Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/46/eabb3461/DC1

Supplementary Materials for

Plasma-derived extracellular vesicle analysis and deconvolution enable prediction and tracking of melanoma checkpoint blockade outcome

Alvin Shi, Gyulnara G. Kasumova, William A. Michaud, Jessica Cintolo-Gonzalez, Marta Díaz-Martínez, Jacqueline Ohmura, Arnav Mehta, Isabel Chien, Dennie T. Frederick, Sonia Cohen, Deborah Plana, Douglas Johnson, Keith T. Flaherty, Ryan J. Sullivan, Manolis Kellis*, Genevieve M. Boland*

*Corresponding author. Email: gmboland@partners.org (G.M.B.); manoli@mit.edu (M.K.)

Published 13 November 2020, *Sci. Adv.* **6**, eabb3461 (2020)
DOI: 10.1126/sciadv.abb3461

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/46/eabb3461/DC1)

Supplementary figure legends

Supplementary figure 1: (a) Outline of both the discovery and validation study cohort. The complete cohort metadata is available in Supplementary Table 1. (b) Outline of the processing and analysis steps undertaken to generate the major results in the paper.

Supplementary figure 2: *In vitro* EV characterization. (a) Electron microscope images of representative cell-line and patient-derived EVs. (b) Nanosight analysis of a representative cell line sample (RPMI) and patient sample. (c) Western blot of 5 cell line/paired EV proteins with no calnexin within EV, but high levels of CD9 within EV as compared to paired cells.

Supplementary figure 3: Characterization of transcriptomic similarities between cell-line derived melanoma samples and their matched EV. (a) Scatter plot visualizing differences between tissue and plasma-derived EV in patient samples. (b) Histogram visualizing the log fold changes between melanoma cell-lines and their EV counterparts. The profiles were compiled using the average of four expression profiles. (c) Concordance analysis across our 4 cell-line samples. Concordance was calculated by using a low expression cutoff as a cut-off for expressed vs. non-expressed status. Genes that were either expressed or not expressed in both tissue and EV compartments are considered concordant. (d) Correlation plot between all 4 cell-lines and their EV counterparts.

Supplementary figure 4: Box-plots and associated p-values for validated MSigDB canonical pathways that differ between responders (green) and non-responders (purple). The visualized points are individual GSVA scores inferred for each pathway. The p-values were generated by comparing responder vs. non-responder GSVA scores via a Mann-Whitney U-test.

Supplementary figure 5: (a-e) Un-normalized time-series plots showing time dynamics for pathways and genes discussed in **Fig. 2b** and **Fig. 2d**. Individual GSVA scores were used to plot the TCR KEGG pathway, while platform-normalized log2 expression values were used to plot the individual gene expression levels.

Supplementary figure 6: (a-d) Normalized and unnormalized time-series plots showing time dynamics for selected pathways. The plotting was performed per methodology previously discussed in captions for **Fig. 2b**, **Fig. 2d** and **SFig 5**.

Supplementary figure 7: Kaplan-Meier progression-free and overall survival curves for selected genes that showed significant or near-significant differences between high-expressed and low-expressed patients.

Supplementary figure 8: Kaplan-Meier progression-free and overall survival curves for selected genes that showed significant or near-significant differences between high-expressed and low-expressed patients.

Supplementary figure 9: Example of per-patient imputed tumor-EV expression from our Bayesian deconvolution model. (a-b) Posterior estimates for two illustrative genes for high immune fraction (CD8A) and high tumor fraction (MIR47888). (c-d) Predicted tumor-derived EV expression from our deconvolution model for CD8A and MIR4788 for a subset of patients.

Supplementary figure 10: Kaplan-Meier progression-free survival (PFS) plots for COSMIC mutational fraction in the validation evRNA-seq cohort. High and low expression classifications were determined for each patient based on whether a particular patient's COSMIC mutational fraction was in the top half (teal) or bottom half (yellow) of the validation cohort's COSMIC mutational fraction distribution.

Supplementary Tables Legends

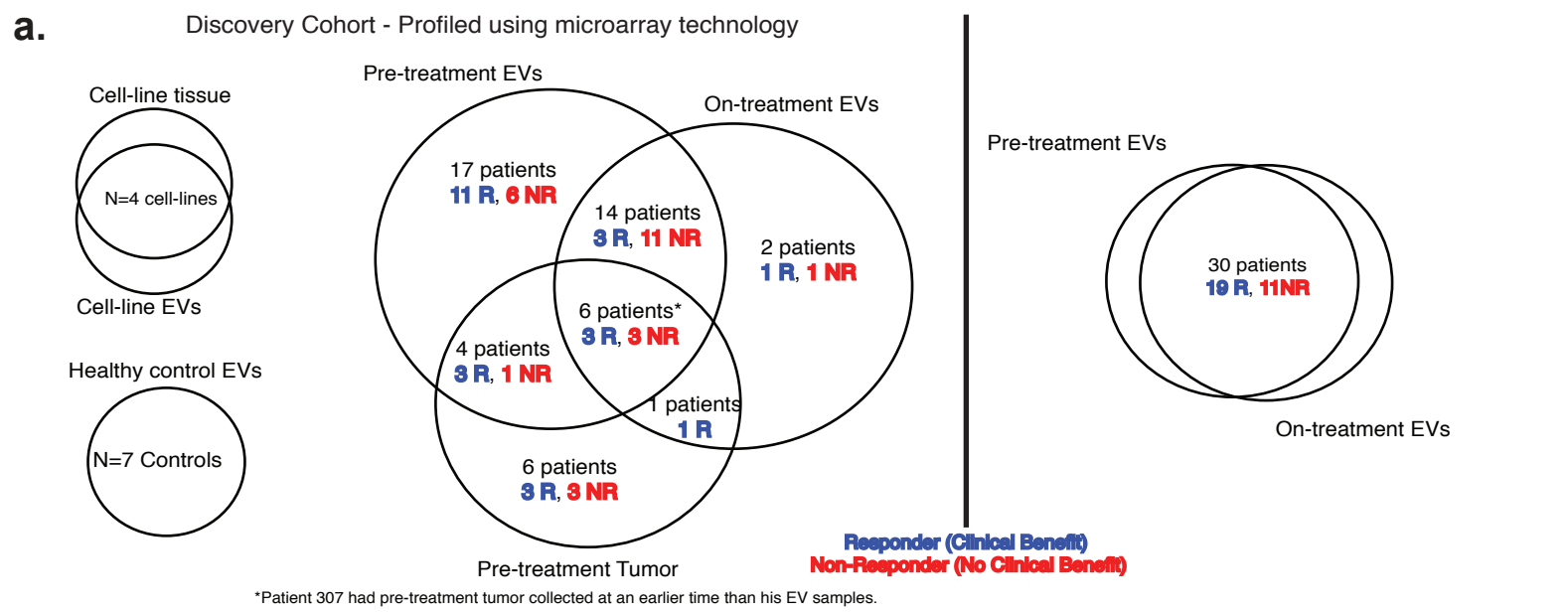
Supplementary Table 1: Worksheet 1: Anonymized metadata of patient cohort from discovery cohort. Worksheet 2: anonymized metadata of patient cohort from validation cohort. Additional columns in worksheet 2 denote driver mutations detected by SNaPShot tumor sequencing and the specific mutations validated in our evRNA-seq mutational analysis.

Supplementary Table 2: GO pathways enriched in patient EV-unique or tumor-unique samples.

Supplementary Table 3: Worksheet 1: Validated DEGs and associated statistics from limma from our on-treatment DEG analysis. Worksheet 2: All DEGs from on-treatment discovery cohort and associated limma statistics. Worksheet 3: All DEGs from on-treatment validation cohort and associated limma statistics.

Supplementary Table 4: Worksheet 1: Validated DEGs and associated statistics from limma from our pre-treatment DEG analysis. Worksheet 2: All DEGs from pre-treatment discovery cohort and associated limma statistics. Worksheet 3: All DEGs from pre-treatment validation cohort and associated limma statistics.

Supplementary Table 5: Estimated mixing fractions from our EV deconvolution model running in multi-gene mode for all genes. Full details regarding the model are available in the supplemental note.



*Patient 307 had pre-treatment tumor collected at an earlier time than his EV samples.

b.

Processing & Analysis Pipelines

*All numbers here refer to the number of samples used in each comparison

Discovery Cohort

Microarray .cel files supplied directly by ThermoFisher (contractor) from Human-Transcriptome

Standardization using 'rma' and within-microarray batch correction using 'ComBat' to produce log2 expression values

Technical replicates averaged and combined to single amalgamted files; removed samples collected from >365 before treatment start

Validation Cohort

evRNA-seq .fastq files from Illumina HiSeq 2000 sequencing from Ting Lab @ MGH (collaborator)

Production of .bam files via STAR alignment against hg38. Count matrix via R library 'featureCounts' against ucsc_hg38.gtf

Count matrix transformed into log2(TPM+1) using custom R script.

