nature genetics

Article

https://doi.org/10.1038/s41588-024-01899-0

Five latent factors underlie response to immunotherapy

Received: 3 May 2024

Accepted: 7 August 2024

Published online: 12 September 2024

Check for updates

Joseph Usset[1,2,3], Axel Rosendahl Huber[1,10,11], Maria A. Andrianova[1,11], Eduard Batlle[1,4,5], Joan Carles[2,6], Edwin Cuppen[3,7], Elena Elez[2,6], Enriqueta Felip[2,6], Marina Gómez-Rey[2], Deborah Lo Giacco[2], Francisco Martinez-Jimenez[2,3], Eva Muñoz-Couselo[2,6], Lillian L. Siu[8], Josep Tabernero[2,5,6], Ana Vivancos[2], Ferran Muiños[1,10], Abel Gonzalez-Perez[1,5,9] ✉ & Nuria Lopez-Bigas[1,4,5,9] ✉

Only a subset of patients treated with immune checkpoint inhibitors (CPIs) respond to the treatment, and distinguishing responders from non-responders is a major challenge. Many proposed biomarkers of CPI response and survival probably represent alternative measurements of the same aspects of the tumor, its microenvironment or the host. Thus, we currently ignore how many truly independent biomarkers there are. With an unbiased analysis of genomics, transcriptomics and clinical data of a cohort of patients with metastatic tumors ($n = 479$), we discovered five orthogonal latent factors: tumor mutation burden, T cell effective infiltration, transforming growth factor-beta activity in the microenvironment, prior treatment and tumor proliferative potential. Their association with CPI response and survival was observed across all tumor types and validated across six independent cohorts ($n = 1{,}491$). These five latent factors constitute a frame of reference to organize current and future knowledge on biomarkers of CPI response and survival.

The development of CPIs has had a tremendous impact on cancer therapy[1]. However, the response of patients with cancer to these agents varies considerably[1–6], and important immune-related adverse events may appear as a result of treatment[7]. Consequently, intense research has been dedicated in recent years to identifying features that influence the response to CPIs[2,4,8–13], leading to the identification of potential biomarkers.

These studies have made it increasingly clear that the response to CPIs is mediated by several characteristics of the tumor, its microenvironment and the host[12], which we may regard as latent factors defining CPI response and survival across patients. However, it is likely that different biomarkers identified across a multitude of studies—often focused on one or a small group of features—represent different versions of the same underlying latent factor. For example, the expression of a number of genes and gene sets previously identified as biomarkers may represent the degree of infiltration of cytotoxic cells in the tumor[14–17]. Furthermore, given that separate research groups independently test different sets of potential biomarkers, there is no effective control of the potential false positives associated with multiple testing. As a result of these problems, it is not clear at present how many such

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [2]Vall d'Hebron Institute of Oncology (VHIO), Vall d'Hebron Barcelona Hospital Campus, Barcelona, Spain. [3]Hartwig Medical Foundation, Amsterdam, Netherlands. [4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [5]Centro de Investigación Biomédica en Red en Cáncer (CIBERONC), Instituto de Salud Carlos III, Madrid, Spain. [6]Medical Oncology Department, Vall d'Hebron University Hospital, Vall d'Hebron Barcelona Hospital Campus, Barcelona, Spain. [7]Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, Netherlands. [8]Division of Medical Oncology & Haematology, Princess Margaret Cancer Centre, University of Health Network, Department of Medicine, University of Toronto, Toronto, Ontario, Canada. [9]Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain. [10]Present address: Centro de Investigación Biomédica en Red en Cáncer (CIBERONC), Instituto de Salud Carlos III, Madrid, Spain. [11]These authors contributed equally: Axel Rosendahl Huber, Maria A. Andrianova. ✉e-mail: abel.gonzalez@irbbarcelona.org; nuria.lopez@irbbarcelona.org

https://doi.org/10.1038/s41588-024-01899-0

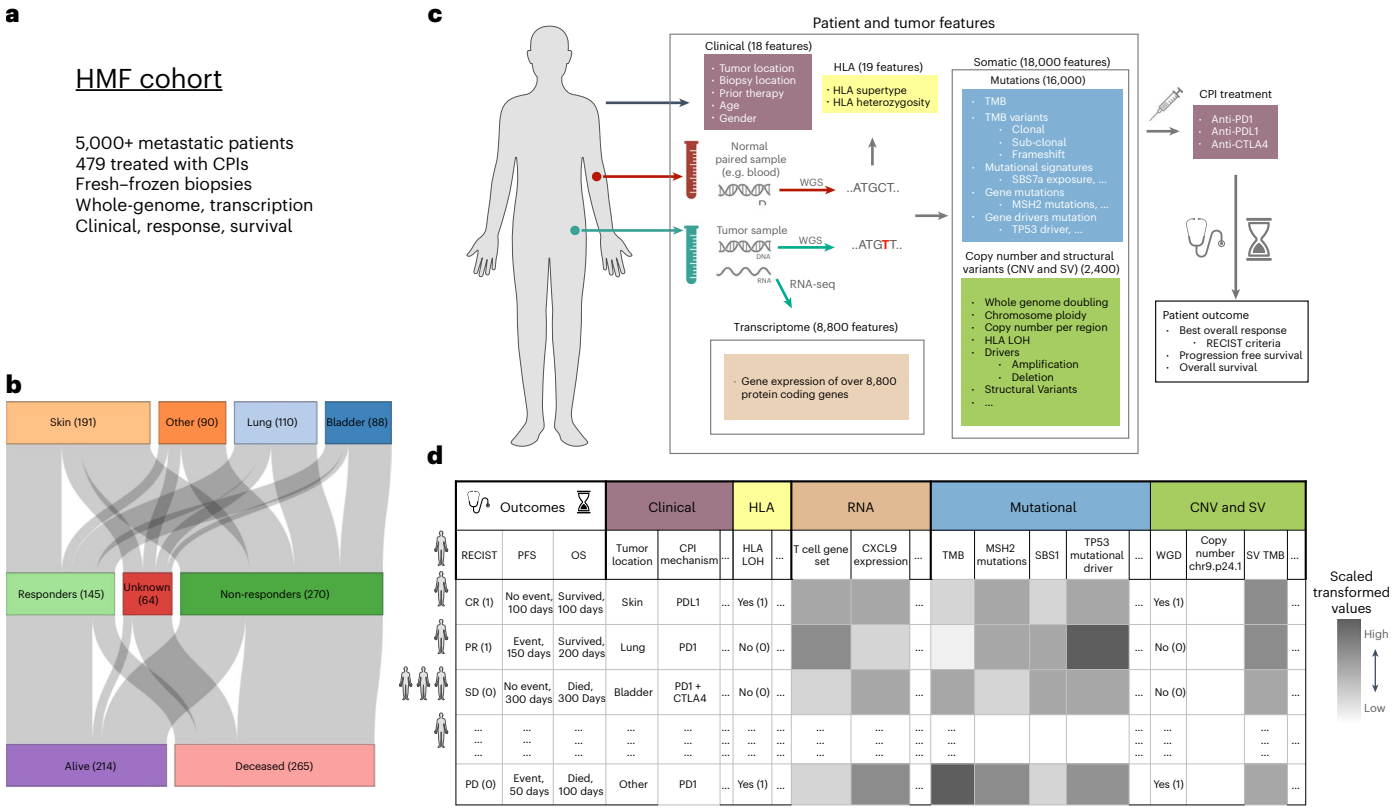


Fig. 1 | Extracting features from the HMF-CPI cohort. a–c, For 479 patients with metastatic cancer in the HMF-CPI database of different cancer types, we obtained 18 clinical features, 19 germline HLA allotype features, 18,382 somatic features (based on single base substitutions, indels, copy number variants and other structural variants affecting specific genomic elements or summaries thereof) and 8,817 transcriptomic features, corresponding to all expressed genes. **d**, Numeric feature values were rescaled and re-normalized (Methods), yielding a large table describing the cohort. LOH, loss of heterozygosity, RECIST, Response Evaluation Criteria, CNV, copy number variant; SV, structural variant; WGD, whole-genome doubling; CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease; OS, overall survival; PFS, progression-free survival.

independent latent factors of CPI response and survival there are, what aspects of the tumor, its microenvironment or the host they represent and whether they are relevant across different tumor types.

To answer these questions, we exploited a richly profiled and annotated cohort of patients with metastatic tumors (fresh–frozen biopsied) treated with CPIs (part of the cohort profiled by the Hartwig Medical Foundation (HMF)[18,19]; $n = 479$). Specifically, we aimed to identify features of the tumors, their microenvironment or the host that appeared to be significantly associated with CPI response and survival, both across the pan-cancer HMF-CPI cohort and all represented cancer types. To this end, we used an exhaustive—not biased by prior knowledge—analysis of thousands of molecular and clinical features to detect their association with CPI response or survival. We discovered that all significantly associated features collapse into one of five independent latent factors that are relevant across all tumor types represented in this cohort. They are the tumor mutation burden (TMB), effective T cell infiltration, whether the patients received any prior treatment, the activity of transforming growth factor-beta (TGF-β) in the tumor microenvironment and the proliferative potential of the tumor. We verified that at the current level of statistical power, there are no other latent factors of CPI response and survival common to all cancer types analyzed. We validated the association of these five latent factors with CPI response and survival in six independent cohorts ($n = 1,491$ patients) spanning six major cancer types; to our knowledge, the largest such validation effort.

Results

Extracting features from a metastatic cancer cohort

Within the HMF[18,19] cohort ($n = 5,288$), 479 patients with metastatic cancer who were part of the Center for Personalized Cancer Treatment study (https://www.cpct.nl/cpct-02) received anti-PD1/PDL1 or a combination of anti-PD1/PDL1 and anti-CTLA4 therapy. We refer to these patients as the HMF-CPI cohort. These include patients who had suffered from primary tumors of the skin (melanomas, $n = 191$), lung ($n = 110$), bladder ($n = 88$) and other cancer types (other; $n = 90$; Fig. 1a,b and Supplementary Table 1). Whole-genome somatic alterations of the metastatic tumors before CPI treatment were identified across all tumors in the HMF-CPI cohort and, for 396 of them, the whole transcriptome of the tumor was also sequenced. Rich clinical data, including treatments received before the diagnosis of their metastatic tumors, response to the CPI therapies, following Response Evaluation Criteria[20] ($n = 467$) and survival information ($n = 479$), were also available (Supplementary Table 1).

To carry out a systematic de novo discovery of biomarkers of CPI response, we computed 27,923 features (Fig. 1c,d and Supplementary Note 1). These included the mutational (single nucleotide variants + indels) status of 15,829 genes, the copy number status of 2,415 genomic regions, 64 aggregated somatic mutation features (for example, TMB, frameshift indel burden, activity of mutational signatures) and the occurrence of known driver structural variants as well as features summarizing the genomic instability (for example, total number of chromosomal fragments, ploidy, whole-genome doubling, and so forth). We also used the expression level of 8,817 genes and clinical characteristics of the patients such as sex, type of treatment received for the primary tumor and age at the time of diagnosis of the metastasis as features. Finally, human leukocyte antigen (HLA) features that can affect the immune response to the tumor were also included, such as their HLA haplotype and the number of somatically lost HLA alleles.

Five latent factors of CPI response and survival

To identify which of the more than 27,000 features computed per patient were significantly associated with CPI response, we performed univariate regressions (adjusted for the site of origin of the tumor, age of the patients, site of biopsy of the metastasis and tumor purity). After controlling for multiple testing[21], we identified several hundred features that appeared to be significantly associated with CPI response (Fig. 2a, Extended Data Fig. 1a–c and Supplementary Note 1).

Then, we asked how these significant features relate to each other and which underlying latent factors of CPI response they represent. To answer these questions, we clustered all significant features based on their pairwise correlations (Fig. 2b). Virtually all (Supplementary Note 1) could be unambiguously assigned to one of three clusters (R1, R2 or R3, encompassing somatic, clinical or transcriptomics features). This implies that only three latent factors associated with CPI response were detected from the more than 27,000 features analyzed.

To understand the nature of cluster R1, we first computed the mean of its integrating features. The single feature in the cluster with the highest correlation to the mean was the overall TMB, with other aggregated mutational features (for example, clonal TMB) also showing a high correlation. Specifically, the increase of TMB is associated with a higher probability of response and also increased survival (Extended Data Fig. 2a and Supplementary Fig. 1). Thus, we named this latent factor TMB, and although it could be measured using any of the features in the cluster, we selected the TMB to represent it. Importantly, the mutation rate of virtually all genes (some of which have been previously associated with CPI response[3,4,10,22–24]) also appear to be highly correlated with the TMB as part of this cluster of features, and indeed, some heavily mutated genes exhibit lower *P* values in the regression analysis than TMB (Supplementary Note 1). This implies that identifying the mutations of individual genes as biomarkers of CPI response independently of the TMB is a very challenging task.

Cluster R2 was integrated by two highly correlated clinical features: prior exposure to systemic therapy[11] and prior exposure to any therapy. These two features appear to be significantly negatively associated with CPI response and survival (Fig. 2b, Extended Data Fig. 2b, Supplementary Fig. 1 and Supplementary Note 1), perhaps owing to increased tumor aggressiveness or deteriorated patient condition. Thus, we named cluster R2 'prior treatment', and for the following analyses, we represented this cluster using exposure to any prior treatment.

Cluster R3 grouped the expression of 48 genes. We reasoned that the expression of gene sets representing biological functions in the tumor or its microenvironment could aid in the interpretation of this cluster. Thus, we computed the mean expression of 255 gene sets (225 representing all Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and cancer hallmarks obtained from the Molecular Signatures Database (MSigDB)[25,26] and 30 collected from the literature[12,27,28]; Supplementary Table 2, Fig. 2c,d and Methods). The mean expression of 13 gene sets was significantly associated with the response to CPI (Fig. 2d, Extended Data Figs. 2c and 3a and Supplementary Fig. 1), and all of them showed a high correlation (Pearson's coefficient of >0.8) with the mean expression of the genes in the cluster. They all represent some aspect of immune infiltration in the tumor; most specifically, T lymphocyte infiltration. Therefore, we named this third latent factor 'effective T cell infiltration' and represented it through the mean expression of all genes in the cluster. The increase in effective T cell infiltration appears significantly associated with a higher probability of CPI response and longer survival.

