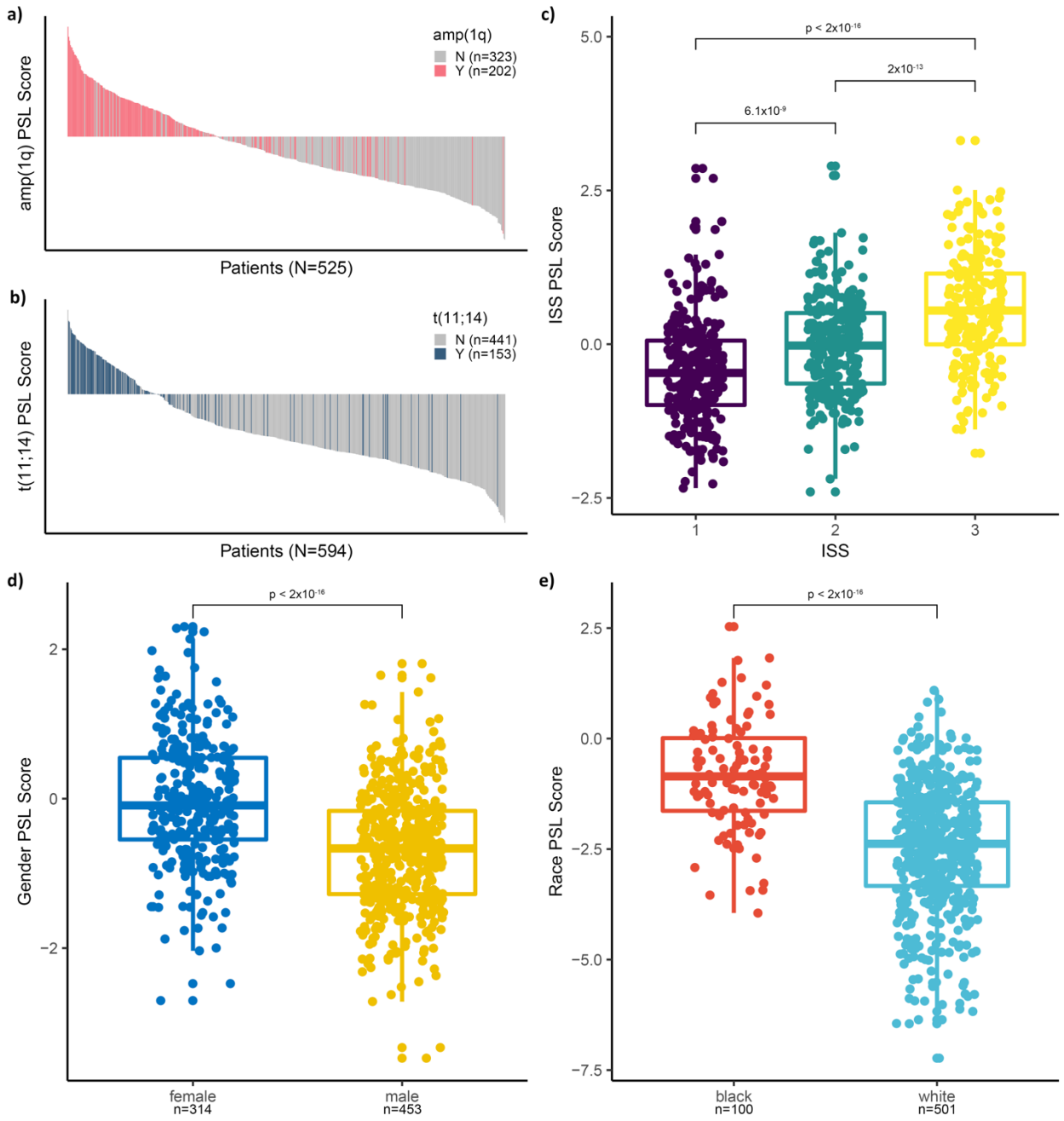
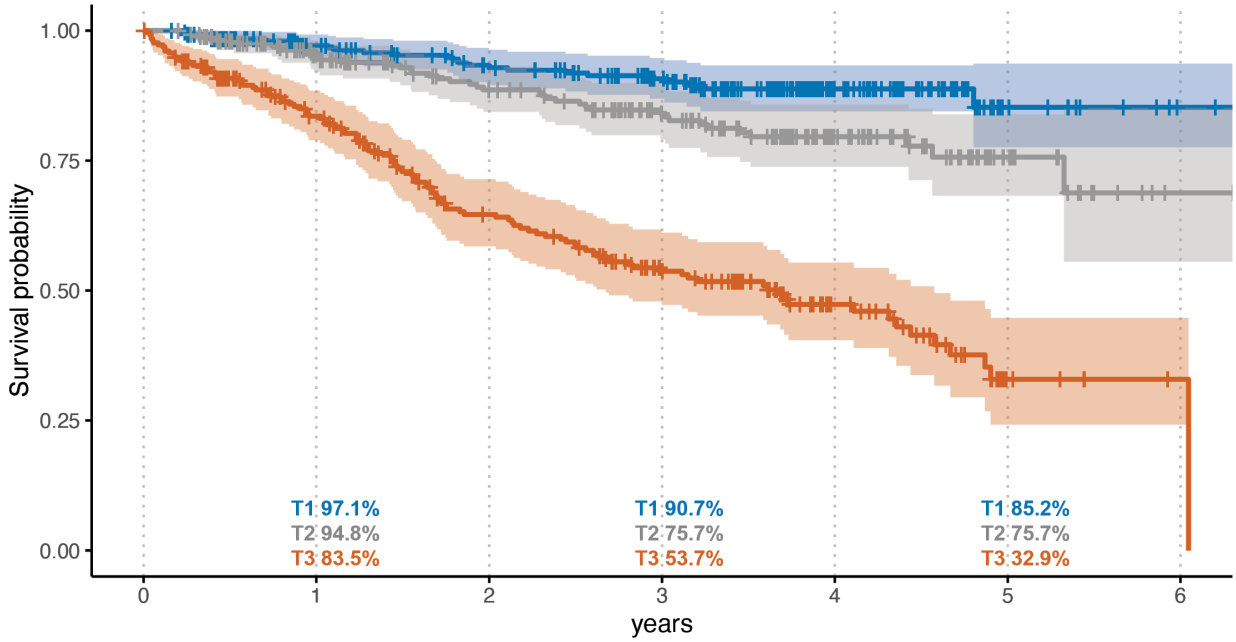


