Deep transcriptome profiling of multiple myeloma with quantitative measures using the SPECTRA approach

Rosalie Griffin Waller,[1,2] Heidi A. Hanson,[1] Brian J. Avery,[1] Michael J. Madsen,[1] Douglas W. Sborov,[1] and Nicola J. Camp[1]*

1. Huntsman Cancer Institute and School of Medicine, University of Utah, Salt Lake City, UT, 84112, USA
2. Computational Biology, Quantitative Health Sciences, Mayo Clinic, Rochester, MN, 55905, USA

Correspondence: *nicola.camp@hci.utah.edu

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 genetic, epigenetic and environmental consequences. SPECTRA is an approach to describe variation in a transcriptome as a set of unsupervised quantitative variables. Spectra variables provide a deep dive into the transcriptome, representing both large (prominent, high-level) and small (deeper, more subtle) sources of variance. Spectra variables are ideal for modeling alongside other variables for any outcome of interest. Each spectrum can also be considered a phenotypic trait, providing new avenues for disease characterization or to explore disease risk. We applied the SPECTRA approach to multiple myeloma (MM), the second most common blood cancer. Using RNA sequencing from malignant CD138+ cells, we derived 39 spectra in 767 patients from the MMRF CoMMpass study. We included spectra in prediction models for several clinical endpoints, compared to established expression-based risk scores, and used descriptive modeling to identify associations with patient characteristics. Spectra-based risk scores added predictive value beyond established clinical risk factors and other expression-based risk scores for overall survival, progression-free survival, and time to first-line treatment failure. We identified significant associations between CD138+ spectra and tumor cytogenetics, race, gender, and age at diagnosis. The SPECTRA approach provides quantitative measures of transcriptome variation to deeply profile tumors. This framework more comprehensively represents signals in the transcriptome and offers greater flexibility to model clinical outcomes and characteristics.

INTRODUCTION

Numerous factors are involved in risk and prognosis in complex disease. Transcriptomes represent the combined effects of inherited, somatic, and epigenetic variation and can provide insight into genetic and environmental risk factors. As a result, gene expression studies are gaining momentum in genomic epidemiology.[1–5] The need to incorporate transcriptome data in multivariable models alongside other risk factors brings new demands for approaches to describe transcriptomes.

Many current transcriptome approaches are focused on immediate biological interpretation and constrained to biological expectations or reduce the data to a single categorical variable (e.g., intrinsic subtypes). While these have advanced knowledge of disease mechanisms[6–8] and identified important high-level differences in disease[9–11], complementary approaches are needed to go deeper and increase flexibility and application. In common disease, the sources of heterogeneity are many, complex, and often poorly understood. Latent variables, focused on capturing signal and not immediate interpretability, provide the potential for new discoveries. Transcriptome characterization using multiple quantitative variables may provide a meaningful deeper dive into the transcriptome.

Here we describe the SPECTRA approach, a data workflow and variable derivation method with principal component analysis (PCA) at its core. PCA has many strengths aligned with our intent, such as providing multiple unsupervised variables that optimize coverage of the global variance. The resulting latent variables are quantitative, uncorrelated, retain integrity to the original data, and have desirable attributes for subsequent multivariable prediction and descriptive modeling. A limitation of PCA is that if data are not curated adequately, spurious variance will be incorporated in the derived variables, including technical artifacts and missing data structures. A key part of the SPECTRA approach is stringent data culling, quality control, zero-handling, and normalization.

For well-curated gene-panel data a PCA-based approach has shown promise. Using a population-based dataset of breast tumors, we previously used PCA to reduce the 50-gene space of the PAM50 gene panel to five quantitative multi-gene expression variables.[12] When implemented as predictor variables in an independent clinical trial dataset, PCA variables were able to predict prognosis and response to paclitaxel; adding value beyond clinical risk factors and outperforming intrinsic subtypes.[13] When utilized as quantitative tumor phenotypes, PCA variables were superior to the standard PAM50 subtypes for gene mapping.[12,14] Here, we extend PCA-based latent variable derivation to the whole transcriptome.

Figure 1 uses a color analogy to illustrate the conceptual framework of SPECTRA and contrast our goal of quantitative variables for direct use in outcome modeling with a more conventional categorization approach using hierarchical clustering. 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 variables to a single categorical variable which may reduce statistical power. For example, in **Figure 1**, x_R cannot be represented by any group ordering in **Figure 1c**, 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 the prediction of a single outcome, it also tethers the groups to the trained outcome and doesn't facilitate comparison to other outcomes.

We illustrate our whole transcriptome SPECTRA approach using bulk RNA sequencing (RNAseq) data from CD138+ sorted myeloma cells from the Multiple Myeloma Research Foundation (MMRF) CoMMpass Study.[15] Multiple myeloma (MM) is a malignancy of the plasma cells with one of the poorest 5-year survival (55.6%) for adult-onset hematological malignancies.[16] It is most frequently diagnosed at ages 65-74 years (median 69 years).[16]

Incidence is higher in men (8.7 men vs. 5.6 women per 100,000) and particularly high in patients self-reporting as African American (AA men 16.3, and AA women 11.9 per 100,000). We use the SPECTRA approach to derive quantitative CD138+ transcriptome variables, referred to as *spectra*, which we use in regression models to predict clinical endpoints, derive latent risk groups that may be clinically meaningful, compare to established expression-based molecular risk scores, and describe associations with clinical and demographic characteristics.

SUBJECTS AND METHODS

SPECTRA: an approach for deep transcriptome profiling with quantitative measures

The motivation is the derivation of well-behaved, quantitative variables from RNAseq data to capture the many sources of intrinsic variation in a transcriptome. PCA is a well-established, unsupervised, data-driven method that, based on the covariance of a dataset, produces a matrix factorization which is a unique solution of linear transformations and transformed values. The linear transformations preserve the variance in the data and provide meaningful comparisons between individuals. The transformed values are orthogonal quantitative variables (linearly uncorrelated), each subsequently going deeper into the global variance. Dimension reduction is achieved by selecting only the first *k* components, for which the proportion of total variance explained can be described. The resulting unsupervised variables can be used as predictors for any outcome, and as novel phenotypes.

To represent meaningful variance and enable deep comparison across individuals, rigorous quality control (QC), normalization, and batch correction must be performed before PCA. The workflow for the SPECTRA approach is described below and in **Figures 2** and **S1**. Details of the matrix factorization (MF) performed by PCA can be found in Supplemental Methods.

Quality control is essential to ensure only meaningful variation across the individuals is retained. Features in the transcriptome likely to be unduly influenced by poor alignment or lacking precision due to sequencing depth are removed as potentials for introducing spurious and unstable variation. We concentrate on autosomal protein-coding genes. Features with inadequate data for precision, defined as more than 5% of samples with fewer than 100 read counts, are removed. We also remove genes known to be unstable across different RNAseq pipelines.[17] After the removal of features, individuals are removed from consideration if more than 10% of the remaining genes have fewer than 100 read counts.

Normalization is required for comparisons across genes and individuals and includes adjustment for gene length, sequencing depth (library size), and RNA composition. We use a robust internal (single sample) normalization to obviate the need for a 'reference' sample and to provide the possibility for portability across datasets. The SPECTRA approach is gene-focused, but the workflow handles both gene-based and transcript-based alignment and quantification; the latter has been suggested to be more accurate.[18] Normalized gene expression values, e_g, are calculated as follows:

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

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. Gene-level read counts are equivalent to m =1. Adjustments for sequencing depth and RNA composition (often referred to as the $size\ factor$) are achieved via the 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;[19] 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 to a five standard deviation threshold from the mean of the normalized gene counts in the relevant direction.

Batch correction is necessary to correct for potential technical artifacts and spurious variation introduced by differing sequencing protocols. We adjust for sequence batch using ComBat[20] as implemented in the SVA R package,[21] with covariates included for patient characteristics that are unbalanced by batch.

PCA is implemented with the covariance matrix. We use singular value decomposition to perform PCA, and 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. The R core function $prcomp$ is used to perform PCA. Eigenvectors contain the gene loadings that define the linear transformation used for each spectrum. A corresponding eigenvalue, λ, indicates the proportion of the global variance represented by the transformed value for an eigenvector. We use a scree test[22] (inflection point of the rank-ordered plot of the λ, or elbow method) to select the k components to retain. The proportion of variance explained by this k-dimensional space ($\sum_{i=1}^{k} \lambda_i / \sum_{\forall i} \lambda_i$) indicates the depth of the dive into the transcriptome data. Spectra variables are the standardized retained components (S_1, …, S_k).

Myeloma CD138+ Spectra

Data from the MMRF CoMMpass Study (release IA14)[15] were downloaded from the MMRF web portal (https://research.themmrf.org/). Clinical data and CD138+ RNAseq were available for 768 patients prior to treatment at study entry (baseline) and 119 follow-up bone marrow samples. Transcript-based expression estimates processed by Salmon (version 0.7.2) were used. The SPECTRA approach was used on baseline samples (n=768) to derive a CD138+ transcriptome framework and SPECTRA variables. While not used for spectra derivation, follow-up samples were batch-corrected alongside the baseline samples. Covariates included in batch correction were age, gender, overall survival, progression-free survival, time to first-line treatment failure, and treatment status. The first 39 components (spectra variables S1, …, S39). were selected based on the scree test.[22]

Predictive Modeling of Clinical Outcomes

Predictive Modeling used Cox proportional hazards regression, implemented in the *survival* package in R.[23,24] An overview of predictive modeling is provided in **Figure S1**. All 39 spectra were considered as predictor variables for each of the three survival outcomes: overall survival (OS), progression-free survival (PFS), and time to first-line treatment failure (TTF). The spectra were standardized in all analyses. For each outcome, a single step variable selection from the all-spectra model was performed to retain only significant spectra ($p<0.05$). The linear predictor from the reduced model was used to define a quantitative spectra risk score for the outcome (i.e., the weighted sum of the spectra retained in the model based on their coefficients: $\sum_j \beta_j S_j$).