There is a clear positive relationship between the correlation of every feature to the representative of its corresponding latent factor (that is, TMB, prior treatment and the mean expression of genes in cluster R3) and the significance of its association with CPI response. The higher the correlation of a feature with the mean of its corresponding cluster, the more significant its association with CPI response (Fig. 2e and Extended Data Fig. 2a–c). These three latent factors are also significantly associated with overall survival and progression-free survival upon CPI treatment (Extended Data Fig. 2a–c).

We then asked whether any other latent factors of the tumor, its microenvironment or the host specifically influence the survival of patients, independently of the previous three latent factors associated with response (given that response is, by itself, a major determinant of survival). To answer this question, we focused on features that appeared to be significantly associated with overall survival, after controlling (in addition to the aforementioned covariates) for the three latent factors previously associated with response (Fig. 3a). Again, to discern how many latent factors were represented by these features, we clustered them based on their pairwise correlations (Fig. 3b).

One of the three clusters was clearly orthogonal to the other two, which exhibited a certain degree of inter-correlation. Thus, we named them clusters S1, S2.1 and S2.2, as they only represent two mutually orthogonal latent factors (Fig. 3b and Supplementary Note 1). To interpret them, we analyzed the correlation of their mean expression with that of 255 gene sets (Supplementary Table 2), as explained above. The mean of cluster S1 showed the highest correlation with a gene set named 'Proliferation potential' (Fig. 3c) and a high correlation with other gene sets representing cell cycle and overall cell proliferation (Extended Data Figs. 2d and 3b,c and Supplementary Fig. 1). We thus named it 'tumor proliferative potential' and represented it through the mean expression of all genes in the cluster.

The mean expression of the genes in cluster S2.1 showed the highest correlation with a gene set representing TGF-β in fibroblasts (Fig. 3d) and a high correlation with other gene sets related to this biological process (Extended Data Figs. 2e and 3b,d–f). As in other cases, we represented this latent factor through the mean expression of the genes in the cluster. Low values of this latent factor (TGF-β activity in the microenvironment) are associated with longer survival of patients upon CPI treatment even without correcting for the effect of the three response-associated latent factors (Fig. 3e).

We next asked whether the latent factors are specifically associated with CPI treatment or whether they represent general elements that influence response and survival upon any type of therapy. To answer these questions, we analyzed the data for 2,497 patients in the HMF cohort who received non-CPI therapies and found that the effects of TMB, effective T cell infiltration and TGF-β activity in the microenvironment are unique to CPI therapies, whereas prior treatment appears to affect the response to both CPI and non-CPI therapies, and the effect of tumor proliferative potential appears even larger across non-CPI than CPI-treated patients (Extended Data Fig. 4). Virtually all features that appear significantly associated with CPI response and/or survival are grouped in one of the five latent factors (Extended Data Fig. 5), indicating that no latent factor remains to be discovered in the HMF-CPI cohort.

In summary, five mutually orthogonal latent factors underlying CPI response and survival across the HMF-CPI cohort (Fig. 3g) emerged from this unbiased analysis. Supplementary Dataset 1 lists the results of the unbiased analysis of features in their association with CPI response and survival. Each of them can be represented through a number of features that are clustered by virtue of their pairwise correlations.

Validation of the five latent factors

Next, we asked whether the five latent factors, identified across HMF-CPI, were of comparable importance in the four groups of tumors with different tissues of origin represented in the cohort. To answer this question, we conducted, for each latent factor, multivariate regressions (adjusted for age, tumor purity, biopsy location and the remaining four latent factors) of their effect on CPI response and survival (Fig. 4). We found that the direction of the association of each factor (with response or survival) was maintained for all tumor types as in the pan-cancer analysis, with small differences in the effect size and the significance of their associations. One exception is the association of effective T cell

https://doi.org/10.1038/s41588-024-01899-0

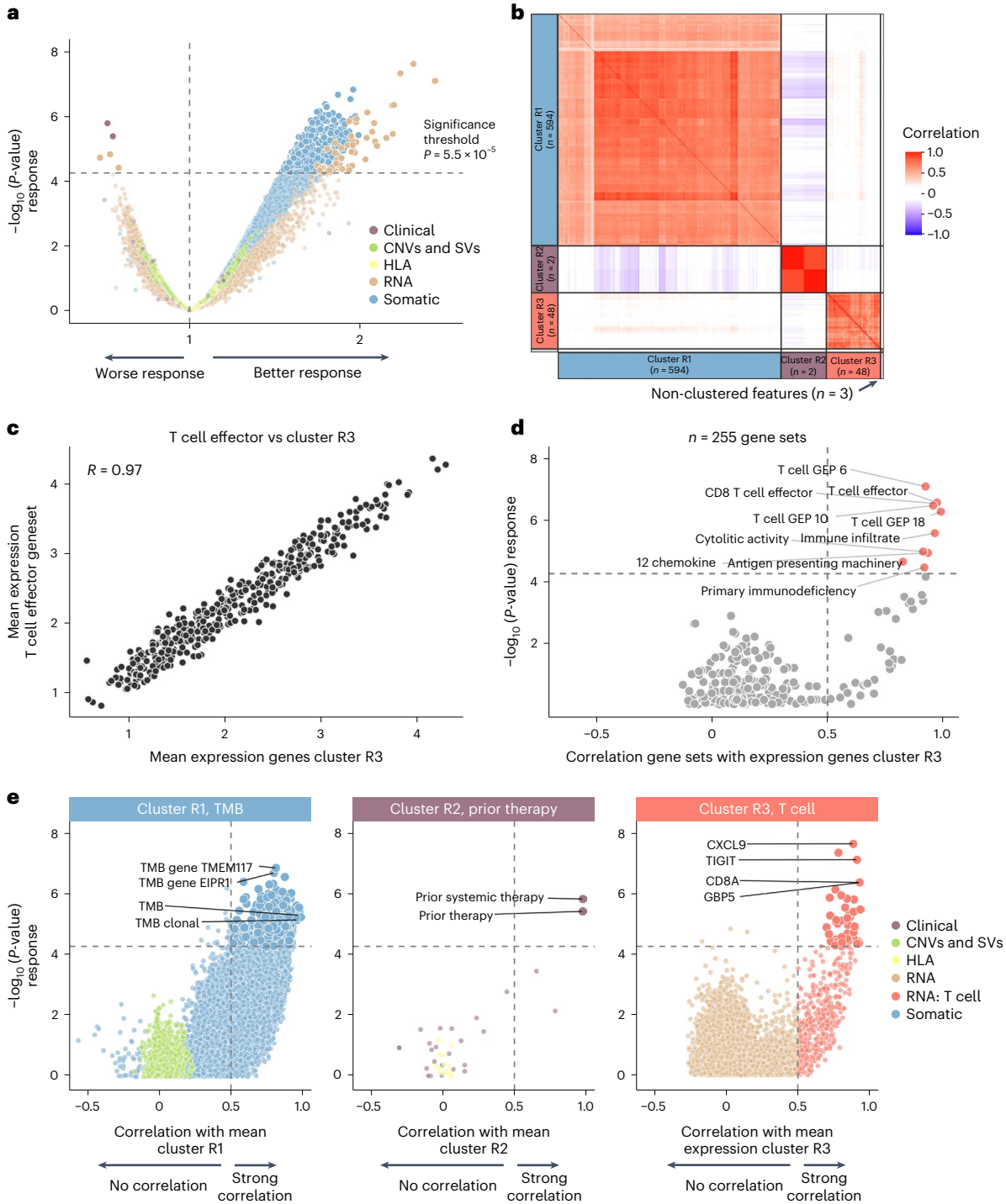


Fig. 2 | Three latent factors associated with CPI response. a, Logistic regression analysis (represented as a volcano plot) identified features significantly associated with CPI response (Methods and Supplementary Note 1). Dots with larger sizes represent significant features, and they are colored following the type of feature. *P* values shown in the plots were computed by logistic regressions. These are, by definition, two-sided. **b**, All significant features were selected and clustered based on their pairwise correlations. The colors denoting the clusters are inherited from the type of feature included in each of them according to the color legend in **a**. **c**, Mean expression values of cluster R3 (x-axis), and the 'T-cell effector' gene set (y-axis), across patients (dots). The Pearson's correlation coefficient is indicated. **d**, To discern the nature of cluster R3, the correlation of its mean to 255 gene sets collected from the literature was computed across patients (as illustrated in panel (**c**)). Dots represent gene sets. **e**, Relationship between the significance of the association with the response (y axis) and the correlation (x axis) to the mean of the cluster of the features in each cluster. *P* values shown in the plots were computed by logistic regressions. These are, by definition, two-sided. Dots in these three panels appear in darker color if they represent features significantly associated with CPI response and with a correlation coefficient above 0.5 with the mean of their respective cluster. In (**a**), (**d**) and (**e**), the horizontal dashed lines represent the significance threshold according to the Benjamini–Yekutieli correction.

https://doi.org/10.1038/s41588-024-01899-0

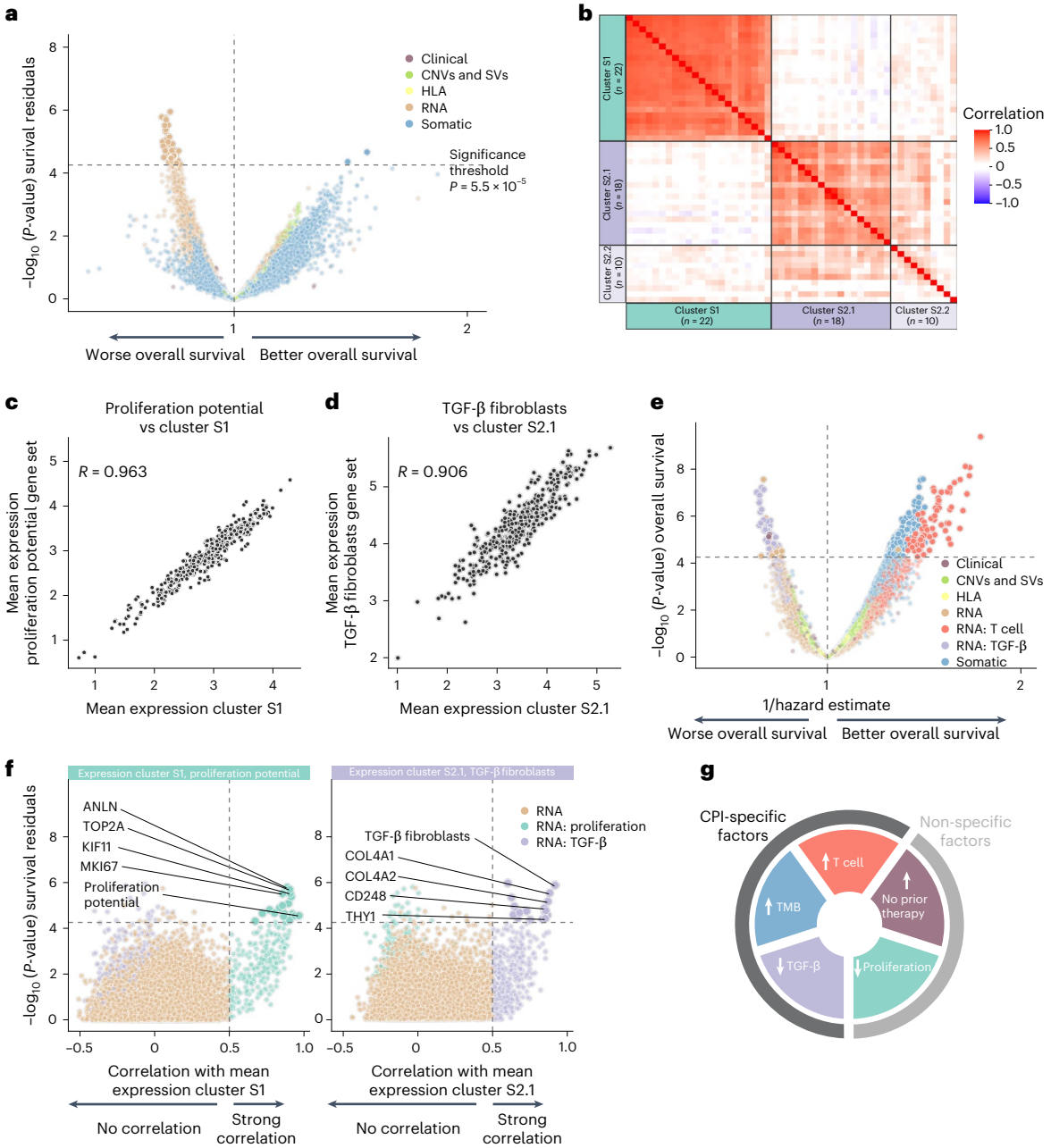


Fig. 3 | Two latent factors associated with survival. a, Features significantly associated with survival residuals, that is, after correction for the three latent factors associated with response. Larger dots represent significant features. Features with high correlation (Pearson coefficient of >0.5) with any of the three previously identified latent factors are removed. *P* values shown in the plots were computed by Cox regressions. These are, by definition, two-sided. **b**, Clusters of features based on their pairwise correlations. **c**,**d**, Cluster S1 and cluster S2.1 are highly correlated with gene sets representing the tumor proliferative potential and the activity of TGF-β in the tumor microenvironment, respectively. Dots represent the mean expression of two gene sets (y axis) and the mean expression of the genes in clusters S1 and S2.1 (x axis) across patients. Pearson's correlation coefficients are indicated. **e**, Features significantly associated with overall survival. Larger dots represent significant features. **f**, Significance of the association with the response (*y* axis) and the correlation (*x* axis) to the mean of the cluster of the features in each cluster. *P* values shown in the plots were computed by Cox regressions. These are, by definition, two-sided. **g**, Depiction of the five latent factors associated with CPI response and survival. Upwards arrows, positive association with response and/or survival; downwards arrows, negative associations. In (**a**), (**e**) and (**f**), the horizontal dashed lines represent the significance threshold according to the Benjamini–Yekutieli correction.

infiltration with response in patients with lung tumors, which was not significant (although the significance in the association with survival is maintained). The other is prior treatment, which does not exhibit a significant association with response across bladder tumors. Very similar results were obtained in cancer type-wise univariate regressions (Extended Data Fig. 6a). In summary, we find that the latent factors, with few exceptions, appear to underlie CPI response and survival across all tumor types represented in the HMF-CPI cohort (Supplementary Note 1).