Mapping with 'bwa' against hg38 reference

Calling all (somatic & germline) mutations against hg38 using GATK HaplotypeCaller

Overlap GATK calls against COSMIC v91 significant genes

Joint-normalization of log2(TPM+1) RNA-seq & log2 microarray expression data with 'ComBat' followed by quantile normalization. This is to remove platform-specific effects & produce reasonable basis joint for analyses requiring both datasets

---

Discovery cohort/Microarray only (micorrary expression was not platform-normalized)

1. Comparing cell-line EVs vs. cell-line (4 melanoma cell-lines vs. 4 melanoma cell-lined derived exosomes, Supplemental Fig 2)

2. CIBERSORT on patient & EV data and comparison of enrichment of specific immune sub-populations (Fig. 1)

3. Deconvolution modeling. Selected genes are displayed in Fig 4a.

Joint validation + discovery cohort analyses using platform-normalized expression data

Discovery Cohort DEG (via limma) & KEGG differential analysis (GAGE)

Validation Cohort DEG calling (via limma) & KEGG differential analysis (GAGE)

Downstream analyses

EV pre-treatment responders vs. non-responders (28R vs. 13NR) + EV pre-treatment responders vs. non-responders (14R vs. 16NR)

- Pre-treatment validation cohort & discovery pathway & DEG overlap (Fig. 3a-b)
- Discovery cohort Random Forest model creation & K=5 internal cross-validation based on replicated DEGs (Fig. 3c, methods)
  - Test performance of Random Forest trained using training set to predict testing set to compute ROC & AUC (Fig. 3c, methods)

EV on-treatment responders vs. non-responders (28R vs. 9NR) + EV on-treatment responders vs. non-responders (14R vs. 16NR)

- On-treatment calidation cohort & discovery KEGG & DEG overlap (Fig. 2a, 2c)
  - Aggregate time dynamics analysis & plotting using GSVA (for KEGG pathways) or normalized gene expression values for validated DEGs (53R samples vs. 21NR samples Fig. 2b, 2d)

Tis pre responders vs. non-responders (16R vs 14NR)

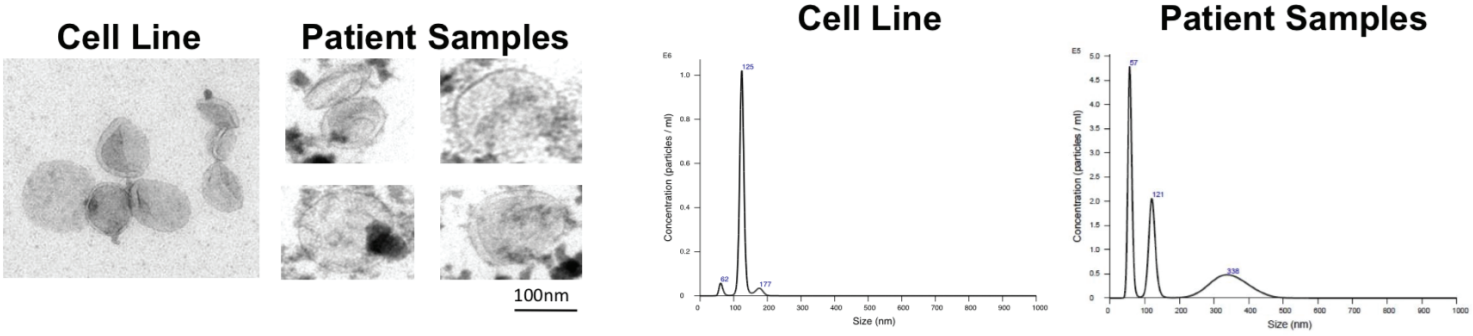
Compared VCF calls to panel sequencing results

Compared per-patient mutational load between responders and non-responders

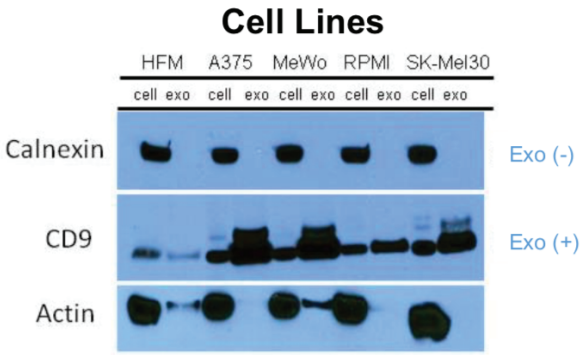
SFig. 1

A) Electron Microscopy

B) NANOSIGHT

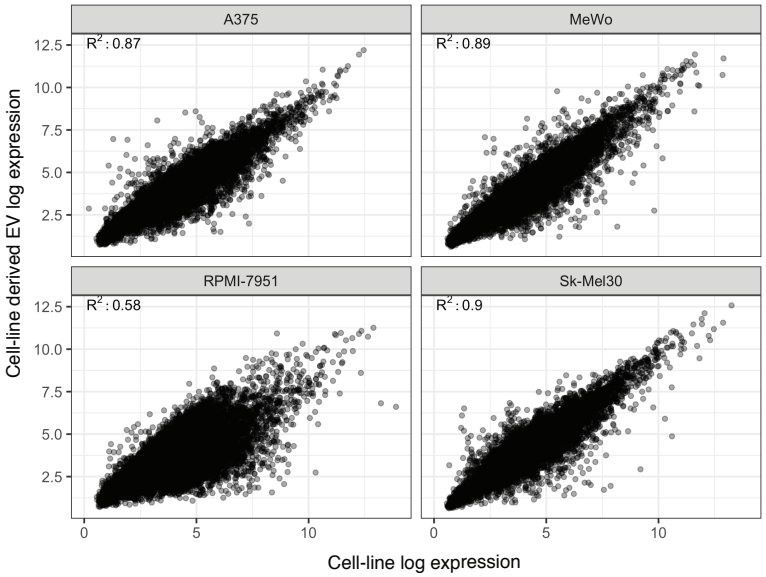


C) Western Blot

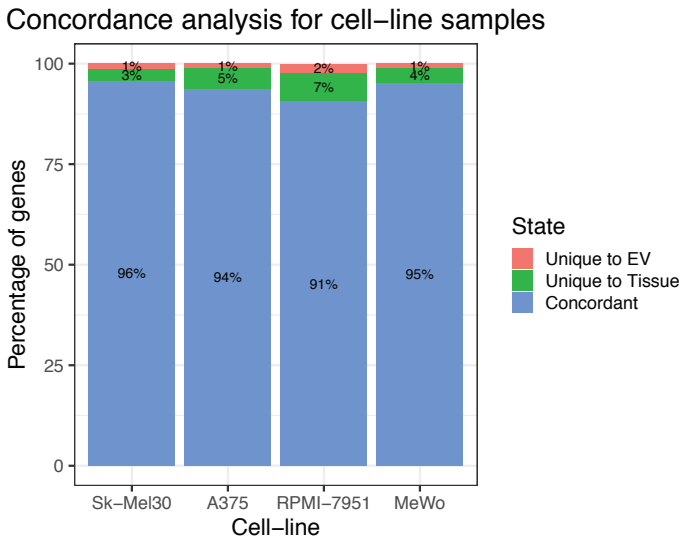


SFig. 2

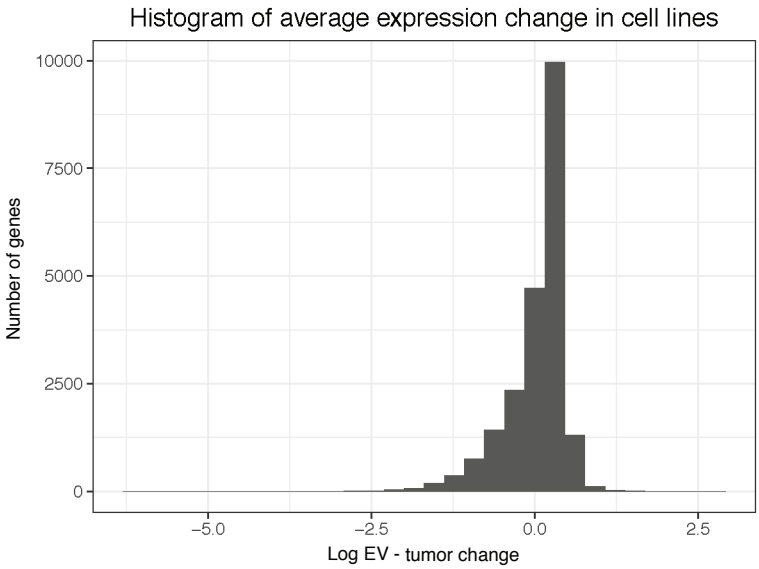
a.


A375
MeWo
RPMI-7951
Sk-Mel30
R^2 : 0.87
R^2 : 0.89
R^2 : 0.58
R^2 : 0.9
Cell-line derived EV log expression
Cell-line log expression

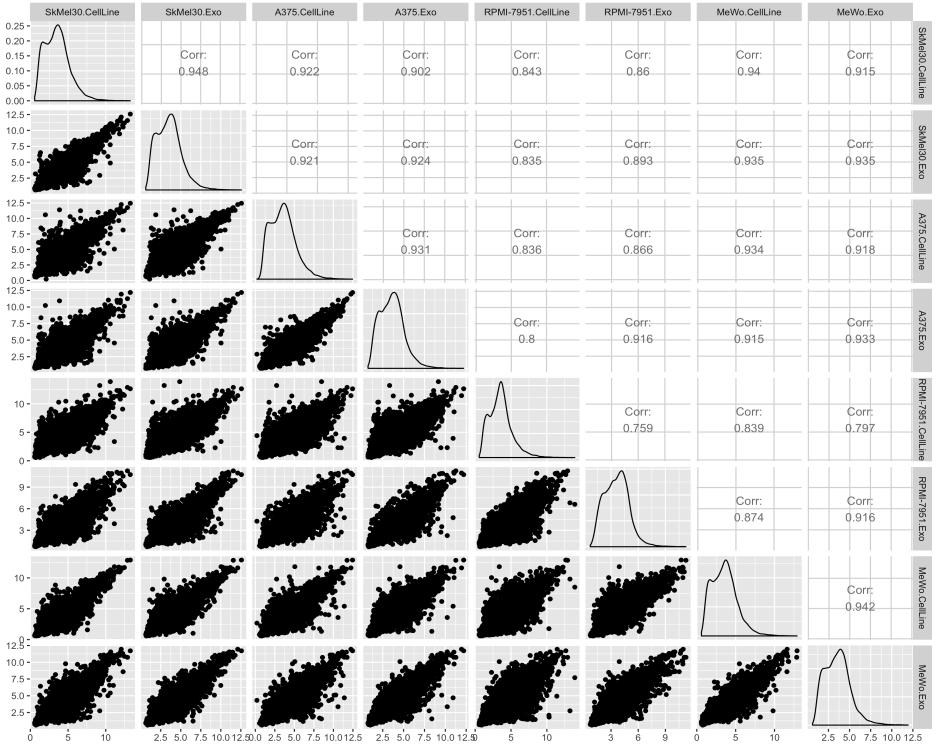
b.


Concordance analysis for cell−line samples
Percentage of genes
Cell−line
Sk−Mel30
A375
RPMI−7951
MeWo
1%
3%
96%
1%
5%
94%
2%
7%
91%
1%
4%
95%
State
Unique to EV
Unique to Tissue
Concordant

c.