To determine if a quantitative spectra risk score displayed evidence of latent risk groups that could be clinically meaningful, the risk score distribution and normal Quantile-Quantile plots were inspected. The *mclust* R package was used to statistically assess evidence for risk groups using density estimation by Gaussian finite Mixture Modeling (GMM)[25,26] assuming equal variances. Bayesian information content (BIC) and bootstrap likelihood ratio tests (LRT) were used to determine the number of risk groups (*mclustBIC, mclustBootstrapLRT*). The hazard ratio (HR) between risk groups was calculated using relative event rates[27] from *survdiff* in the *survival* package in R.[24]

To address overfitting and lack of a replication sample, bootstrap internal validation was used.[28] Bootstrap internal validation involves replicating the entire modeling process on bootstrap resamples of the data to determine and adjust for over-fitting. Here, the process includes both the single step variable selection, GMM density estimation for risk group designation, and adjustment of HRs, C-index and bootstrap confidence limits.[29] Initial estimates from a model are referred to as 'apparent'. The measure of overfitting is referred to as

'optimism'. The corrected estimates are referred to as 'adjusted' values. Adjusted estimates are more reasonable assessment of effect size. If adjusted confidence intervals do not contain 1.0, this indicates validation.

Comparison with existing transcriptome risk scores. We compared spectra-based risk to two previously published risk scores: 1) the first and most widely adopted supervised expression risk score in myeloma, from the University of Arkansas for Medical Sciences (UAMS);[11] and 2) a recent supervised risk score, from the Shahid Bahonar University of Kerman (SBUK).[30] The UAMS risk score was developed in microarray data and tested 54,657 probes for association with disease-related survival.[11] A total of 70 genes were selected (19 under and 51 overexpressed prognostic genes). The UAMS risk score is the ratio of mean expression of the up-regulated to down-regulated genes, with k-means clustering to determine a cutoff for 'high-risk' classification.[11] The SBUK prognosis score was developed in the CoMMpass RNAseq data. All genes were tested for association with survival, followed by a multi-step process, including univariate Cox analysis, the intersection in six Class Prediction algorithms, and multivariate Cox analysis, was used to select 17 genes consistent across multiple methods.[30] These 17 were entered in a multivariable Cox regression to define the SBUK score, with the 75th percentile used to classify patients into high-risk and low-risk categories.[30] We calculated each patient's UAMS and SBUK risk scores and their risk status (low or high) using the CoMMpass RNAseq data.

Added predictive value. Predictive value beyond established clinical risk factors is important to establish. We calculated the added predictive value (APV)[28,31,32] of each expression risk score (spectra, UAMS, SBUK) beyond risk predicted from established clinical predictors. We used quantitative risk scores, and not risk group status, as these contain more complete risk information and because each gene-expression method employed a different thresholding approach making risk status comparisons potentially misleading. The APV compares the LRT statistic for a model including established clinical predictors to a model which also includes each

expression risk score. We also determined whether the spectra risk score provided value beyond known clinical predictors and UAMS. APV=0.0 indicates no additional prediction information, i.e., all predictive power is duplicative with other factors already in the model. With larger independent predictive contributions, APV increases, to a maximum of 1.0.

CD138+ Spectra in Follow-Up Samples

To illustrate the potential to track changes over time, OS spectra risk scores were calculated in longitudinal follow-up samples. For these samples, normalized and batch-corrected gene expression values were centered to the derivation set. Spectra values (S_j) were determined using the previously derived linear transformation and scaled to the derivation set. The OS risk score was calculated using the previously established linear combination across spectra ($\sum_j \beta_j S_j$). Spectra variables in all samples (baseline and follow-up) are provided (see Data and Code Availability).

Descriptive Modeling of Clinical and Demographic factors

We used descriptive modeling to illustrate associations between the CD138+ spectra and clinical and demographic factors with elevated myeloma risk: tumor aberrations (high risk del(17p), t(14;16), gain 1q, and t(4;14), and standard risk t(11;14)),[33] age, gender, race. Linear or logistic regression was used with all 39 CD138+ spectra modeled as independent variables. Spectra were noted if significant in the descriptive model ($p < 0.05$).

RESULTS

CD138+ Spectra

We applied the SPECTRA approach to RNAseq from treatment naïve CD138+ cells collected at diagnosis from 768 patients in the MMRF CoMMpass study.[15] We used transcript-based expression estimates generated by the CoMMpass study with Salmon.[34] After quality control, gene expression values for 7,449 genes (56,339 transcripts) in 767 patients were normalized and batch corrected. After PCA, the first 39 components were selected based on a scree test and standardized to create spectra S1-S39. Together these explain 65% of the global expression variation (**Figure 3A**). As linearly uncorrelated variables each of the 39 CD138+ spectra capture a different source of variance – each spectrum has the potential to describe patient differences. A patient's CD138+ transcriptome spectra variables can be visualized with a 39-spectra barcode (**Figure 3B**).

Predictive Modeling – Overall Survival (OS)

In Cox regression of OS (179 events), nine spectra were significant ($p < 0.05$) and selected for the predictive model (apparent C-index = 0.70). Internal validation (1000 bootstraps) estimated C-optimism = 0.04, leading to an adjusted C-index = 0.66 (0.65-0.74). Model coefficients are given in **Table 1**. Gaussian mixture modeling on the OS spectra risk scores identified two risk groups ($p = 0.001$) (**Figure 4**) with 88 patients in the high-risk group (58 events, median OS 20.1 months) and 676 patients in the low-risk group (121 events, median OS not reached after 78 months) (**Figure 4C**). After internal validation, the optimism adjusted hazard ratio (HR_{adj}) for the high- compared to low-risk patients was 4.27 (2.31-12.69) (**Table 2**). This compared to UAMS OS HR_{adj} = 3.89 (2.53-5.45) and SBUK OS HR_{adj} = 2.54 (1.91-3.47) in CoMMpass data.

The added predictive value (APV) of the quantitative spectra risk score, beyond Revised International Staging System (R-ISS) and age at diagnosis was 0.612 (**Table 3**). In parallel analyses, UAMS APV was 0.576 and SBUK APV = 0.502, indicating each contains substantial predictive value beyond clinical factors. Compared to a model including R-ISS, age at diagnosis, and UAMS, however, SBUK showed very little added value (APV=0.006, **Table 3**), indicating its predictive gene expression information is almost all duplicative with UAMS. In contrast, the spectra score APV = 0.109 beyond R-ISS, age at diagnosis, and UAMS (**Table 3**) - also significant by LRT ($p = 7.4x10^{-5}$) - indicating that the spectra risk score contains predictive information beyond clinical factors and the previously established UAMS risk score.

Eleven patients had at least three follow-up CD138+ samples, spanning up to six years. To illustrate a potential to track changes over time, the OS spectra risk score was graphed for the initial and longitudinal follow-up samples (**Figure 5**).

Spectra Predictive Modeling – Progression-Free Survival (PFS)

As the spectra are unsupervised, the same 39 spectra variables can be used to model any outcome. We repeated the same predictive modeling procedure for PFS. Cox regression of PFS (392 events) selected nine spectra in a model with C_{adj} = 0.60 (0.60-0.67). Four of the retained spectra were distinct from the OS spectra model (**Table 1**). Gaussian mixture modeling identified two risk groups (p = 0.001) (**Figure S2**). Median time to progression was 9.7 months in the high-risk patients (n = 60, 50 events) and 35.7 months in the low-risk patients (n = 707, 342 events) (**Figure 6A**). Between high and low risk groups, spectra PFS HR_{adj} = 3.08 (1.67-8.44) while UAMS PFS HR_{adj} = 2.40 (1.70-3.31) and SBUK PFS HR_{adj} = 1.92 (1.59-2.43) (**Table S1**). Spectra added information beyond R-ISS, age at diagnosis, and UAMS (APV = 0.24, LRT p = $8.2x10^{-8}$) in predicting PFS (**Table S2**).

Spectra Predictive Modeling – Time to First-Line Treatment Failure (TTF)

In Cox regression modeling to predict TTF (369 events) ten spectra were selected with the model C_{adj} = 0.60 (0.59-0.66). Four of the retained spectra were distinct from the OS and PFS models (**Table 1**). Two risk groups were identified in the TTF spectra score with GMM (p = 0.008) (**Figure S3**). Patients in the spectra high-risk TTF (n = 31, 25 events) had median TTF of 9.2 months compared to low-risk patients (n = 736, 344 events) with median TTF of 32.8 months (**Figure 6B**). Hazards ratios between high- and low-risk groups using TTF spectra risk groups (HR_{adj} = 3.10 [1.31-5.46]) outperformed UAMS (TTF HR_{adj} = 1.98 [1.48-2.68]) and SBUK (TTF HR_{adj} = 1.85 [1.52-2.36]) (**Table S3**). Spectra provided additional information in predicting TTF beyond R-ISS, age at diagnosis, and UAMS (APV = 0.29, LRT p = 2.6×10^{-6}) (**Table S4**).