We next asked whether the five latent factors are validated in independent cohorts of the same and other tumor types representing the wide diversity of approaches of sample processing and tumor profiling used in the clinic. To this end, we collected data from the literature for five independent cohorts or metacohorts (INSPIRE[29], Lyon[30], MARIATHASAN[27], PARKER ICI[31], RAVI[32]) and obtained the data from another cohort of patients treated at the Vall d'Hebron Institute of Oncology (VHIO). These validation cohorts comprised 1,491 patients with primary

https://doi.org/10.1038/s41588-024-01899-0

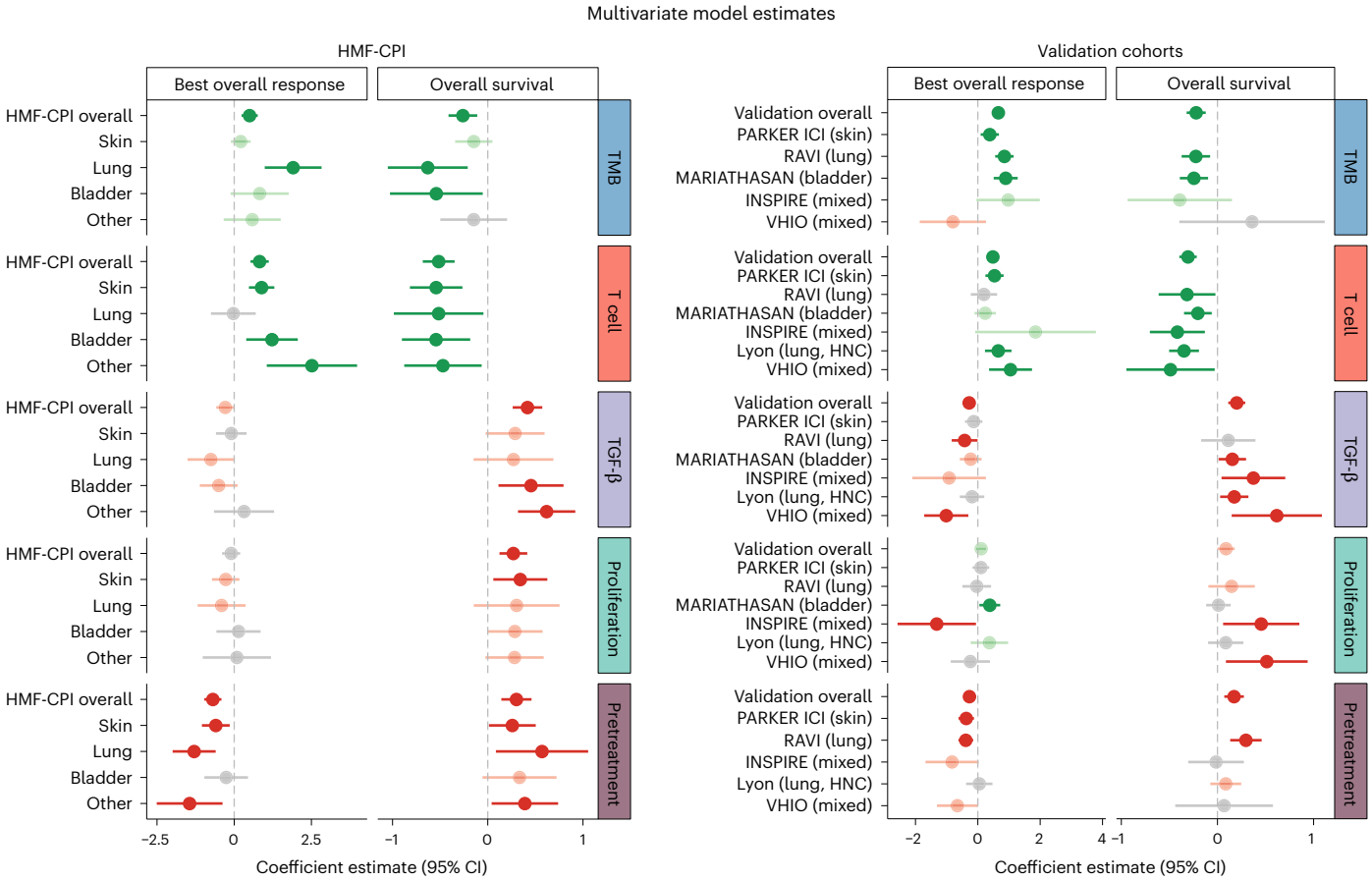


Fig. 4 | Validation of the latent factors across independent cohorts. Forest plots illustrating the association of latent factors across groups of tumors with different origins in the HMF-CPI cohort (left) and across six independent cohorts (right) with CPI response and overall survival. The value of each latent factor was computed as the mean of the cluster of features obtained in the HMF-CPI cohort across each validation cohort, except in the VHIO cohort, where the transcriptomics latent factors were estimated from alternative sets of genes (Methods). In the forest plots, the dots represent the strength (coefficients estimated through multivariate logistic or Cox regression) of the association between the latent factor and response or survival across cohorts. The horizontal bars denote the 95% confidence intervals. Gray dots represent latent factors whose estimates are within one standard error at either side of 0, dots with a light color (green or red) represent non-significant associations with coefficient estimates above (or below) one standard error of 0 and dark-colored dots represent significant associations. Green dots represent positive associations with improved outcomes (higher response odds or lower hazard ratio), while red dots represent negative associations (lower response or higher hazard ratio). Mixed denotes cohorts integrated by patients with multiple tumor types.

or metastatic tumors of different organs (Supplementary Table 1 and Supplementary Note 1). For 339 of these patients, we obtained sufficient information to compute the five latent factors, while for the remaining 1,152, we could compute only four or three latent factors. Using the available clinical information, we were able to evaluate the association of the latent factors with CPI response for 1,294 patients across all cohorts, while the association with overall survival could be computed for 1,165 patients across five cohorts. Unlike in the case of the HMF-CPI cohort, most of these studies (except INSPIRE) started from formalin-fixed, paraffin-embedded samples. The approaches used to identify somatic mutations range from whole-exome tumor-normal paired sequencing to tumor-only sequencing of a panel of 432 genes. The expression of genes was measured by whole-transcriptome, targeted RNA sequencing (RNA-seq) or a panel of 170 genes using the nCounter (NanoString) platform.

In a multivariate analysis, pooling all external cohorts, the associations were consistent between each of the five latent factors and CPI response or survival, with all except tumor proliferative potential reaching significance (Fig. 4). In some individual cohorts, the association of a particular latent factor with CPI response or survival could not be verified, such as the TMB in the VHIO cohort. In this case, owing to the lack of a control sample to reliably call somatic mutations, the calculation of TMB is probably not reliable (Supplementary Note 1). Nevertheless, despite the differences in cohorts, profiling and sample collections, the associations observed in the HMF-CPI cohort for the five latent factors were, overall, reproduced across the validation cohorts. T cell effective infiltration was positively associated with CPI response across five validation cohorts (three significantly), TGF-β activity in the microenvironment was negatively associated with survival in the five cohorts in which it could be evaluated (four significantly) and tumor proliferative potential was negatively associated with survival in four out of five cohorts (two significantly; Fig. 4). The association with prior treatment was validated in all (two significantly) but one cohort (Fig. 4 and Supplementary Note 1). Genes closer to the mean of clusters R1 (T cell effective infiltration), S2.1 (TGF-β activity in the microenvironment) and S1 (tumor proliferative potential) in HMF-CPI also tend to correlate better with one another across the four validation cohorts with transcriptomics data (Methods; Extended Data Fig. 6b).

In summary, despite the wide differences in tumor sample processing and profiling, many of the associations between the five latent factors and CPI response or survival previously discovered in the HMF-CPI are also observed across six independent cohorts.

Multivariate models to predict CPI response and survival

We next asked how the effects of the five latent factors combine (through accumulation or interaction) to influence CPI response and survival.

https://doi.org/10.1038/s41588-024-01899-0

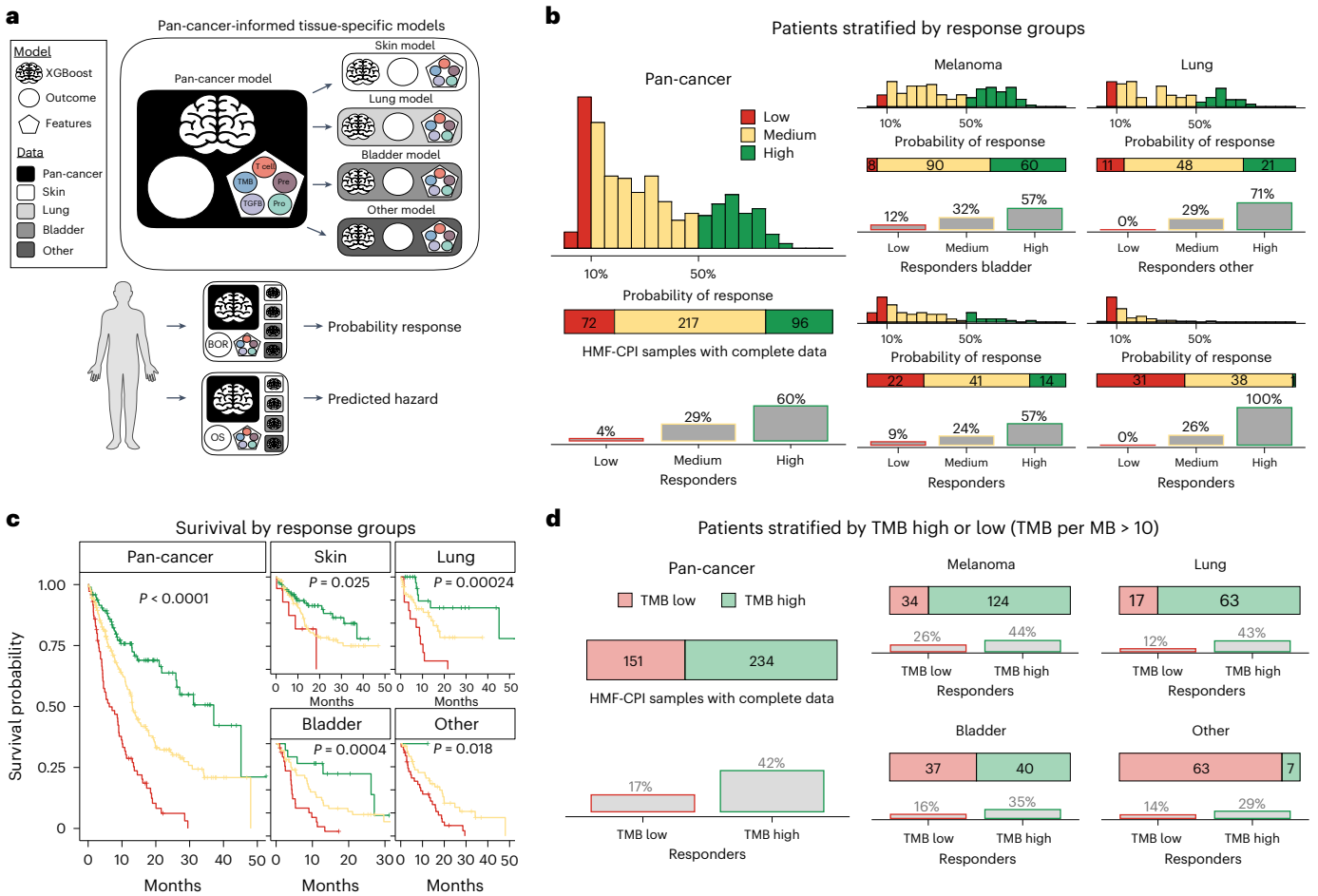


Fig. 5 | Multivariate models to predict patients' response and survival. a, The values of the representative biomarkers of the five latent factors across patients in the HMF-CPI cohort were used to train hybrid (pan-cancer-informed tumor type-specific) gradient-boosting models to predict CPI response and survival. The performance of the models was assessed through cross-validation (Methods and Supplementary Note 1). **b**, Stratifying patients based on model predictions. We separated the patients in the HMF-CPI cohort into three groups based on their predicted probability of response (histograms) and the three-segment bar below. We then calculated the fraction of responders within each group (bar plots below each histogram). **c**, Differences in overall survival between the three groups of patients are represented by Kaplan–Meier curves. The *P* value for each cohort (annotated in the plot) was calculated with a one-sided log-rank test. The line colors correspond to the three groups of patients defined in **a**. **d**, The TMB for each patient in the HMF-CPI cohort (with complete data for all five latent factors) was computed with a measure commonly used in the clinic: the number of mutations per genomic megabase. Tumors were classified as low-TMB or high-TMB based on a simple cutoff (10 mutations per megabase). The bars are colored according to the fraction of patients with high or low TMB in each of them. Interestingly, a number of patients with high-TMB tumors are predicted to have a low probability of response, whereas some patients with low-TMB tumors appear in the high probability of response group. The bottom bar plots present the percentage of patients in the low-TMB and high-TMB groups that showed clinical response to CPIs. OS, overall survival; BOR, best overall response according to RECIST; MB, megabase.

To that end, we trained multivariate machine-learning (tree-based gradient-boosting) models[33] to predict the response, overall survival or progression-free survival of patients in the HMF-CPI cohort. To exploit the higher statistical power provided by the full cohort and the specificity inherent in the response across cancer types, we first constructed pan-cancer models and then used them as the base to obtain hybrid models; that is, subjecting the pan-cancer models to added rounds of training on the data corresponding to each tumor type (Fig. 5a and Supplementary Note 1). The hybrid models trained on the five latent factors outperformed models trained solely on tumor type-specific data (Supplementary Fig. 2a) as well as equivalent models trained solely on values of TMB and PDL1 expression (Supplementary Fig. 2b) within the HMF-CPI cohort. Models trained on different representations of the five latent factors showed comparable performance, supporting the idea that the features of each cluster constitute alternative representations of the latent factors (Supplementary Fig. 3a,b). The variability in the influence of the five latent factors across different tumor types in the HMF-CPI cohort observed in the multivariate regression analysis described above is verified through a survey of their relative importance on the prediction cast by the multivariate machine-learning models of CPI response and overall survival (Methods; Extended Data Fig. 7a,b).