Histogram of average expression change in cell lines
Number of genes
Log EV - tumor change

d.

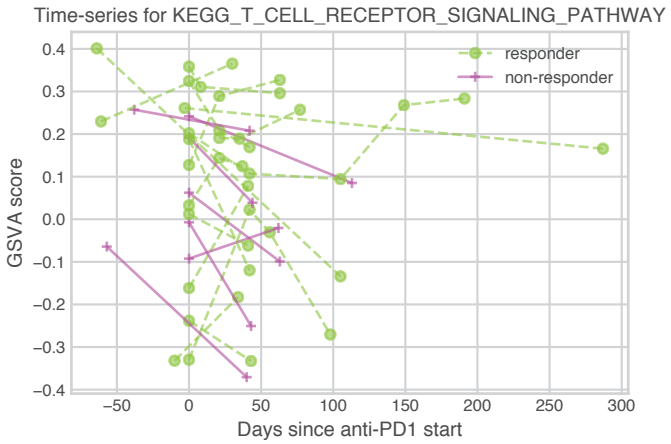

SkMel30.CellLine
SkMel30.Exo
A375.CellLine
A375.Exo
RPMI-7951.CellLine
RPMI-7951.Exo
MeWo.CellLine
MeWo.Exo
Corr: 0.948
Corr: 0.922
Corr: 0.902
Corr: 0.843
Corr: 0.86
Corr: 0.94
Corr: 0.915
Corr: 0.921
Corr: 0.924
Corr: 0.835
Corr: 0.893
Corr: 0.935
Corr: 0.935
Corr: 0.931
Corr: 0.836
Corr: 0.866
Corr: 0.934
Corr: 0.918
Corr: 0.8
Corr: 0.916
Corr: 0.915
Corr: 0.933
Corr: 0.759
Corr: 0.839
Corr: 0.797
Corr: 0.874
Corr: 0.916
Corr: 0.942

SFig. 3

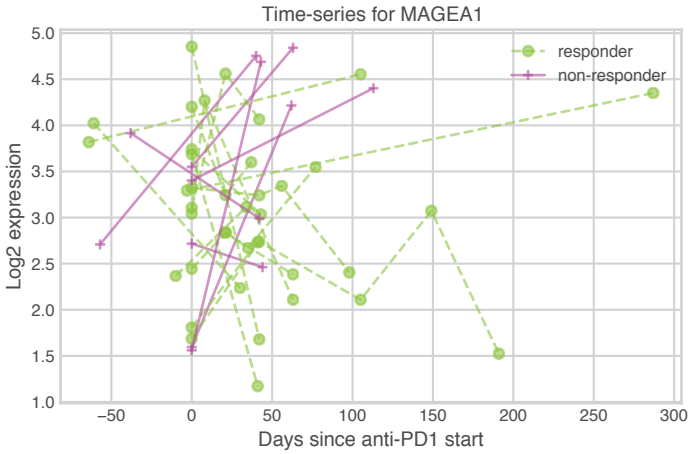
On-treatment validated differential pathways

SFig. 4

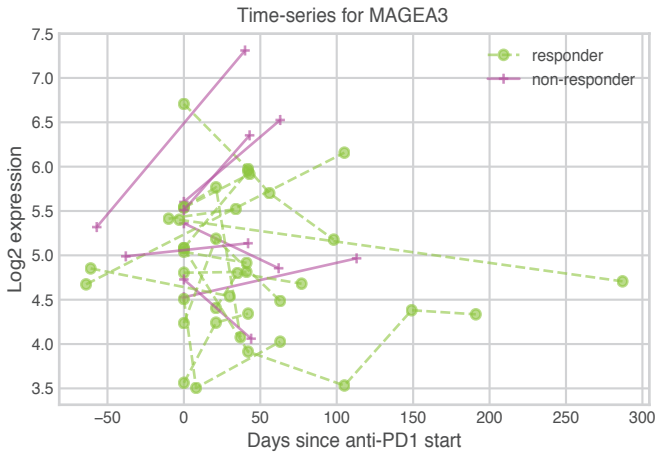
a.


Time-series for KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY
GSVA score
Days since anti-PD1 start
responder
non-responder

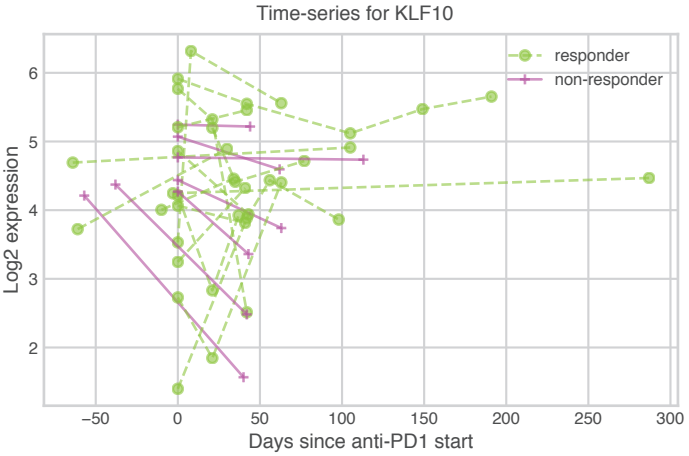
b.


Time-series for MAGEA1
Log2 expression
Days since anti-PD1 start
responder
non-responder

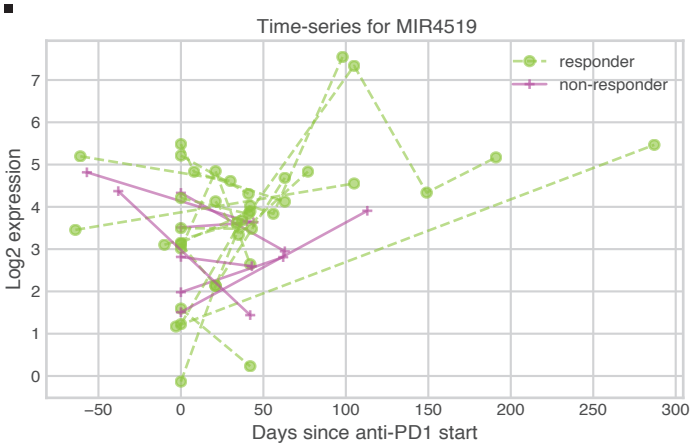
c.


Time-series for MAGEA3
Log2 expression
Days since anti-PD1 start
responder
non-responder

d.

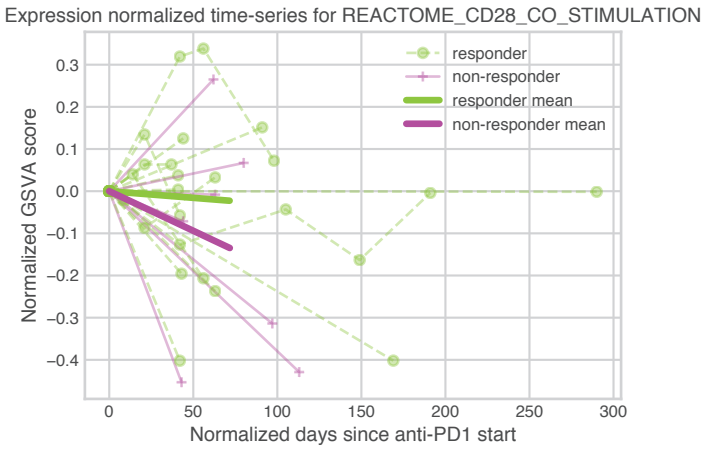

Time-series for KLF10
Log2 expression
Days since anti-PD1 start
responder
non-responder

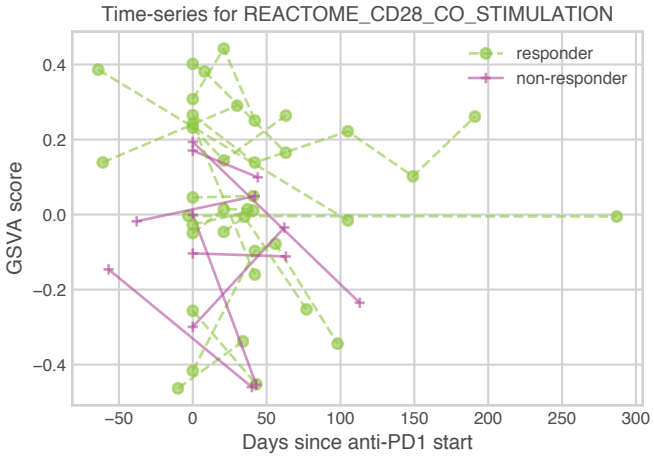
e.


Time-series for MIR4519
Log2 expression
Days since anti-PD1 start
responder
non-responder

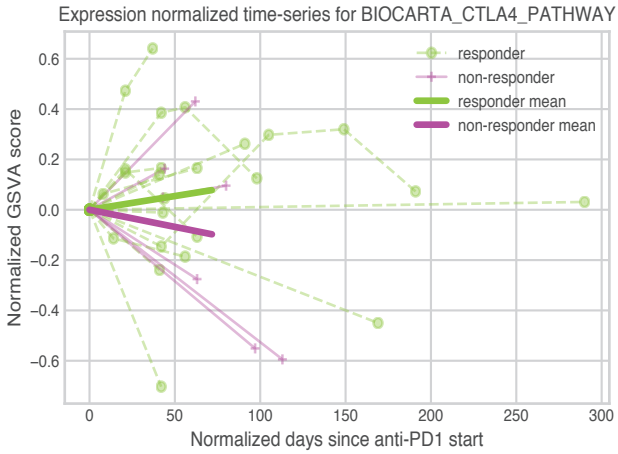
SFig. 5

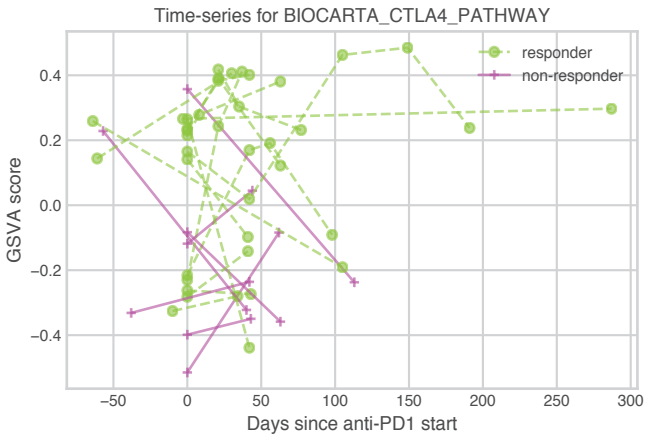
a.