Spectra Descriptive Modeling

Figure 7 illustrates associations between the CD138+ spectra and tumor aberrations or demographic groups with elevated myeloma risk. For descriptive modeling a model containing all 39 spectra was fit for each characteristic. In logistic regression models, each tumor aberration showed different significant spectra in the model, with some spectra unique to only one aberration. Linear regression with age at diagnosis as a quantitative outcome highlighted associations with thirteen spectra. Seven spectra were significantly associated with gender and thirteen spectra significant with race (self-reported black or white; other racial categories too small to consider).

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.[35] Gene expression in diseased tissue is an established source of heterogeneity.[36] Tools that can take a deeper dive and better characterize expression heterogeneity will be important to advance the promise of personalized medicine. Clinical and epidemiologic studies that wish to model multiple sources of risk in a population, transcriptome variables that can be easily incorporated with other variables and across endpoints of interest are advantageous. 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 quantitative molecular phenotypes. The SPECTRA approach identifies quantitative, orthogonal (non-correlated) variables that capture sources of transcriptome variation for use in subsequent predictive or descriptive modeling or as quantitative phenotypes.

Data quality and processing are paramount in the quest to derive informative agnostic variables. PCA is a simple procedure that provides linear transformations of the data to represent variance. If the data have technical artifacts, batch effects, unstable or non-comparable expression measures, the noise can overwhelm authentic variance. Accordingly, the SPECTRA approach intentionally includes strict quality control, careful zero-handling, robust normalization, and batch correction (**Figure 2**). Without these steps, PCA can fail to provide informative variables. 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. The limitation of an agnostic approach is reduced biological interpretation or mechanistic insight of the variables themselves, prior to modeling. However, there are already many approaches that take this alternate goal of intermediate interpretation,[37] whose limitations

are instead the reduced flexibility of the variables they produce. Hence, SPECTRA is a complementary approach to the current toolset available for all fields.

Other potential advantages of SPECTRA include its unique solution within a dataset. Dimension reduction is post-hoc and does not influence the definition of the components. As a statistically rigorous technique, it also provides a measured dive into the transcriptome. Each spectrum iteratively moves quantifiably deeper into the variance of the data (measurable by its corresponding λ). Methods that iteratively find independent components (PCA and independent component analysis) have previously been shown to provide superior coverage of transcriptome data.[8] The importance of a deeper dive is illustrated in the MM predictive modeling. Spectra representing variance deeper in the data were important in predicting OS, PFS, and TTF (**Table 1**). For example, S32 (which explains 0.5% variance in the transcriptome) provided the same weight to the PFS risk score as S1 (8.7% variance). As another illustration, S19 and S26 (both with variance < 1%) are in the OS spectra risk score. Further, twelve patients with poor survival – 14% of the high-risk group – would not have been classified in the high-risk group if their S19 and S26 spectra values had been at the population mean. In our MM study, retention of components deep in the data, representing small variances (i.e., deep dives) improved prediction. We suggest that deep transcriptome characterization is a new tool with the potential to identify the few patients that respond to a drug, or small groups of individuals with large effects in outcome studies, both relevant to precision medicine. Furthermore, as shown previously for the PAM50 gene panel, superior coverage of data may identify previously overlooked expression differences between familial and sporadic tissues, identifying potential for familial components,[12] and providing new avenues for gene discovery, exposure and gene by environment studies.

In the SPECTRA approach negative gene loadings are embraced. The allowance of negative values in its MF is often given as a criticism of PCA,[6,38] argued as a conceptual source

of its lack of biological interpretability; the suggestion 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 matrix (equivalent to the eigenvector matrix) and pattern matrix (equivalent to the transformed component values in PCA) to be non-negative. Reasoned as beneficial and a natural restriction because gene expression cannot be negative and because the non-negativity leads to NMF transformed values that are proportions of each factor, providing simple interpretation. However, non-negative matrix values may not be a natural restriction to *systems of genes*, and over-simplicity may not adequately represent true complexity. Genes in a system may act in opposite directions producing both surplus and deficits of that system. With the non-negativity restriction, NMF is limited to the identification of groups of over-expressed genes,[6,38] and thus models only neutrality and surplus. Deficits may also be important. 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 matrix values, PCA can also capture gene systems in deficit, which may be more difficult to interpret, but important to recognize. These differences underscore SPECTRA as a valuable and complementary tool to existing approaches.

In our MM study, we illustrated the implementation of 39 CD138+ spectra in predictive modeling for OS, PFS, and TTF. In particular, we showed added value beyond established risk scores (UAMS, SBUK) and clinical risk factors (R-ISS, age at diagnosis). The framework of 39 quantitative variables has the potential to provide a bridge to compare many patient or tissue characteristics (**Figure 7**) and could be used to compare existing categorizations (such as subtypes) of patients, even when no genes overlap in their signatures, or they predict different outcomes.[39] In this way, spectra provide an alternate to categorical intrinsic subtyping, a well-established practice for many cancers.[40]

Clinically-relevant stratification may be better represented using risk groups within a transcriptome framework.[13] To this end, we have described a rigorous strategy to define spectra-based risk-groups using GMM (**Figure 2**, **Figure 4, Table 2**). Using our approach to identify risk groups for OS, we classified 27 patients as high risk that UAMS classified as low risk (**Figure S4**). These 27 patients had median survival of 26.7 months. Conversely, the 35 patients classified as low risk by spectra score but high risk by UAMS did not reach median survival within the study timeframe. SPECTRA outperform existing classification strategies in MM, while also allowing the flexibility of modeling with quantitative variables and multiple outcomes.

Descriptive modeling showed significant associations between spectra and tumor cytogenetics, race, gender, and age a diagnosis (**Figure 7**). As expected for a framework of agnostically derived variables, not all spectra are relevant to every dependent variable. Across the five tumor abnormalities, the number of associated spectra ranged from 2 to 15 with no single spectra associated with all abnormalities (**Figure 7**). Importantly, these examples show the flexibility of the framework as well as how it can support comparisons across different models and outcomes.

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). A prior CoMMpass study whose goal was to identify differences by race for myeloma tumors used the UAMS score to compare transcriptomes by race and did not identify significant differences ($p = 0.662$).[41] In contrast, our framework of 39 CD138+ spectra was able to identify 13 spectra that differed significantly by race (**Figure 7**). 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. 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 have provided the variable framework (gene transformations) and the spectra variables for the CoMMpass patients (see Data and Code Availability) to enable further study of spectra in other CoMMpass studies, as well as in other myeloma studies.

As for any approach, there are limitations. A key question is 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,[42,43] 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.

Another limitation is, as a data driven technique, the PCA step will model all sources of variance in the dataset. Also, the complete PCA decomposition will be overfit to the derivation dataset. To minimize this issue of modeling artifact and noise, 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.[17]

To limit overfitting in spectra derivation, we used dimension reduction to focus only on the first *k* spectra (largest *k* components of variation), selected using a scree test[22].

In conclusion, we present a new approach, 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 variables for flexible modeling, along with other covariates, to better differentiate individuals for any outcome. Spectra may also be used as novel transcriptome phenotypes. 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 prognostic outcomes and provide new insight into potential differences between tumors and patients from demographic groups. Fundamentally, the technique shifts from characterizing a transcriptome using categories to multiple quantitative measures. The SPECTRA approach provides a new paradigm and tool for exploring transcriptomes that hold promise for discoveries to advance precision screening, prevention, intervention, and survival studies.

ACKNOWLEDGMENTS

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), and 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.

DECLARATION OF INTERESTS

None.

DATA AND CODE AVAILABILITY

Processed RNAseq data from the CoMMpass Study can be downloaded from https://research.themmrf.org/. Code used to derive the CD138+ transcriptome spectra and generate the myeloma results is freely available on GitHub: https://github.com/njcamp-lab/MM_spectra. 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 are also provided. Spectra for individuals in the IA14 CoMMpass data can be downloaded from https://github.com/njcamp-lab/MM_spectra/SpectraData.

REFERENCES

1. Sweeney, C., Bernard, P.S., Factor, R.E., Kwan, M.L., Habel, L.A., Quesenberry, C.P., Shakespear, K., Weltzien, E.K., Stijleman, I.J., Davis, C.A., et al. (2014). 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.

2. Stopsack, K.H., Ebot, E.M., Downer, M.K., Gerke, T.A., Rider, J.R., Kantoff, P.W., and Mucci, L.A. (2018). Regular aspirin use and gene expression profiles in prostate cancer patients. Cancer Causes Control *29*, 775–784.

3. Allott, E.H., Shan, Y., Chen, M., Sun, X., Garcia-Recio, S., Kirk, E.L., Olshan, A.F., Geradts, J., Earp, H.S., Carey, L.A., et al. (2020). Bimodal age distribution at diagnosis in breast cancer persists across molecular and genomic classifications. Breast Cancer Res. Treat. *179*, 185–195.

4. López, C., Kleinheinz, K., Aukema, S.M., Rohde, M., Bernhart, S.H., Hübschmann, D., Wagener, R., Toprak, U.H., Raimondi, F., Kreuz, M., et al. (2019). Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. Nat Commun *10*, 1459.

5. Zhang, M., Lykke-Andersen, S., Zhu, B., Xiao, W., Hoskins, J.W., Zhang, X., Rost, L.M., Collins, I., Bunt, M. van de, Jia, J., et al. (2018). Characterising cis-regulatory variation in the transcriptome of histologically normal and tumour-derived pancreatic tissues. Gut *67*, 521–533.