We then stratified 396 patients in the HMF-CPI cohort with all data types (jointly and separately by tumor type) into three groups of low (below 0.1), medium (between 0.1 and 0.5) and high (greater than 0.5) predicted probability of response to CPI. Only 2 (3%) of the 67 patients in the low-probability group actually responded to CPI treatment, compared with 61 out of 97 (63%) patients in the high probability of response group (Fig. 5b). This stratification also significantly separated patients in the HMF-CPI cohort based on their survival (Fig. 5c). Stratifying the patients based on a threshold of TMB used in the clinical practice (ten mutations per Mbp[34,35]) to separate high-TMB and low-TMB tumors is less optimal, with 17% of responders among patients with low-TMB tumors and 42% among those with high-TMB tumors (Fig. 5d). Interestingly, patients in the group with low probability of response exhibit a range of predicted hazards according to the overall survival pan-cancer model (Extended Data Fig. 8a–g and Supplementary Note 1). Across the VHIO

and INSPIRE cohorts, the stratification based on the predicted probability of response produced a perfect identification of patients with a low probability of response, while results were less accurate across the RAVI cohort (Extended Data Fig. 9a and Supplementary Fig. 4).

When applied to patients in the HMF cohort who did not receive CPIs, the multivariate models of response identified an important fraction of the patients with skin (35%), bladder (42%) and lung (16%) tumors with a high likelihood of response to the treatment (Extended Data Fig. 10). Interestingly, patients suffering from other metastatic malignancies (some not usually considered as candidates for CPI) were also identified as potentially good responders. For example, 18 (4%) patients with breast cancer, 10 (3%) patients with colorectal cancer, 10 (19%) patients with kidney tumors and 5 (15%) patients with liver cancer exhibited high probability of response to CPI.

In summary, we illustrate that multivariate models combining the five latent factors produce a more accurate stratification of patients according to their predicted probability of response than the TMB alone.

Discussion

In this work, we followed a completely unbiased approach to discover genomics, transcriptomics and clinical features associated with CPI response and survival across patients with cancer. We aimed to answer how many and which aspects of the tumor, its microenvironment and the host influence the response to CPIs across patients (that is, latent factors), in an effort to provide a framework of reference to existing copious reports of biomarkers. First, through univariate logistic and Cox regressions, we identified a few hundred features that are significantly associated with CPI response and/or survival. Five latent factors emerge when these significant features are clustered based on their pairwise correlation. These represent mutually independent aspects of the tumor, its microenvironment and the host that influence the response of a patient to CPI and their hazard after the treatment.

Although the fact that some genomics and transcriptomics features may represent the same aspect of tumors, their microenvironment or the host had been reported before, here we show that an array of different, highly intercorrelated features (for example, expression of genes related to T cell function) actually represent different measurements of the same latent factor. This is particularly striking in the case of TMB: the mutation rate of hundreds of genes (including cancer driver genes) appears to be highly correlated with TMB, suggesting that the association of mutations in a given gene with CPI response rather than an independent biomarker is just an alternative proxy measurement of TMB. Although the mutation status of some genes may still be bona fide biomarkers of CPI response, independently of TMB, any analysis to identify them should account for the confounding factor of their correlation with TMB. We also demonstrate that the associations of the latent factors with CPI response and survival are observed across different tumor types, and we validated them across six independent cohorts of patients. Most of the associations discovered in the HMF-CPI cohort were corroborated across these independent cohorts, despite differences in sample collection and processing procedures and profiling methods between these cohorts and the HMF-CPI. This indicates that these latent factors are mostly universally associated with CPI response and survival. They may thus be potentially used in the future within the clinical practice, despite the heterogeneity of sample processing and tumor profiling approaches used. The variability observed across tumors of different origins in more than one cohort (for example, the smaller association of effective T cell infiltration with the response of lung tumors) may point to findings that could be pursued further. To our knowledge, this constitutes the most extensive exploration, to date, of the biomarkers of CPI response and survival across cohorts with tumors from different organs.

Importantly, no features other than these five latent factors were significantly associated with CPI response or survival. That is, virtually all significant features cluster within one of them. However, some relevant features may still lay below the statistical power of the HMF-CPI cohort or appear significantly associated with CPI response or survival in only one tumor type. This is particularly important for features that may be relevant for a fraction of patients, such as mechanisms of immune escape (for example, *B2M* deletion, an event that we observe as significant across melanomas but not other tumors in the HMF-CPI cohort; see Supplementary Note 1)[8,22,24,36]. Other examples of such features relevant for specific groups of patients may include common polymorphisms that affect the immune response[37] and the heterozygosity at HLA loci[38]. Particularities of the tumor types not represented in the HMF-CPI cohort are, of course, also absent from our current catalog of proxy biomarkers. These will be discovered when larger CPI cohorts are analyzed; it is not inconceivable that even more latent factors will become apparent then. Our discovery of latent factors is also limited by the profiling technologies used, which rely on deconvolution of the immune infiltrate based on bulk RNA-seq data. More detailed studies of this infiltrate—based on fine mapping of immune populations and their interactions with other cells in the microenvironment—will contribute in the future to refine the landscape of biomarkers of CPI response and survival.

A test of the application of multivariate models combining the five latent factors produced a stratification of the patients in the HMF-CPI cohort based on their predicted response probabilities that discriminates better between responders and non-responders than the threshold of TMB frequently used in the clinic. This could be regarded as a proof-of-principle for application of the five latent factors to clinical practice. Being able to identify patients with a very low probability of response would be relevant to spare them the potential side effects of the therapy[7]. Additionally, it may aid in reducing the financial burden on healthcare providers[39]. There is also the possibility—illustrated through the analyses described above—to use such multivariate models to identify patients with tumors that are not usually considered suitable candidates for CPI who have a high probability to respond. In the future, when these types of models can be used in support of clinical decision-making, this could potentially contribute to expanding the therapeutic options for patients suffering from these malignancies.

In summary, we envision that the results of this work can provide a frame of reference to the research of biomarkers of CPI response and survival, resulting in the classification of all identified significant features falling into one of these five latent factors, or a completely independent one.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01899-0.

References

1. Korman, A. J., Garrett-Thomson, S. C. & Lonberg, N. The foundations of immune checkpoint blockade and the ipilimumab approval decennial. *Nat. Rev. Drug Discov.* **21**, 509–528 (2022).
2. Kalbasi, A. & Ribas, A. Tumour-intrinsic resistance to immune checkpoint blockade. *Nat. Rev. Immunol.* **20**, 25–39 (2020).
3. Havel, J. J., Chowell, D. & Chan, T. A. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat. Rev. Cancer* **19**, 133–150 (2019).
4. Keenan, T. E., Burke, K. P. & Van Allen, E. M. Genomic correlates of response to immune checkpoint blockade. *Nat. Med.* **25**, 389–402 (2019).
5. Sanmamed, M. F., Berraondo, P., Rodriguez-Ruiz, M. E. & Melero, I. Charting roadmaps towards novel and safe synergistic immunotherapy combinations. *Nat. Cancer* **3**, 665–680 (2022).

6. Kubli, S. P., Berger, T., Araujo, D. V., Siu, L. L. & Mak, T. W. Beyond immune checkpoint blockade: emerging immunological strategies. *Nat. Rev. Drug Discov.* **20**, 899–919 (2021).
7. Martins, F. et al. Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance. *Nat. Rev. Clin. Oncol.* **16**, 563–580 (2019).
8. Bruni, D., Angell, H. K. & Galon, J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nat. Rev. Cancer* **20**, 662–680 (2020).
9. Chan, T. A. et al. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann. Oncol.* **30**, 44–56 (2019).
10. Saigi, M., Alburquerque-Bejar, J. J. & Sanchez-Cespedes, M. Determinants of immunological evasion and immunocheckpoint inhibition response in non-small cell lung cancer: the genetic front. *Oncogene* **38**, 5921–5932 (2019).
11. Chowell, D. et al. Improved prediction of immune checkpoint blockade efficacy across multiple cancer types. *Nat. Biotechnol.* **40**, 499–506 (2022).
12. Litchfield, K. et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596–614.e14 (2021).
13. Chen, X. et al. $CD8^+$ T effector and immune checkpoint signatures predict prognosis and responsiveness to immunotherapy in bladder cancer. *Oncogene* **40**, 6223–6234 (2021).
14. Charoentong, P. et al. Pan-cancer immunogenomic analyses reveal genotype–immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* **18**, 248–262 (2017).
15. Bindea, G. et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782–795 (2013).
16. Angelova, M. et al. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* **16**, 64 (2015).
17. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
18. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
19. Martínez-Jiménez, F. et al. Pan-cancer whole genome comparison of primary and metastatic solid tumors. *Nature* **618**, 333–341 (2023).
20. Schwartz, L. H. et al. RECIST 1.1—update and clarification: from the RECIST committee. *Eur. J. Cancer* **62**, 132–137 (2016).
21. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
22. Martínez-Jiménez, F. et al. Genetic immune escape landscape in primary and metastatic cancer. *Nat. Genet.* **55**, 820–831 (2023).
23. Ma, X. et al. Functional landscapes of POLE and POLD1 mutations in checkpoint blockade-dependent antitumor immunity. *Nat. Genet.* **54**, 996–1012 (2022).
24. Tamborero, D. et al. A pan-cancer landscape of interactions between solid tumors and infiltrating immune cell populations. *Clin. Cancer Res.* **24**, 3717–3728 (2018).
25. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci.* **102**, 15545–15550 (2005).
26. Liberzon, A. A description of the molecular signatures database (MSigDB) web site. *Methods Mol. Biol.* **1150**, 153–160 (2014).
27. Mariathasan, S. et al. TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* **554**, 544–548 (2018).
28. Tauriello, D. V. F. et al. TGFβ drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature* **554**, 538–543 (2018).
29. Cindy Yang, S. Y. et al. Pan-cancer analysis of longitudinal metastatic tumors reveals genomic alterations and immune landscape dynamics associated with pembrolizumab sensitivity. *Nat. Commun.* **12**, 5137 (2021).
30. Foy, J.-P. et al. Datasets for gene expression profiles of head and neck squamous cell carcinoma and lung cancer treated or not by PD1/PD-L1 inhibitors. *Data Brief.* **44**, 108556 (2022).
31. Campbell, K. M. et al. Prior anti-CTLA-4 therapy impacts molecular characteristics associated with anti-PD-1 response in advanced melanoma. *Cancer Cell* **41**, 791–806.e4 (2023).
32. Ravi, A. et al. Genomic and transcriptomic analysis of checkpoint blockade response in advanced non-small cell lung cancer. *Nat. Genet.* **55**, 807–819 (2023).
33. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
34. Addeo, A., Friedlaender, A., Banna, G. L. & Weiss, G. J. TMB or not TMB as a biomarker: that is the question. *Crit. Rev. Oncol. Hematol.* **163**, 103374 (2021).
35. Addala, V. et al. Computational immunogenomic approaches to predict response to cancer immunotherapies. *Nat. Rev. Clin. Oncol.* **21**, 28–46 (2024).
36. McGranahan, N. et al. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* **171**, 1259–1271.e11 (2017).
37. Sears, T. et al. Integrated germline and somatic features reveal divergent immune pathways driving ICB response. Preprint at https://doi.org/10.1101/2024.01.12.575430 (2024).
38. Krishna, C. et al. An immunogenetic basis for lung cancer risk. *Science* **383**, eadi3808 (2024).
39. Verma, V. et al. A systematic review of the cost and cost-effectiveness studies of immune checkpoint inhibitors. *J. Immunother. Cancer* **6**, 128 (2018).

Methods

Discovery cohort

Whole-genome somatic mutations, copy number and other structural variants across metastatic tumors from 4,484 patients in the HMF cohort were obtained from the HMF database[18,19] (version DR-263_update1). Of these, 479 subsequently received CPI therapy (HMF-CPI cohort), for which all somatic variation information was available. Whole transcriptome from RNA-seq was available from the same source for a subset of 396 patients in the HMF-CPI cohort. Several features computed by the HMF pipeline from this data for each tumor (for example, number of neoepitopes, activity of mutational signatures) as well as the patients' germline features (such as HLA allotypes) were obtained as part of the dataset. It also included all relevant clinical information regarding exposure to treatment for their primary malignancy, the subsequent treatment regimen for the metastasis and longitudinal measurements of the outcome (Supplementary Note 1 and Supplementary Table 1). The ethical approval to use this data in research has been obtained by the HMF.

Validation cohorts

INSPIRE. Whole-exome somatic mutations, the whole-transcriptome RNA-seq gene expression of tumors and all clinical data pertaining to prior treatment as well as outcome upon treatment with pembrolizumab within the INSPIRE basket trial (NCT02644369)[29] were obtained for 64 patients from https://github.com/pughlab/inspire-genomics.

Lyon. Targeted RNA-seq (2,559 genes) and clinical data from 315 patients treated at several hospitals in Lyon and Paris[30] were obtained from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE159067, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161537, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162519 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162520.

MARIATHASAN. Whole-exome somatic mutations, the whole-transcriptome RNA-seq gene expression of tumors and all clinical data in a previously published study[27] of 348 patients were obtained from http://research-pub.gene.com/IMvigor210CoreBiologies.

PARKER ICI. Whole-exome somatic mutations, the whole-transcriptome RNA-seq gene expression of tumors and all clinical data of several cohorts of tumors (including those within clinical trials CheckMate 038 and CheckMate 067 and two cohorts published within other studies) compiled in a previous publication[31] totaling 315 patients were obtained from https://github.com/ParkerICI/MORRISON-1-public.

RAVI. Whole-exome somatic mutations, the whole-transcriptome RNA-seq gene expression of tumors and all clinical data of the SU2C-MARK cohort from a previous publication[32] comprising 352 patients were obtained from https://zenodo.org/records/7625517.

VHIO. The estimated TMB and the expression (via NanoString) of 170 genes across the tumors of 74 patients with cancer profiled and treated at the VHIO Hospital in Barcelona were obtained directly from the Cancer Genomics Group. Clinical data of these patients were provided by attending oncologists at VHIO.

Details of all cohorts appear in Supplementary Table 1 and Supplementary Note 1.

Ethical approval to use the data of the first five validation cohorts in research was obtained by the original institutions, who obtained written informed consent from patients and made the data available through scientific publications. The Vall d'Hebron University Hospital Ethics Committee of Clinical Research approved the study according to local guidelines and regulations, and written consent was obtained from all the patients included in this study.