Expression normalized time-series for REACTOME_CD28_CO_STIMULATION
responder
non-responder
responder mean
non-responder mean
Normalized GSVA score
Normalized days since anti-PD1 start

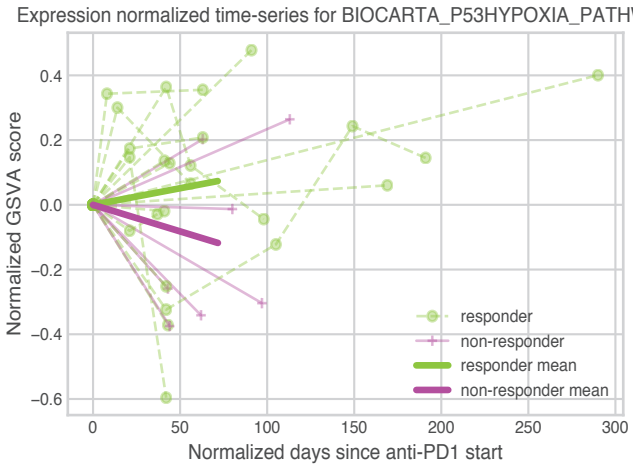

Time-series for REACTOME_CD28_CO_STIMULATION
responder
non-responder
GSVA score
Days since anti-PD1 start

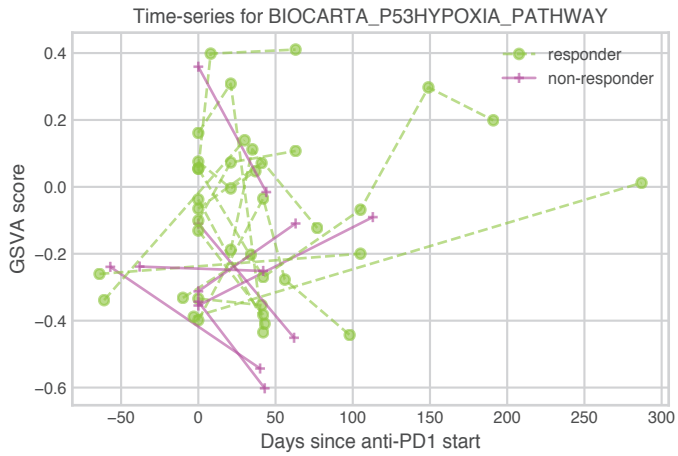
b.


Expression normalized time-series for BIOCARTA_CTLA4_PATHWAY
responder
non-responder
responder mean
non-responder mean
Normalized GSVA score
Normalized days since anti-PD1 start

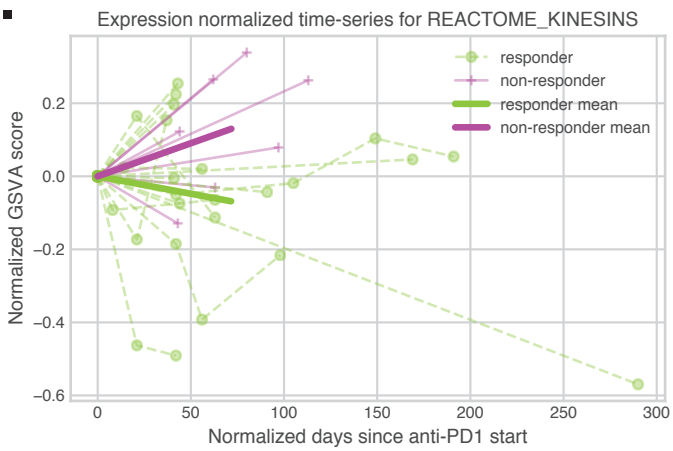

Time-series for BIOCARTA_CTLA4_PATHWAY
responder
non-responder
GSVA score
Days since anti-PD1 start

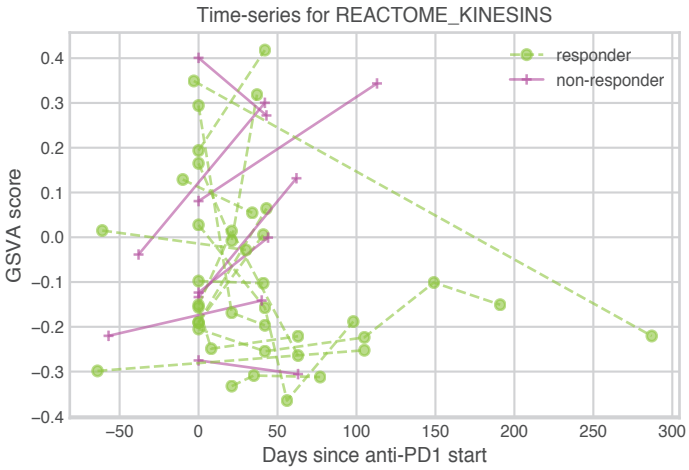
c.


Expression normalized time-series for BIOCARTA_P53HYPOXIA_PATHWAY
responder
non-responder
responder mean
non-responder mean
Normalized GSVA score
Normalized days since anti-PD1 start


Time-series for BIOCARTA_P53HYPOXIA_PATHWAY
responder
non-responder
GSVA score
Days since anti-PD1 start

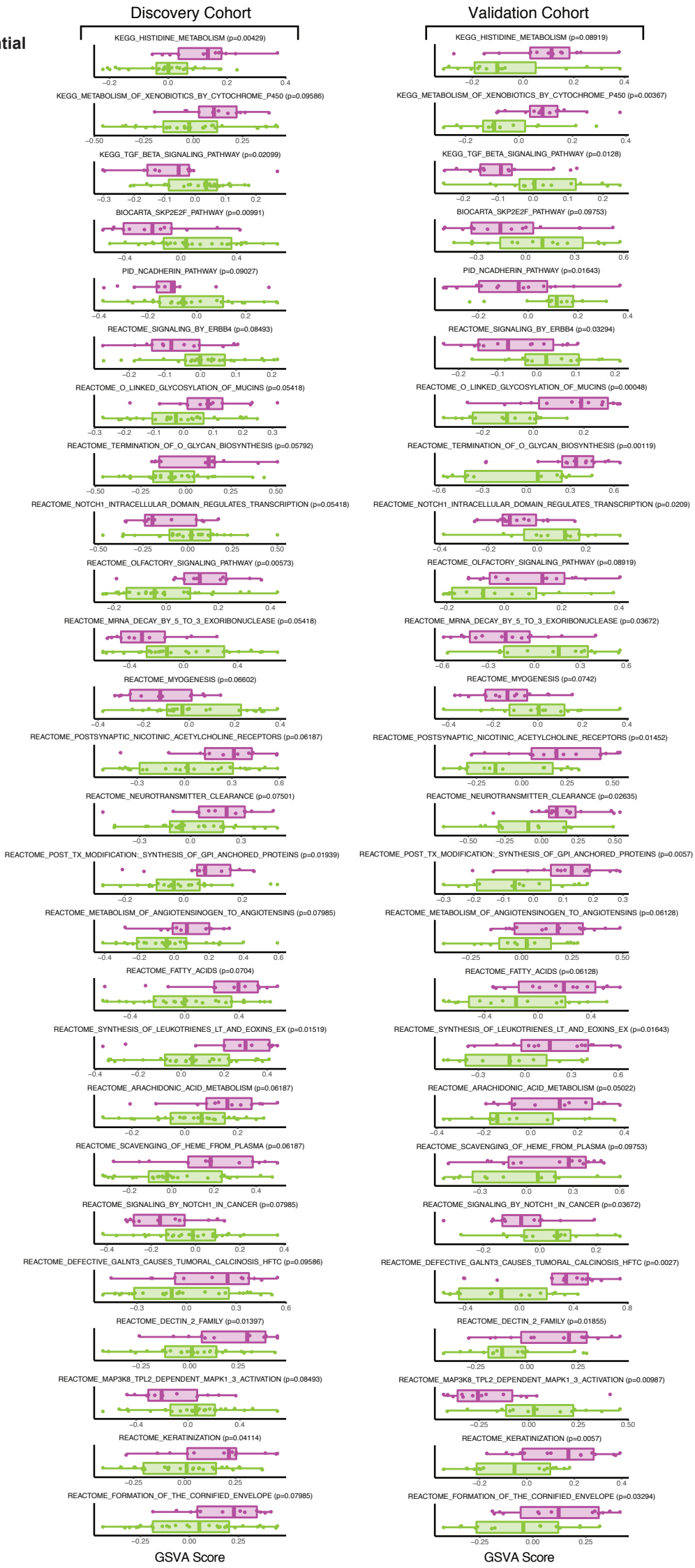
d.