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

7. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. (1999). 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.

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

9. Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.

10. Lapointe, J., Li, C., Higgins, J.P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., et al. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proceedings of the National Academy of Sciences *101*, 811–816.

11. Shaughnessy, J.D., Zhan, F., Burington, B.E., Huang, Y., Colla, S., Hanamura, I., Stewart, J.P., Kordsmeier, B., Randolph, C., Williams, D.R., et al. (2007). 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.

12. Madsen, M.J., Knight, S., Sweeney, C., Factor, R., Salama, M., Stijleman, I.J., Rajamanickam, V., Welm, B.E., Arunachalam, S., Jones, B., et al. (2018). 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.

13. Camp, N.J., Madsen, M.J., Herranz, J., Rodríguez-Lescure, Á., Ruiz, A., Martín, M., and Bernard, P.S. (2019). 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.

14. Hanson, H.A., Leiser, C.L., Madsen, M.J., Gardner, J., Knight, S., Cessna, M., Sweeney, C., Doherty, J.A., Smith, K.R., Bernard, P.S., et al. (2020). 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.

15. Keats, J.J., Craig, D.W., Liang, W., Venkata, Y., Kurdoglu, A., Aldrich, J., Auclair, D., Allen, K., Harrison, B., Jewell, S., et al. (2013). 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.

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

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

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

19. Shahriyari, L. (2019). 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.

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

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

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

23. Terry M. Therneau and Patricia M. Grambsch (2000). Modeling Survival Data: Extending the Cox Model (New York: Springer).

24. Therneau, T.M. (2021). A Package for Survival Analysis in R.

25. Fraley, C., and Raftery, A.E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association *97*, 611–631.

26. Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J *8*, 289–317.

27. (2002). Statistical Methods in Medical Research (Oxford, UK: Blackwell Science Ltd).

28. Harrell, J. (2015). Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis (Cham: Springer International Publishing : Imprint: Springer).

29. Noma, H., Shinozaki, T., Iba, K., Teramukai, S., and Furukawa, T.A. (2021). Confidence intervals of prediction accuracy measures for multivariable prediction models based on the bootstrap-based optimism correction methods. ArXiv:2005.01457 [Stat].

30. Zamani-Ahmadmahmudi, M., Nassiri, S.M., and Soltaninezhad, F. (2020). Development of a RNA sequencing-based prognostic gene signature in multiple myeloma. Br J Haematol bjh.16744.

31. Al-Radi, O.O., Harrell, F.E., Caldarone, C.A., McCrindle, B.W., Jacobs, J.P., Williams, M.G., Van Arsdell, G.S., and Williams, W.G. (2007). Case complexity scores in congenital heart surgery: A comparative study of the Aristotle Basic Complexity score and the Risk Adjustment in Congenital Heart Surgery (RACHS-1) system. The Journal of Thoracic and Cardiovascular Surgery *133*, 865–875.

32. Califf, R.M., Phillips, H.R., Hindman, M.C., Mark, D.B., Lee, K.L., Behar, V.S., Johnson, R.A., Pryor, D.B., Rosati, R.A., Wagner, G.S., et al. (1985). Prognostic value of a coronary artery jeopardy score. Journal of the American College of Cardiology *5*, 1055–1063.

33. Rajkumar, S.V., and Kumar, S. (2020). Multiple myeloma current treatment algorithms. Blood Cancer J. *10*, 94.

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

35. Ramón y Cajal, S., Sesé, M., Capdevila, C., Aasen, T., De Mattos-Arruda, L., Diaz-Cano, S.J., Hernández-Losa, J., and Castellví, J. (2020). Clinical implications of intratumor heterogeneity: challenges and opportunities. J Mol Med *98*, 161–177.

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

37. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. *102*, 15545–15550.

38. Stein-O'Brien, G.L., Arora, R., Culhane, A.C., Favorov, A.V., Garmire, L.X., Greene, C.S., Goff, L.A., Li, Y., Ngom, A., Ochs, M.F., et al. (2018). Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet. *34*, 790–805.

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

40. Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., and Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res *5*, 2929–2943.

41. Manojlovic, Z., Christofferson, A., Liang, W.S., Aldrich, J., Washington, M., Wong, S., Rohrer, D., Jewell, S., Kittles, R.A., Derome, M., et al. (2017). Comprehensive molecular profiling of 718 Multiple Myelomas reveals significant differences in mutation frequencies between African and European descent cases. PLoS Genet *13*, e1007087.

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

43. Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.

TABLES

Table 1. Cox regression beta coefficients for overall survival (OS), progression free survival (PFS), and time to treatment failure (TTF). Beta coefficients are by per spectra standard deviation. The quantitative spectra risk score for an outcome is the weighted sum of the spectra retained in its model based on their beta coefficients: $\sum_j \beta_j S_j$.

Spectra	OS	PFS	TTF
1		-0.132	-0.218
2	-0.335	-0.241	-0.212
3	0.243		
4	0.461	0.282	0.182
5	-0.324	-0.271	-0.257
9	-0.335	-0.182	-0.176
10			-0.138
11		0.083	
12	-0.179		
13	0.174		
18			0.131
19	0.111		
24		0.178	0.118
26	0.184	0.101	
31			0.102
32		0.121	0.117

Table 2. Overall survival hazard ratios between high- and low-risk patients using spectra, UAMS, or SBUK to define patient groups.

Risk Score	Apparent HR	Adjusted HR	Optimism
OS Spectra	6.44 (4.47-14.86)	4.27 (2.31-12.69)	2.16
UAMS	3.98 (2.62-5.54)	3.89 (2.53-5.45)	0.09
SBUK	2.58 (1.95-3.51)	2.54 (1.91-3.47)	0.04

Table 3. Overall survival Cox models including covariates and added predictive values.

Variable	Base Model	Δ AIC	Δ BIC	APV	LRT *P* Value
UAMS	R-ISS + Age	-71.7	-68.9	0.576	9.20×10^{-18}
SBUK	R-ISS + Age	-52.7	-49.9	0.502	1.42×10^{-13}
OS Spectra	R-ISS + Age	-83.5	-80.7	0.612	2.36×10^{-20}
UAMS	R-ISS + Age + SBUK	-17.8	-15.0	0.154	8.58×10^{-6}
UAMS	R-ISS + Age + OS Spectra	-1.9	0.8	0.027	0.048
SBUK	R-ISS + Age + UAMS	1.2	4.0	0.006	0.371
SBUK	R-ISS + Age + OS Spectra	1.1	3.8	0.007	0.333
OS Spectra	R-ISS + Age + UAMS	-13.7	-10.9	0.109	7.39×10^{-5}
OS Spectra	R-ISS + Age + SBUK	-29.7	-27.0	0.226	1.77×10^{-8}
OS Spectra	R-ISS + Age + UAMS + SBUK	-12.9	-10.2	0.104	1.13×10^{-4}

<u>Key</u>

R-ISS: Revised international staging system

Age at diagnosis

Δ AIC: Change in Akaike's Information Criterion by adding the variable

Δ BIC: Change in Bayesian information criterion by adding the variable

APV: Variable added predictive value of the variable

LRT *P* Value: Likelihood ratio test p-value between the base model with and without the variable

FIGURES

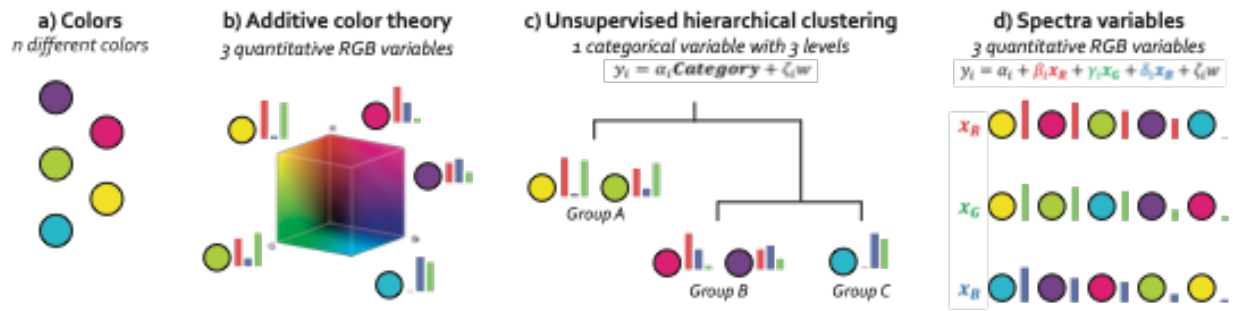

Figure 1. **A color analogy to illustrate the advantages of spectra variables for modeling**. a) Individual observations of color. b) Dimension Reduction (additive color theory), all colors can be represented using 3 quantitative RGB variables. c) Standard-use, modeling on the 3 RGB variables used to 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 spectra 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.