Figure 6. CD138+ spectra and clinical or demographic risk. Waterfall plots for PSL scores for models of **a)** tumor amplification chr1q, **b)** tumor translocation chr11;16. Box and Whisker plots for PSL scores for models of **c)** international tumor stage at diagnosis, **d)** gender, and **e)** self-reported black or white race.

a) Overall Survival, 179 events, $\chi^2_{logrank} = 120$, $p = 8.6 \times 10^{-27}$



b) Time to first-line treatment failure, 369 events, $\chi^2_{logrank} = 95.7$, $p = 1.7 \times 10^{-21}$

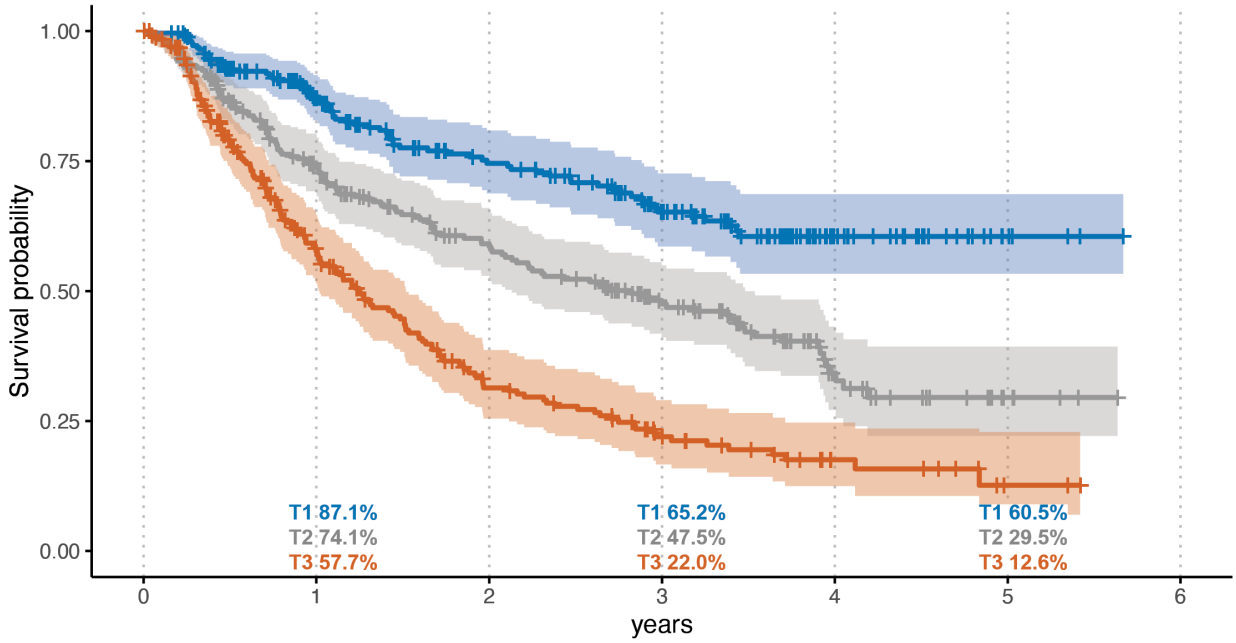


Figure 7. CD138+ spectra and disease course. Cox proportional hazards models were generated for overall survival (OS) and