Expression normalized time-series for REACTOME_KINESINS
responder
non-responder
responder mean
non-responder mean
Normalized GSVA score
Normalized days since anti-PD1 start

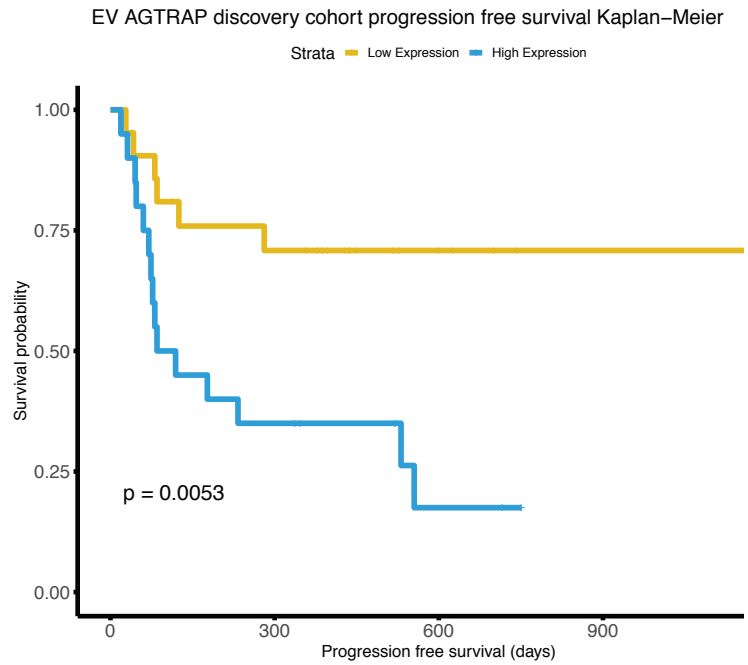

Time-series for REACTOME_KINESINS
responder
non-responder
GSVA score
Days since anti-PD1 start

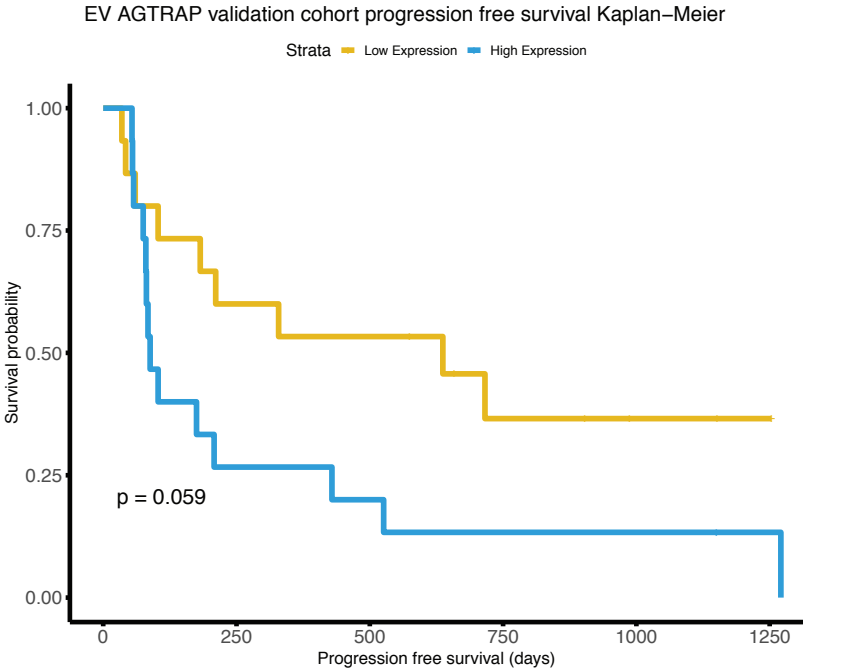
SFig. 6

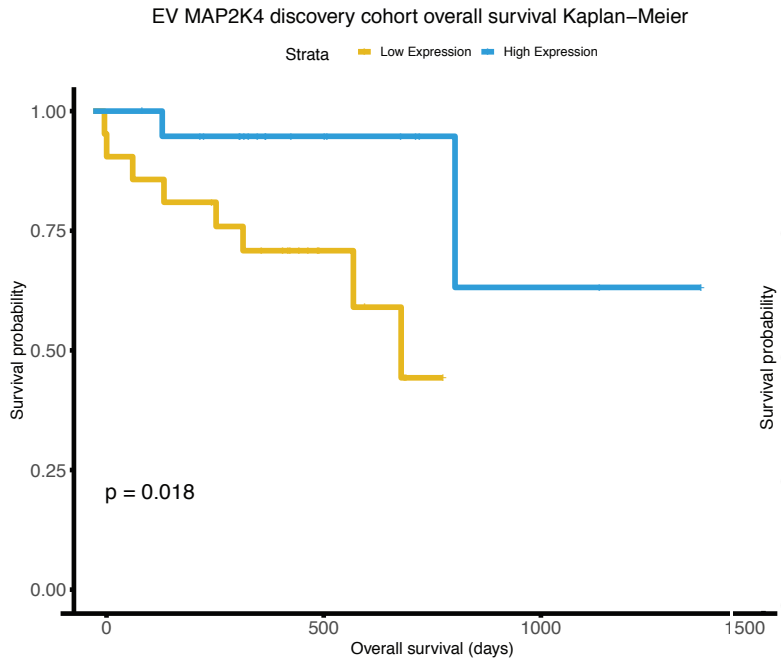
Pre-treatment validated differential pathways

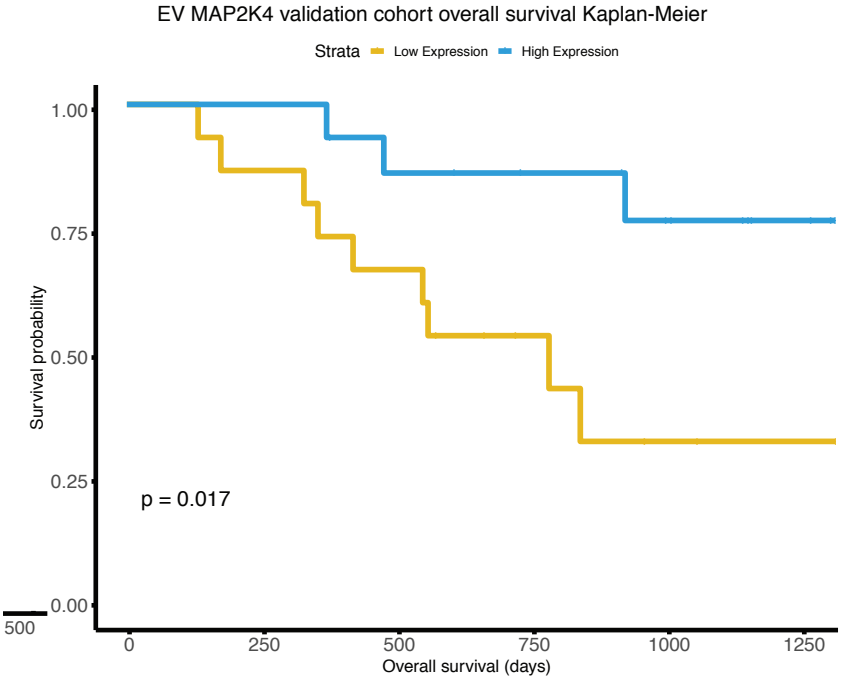


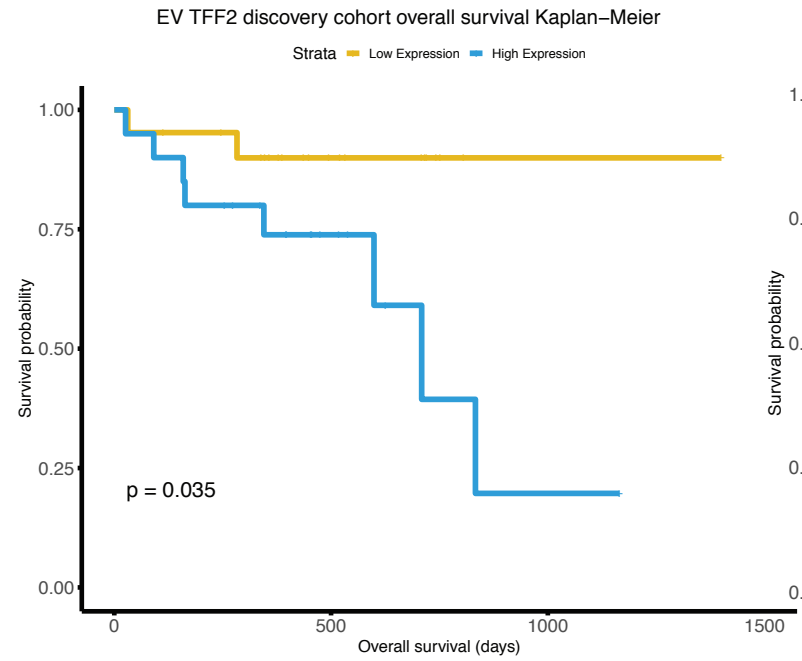
SFig. 7

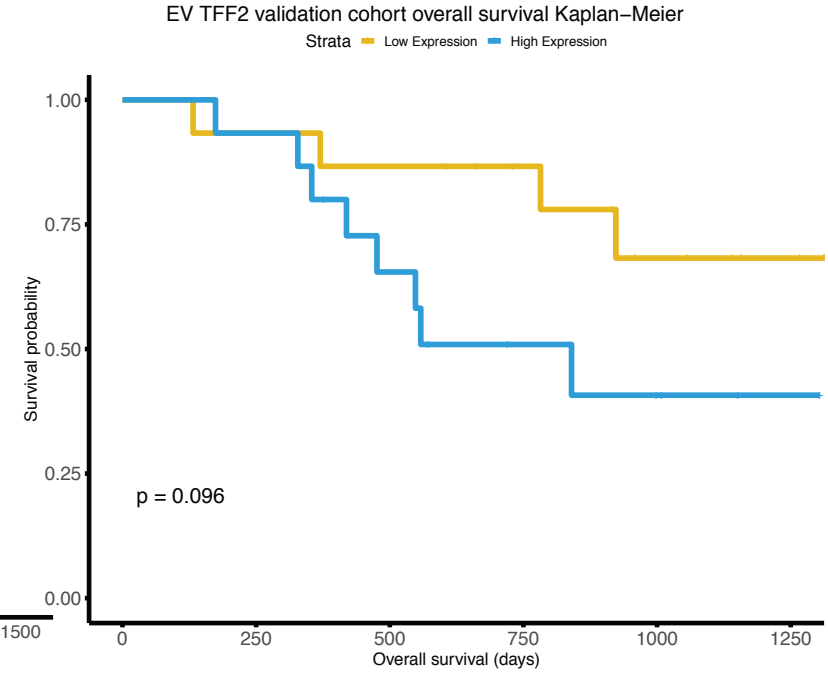

EV AGTRAP discovery cohort progression free survival Kaplan–Meier
Strata
Low Expression
High Expression
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.0053
0
300
600
900
Progression free survival (days)