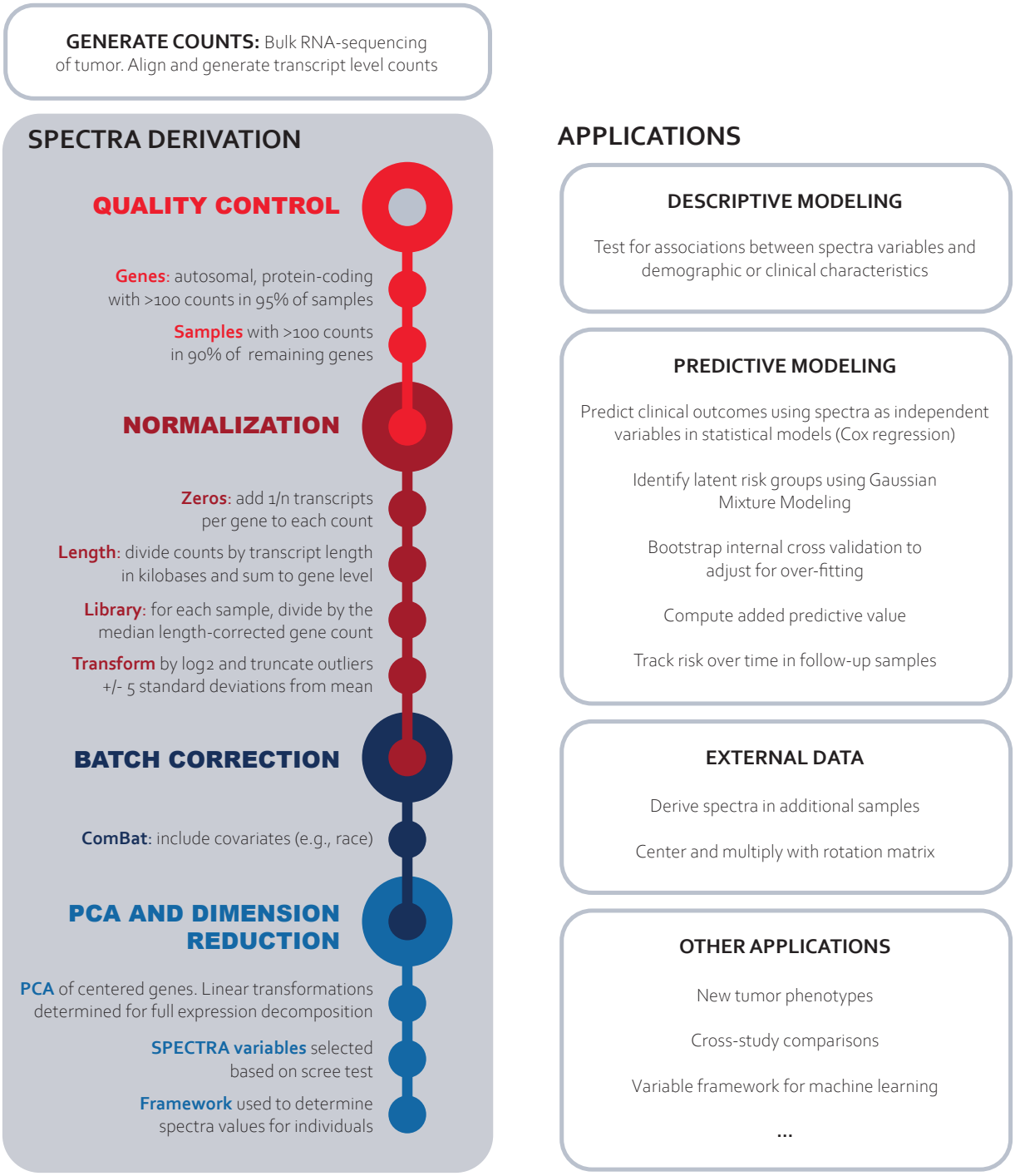

Figure 2. Overview of SPECTRA framework to derive spectra variables and example applications.

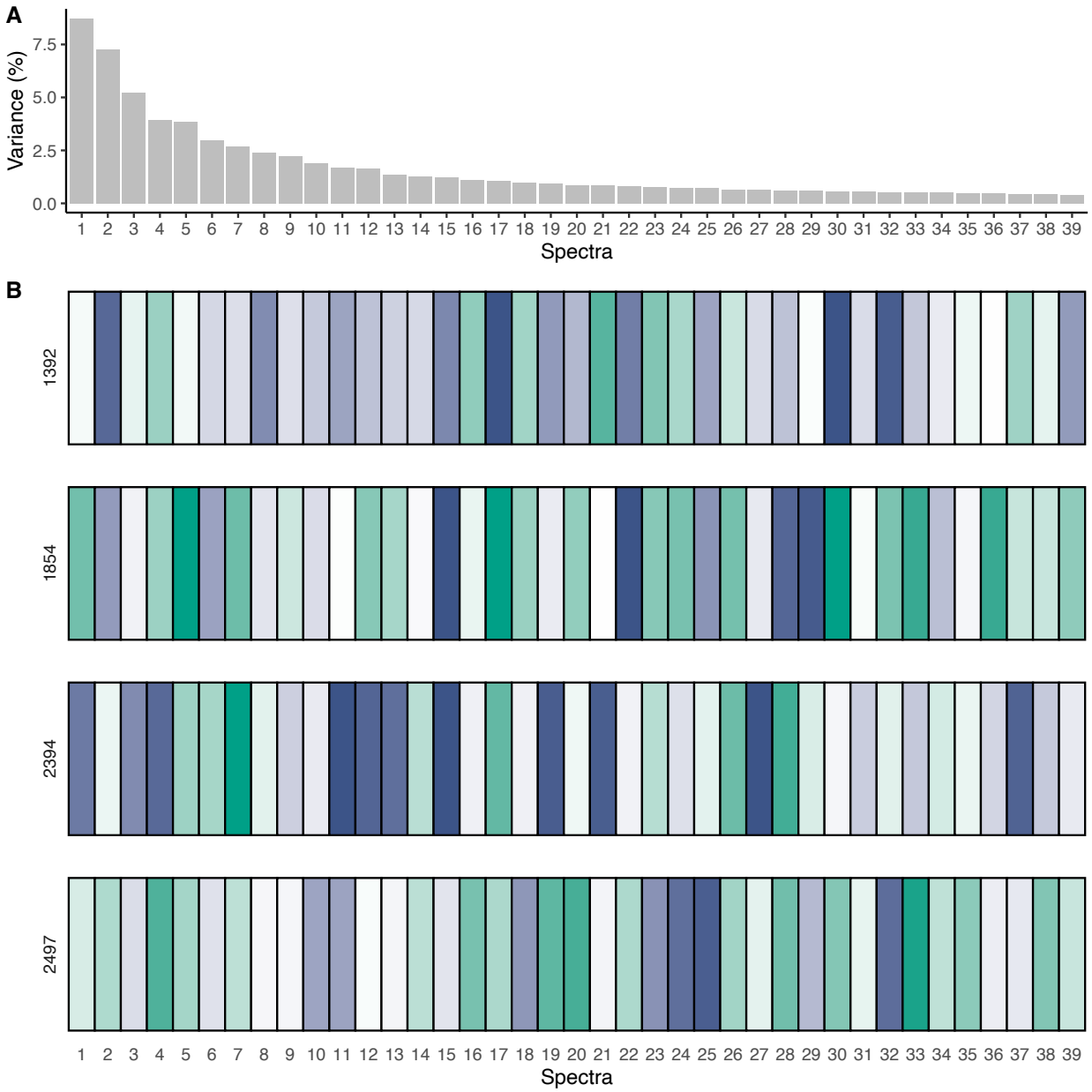

Figure 3. Spectra barcodes in four CoMMpass patients. A) Percent of transcriptome-wide variance between samples captured by each spectrum. Together the spectra capture 65% of the transcriptome-wide variance. B) For each patient, all 39 spectra are illustrated with the value represented by the bar intensity. The color indicates if the patient's spectra value is positive (blue) or negative (green). 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. For spectrum S1, 2497 is more like 1392, for spectrum S15, 1854 is more like 2394.