Feature extraction for systematic analysis

Somatic features (18,382) representing single nucleotide variants, indels, copy number variants and other structural variants were extracted from files directly downloaded from the HMF database. These included the list of variants as well as summary statistics, such as TMB, burden of structural variants, predicted neoantigen burden, and so on. Although some features were obtained directly from the files, others were derived. The level of expression (transcripts per million) of 8,817 genes (measured through whole-transcriptome RNA-seq) was also obtained from files downloaded from the HMF database after pre-processing (see below). Other RNA-seq features were derived from these values, mainly through the summarization of the expression of genesets in separate features[12,27,40,41], or by Cibersort[17] derivation of immune cell populations from gene expression data (Supplementary Table 2). The HLA allotypes of HMF-CPI patients were directly obtained from files downloaded from the HMF database, while somatic HLA loss of heterozygosity in the tumors was estimated using the LILAC tool[22]. Clinical information regarding courses of treatment before the biopsy of the metastasis and the subsequent outcome of CPI treatment was also obtained from files downloaded from the HMF database. Again, part of that information was directly converted into features, while other features, such as the time elapsed between the end of the prior treatment and the biopsy of the metastasis, were derived from these data. Some of these clinical features were converted into outcomes of the analysis (best overall response to CPI, overall survival and progression-free survival upon CPI), while others were maintained as potentially predictive features. A detailed description of the strategy followed for the extraction of features in the HMF-CPI cohorts appears in Supplementary Note 1.

Pre-processing

All outcomes and features were computed across all samples in the HMF database. Finally, the data was joined based on the sample identifiers to produce a data frame ready for statistical analyses. Before systematic analyses, several pre-processing steps were performed. First, to reduce multiple testing, we applied filters to remove features with little chance of providing meaningful associations. For somatic mutations by gene, only genes with at least one mutation per 20 samples were kept for the analyses. For RNA expression, only coding genes with a mean and standard deviation of adjusted transcript per million values greater than 0.5 were considered. For the driver features, only driver genes mutated in at least one in 30 samples were included. Similarly, for mutational signatures, only signatures with exposure greater than 0.02 for at least one in 20 samples were included. Second, all features were standardized to have a mean of zero and a standard deviation of one across the CPI samples. This standardization allowed for fair comparisons of estimated effect sizes. Outside of the primary tissue location, all features in the analyses were numeric or ordinal.

Systematic analyses

Each feature was tested individually for the strength of association to best overall response, progression-free survival and overall survival. Generalized linear models and their native maximum-likelihood-based tools were used for all estimation, standard error calculation and hypothesis testing.

The best overall response was modeled with logistic regression, in which we assumed that the probability of response followed a Bernoulli distribution with mean p. For each feature X, we accounted for primary tissue, biopsy location, tumor purity and age as model covariates. Formally, let I_j represent the covariate indicator functions for primary tissue (skin, lung, bladder, other tissue), let I_k represent the indicator function for biopsy location (lung, liver, lymph node, primary, skin, other tissue), let X_{age} represent patient age and X_{purity} represent the tumor purity. The full and reduced models were fit as follows.

Full model:

$$Logit(p) = \beta_0 + X\beta_X + \sum_{j \in tissue} I_j\beta_j + \sum_{k \in biopsy} I_k\beta_k + X_{purity}\beta_{purity} + X_{age}\beta_{age}$$

Reduced model:

$$Logit(p) = \beta_0 + \sum_{j \in tissue} I_j\beta_j + \sum_{k \in biopsy} I_k\beta_k + X_{purity}\beta_{purity} + X_{age}\beta_{age}$$

The models were fitted with the base R *glm* function.

Progression-free and overall survival outcomes were modeled for each feature with Cox proportional hazards models. The hazard rates, denoted $h(t)$, were modeled as follows.

Full model:

$$\log(h(t)) = X\beta_X + \sum_{j \in tissue} I_j\beta_j + \sum_{k \in biopsy} I_k\beta_k + X_{purity}\beta_{purity} + X_{age}\beta_{age}$$

Reduced model:

$$\log(h(t)) = \sum_{j \in tissue} I_j\beta_j + \sum_{k \in biopsy} I_k\beta_k + X_{purity}\beta_{purity} + X_{age}\beta_{age}$$

Survival models were fitted using the *coxph* function from the survival package in R.

For all analyses, *P* values were computed based on the likelihood ratio tests comparing the full and reduced models.

For the main analyses of best overall response, progression-free survival and overall survival, the covariates included were the indicators for primary tissue, the age of patients, the site of biopsy of the metastasis and the tumor purity. For the overall survival residuals analysis, the covariates additionally included the representative biomarkers of the three latent factors explaining response: TMB, T cell effective infiltration and pretreatment. For all model–feature–covariate combinations, the *P* values were calculated from the likelihood ratio test comparing the full model to the reduced model (with the feature of interest removed). Effect sizes (log odds ratio for the logistic regression and hazard ratios for the Cox regression) and standard errors were estimated with maximum likelihood from the full models. All effect sizes, standard errors and corresponding *P* values were stored for further analysis. Given the large dependency in tests, we used the Benjamini–Yekutieli multiple testing threshold to control the false discovery rate[21]. Several exhaustive analyses were run, with different sets of covariates each producing similar conclusions. Full documentation of all exhaustive analyses can be found in Supplementary Note 1.

Identification of latent factors

Latent factors were defined as the independent biological mechanisms underlying the features most predictive of CPI response and survival. To label latent factors, we first focused on features passing the Benjamini–Yekutieli multiple test significance threshold. From these significant features, we computed their pairwise Pearson correlations and identified clusters using hierarchical clustering (hclust() in R with the Ward.D2 algorithm). The optimal number of clusters was defined using the R package 'factoextra' function fviz_nbclust, using silhouette', 'wss' and 'gap_stat' options.

To label transcriptomics clusters, we computed the expression of 255 gene sets reported in the literature. The gene sets (Supplementary Table 2) were collected by downloading the Hallmark and KEGG annotated gene sets from MSigDB[25,26] (version 2023.1.Hs). These genesets were further complemented by others, obtained from previous publications[27], including a paper describing the CPI-1000 analyses[12]. Genesets with a Pearson correlation of >0.8 with the mean of a specific cluster and passing the multiple test *P* value threshold of association with CPI response or survival were considered cluster-specific and thus used to discern the nature of the cluster.

Stability of transcriptomics latent factors

For each gene in each transcriptomics cluster, we calculated the silhouette score[42] using the silhouette() function from the package 'cluster' in R. This score reflects how close the particular data point is to the cluster of assignment and how far it is from other clusters. The matrix of distances between genes used to calculate silhouette scores was obtained from the correlation matrix

$$d = \sqrt{1 - |cor|},$$

where *cor* is the correlation matrix of gene expression levels.

We then calculated silhouette scores for each gene in every other cohort with available expression data (INSPIRE, MARIATHASAN, PARKER ICI, RAVI) using gene expression levels from the corresponding dataset but keeping the initial clustering obtained in the HMF cohort. Aggregation of silhouette scores across datasets was performed using the aggregateRanks (method = 'stuart') function from the 'RobustRankAggreg' R package[43].

Multivariate machine-learning models

Multivariate models were fitted using the Extreme Gradient Boosting (XGBoost) package in R[44]. The training in all cases sought to find the tree function T (sum of trees) that minimizes the expected loss between the observed and predicted response values; that is:

$$\hat{T} = argmin_T E_{X,Y} L(Y, T(X))$$

where X and Y are the feature and response data, respectively, and L is the loss function of choice. In our setting, the loss function was chosen to be a negative likelihood compatible with the typical distributional assumptions for each type of data. Specifically, the best overall response was modeled using the logistic regression likelihood, while progression-free survival and overall survival were modeled using the Cox proportional hazards likelihood.

We trained three pan-cancer models (one per outcome) incorporating all available training data (479 patients). We also trained 12 hybrid models based on the three pan-cancer models followed by further cycles of training on patients of each tumor type, but maintaining exactly the same loss function and all hyperparameters. Patients suffering from malignancies other than skin melanomas and lung or bladder tumors were pooled within a group labeled as 'other' tumor types. This model fit procedure found a compromise between low variance but high bias from the pan-cancer models and low bias but high variance in pure tumor type-specific models. Finally, we also trained pure tumor type-specific models, starting from the patients in each of the four groups separately (details in Supplementary Note 1).

The XGBoost models require many tuning parameters (learning rate, depth, sub-sampling, minimum tree leaf size) that guide the internal model fitting. Initially, our model building used grid searches to select optimal internal tuning parameters. However, in our cross-validation study, we found that simple additive models (depth, 1; fast learning rate, 0.05; minimum leaf size, 5; sub-sampling, 0.75) had the best performance.

For the best overall response, the model casts the prediction outputs as log odds ratio scores that can then be recast into probability scores (continuous values between 0 and 1). For progression-free survival and overall survival, the models cast the prediction outputs as log hazard ratios that can then be recast into hazard ratios (continuous and positive).

Separate models were trained solely on TMB and PDL1 expression (the continuous value reported in the HMF-CPI cohort by whole-transcriptome RNA-seq). These models were used to represent the predictive power of clinically approved biomarkers across analyses of the performance of multivariate models.

Calculation of Shapley values

Given that the final tree-based models were additive, the calculation and extraction of Shapley values was straightforward. For each feature, for a given additive model and individual sample, there was 1-to-1 mapping from the feature values and the Shapley values. This relationship between feature and Shapley values is visualized by the marginal dependence plots in Extended Data Figure 7a,b. In R, using the predict function applied to the XGBoost output, we set the argument contribution = TRUE to extract the Shapley values. The extracted Shapley values measure additive feature contribution to the log odds ratio for response models and the log hazard ratio for the Cox survival models.

Proxy biomarkers in the VHIO cohort

In the VHIO cohort, the TMB was estimated from the mutations detected using a 432-gene hybrid capture-based panel[45]. The expression of 170 genes was measured using the nCounter (NanoString) platform[46]. Normalized NanoString counts were log transformed and standardized, and proxy biomarkers were selected based on their correlation with the representative biomarkers of the five latent factors in the HMF-CPI cohort. For the T-cell effective infiltration gene set, *CXCL9*, *CXCL10*, *CXCL11*, *GZMA*, *GZMB* and *IFNG* were selected. Overall, this gene set was strongly correlated with the original T-cell effective infiltration gene set ($\rho = 0.97$) and showed high statistical significance in the exhaustive analysis ($P = 7.0 \times 10^{-8}$). To select a set of genes to represent the latent factors of TGF-β activity in the tumor microenvironment and tumor proliferative potential, we selected genes with a correlation of >0.5 to the respective gene set. This process yielded *BRCA1*, *BRCA2* and *TUBB* for the tumor proliferative potential gene set. Although none of these genes were included in the representative biomarker obtained from the HMF-CPI cohort, they all showed a strong correlation to this gene set. The proxy gene set also showed a statistically significant association with overall survival residuals. The aforementioned process, in the case of the VHIO TGF-β gene set, yielded *DLL4*, *HEYL*, *NOTCH3*, *NOTCH4*, *SERPINE1*, *TGFB1* and *TGFB3*. This gene set also showed a very strong correlation to the representative TGF-β activity in the tumor microenvironment biomarker (Supplementary Note 1).

Statistics and reproducibility

The systematic analysis to identify features associated with CPI response and survival was carried out through logistic and Cox regressions, and the results were filtered for multiple testing as described in the Methods and Supplementary Information. These features were grouped into latent factors based on their pairwise correlations. Standard statistical approaches, such as univariate and multivariate regressions or Kaplan–Meier analysis, were used downstream for the analysis of the latent factors across validation cohorts. No statistical method was used to determine sample size for the analysis. All available samples from the discovery and validation cohorts were used; none were excluded from the analysis. Given that the study consisted entirely of the analysis of existing data, it was not randomized and the investigators were not blinded, as no allocation of samples in groups was carried out. All data used in this study are publicly available (see below) and the code used to reproduce the analysis described in the paper has been deposited in public repositories (see below).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Access to the HMF-CPI data can be obtained through a request to the HMF (https://www.hartwigmedicalfoundation.nl/en/data/data-acces-request)[18,19]. The validation datasets can be obtained as follows: INSPIRE[29], INSPIRE github repository (https://github.com/pughlab/inspire-genomics); Lyon[30], GEO GSE159067, GEO 161537, GEO GSE162519 and GEO 162520; MARIATHASAN[27], public repository (http://research-pub.gene.com/IMvigor210CoreBiologies); PARKER ICI[31], PARKER ICI github repository (https://github.com/ParkerICI/MORRISON-1-public); RAVI[32], zenodo repository: 7625517; VHIO, this study github repository (https://github.com/bbglab/immunebiomarkers).

Code availability

All code necessary to carry out the extraction of the features from the HMF-CPI provided files (version DR-263_update1) and to generate the data frame needed for analysis is freely available in a public repository (https://github.com/bbglab/hartwig_biomarkers). The code to reproduce all analyses is also publicly available (https://github.com/bbglab/immunebiomarkers).

References

40. Pabla, S. et al. Proliferative potential and resistance to immune checkpoint blockade in lung cancer patients. *J. Immunother. Cancer* **7**, 27 (2019).
41. McDermott, D. F. et al. Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. *Nat. Med.* **24**, 749–757 (2018).
42. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
43. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012).
44. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016). https://doi.org/10.1145/2939672.2939785
45. Saura, C. et al. Early-stage breast cancer detection in breast milk. *Cancer Discov.* **13**, 2180–2191 (2023).
46. Hernando-Calvo, A. et al. A pan-cancer clinical platform to predict immunotherapy outcomes and prioritize immuno-oncology combinations in early-phase trials. *Med* **4**, 710–727.e5 (2023).

Acknowledgements

We wish to acknowledge the contribution of patients, families and biomedical researchers who shared, processed and sequenced the data used in the study. This publication and the underlying study have been made possible partly based on the data that HMF has made available. We also acknowledge the use of data from the INSPIRE clinical trial (NCT02644369) and data obtained from patients presented at the VHIO. N.L.-B. acknowledges funding from the European Research Council (ERC; consolidator grant 682398). This project has received funding from the European Union's Horizon program HORIZON-HLTH-2021-CARE-05-02 for the project CGI-Clinics under grant agreement no. 101057509. E.B. receives support from Generalitat de Catalunya (2021 SGR 001278) and ERC Advanced Grant 884623. The Institute for Research in Biomedicine Barcelona is a recipient of a Severo Ochoa Centre of Excellence Award from the Spanish Ministry of Economy and Competitiveness (MINECO; Government of Spain) and an Excellence Institutional grant by the Asociacion Española contra el Cancer and is supported by CERCA (Generalitat de Catalunya).