EV AGTRAP validation cohort progression free survival Kaplan–Meier
Strata
Low Expression
High Expression
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.059
0
250
500
750
1000
1250
Progression free survival (days)

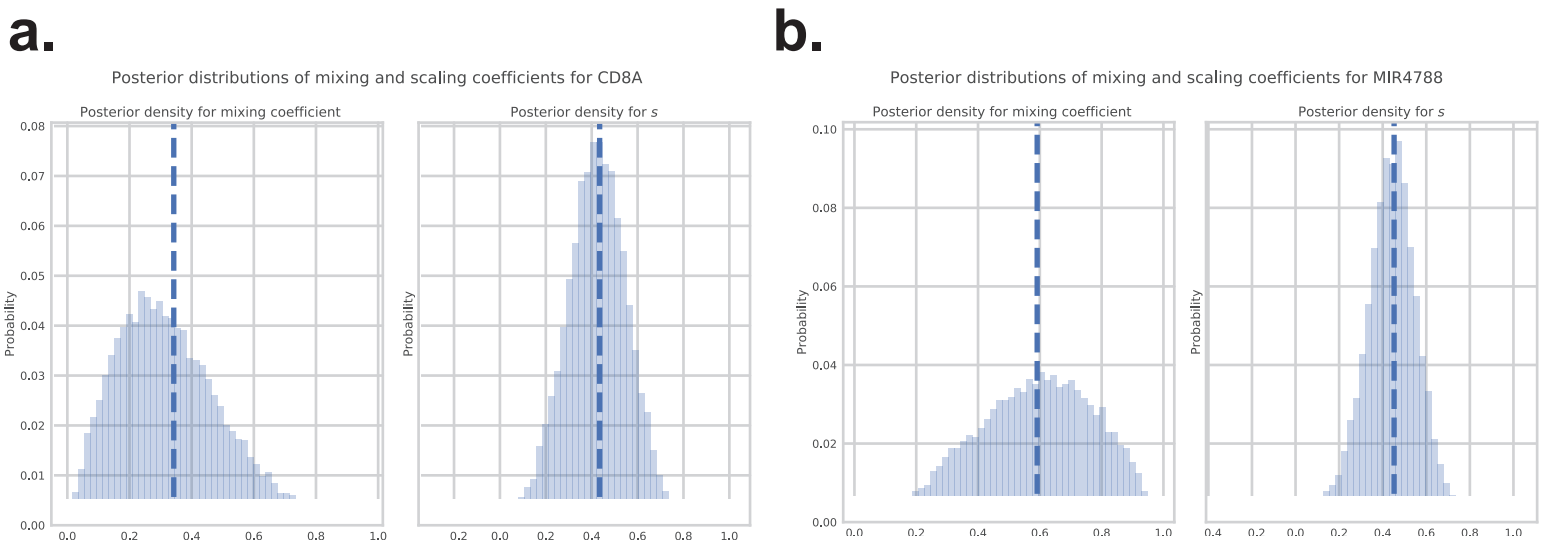

EV MAP2K4 discovery cohort overall survival Kaplan–Meier
Strata
Low Expression
High Expression
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.018
0
500
1000
1500
Overall survival (days)

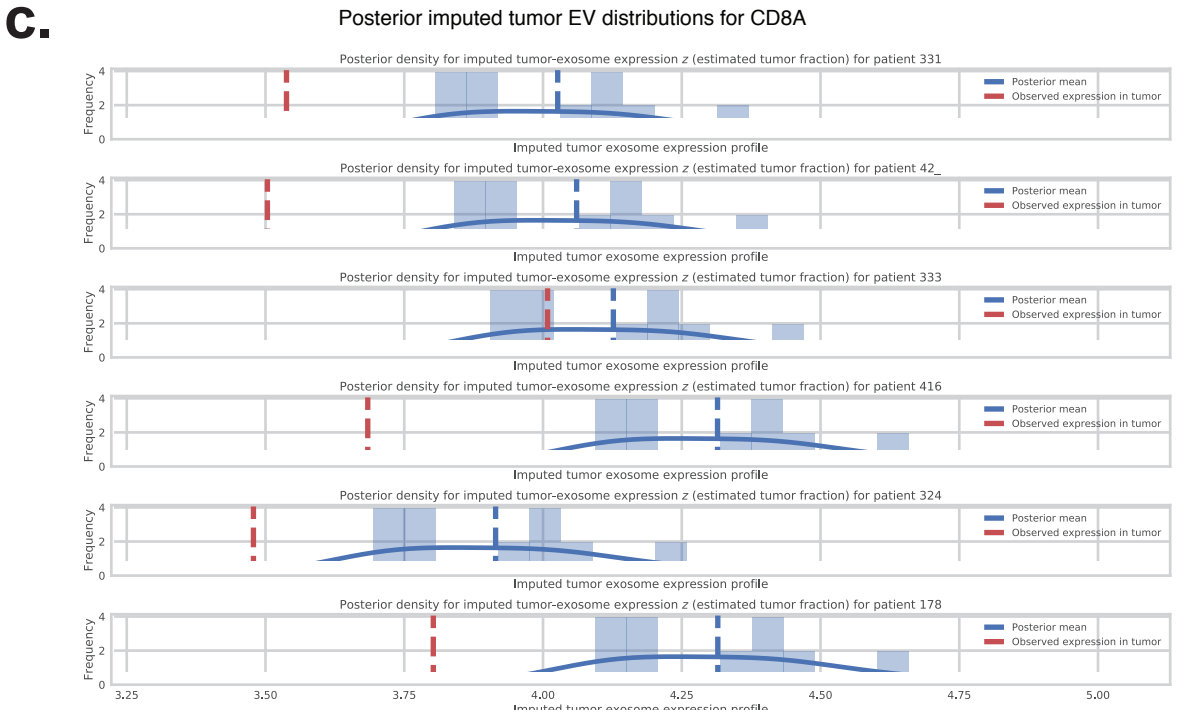

EV MAP2K4 validation cohort overall survival Kaplan-Meier
Strata
Low Expression
High Expression
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.017
0
250
500
750
1000
1250
Overall survival (days)

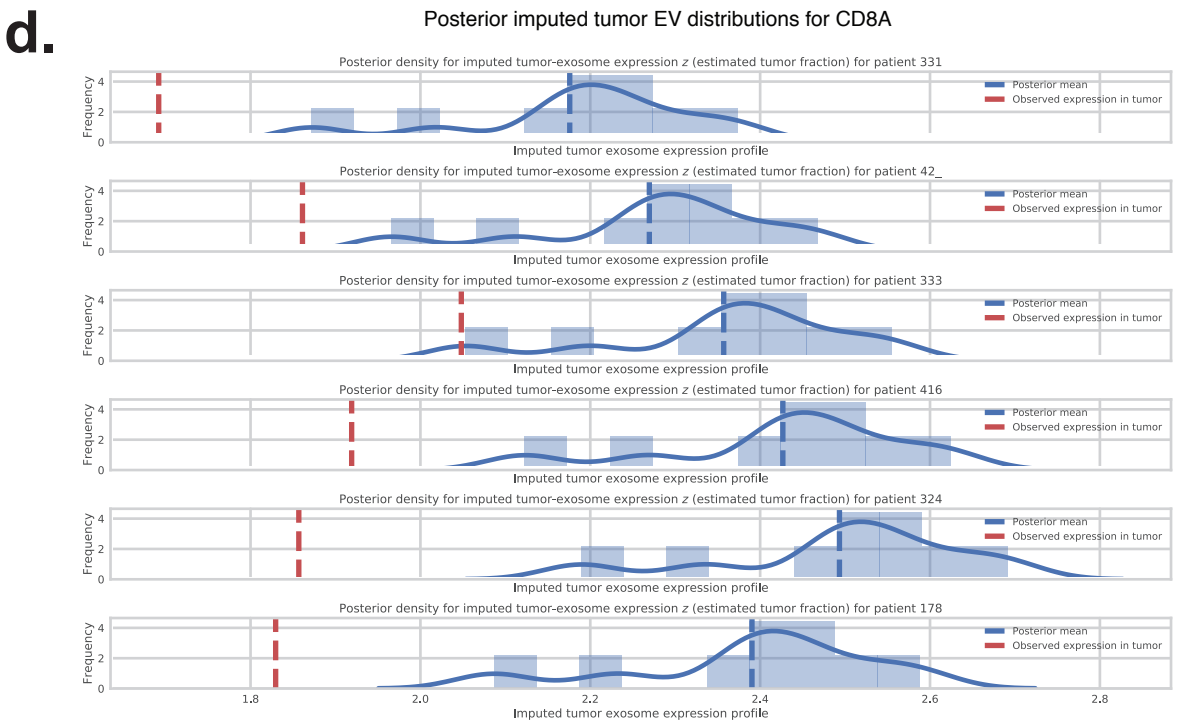

EV TFF2 discovery cohort overall survival Kaplan–Meier
Strata
Low Expression
High Expression
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.035
0
500
1000
1500
Overall survival (days)


EV TFF2 validation cohort overall survival Kaplan–Meier
Strata
Low Expression
High Expression
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.096
0
250
500
750
1000
1250
Overall survival (days)