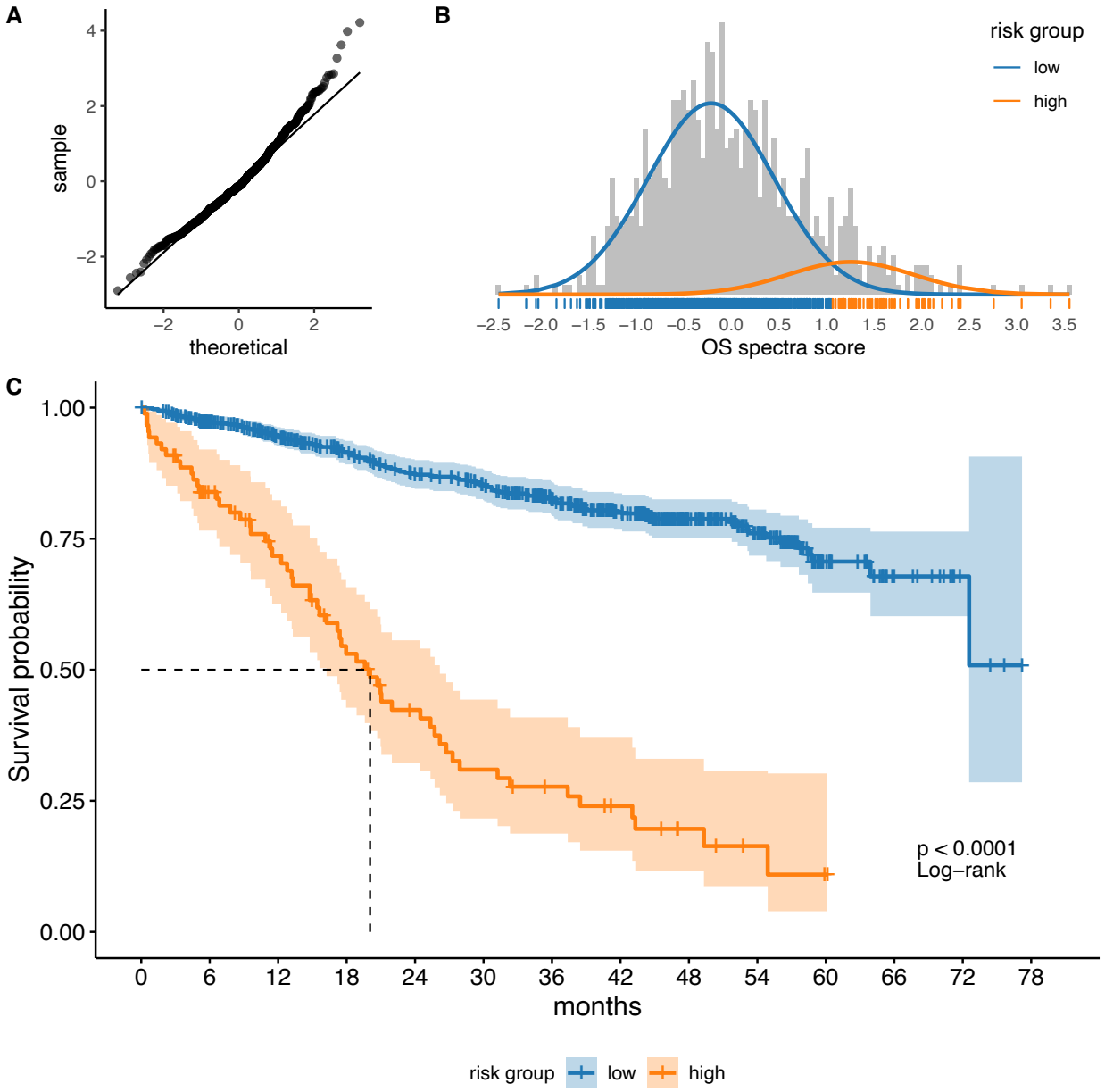

Figure 4. Overall survival spectra predictive modeling. A) Quantile-quantile plot of the actual spectra overall survival risk score (Cox linear predictors) compared to theoretical values. Elevated tail suggests high-risk group exists. B) Gaussian mixture modeling identified two distributions in the spectra score. Each patient was assigned into an OS risk group based on distribution probability. The blue bars on the x-axis represent patients in the low-risk group and the orange bars are patients in the high-risk group. C) Kaplan-Meier curve of the high/low risk groups. High-risk patients (n = 88, 58 events) had median

survival of 20.1 months. Low-risk patients (n = 679, 121 events) did not reach median survival in the study timeframe.

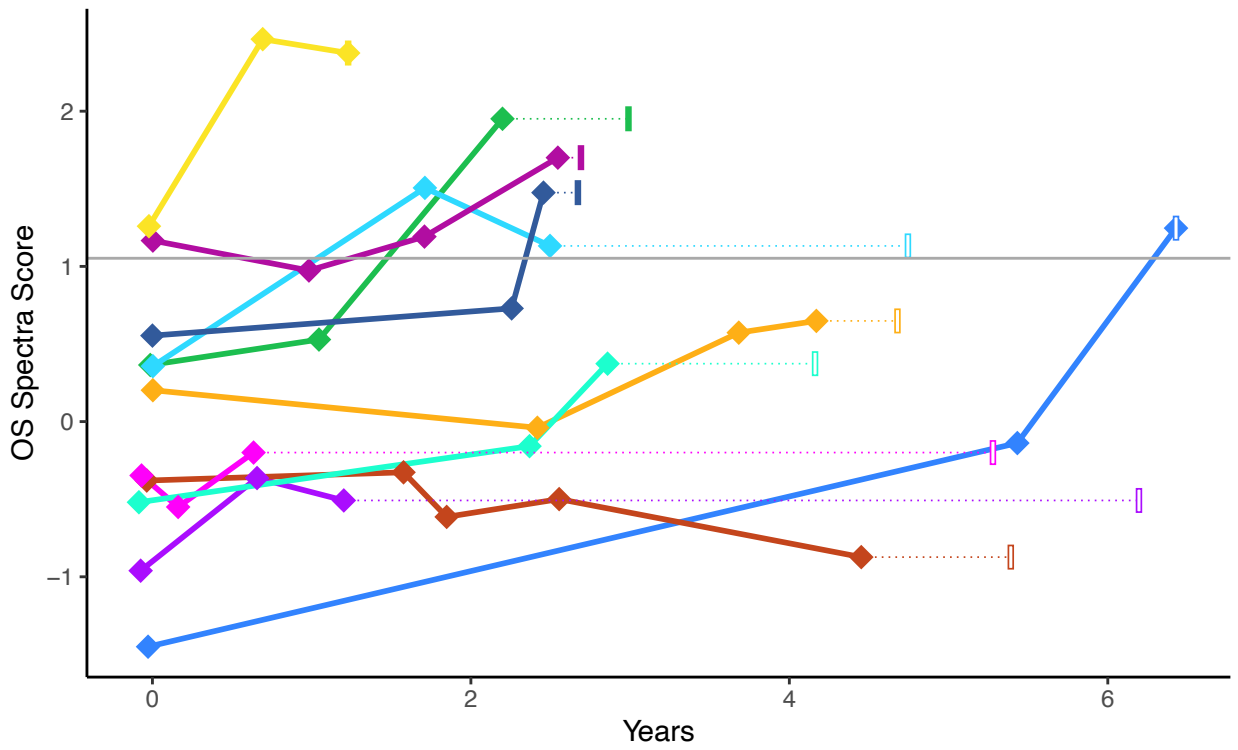

Figure 5. Aggregate effect of CD138+ spectra for overall survival over time. Spectra scores for overall survival are shown in eleven patients with RNAseq data at multiple time points. Diamonds indicate sequencing events and show the spectra score at that timepoint. The final narrow rectangle indicates timepoint the patient died (filled) or was last known alive (open). Gray horizontal line shows the high/low risk group threshold.

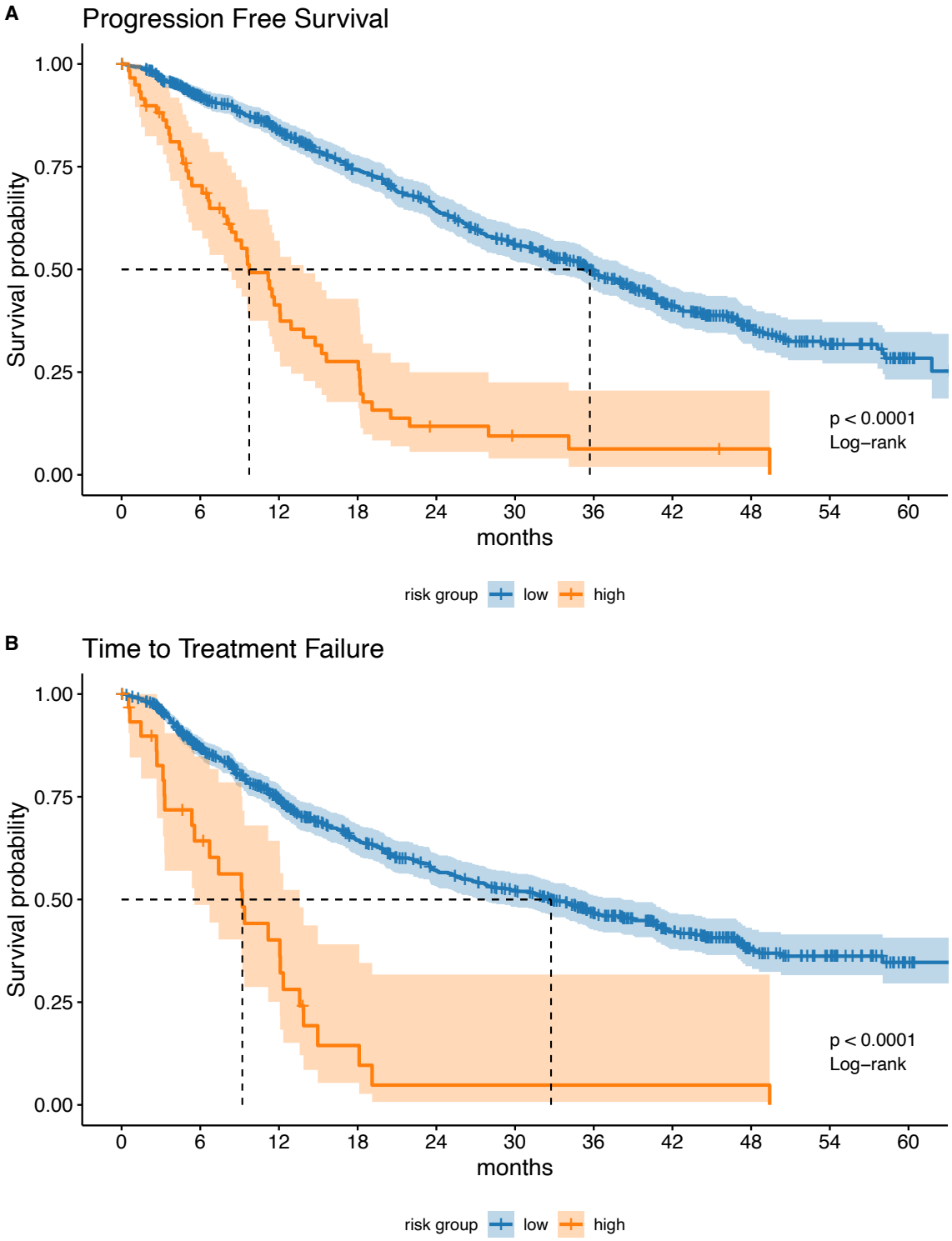

Figure 6. Kaplan-Meier curves of progression free survival and time to first line treatment failure.