Author contributions

A.G.-P. and N.L.-B. conceptualized the project. J.U. designed and carried out the exhaustive analysis, training, testing and validating multivariate methods, and performed most other analyses presented in the manuscript and supplement and participated

in the discussion of results. A.R.H. and M.A.A. participated in the analysis and interpretation of latent factors. F.M. participated in the conceptualization of the analyses, provided support to perform them and participated in the discussion of results. A.G.-P. and N.L.-B. supervised the project, participated in the discussion of results and wrote the first draft of the manuscript. D.L.G., M.G. and A.V. provided the molecular data of patients in the VHIO cohort. E.F., E.E., J.C., E.M.-C. and J.T. provided the clinical data of patients in the VHIO cohort. A.V., E.F., J.T., F.M.-J., E.B., E.C. and L.L.S. contributed ideas to the design of analyses and participated in the discussion of results. All authors have read and approved the manuscript.

Competing interests

E.M.-C. reports a consultant or advisory role for Bristol Myers Squibb, Merck Sharp & Dohme, Novartis, Pierre Fabre, Roche and Sanofi; research funding from MSD, Sanofi and BMS; speaking engagements for Amgen, Bristol Myers Squibb, Merck Sharp & Dohme, Novartis and Pierre Fabre; clinical trial participation (as principal investigator) for Amgen, Bristol Myers Squibb, GlaxoSmithKline, Merck Sharp & Dohme, Novartis, Pierre Fabre, Roche and Sanofi. L.L.S. has a consultant/advisory role for Pfizer, AstraZeneca, Roche, GlaxoSmithKline, Voronoi, Arvinas, Navire, Relay, Marengo, Daiichi Sankyo, Amgen, Medicenna, LTZ Therapeutics, Tubulis, Nerviano, Pangea, Incyte and Gilead; received grant/research support (Institution—for clinical trials) from Merck, Novartis, Bristol Myers Squibb, Pfizer/SeaGen, Boerhinger-Ingelheim, GlaxoSmithKline, Roche, Genentech, AstraZeneca, Bayer, Abbvie, Amgen, Symphogen, EMD Serono, 23Me, Daiichi Sankyo, Gilead, Marengo, Incyte, LegoChem, Loxo/Eli Lilly, Medicenna and Takara; reports a leadership position (spouse) at Treadwell Therapeutics (founder) and stock ownership (spouse) in Agios. E.B. is the author of a patent related to TGF-β inhibitors, a patent describing bispecific antibodies to target cancer stem cells; E.B.'s lab has received research funding from MERUS, INCYTE and Revolution Medicines; and received honoraria for consulting from Genentech. J.T. reports personal financial interest in the form of scientific consultancy role for Alentis Therapeutics, AstraZeneca, Aveo Oncology, Boehringer Ingelheim, Cardiff Oncology, CARSgen Therapeutics, Chugai, Daiichi Sankyo, F. Hoffmann–La Roche, Genentech, hC Bioscience, Ikena Oncology, Immodulon Therapeutics, Inspirna, Lilly, Menarini, Merck Serono, Merus, MSD, Mirati, Neophore, Novartis, Ona Therapeutics, Ono Pharma USA, Orion Biotechnology, Peptomyc, Pfizer, Pierre Fabre, Samsung Bioepis, Sanofi, Scandion Oncology, Scorpion Therapeutics, Seattle Genetics, Servier, Sotio Biotech, Taiho, Takeda Oncology and Tolremo Therapeutics; stocks in Oniria Therapeutics, Alentis Therapeutics, Pangaea Oncology and 1TRIALSP; and an educational collaboration with Medscape Education, PeerView Institute for Medical Education and Physicians Education Resource (PER). E.E. has received personal honoraria from Amgen, Bayer, BMS, Boehringer Ingelheim, Cure Teq AG, Hoffman–La Roche, Janssen, Lilly, Medscape, Merck Serono, MSD, Novartis, Organon, Pfizer, Pierre Fabre, Repare Therapeutics, RIN Institute, Sanofi, Seagen International, Servier and Takeda. E.F. reports a consulting or advisory role with Abbvie, Amgen, AstraZeneca, Bayer, Boehringer Ingelheim, BMS, Daiichi Sankyo, F. Hoffmann–La Roche, Genmab, Gilead, GSK, Janssen, Johnson & Johnson, Merck Serono, MSD, Novartis, Peptomyc, Pfizer, Regeneron, Sanofi, Takeda; and speakers' bureau for Amgen, AstraZeneca, BMS, Daiichi Sankyo, Eli Lilly, F. Hoffmann–La Roche, Janssen, Medical Trends, Medscape, Merck Serono, MSD, Peervoice, Pfizer Regeneron, Seagen, Touch Oncology; board of directors role with Grifols; principal investigator in trials (institutional financial support for clinical trials) sponsored by AstraZeneca, Abbvie, Amgen, Bayer, Beigene, Boehringer Ingelheim, BMS, Daiichi Sankyo, Exelixis, F. Hoffmann-La Roche, Genentech, GSK, Janssen, MSD, Merck KGAA, Mirati, Novartis, Nuvalent, Pfizer and Takeda. J.C. reports a role with the advisory board of Astellas Pharma, AstraZeneca, Bayer, Bristol Myers Squibb, Exelixis, Ipsen, Johnson & Johnson, MSD Oncology, Novartis (AAA), Pfizer and Sanofi; institutional funding received from Janssen-Cilag International NV, Lilly, S.A, Medimmune, Novartis Farmacéutica, S.A. and Sanofi-Aventis, S.A.; other role as Member of the Comission Catalan Program of Ambulatory Medication Comission (CAHMDA). The remaining authors declare no competing interests.

Additional information

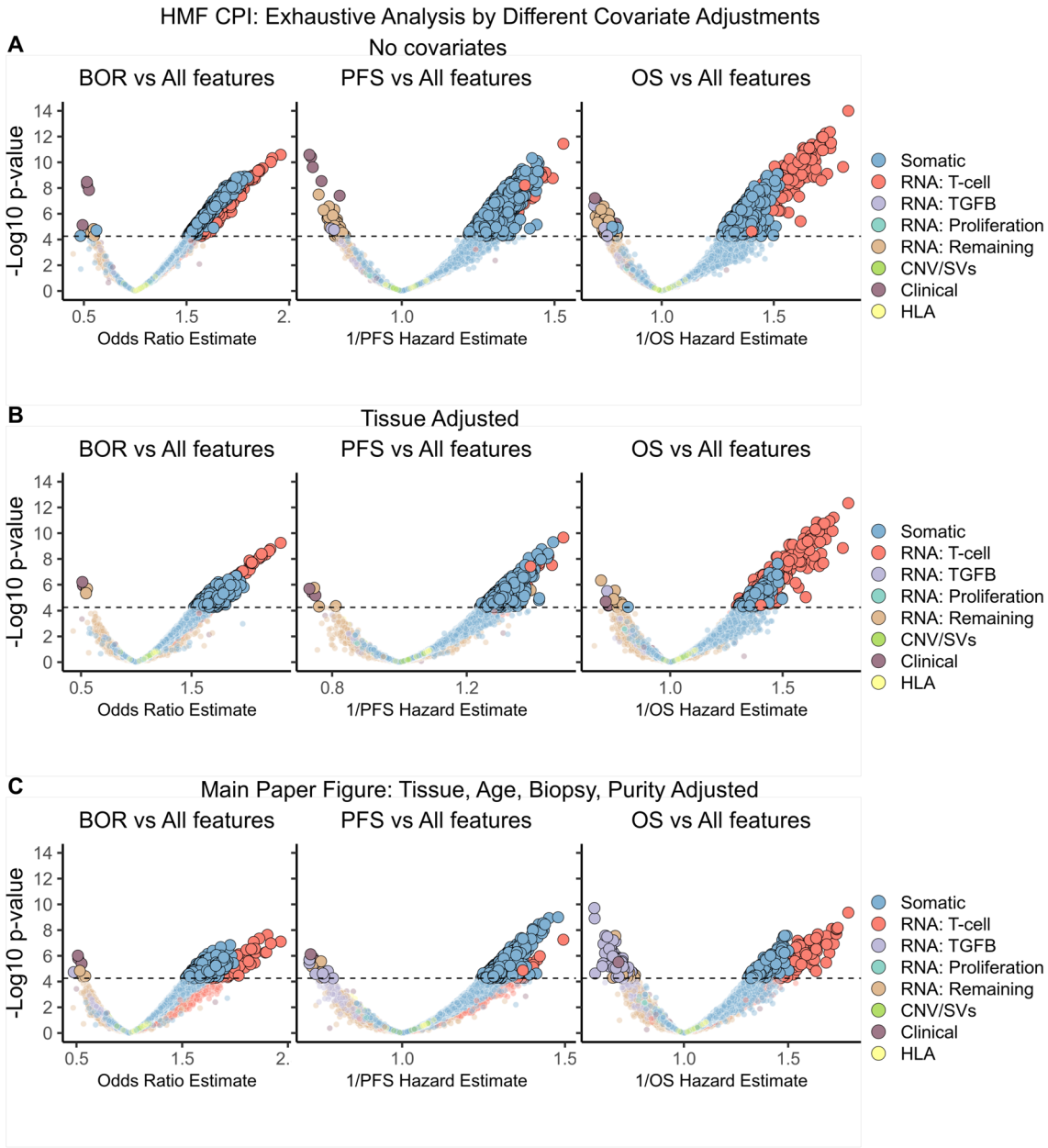
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01899-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01899-0.

Correspondence and requests for materials should be addressed to Abel Gonzalez-Perez or Nuria Lopez-Bigas.

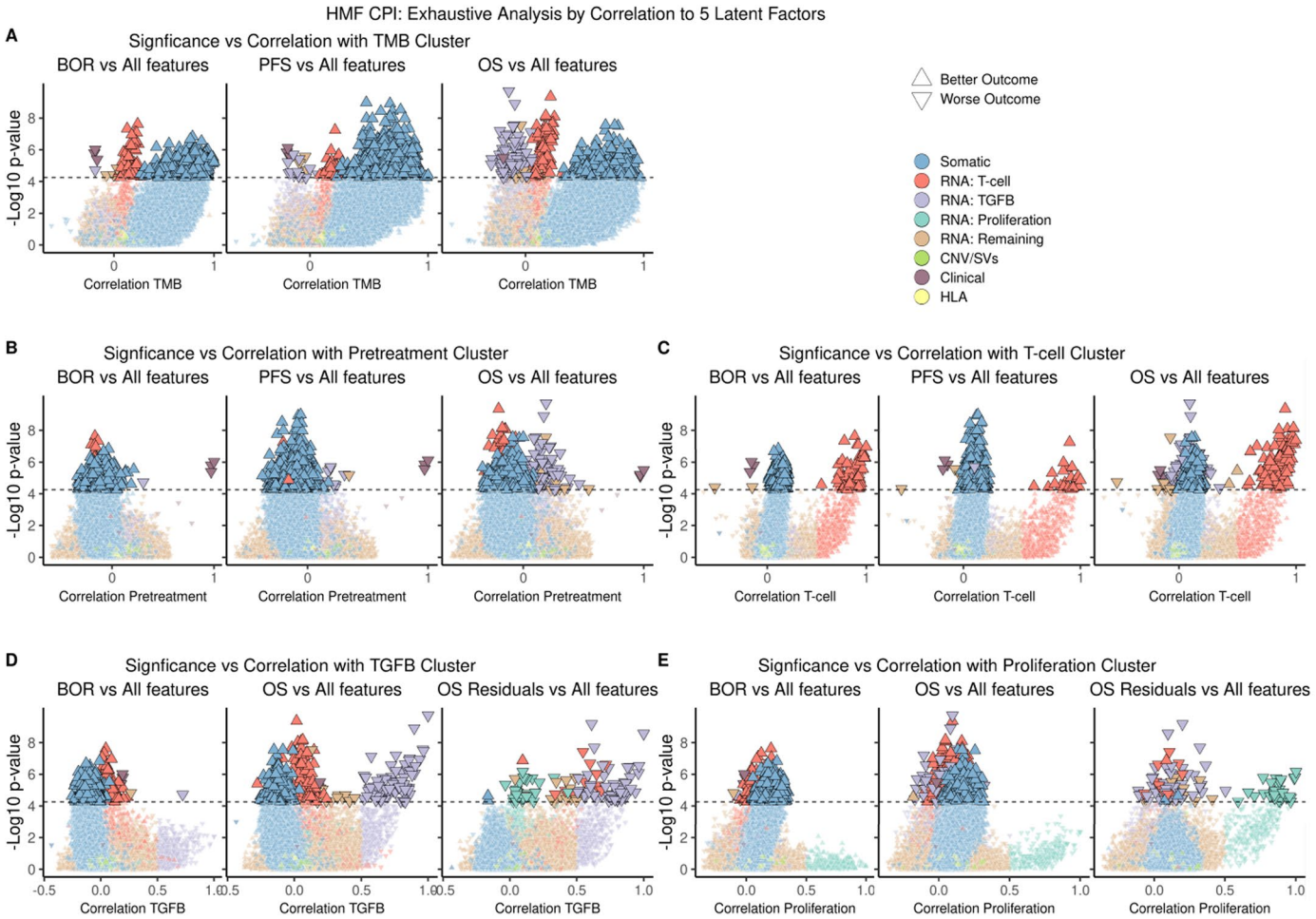
Peer review information *Nature Genetics* thanks Kevin Litchfield for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-024-01899-0