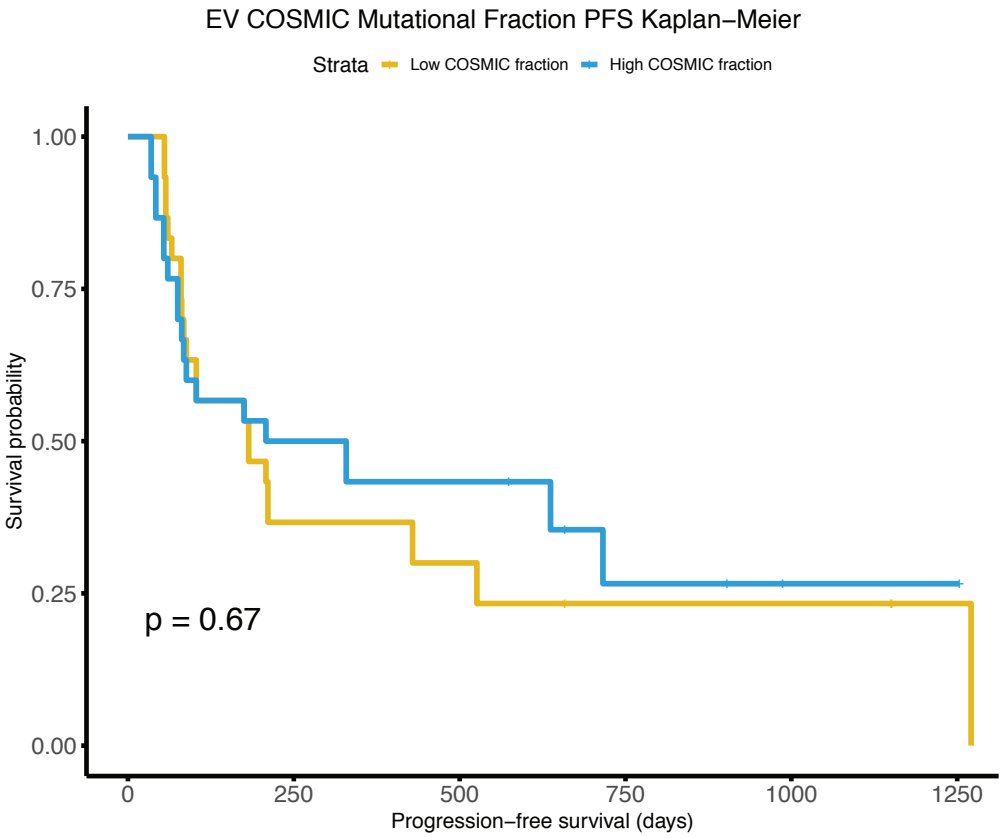
SFig. 8


a.
Posterior distributions of mixing and scaling coefficients for CD8A
Posterior density for mixing coefficient
Posterior density for s
Probability
b.
Posterior distributions of mixing and scaling coefficients for MIR4788
Posterior density for mixing coefficient
Posterior density for s
Probability


c.
Posterior imputed tumor EV distributions for CD8A
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 331
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 42_
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 333
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 416
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 324
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 178
Posterior mean
Observed expression in tumor
Frequency
Imputed tumor exosome expression profile


d.
Posterior imputed tumor EV distributions for CD8A
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 331
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 42_
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 333
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 416
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 324
Posterior density for imputed tumor-exosome expression z (estimated tumor fraction) for patient 178
Posterior mean
Observed expression in tumor
Frequency
Imputed tumor exosome expression profile

SFig. 9


EV COSMIC Mutational Fraction PFS Kaplan–Meier
Strata
Low COSMIC fraction
High COSMIC fraction
1.00
0.75
0.50
0.25
0.00
Survival probability
p = 0.67
0
250
500
750
1000
1250
Progression–free survival (days)

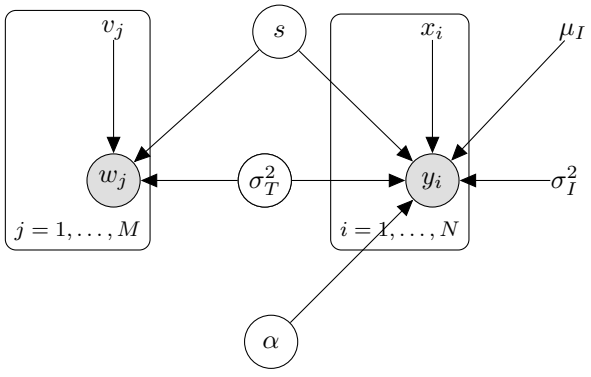
SFig. 10

Supplemental Note: Plasma-derived EV deconvolution model

In this supplemental note, we provide the mathematical background underlying our deconvolution model, as described in Figure 4a of "Plasma-derived extracellular vesicles analysis and deconvolution enables prediction and tracking of melanoma checkpoint blockade outcome". The rationale for the model is introduced in the main text. References in the Supplemental Note will also reference the main text. To summarize, we want to infer the contribution of the tumor-derived EV component and non-tumor derived (interchangeably referred to as "immune" and "non-tumor") component to the observed plasma-derived EV transcriptomic profile. In contrast to existing deconvolution models designed for bulk deconvolution (e.g., CIBERSORT[12]), our model explicitly models the changes in transcript abundance as a result of export/packaging from the transcript abundance in the tumor to the transcript abundance. All of the data shown both here and in the main-figures related to deconvolution utilized only non-platform-corrected discovery cohort microarray data, since the discovery cohort exoRNA-seq included only plasma-derived EV samples and thus were not suitable inputs for our deconvolution model (see Supplementary Figure 1b for more detailed information regarding our analysis methodology). We created two versions of the deconvolution probabilistic model. In **single-gene mode**, the probabilistic model is fully fitted using the No-U-Turn Sampler (NUTS) Hamiltonian Monte Carlo (HMC) algorithm[46] and full posterior estimates for all the relevant parameters are returned. This model is fitted using the probabilistic programming language Stan[45]. This mode is designed for in-depth analysis of a single gene (or few genes), or situations where inferred parameters (e.g., scaling parameter, patient inferred tumor-EV expression) of interest requires a full posterior estimate. Due to the time and resource intensive nature of the fitting process, it is impractical to perform full MCMC inference when we want to analyze the deconvolution profiles for tens of thousands of genes. Thus, we included a second mode, a **multi-gene mode**, in which we fit a simplified version of the single-gene model using Scipy's implementation sequential least squares programming (SLSQP) to return a point estimate of the mixing coefficient. The single-gene model not only returns posterior distribution mixing fraction, but also the full posterior distribution of the scaling coefficient, which allows per-patient imputation of the tumor-derived EV fractions (Supplemental Figure 9); however, the multi-gene model only returns a single maximum a posterori (MAP) estimate of the mixing fraction. We envision the usage of the single-gene model in cases when a specific gene needs to carefully dissected and more robust inference is required, whereas the multi-gene model can be used on large-scale transcriptomic datasets to infer population-wide mixing fractions.

1 Single-gene Bayesian probabilistic model

- N: number of patient derived tumor profiles and tumor EV profiles
- M: number of cell-line tumor and tumor EV profiles
- x_i: the ith patient's observed tumor expression for the current gene
- y_i: the ith patient's observed peripheral-blood derived expression for the current gene
- w_j: the jth cell-line's
- α: mixing fraction between tumor-component and immune-component
- σ_T^2: variance component
- μ_I: Immune component mean (fixed parameter)
- σ_I^2: Immune variance component (fixed parameter)



Prior specification

$$s \sim \mathcal{N}(0, 2)$$
$$\sigma_T^2 \sim \mathcal{IG}(1, 1)$$
$$\alpha \sim \mathcal{B}(2, 2)$$

Where $\mathcal{B}$ denotes the Beta distribution and $\mathcal{IG}$ denotes the Inverse-Gamma distribution.

Data likelihood

$$p(y_i|\alpha, \sigma_T^2; \mu_I, \sigma_I^2) = \underbrace{\alpha\mathcal{N}(y_i|x_i + s, \sigma_T^2)}_{\text{Tumor EV component}} + \underbrace{(1-\alpha)\mathcal{N}(y_i|\mu_I, \sigma_I^2)}_{\text{Immune/background EV component}} \tag{1}$$

$$p(w_j|s, v_j, \sigma_T^2) = \mathcal{N}(w_j|s + v_j, \sigma_T^2) \tag{2}$$

$$p(\mathbf{y}, \mathbf{w}|\mathbf{x}, \mathbf{v}, s, \alpha, \sigma_T^2; \mu_I, \sigma_I^2) = \prod_{i=1}^{N} \left[\alpha\mathcal{N}(y_i|x_i + s, \sigma_T^2) + (1-\alpha)\mathcal{N}(y_i|\mu_I, \sigma_I^2)\right] \prod_{j=1}^{M} \mathcal{N}(w_j|s + v_j, \sigma_T^2) \tag{3}$$

Full Posterior

$$p(\mathbf{y}, \mathbf{w}|\mathbf{x}, \mathbf{v}, s, \alpha, \sigma_T^2; \mu_I, \sigma_I^2) \propto p(\mathbf{y}, \mathbf{w}, \mathbf{x}, \mathbf{v}, s, \alpha, \sigma_T^2; \mu_I, \sigma_I^2) p(s, \sigma_T^2, \alpha) \tag{4}$$

$$p(\mathbf{y}, \mathbf{w}|\mathbf{x}, \mathbf{v}, s, \alpha, \sigma_T^2; \mu_I, \sigma_I^2) \propto \prod_{i=1}^{N} \left[\alpha\mathcal{N}(y_i|x_i + s, \sigma_T^2) + (1-\alpha)\mathcal{N}(y_i|\mu_I, \sigma_I^2)\right] \prod_{j=1}^{M} \mathcal{N}(w_j|s + v_j, \sigma_T^2) \tag{5}$$

$$p(s, \alpha, \sigma_T^2|\mathbf{x}, \mathbf{v}, \mathbf{y}, \mathbf{w}; \mu_I, \sigma_I^2) \propto \prod_{i=1}^{N} \left[\alpha\mathcal{N}(y_i|x_i + s, \sigma_T^2) + (1-\alpha)\mathcal{N}(y_i|\mu_I, \sigma_I^2)\right] \prod_{j=1}^{M} \mathcal{N}(w_j|s + v_j, \sigma_T^2) p(s, \alpha, \sigma_T^2) \tag{6}$$

2 Multi-gene probabilistic model

For the multi-gene model, we utilize a simplified version of the model from the single gene model to reduce the time and computational intensiveness of the model. The multi-gene model only returns the mixing fraction α for the gene-of-interest, and as a result, it runs significantly faster and is able to process transcriptome-wide datasets on the order of hours instead of days on a single computer, unlike the more computationally intensive single-gene model.

The multi-gene model differs from the single gene model in several key respects:

- The inference of the scaling coefficient s is not done jointly with the inference of α and other parameters of interest. Instead, it is precomputed for each gene. To compute s, we take the mean difference between v_j and w_j for each gene and use this as a fixed constant.
- Thus, the inference of α is simplified into a much simpler problem that only depends on the right plate in the graph model depicted in section 1. Since we are interested in only in α, we optimize the log posterior formula using Sequential Least Squares Programming (SLSQP) from the python SciPy library with the appropriate constraints for α. The $\mathcal{B}(2, 2)$ prior for α is retained.