A) PFS high-risk patients (n = 60, 50 events) had median survival of 9.7 months. PFS low-risk patients (n

= 707, 342 events) had median survival of 35.7 months. B) Patients in the spectra high-risk TTF (n = 31, 25 events) had median TTF of 9.2 months compared to low-risk patients (n = 736, 344 events) with median TTF of 32.8 months.

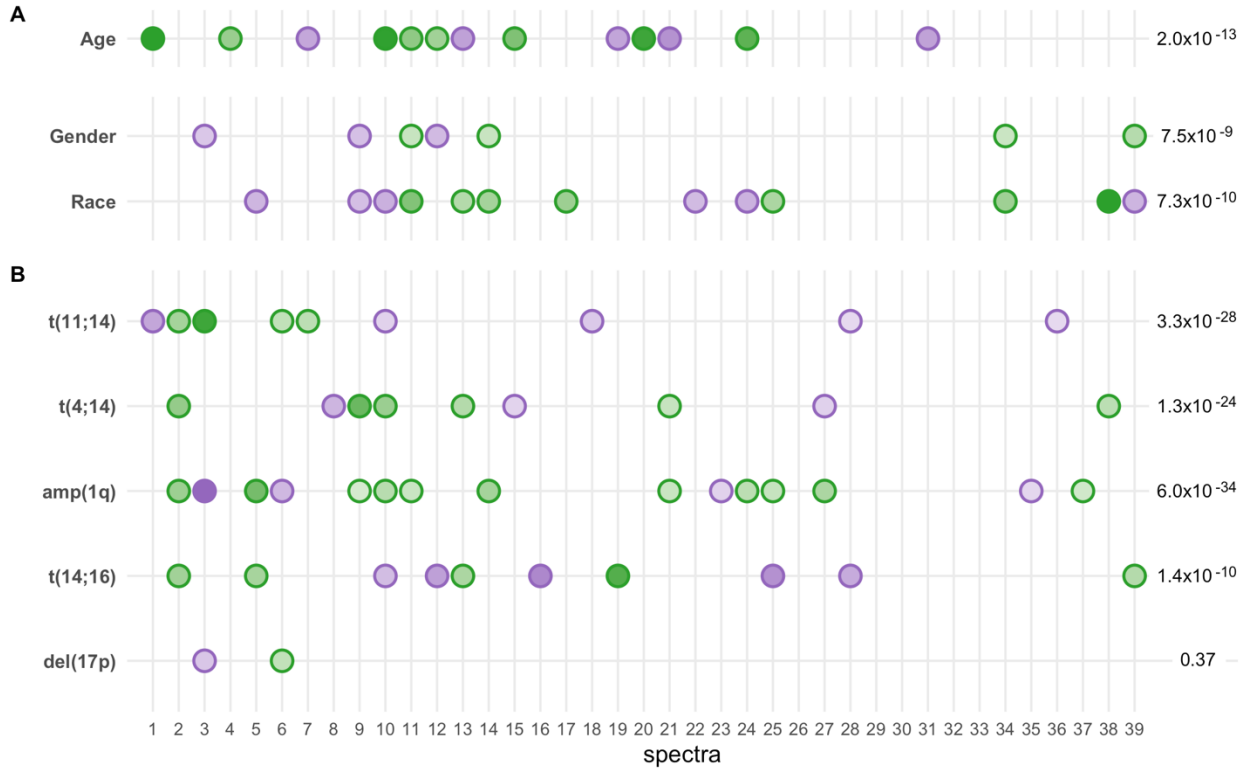

Figure 7. Overview of CD138+ spectra descriptive modeling. Multivariable modeling results for A) demographic characteristics and B) tumor FISH aberrations. The outcome of interest (left y-axis) and the significance of the full 39-spectra model (right y-axis). Spectra variables are illustrated on the x-axis. The effect size (intensity) and direction of each spectrum are indicated: green negative beta coefficient, purple positive beta coefficient. Effect size (intensity) is scaled to each plot (separated by a horizontal white space). No dot is shown if a spectrum was not significantly associated at $p < 0.05$ level.

SUPPLEMENTAL METHODS

Principal component analysis (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 referred to as *SPECTRA variables*, or *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.

Here we establish the matrix factorization (**MF**) natural for individual-based outcome modeling. Data matrices, $\boldsymbol{X}$ and $\boldsymbol{T}$, are *oriented with individuals as subjects (n rows) and genes as variables (g columns)*. 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, 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.

SUPPLEMENTAL TABLES

Table S1. Hazard ratios of high vs low risk of progression free survival.

Risk Score	Apparent HR	Adjusted HR	Optimism
PFS Spectra	4.22 (2.81-9.59)	3.08 (1.67-8.44)	1.14
UAMS	2.44 (1.74-3.35)	2.40 (1.70-3.31)	0.04
SBUK	1.94 (1.61-2.45)	1.92 (1.59-2.43)	0.01

Table S2. Progression free survival covariates and added predictive values.

Variable	Base Model	Δ AIC	Δ BIC	APV	LRT *P* Value
UAMS	R-ISS + Age	-49.6	-46.1	0.566	6.64×10^{-13}
SBUK	R-ISS + Age	-54.7	-51.1	0.589	5.12×10^{-14}
PFS Spectra	R-ISS + Age	-77.7	-74.2	0.668	4.28×10^{-19}
UAMS	R-ISS + Age + SBUK	-1.9	1.7	0.039	0.049
UAMS	R-ISS + Age + PFS Spectra	1.3	4.9	0.005	0.419
SBUK	R-ISS + Age + UAMS	-6.9	-3.4	0.089	0.003
SBUK	R-ISS + Age + PFS Spectra	-1.3	2.3	0.027	0.070
PFS Spectra	R-ISS + Age + UAMS	-26.7	-23.2	0.240	8.26×10^{-8}
PFS Spectra	R-ISS + Age + SBUK	-24.3	-20.8	0.215	2.87×10^{-7}
PFS Spectra	R-ISS + Age + UAMS + SBUK	-20.5	-16.99	0.184	2.07×10^{-6}

Table S3. Hazard ratios of high vs low risk early treatment failure.

Risk Score	Apparent HR	Adjusted HR	Optimism
TTF Spectra	3.93 (2.13-6.28)	3.10 (1.31-5.46)	0.82
UAMS	2.00 (1.50-2.70)	1.98 (1.48-2.68)	0.02
SBUK	1.86 (1.52-2.36)	1.85 (1.52-2.36)	0.01

Table S4. Time to first-line treatment failure covariates and added predictive values.

Variable	Base Model	Δ AIC	Δ BIC	APV	LRT *P* Value
UAMS	R-ISS + Age	-33.5	-30.1	0.478	2.49×10^{-9}
SBUK	R-ISS + Age	-40.6	-37.1	0.523	6.73×10^{-11}
TTF Spectra	R-ISS + Age	-55.1	-51.6	0.595	4.10×10^{-14}
UAMS	R-ISS + Age + SBUK	0.6	4.1	0.017	0.240
UAMS	R-ISS + Age + TTF Spectra	1.5	5.0	0.005	0.474
SBUK	R-ISS + Age + UAMS	-6.4	-2.9	0.102	0.004
SBUK	R-ISS + Age + TTF Spectra	-0.4	3.1	0.024	0.124
TTF Spectra	R-ISS + Age + UAMS	-20.1	-16.6	0.229	2.61×10^{-6}
TTF Spectra	R-ISS + Age + SBUK	-14.9	-11.4	0.172	3.96×10^{-5}
TTF Spectra	R-ISS + Age + UAMS + SBUK	-13.6	-10.1	0.158	7.98×10^{-5}