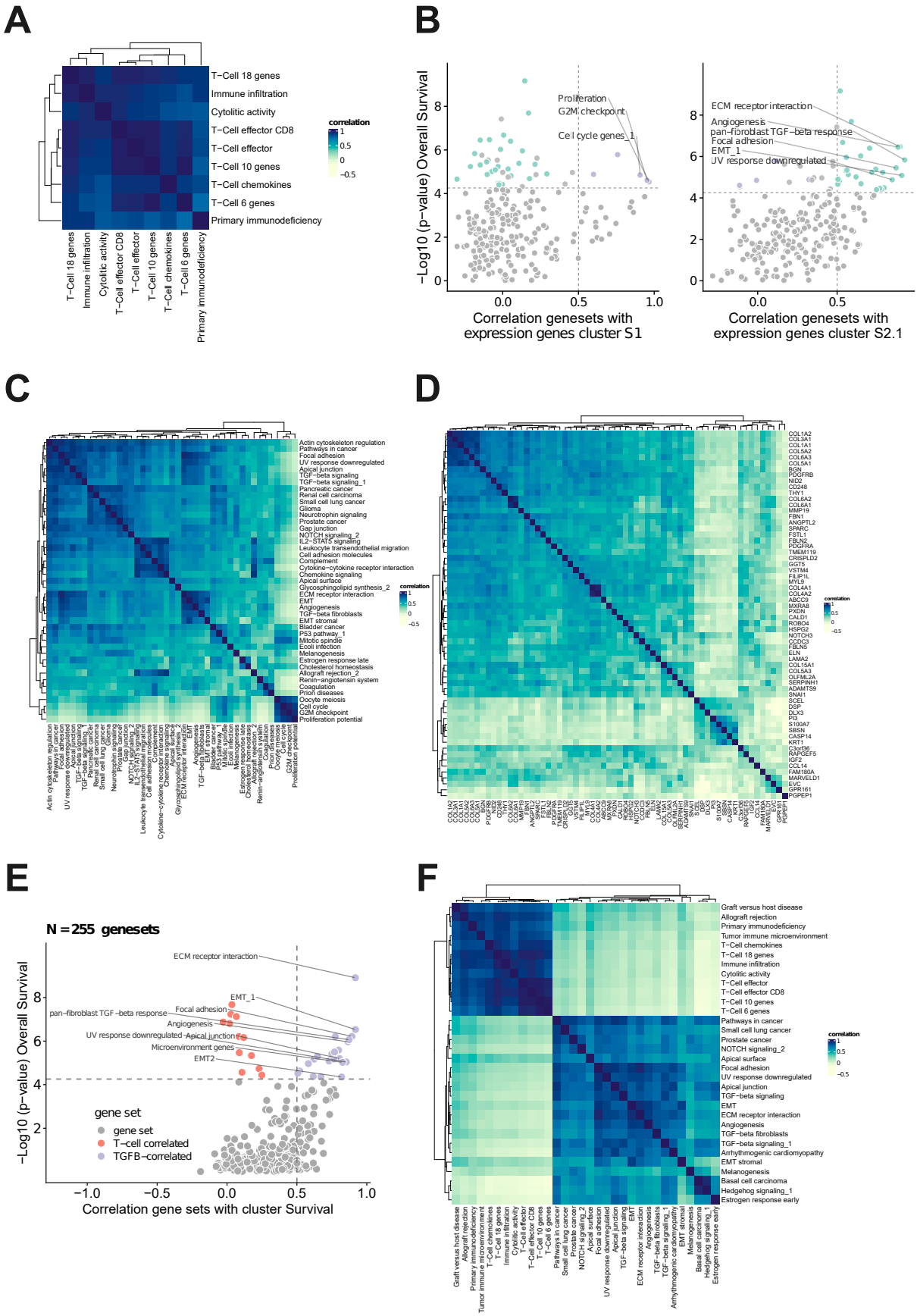
Extended Data Fig. 1 | Identification of latent factors associated with CPI response and survival across the HMF-CPI cohort. The figure provides a broad comparison of the landscape of features identified as significantly associated with CPI response (BOR), Progression Free Survival (PFS) and Overall Survival (OS) through the systematic use of univariate regression models corrected with different sets of covariables (see main manuscript and Supplementary Note 1). The three panels illustrate the results of the systematic analysis using no covariables (**a**), only the tissue as covariable (**b**), or the tissue, age, biopsy site and tumor purity as covariables (**c**) for the regressions. All analyses described in the main manuscript were carried out taking into account all covariables described in c. Features of different nature are colored following the same legend as in the main Figures. All p-values shown in the plots were computed via logistic (response) or Cox (survival) regressions, as in Figs. 2 and 3 of the main manuscript. These are, by definition, two-sided, denoted by positive or negative odds ratios (logistic regressions) or the reverse of hazard estimates (Cox regressions). OS: overall survival; PFS: progression-free survival; BOR: best overall response according to RECIST.

https://doi.org/10.1038/s41588-024-01899-0



Extended Data Fig. 2 | The five latent factors are integrated by highly correlated and significant features, and are mutually orthogonal. All graphs present the relationship between the significance of the association between individual features with CPI response or survival and their correlation to the mean of the clusters of features representing each latent factor. **a**) TMB cluster. Features integrating this latent factor are significantly associated with CPI response and survival. **b**) Pretreatment cluster. Only very few features, all capturing different treatments, appear correlated with the mean of this cluster. Their association is also apparent with CPI response and survival. **c**) Effective T-cell infiltration cluster. Features integrating this latent factor are significantly associated with CPI response and survival. **d**) TGF-β activity in the microenvironment cluster. Features included in this cluster are highly correlated with the mean of the cluster, while some features included in the effective T-cell infiltration cluster show a moderate correlation (~0.5). These features are only significantly associated with CPI survival (including survival residuals), but not with response. **e**) Proliferative potential cluster. These features are only significantly associated with CPI survival residuals. Features of different nature are colored following the same legend as in the main Figures. All p-values shown in the plots were computed via logistic (response) or Cox (survival) regressions, as in Figs. 2 and 3 of the main manuscript. These are, by definition, two-sided. OS: overall survival; PFS: progression-free survival; BOR: best overall response according to RECIST.

https://doi.org/10.1038/s41588-024-01899-0

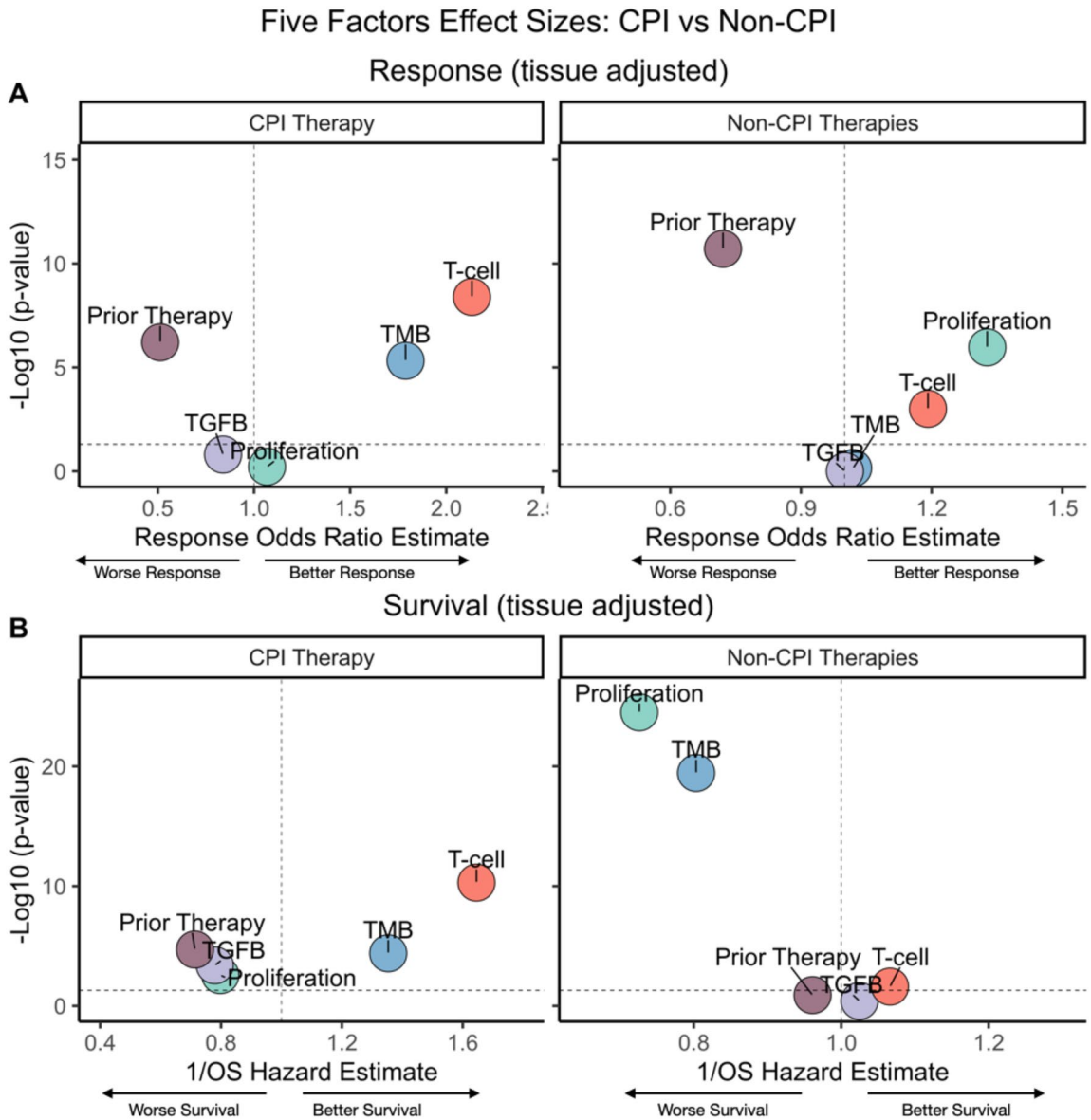


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41588-024-01899-0

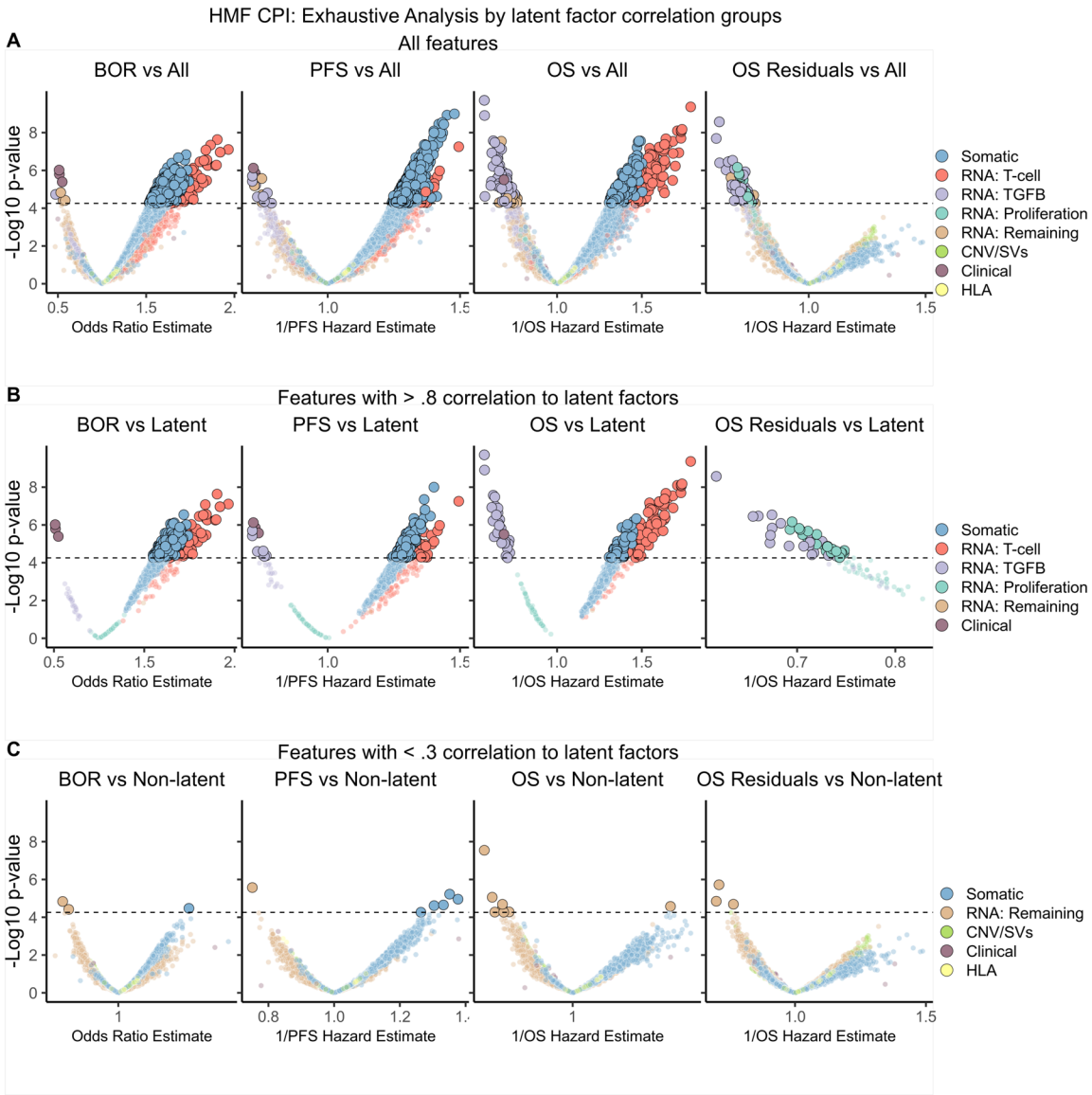
Extended Data Fig. 3 | Interpretation of significant expression features using genesets. a) Heatmap representing the pairwise correlation between genesets highlighted in Fig. 2 of the main paper. **b**) Significance of the association of 255 genesets with CPI survival residuals and their correlation with the mean of cluster S1 (left) and S2.1 (right). Significant genesets and correlation above 0 are highlighted. **c**) Heatmap representing the pairwise correlations between genesets that appear significantly associated with CPI survival residuals and correlated with cluster S2.1. **d**) All significant features from the volcano plot represented in Fig. 3e which do not belong to any of the response clusters previously identified (TMB, T-cell effective infiltration, prior treatment) were selected and clustered based on their pairwise correlations. One large cluster (along a few unclustered features) is apparent, called cluster Survival. **e**) We computed the correlation of the mean value of the Survival cluster with 255 genesets. It was highly correlated with genesets representing the activity of TGF-β in the tumor microenvironment (purple dots). Other significant genesets (uncorrelated with cluster Survival) represent T-cell effective infiltration (red dots). **f**) Pairwise correlations between all genesets that appear significantly associated with CPI overall survival not corrected by TMB, T-cell effective infiltration and prior treatment. Two clusters are apparent. One of them represents T-cell effective infiltration. The other represents TGF-β activity in the microenvironment. P-values shown in the plot were computed via logistic (response) or Cox (survival) regressions, as in Figs. 2 and 3 of the main manuscript. These are, by definition, two-sided. OS: overall survival; PFS: progression-free survival; BOR: best overall response according to RECIST.