3 Validation of our deconvolution model using CIBERSORTx

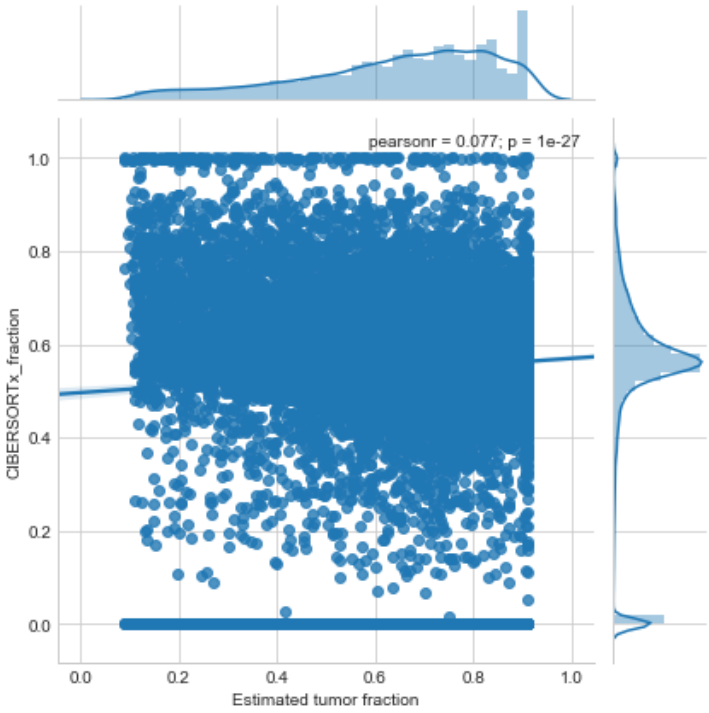


Figure 1: Scatterplot with line-of-best-fit and correlation between inferred CIBERSORTx and our deconvolution model inferred tumor fraction estimates on discovery cohort data

In order to provide evidence that our model is correctly partitioning genes into tumor and non-tumor components, we utilized CIBERSORTx[13]- a recently published bulk deconvolution program from Newman et al. that attempts to separate bulk transcriptomic profiles into component cell-type-specific profiles. In order to run the program, we utilized first inputted our discovery cohort (non-cross-platform corrected) microarray data matrix into the online CIBERSORTx web portal and utilized the melanoma reference profiles from Tirosh et al.'s *Science* 2016 single-cell dissection of metastatic melanoma that is provided by CIBERSORTx's default online profiles[12,13]. Since the algorithm generates estimates for all component cell-types instead of tumor-only profiles, we averaged the non-tumor cell-types in order to make it comparable to our non-tumor component estimation. A direct comparison of our values can be found in Supplemental Note Figure 1. We see that there's a slight but significant correlation between CIBERSORTx inferred-tumor fraction and tumor fraction inferred from our deconvolution model; however, it is clear from the CIBERSORTx density plot (y-axis) that the inferred tumor fraction is roughly normally distributed, an assumption that our model does not make (see density plot on x-axis). This continuous coding of tumor fractions hinders direct comparison of model predictions between CIBERSORTx and our model; thus, in order to better compare our cell-type predictions, we binarized model predictions for each gene as either tumor-derived or non-tumor derived using a cut-off of 0.5 as the threshold between tumor and non-tumor (same threshold used in the main manuscript). Using this cutoff, we can generate the confusion matrix found in Supplementary Note Table 1.

	CIBERSORTx tumor	CIBERSORTx non-tumor
EV deconvolution tumor	12984	2319
EV deconvolution non-tumor	3719	1244

Table 1: Confusion matrix between CIBERSORTx and our deconvolution model using 0.5 tumor fraction as a cutoff between tumor and non-tumor binary classification of genes

We can assess the concordance between binary predictions generated by CIBERSORTx and our model using values from the confusion matrix. This corresponded to the following binary classification statistics shown in Supplementary Note Table 2, using CIBERSORTx tumor predictions as "ground"

truth and our deconvolution model estimates as predictions. We see that overall accuracy (0.70), sensitivity (0.78), precision (0.85), and F1-score (0.81) all support the ability of our deconvolution model to properly classify CIBERSORTx predicted tumor-derived genes. However, the two models' predictions diverge significantly in terms of specificity (0.35) and negative predictive value (0.25), suggesting that the two models differ significantly in the overall prediction of non-tumor derived genes, with our model predicting a higher fraction of tumor-derived genes relative to CIBERSORTx. This is likely a result of the different distributional assumptions regarding tumor vs. non-tumor distributions between the two models (see Supplementary Note Figure 1). We reason that our model is likely to approximate reality more closely, based on known literature regarding significant increases in both overall and tumor-derived EV load in plasma during progression[10]. Furthermore, as mentioned in the main text, our model explicitly accounts for EV-specific characteristics - such as the differential EV packaging process - that bulk deconvolution techniques like CIBERSORTx does not account for. Additionally, our the underlying reference profiles is trained directly or inferred utilizing EV data, which is likely a far better approximation of the underlying mixture profiles in the context of cell-type deconvolution than bulk references. Though *in silico* independent validation via CIBERSORTx provides evidence for the validity of our deconvolution model predictions, particularly the prediction of tumor-derived genes, *in vivo* experimental evidence gathered via tumor vs. non-tumor derived EV selection/enrichment remains the gold standard to validate our model.

Metrics	**Value**
Accuracy	0.70
Sensitivity	0.78
Specificity	0.35
Precision	0.85
Negative Predictive Value	0.25
False Positive Rate	0.65
F1 Score	0.81

Table 2: Binary classification performance metrics generated from the confusion matrix in Table 1

.

4 Source Code

4.1 Stan code for the single-gene model:

```
data {
        int<lower = 0> N; // number of patient samples
        int<lower = 0> M; // number of cell line samples/controls
        real<lower = 0> L; //lower bound on truncated normal
        real<lower = 0> mu_i; // non-tumor observed mean
        real<lower = 0> sigmaSq_i; // non-tumor observed variance
        vector[N] y; // observed EV mixture
        vector[N] x; // observed tumor
        vector[M] ct; // observed cell line cancer
        vector[M] ce; // observed cell line EV
}

parameters {
        real s; // tumor to EV
        //real<lower = 0> sigmaSq_e; // variance of tumor EV
        real<lower = 0> sigmaSq_t; // variance of tumor
        real<lower = 0, upper=1> alpha; //mixing coefficient
        // real<lower = 0> sigma_common;
        // real<lower = 0> sigmaSq_immune;
}

transformed parameters{
        vector[N] z;
```

```
        //real<lower = 0> sigma_immune;
        real<lower = 0> sigma_t;
        real<lower = 0> sigma_i;
        real<lower = 0> sigma_e;
        for (n in 1:N)
        z[n] = s+x[n];
        // sigma_immune = sqrt(sigmaSq_immune);
        sigma_t = sqrt(sigmaSq_t);
        //sigma_e = sqrt(sigmaSq_e);
        sigma_i = sqrt(sigmaSq_i);
}

model {
        s ~ normal(0,1);
        sigma_t ~ inv_gamma(1,1);
        //sigma_e ~ inv_gamma(1,1);
        alpha ~ beta(2,2);
        for (m in 1:M)
        target += normal_lpdf(ce[m] | s+ct[m], sigma_t);
        for (n in 1:N)
        target += log_mix(alpha, normal_lpdf(y[n] | z[n], sigma_t),
        normal_lpdf(y[n] | mu_i, sigma_i));
}
```

Inference was performed using pystan's MCMC module with 10000 iterations and 4 chains, and all other parameters at default. The default pystan inference algorithm is the No-U-Turn Hamiltonian Monte Carlo (NUTS-HMC) sampler.

4.2 Python inference code for the simplified multi-gene model MAP inference procedure.

```
from scipy.optimize import minimize
from scipy.stats import beta, norm
import numpy as np
import math

def log_loss(theta,x,y,s,mu_i,sigmasq_i,N):
        alpha = theta[0]
        sigmasq = theta[1]
        log_loss = 0
        for i in range(N):
                log_loss += np.log(alpha*norm.pdf(y[i], x[i]+s, sigmasq)
                        + (1-alpha)*norm.pdf(y[i], mu_i,sigmasq_i))
        log_loss += beta.logpdf(alpha, 2,2)
        return -1*log_loss


def SimpleInferenceSLSQP(gene):
        N = tumor_shared_df.shape[1]
        M = cell_line_exp_mat.shape[1]
        mu_i = healthy_exp_mean.loc[gene]
        sigmasq_i = healthy_exp_var.loc[gene]
        y = exosome_shared_df.loc[gene,:]
        x = tumor_shared_df.loc[gene,:]
        ct = cell_line_exp_df.loc[gene,:]
        ce = exosome_exp_df.loc[gene,:]

        exosome_shared_mean = np.mean(y) #healthy
```

```
    cell_line_exosome = np.mean(ce)
    cell_line_tissue = np.mean(ct)
    cell_line_diff = cell_line_tissue - cell_line_exosome
    imputed_tumor = x + cell_line_diff
    imputed_immune = np.repeat(mu_i, N)

    design_matrix = np.vstack([imputed_tumor, imputed_immune]).T

    alpha0 = [0.5, 1]
    xargs = (x, y, cell_line_diff,mu_i, sigmasq_i, N)
    bounds = [(0,1), (None, None)]
    SLSQP = minimize(log_loss_two, alpha0,args=xargs,
            method='SLSQP', bounds=bounds, tol = 1e-9)
    return SLSQP.x[0]
```

MAP inference was performed using the SLSQP package from the Python scipy.optimize library using default parameters. Note that the MAP estimates for α can and often differ from the single-gene probabilistic model - this is primarily due to the (1) the simplification of the probabilistic model for computational ease and (2) less robust/advanced inference algorithm relative to the NUTS-HMC used in the single-gene model.