SUPPLEMENTAL FIGURES

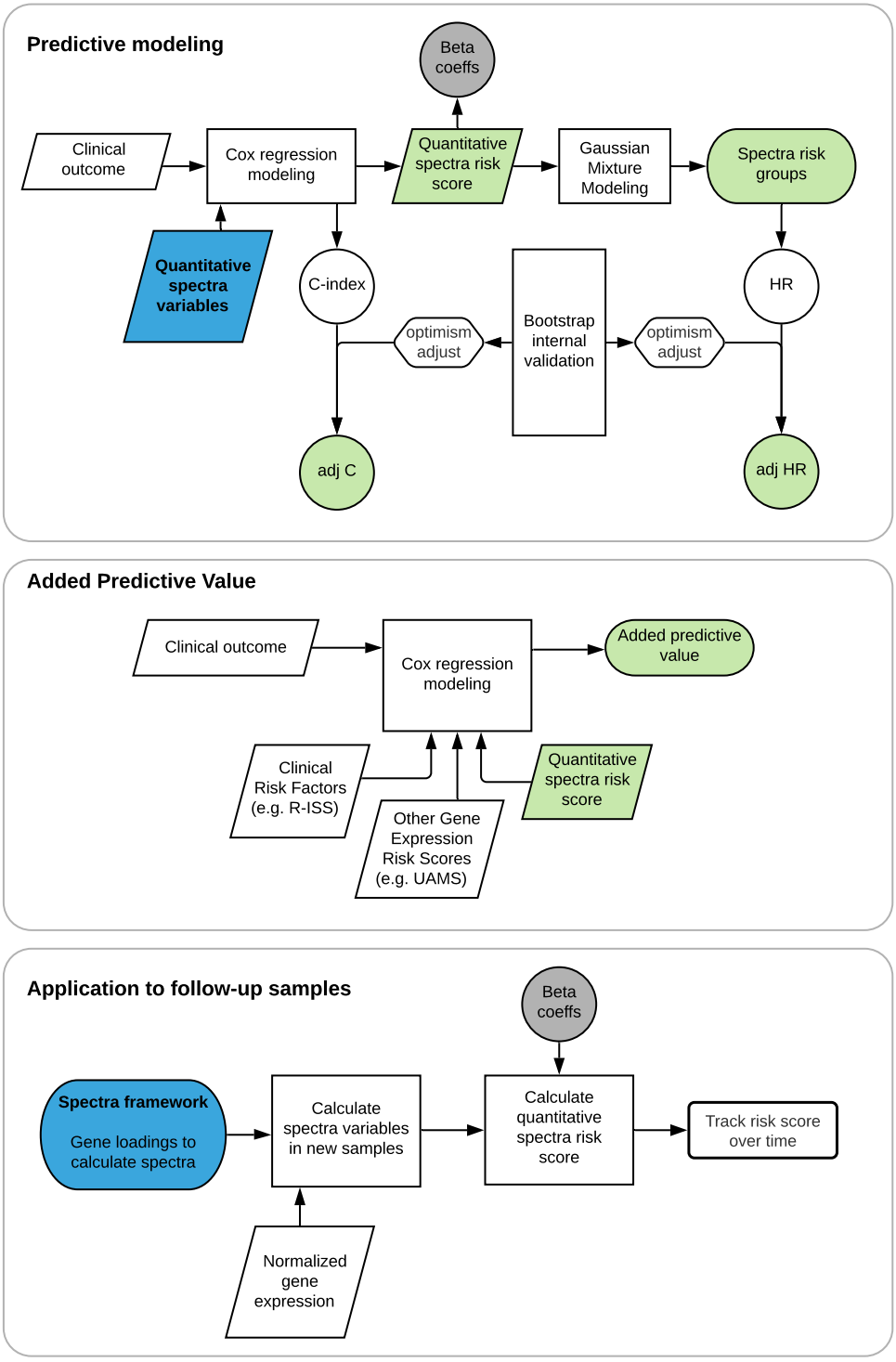

Figure S1. Diagram of predictive modeling.

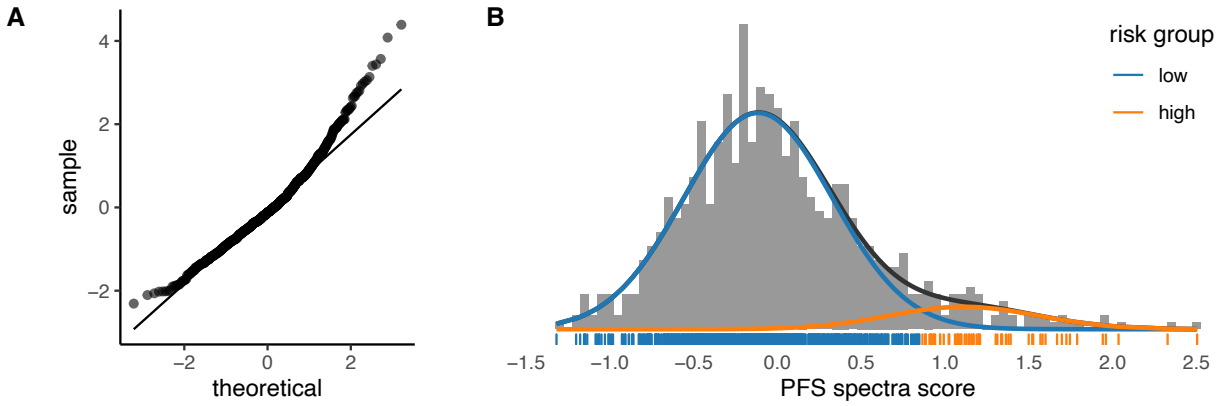

Figure S2. Progression free survival spectra score. A) Quantile-quantile plot. B) Gaussian mixture modeling to identify patients at high and low risk of progression.

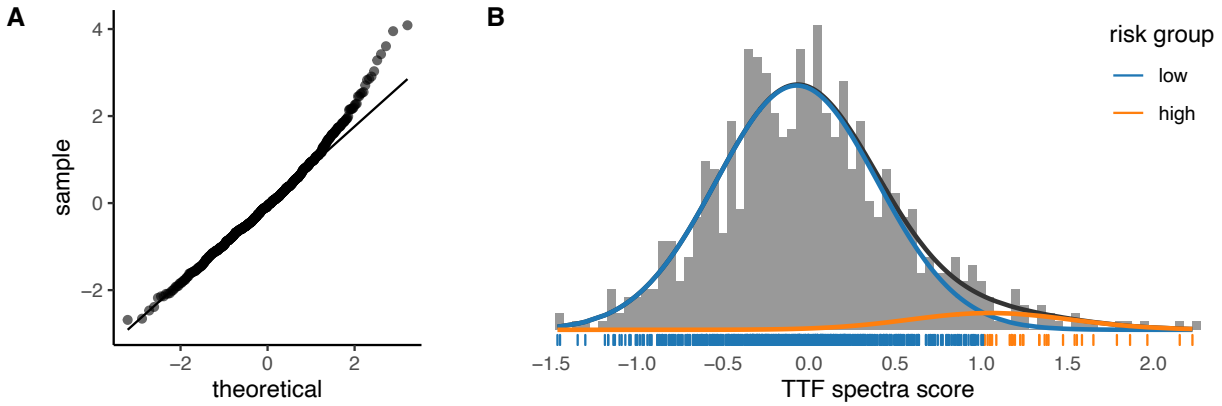

Figure S3. Time to first line treatment failure spectra score. A) Quantile-quantile plot. B) Gaussian mixture modeling to identify patients at high and low risk of early treatment failure.

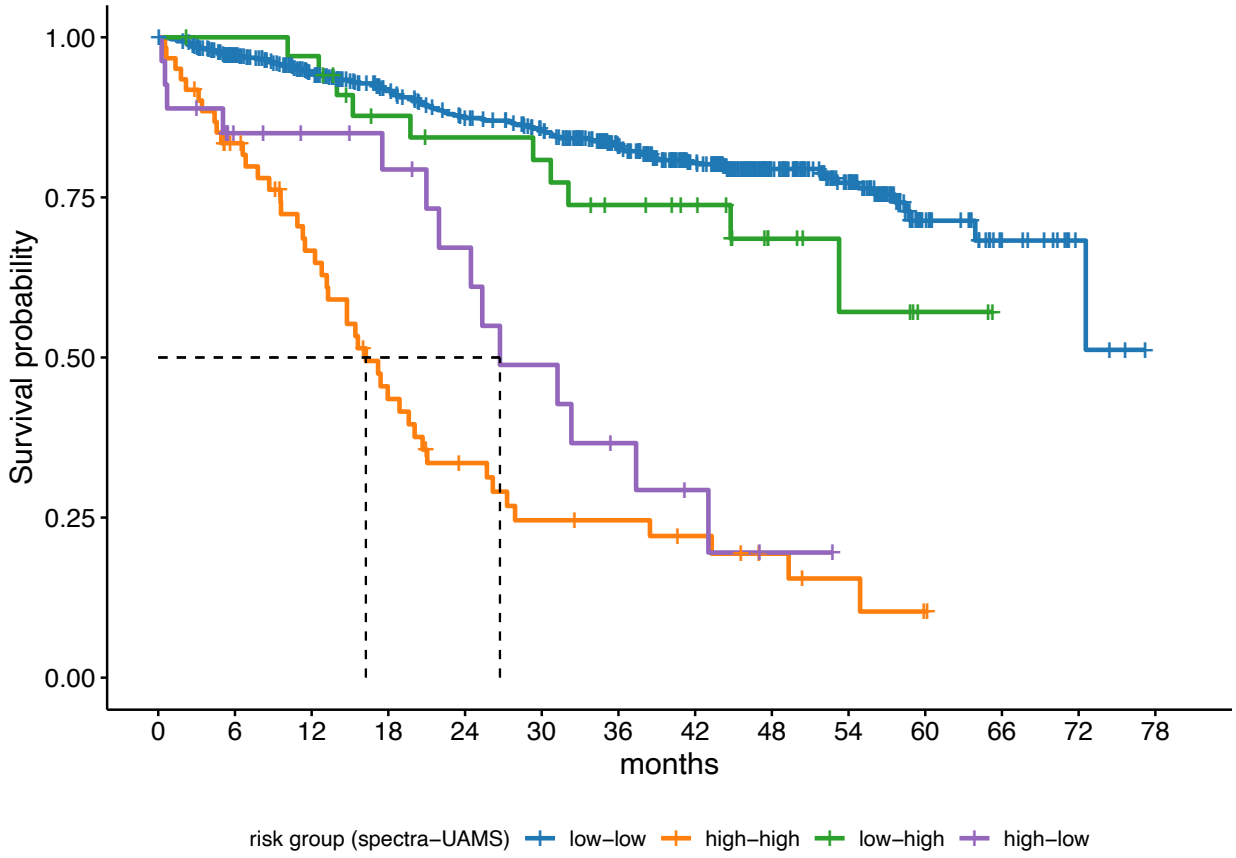

Figure S4. Kaplan-Meier curves of overall survival by spectra and UAMS risk-groups. Spectra identified patients as high-risk with worse survival that UAMS classified as low risk.