time to first-line treatment failure (TTF). From the models, PSL scores were generated and split into three equal tertiles. Kaplan-Meier curves of the PSL scores by tertile are shown for **a)** overall survival and **b)** time to first-line treatment failure.

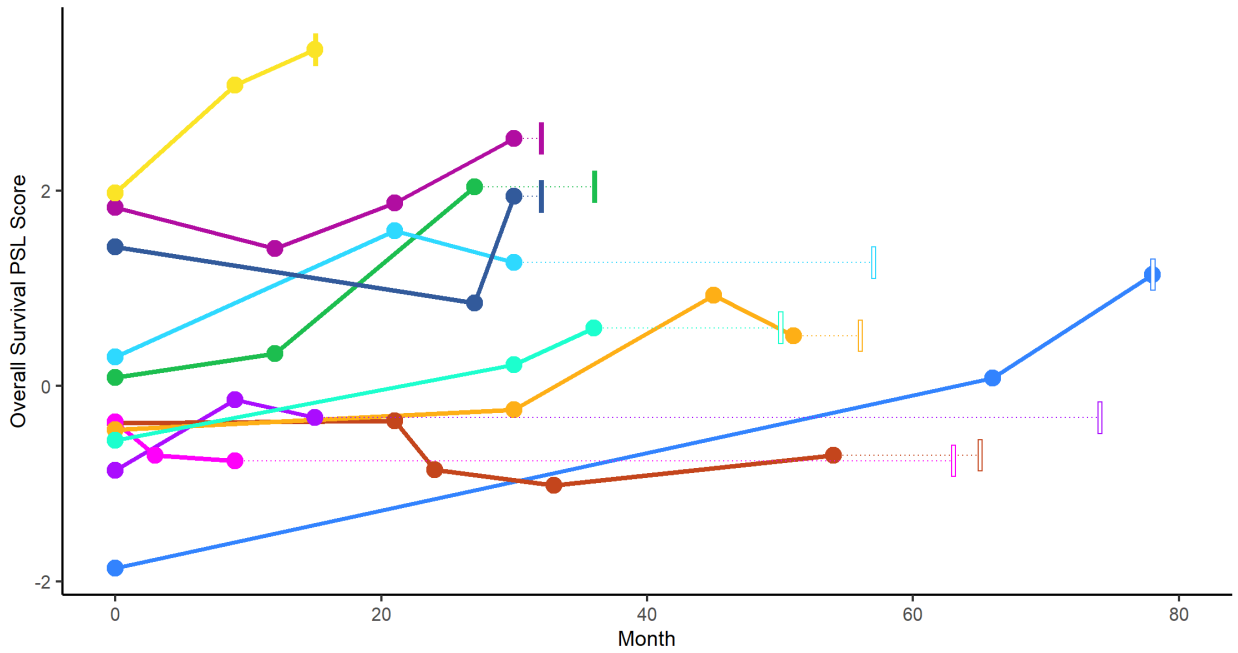


Figure 8. Aggregate effect of CD138+ spectra for overall survival over time. Poly-spectra liability scores for overall survival are shown for eleven patients with RNAseq data at multiple time points. Dots indicate sequencing events and show the PSL score at that timepoint. The final narrow rectangle indicates the month after diagnosis the patient died (filled) or was last known alive (open).