https://doi.org/10.1038/s41588-024-01899-0



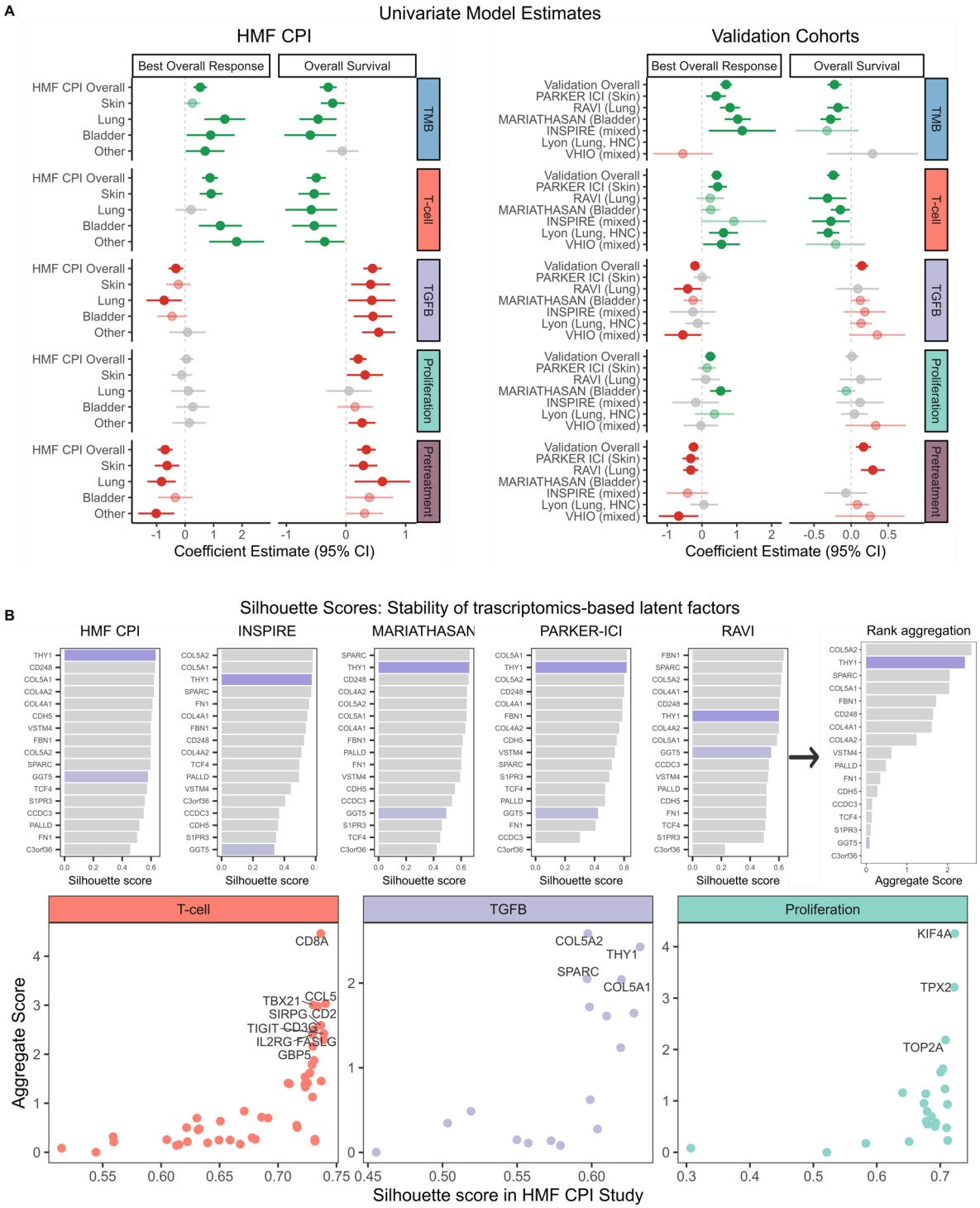
Extended Data Fig. 4 | Association of the five latent factors with anti-cancer systemic therapies other than CPI. Association of the five latent factors with the response to treatment (**a**) and overall survival (**b**) of patients in the HMF cohort who received CPI (left) or other therapies (right). All patients with an annotation of having received a treatment (other than CPI) for the metastatic tumor and for which an annotation of the organ of origin of the primary tumor was available were included in this group (N = 2,497). In each of the graphs the horizontal dotted line represents the threshold of statistical significance, while the vertical dotted line separates the positive (increased response or survival) and negative (decreased response or survival) effects. The association of each of the latent factors with CPI response or survival has been assessed using a univariate regression (on the values of the representative of the latent factor computed across tumors). Hence, a circle in the top right quadrant denotes a latent factor significantly associated with a positive outcome (increased response or survival); a circle in the top left quadrant represents a latent factor associated with a negative outcome (decreased response or survival). A circle in either of the two bottom quadrants represents a latent factor not significantly associated with the outcome measured. P-values shown in the plots were computed via logistic (response) or Cox (survival) regressions, using as independent variable, in each case, the estimator of each latent factor. These are, by definition, two-sided, denoted by positive or negative odds ratios (logistic regressions) or the reverse of hazard estimates (Cox regressions).

https://doi.org/10.1038/s41588-024-01899-0



Extended Data Fig. 5 | The five latent factors capture all the signal of features associated with CPI response and survival. a) Features of different types significantly associated with CPI response or survival. The three first graphs correspond to Extended Data Figure 1C. The fourth graph presents the regression of survival residuals (that is, controlling for the features identified as associated with response) on all features. **b**) Volcano plots resulting from the regression analyses presented in panel A, including only features with correlation coefficient above 0.8 with the mean of any latent factor. Significant features from all regression analyses show high correlation to the clusters' mean (as the clusters are precisely constructed from them). Other non-significant features show equally high correlation with the clusters. **c**) Volcano plots as in panels A and B, but showing only features with correlation coefficient below 0.3 to the mean of the clusters defining the latent factors. Only scattered features uncorrelated to the five latent factors appear significantly associated with CPI response or survival, indicating the absence of any other mutually orthogonal latent factor in the HMF-CPI cohort at the level of statistical significance set by the stringent False Discovery Rate used. Features of different nature are colored following the same legend as in the main Figures. The p-values and effect sizes shown result from logistic or Cox regressions. P-values shown in the plots were computed via logistic (response) or Cox (survival) regressions, as in Figs. 2 and 3 of the main manuscript. These are, by definition, two-sided. OS: overall survival; PFS: progression-free survival; BOR: best overall response according to RECIST.

https://doi.org/10.1038/s41588-024-01899-0

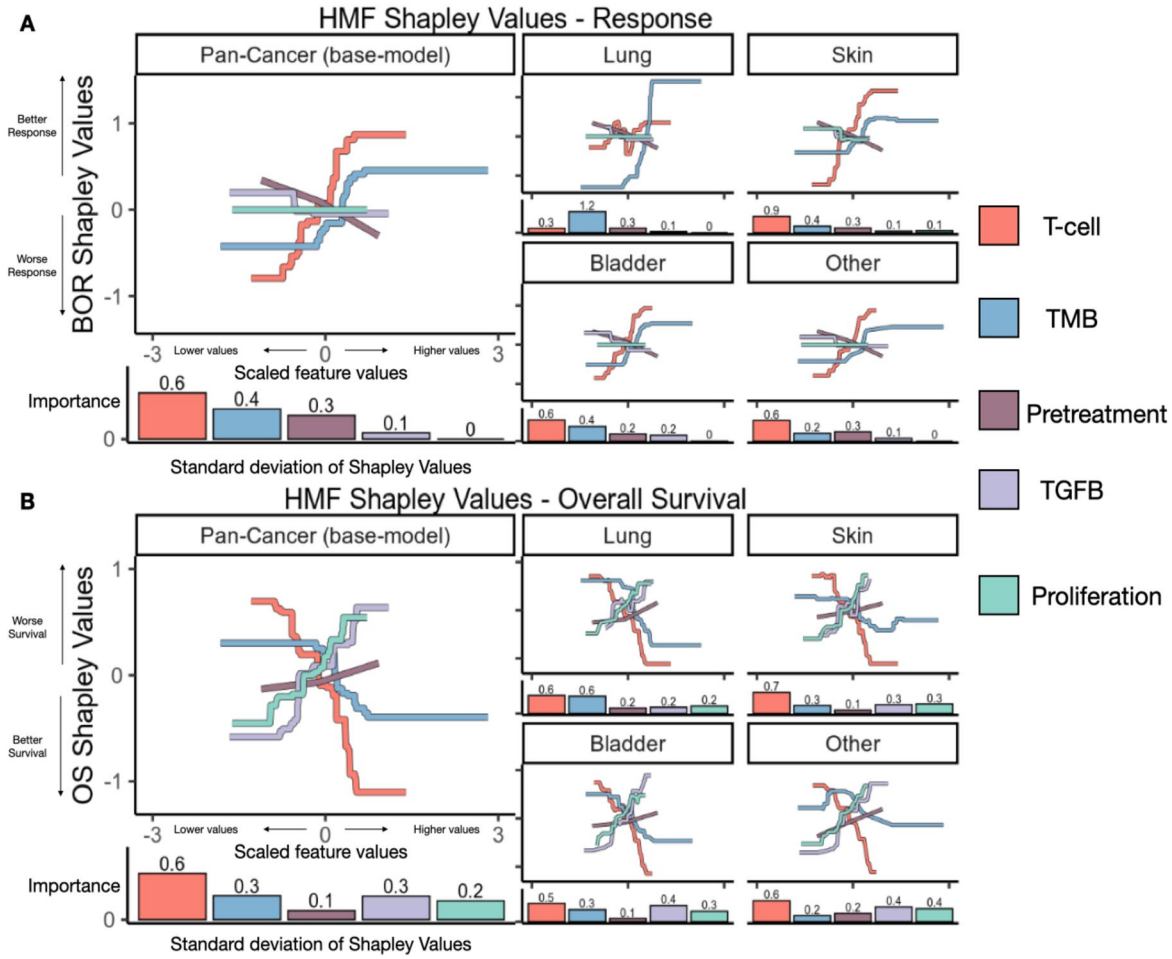


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41588-024-01899-0

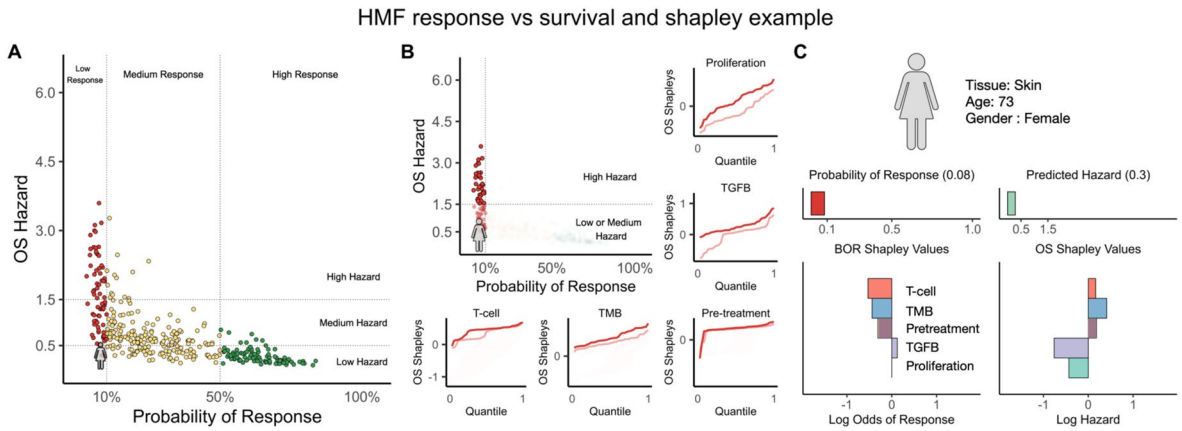
Extended Data Fig. 6 | Univariate analyses reveal the association of latent factors with CPI response across different tissues in the HMF-CPI cohort and six validation cohorts. a) Left panel: Forest plot illustrating the association (calculated through univariate regression models) of the five latent factors with CPI response and survival across groups of patients with different types of tumors in the HMF-CPI cohort. Right panel: Idem across six validation cohorts. Red or green dots denote clear association (regression coefficients estimate more than 1 (light) / 1.96 (dark) standard errors from 0) of a latent factor with response or survival, while gray dots denote lack of association. Dark color denotes significance of the association, while light color represents non-significant associations. In the forest plots, the dots represent the strength (coefficients estimated through multivariate logistic or Cox regression) of the association between the latent factor and response or survival across cohorts. The horizontal bars across dots denote the 95% confidence intervals. Gray dots represent latent factors whose estimates are within one standard error of 0, dots with light color (green or red) represent non-significant associations with coefficient estimates above (or below) one standard error of the 0, while dark colored dots represent significant associations. Green dots represent positive associations with improved outcomes (higher response odds or lower hazard ratio), while red dots represent negative associations (lower response or higher hazard ratio). **b**) Stability of transcriptomics latent factors across validation cohorts. We computed the relationship between the distance of each feature to all the members of its cluster (defined in the HMF-CPI cohort) and all members of other clusters (silhouette score; Methods). The silhouette scores thus computed for genes in the TGF-beta activity in the microenvironment across HMF-CPI and four validation cohorts are represented in the first five bar plots in the top panel. Two genes, one with relatively high silhouette score, and another showing more variability across all cohorts appear highlighted. The ranks of the genes (sorted according to their silhouette scores) are aggregated across all cohorts, and a significance score (reflecting genes that are ranked consistently better than expected) is computed (right-hand bar plot). The three graphs at the bottom of the panel represent the relationship between the silhouette score of the genes in each transcriptomics latent factor in the HMF-CPI cohort (x-axis) and their aggregated score (y-axis). Sample sizes for all datasets tested can be found in Supplementary Table 1.

https://doi.org/10.1038/s41588-024-01899-0



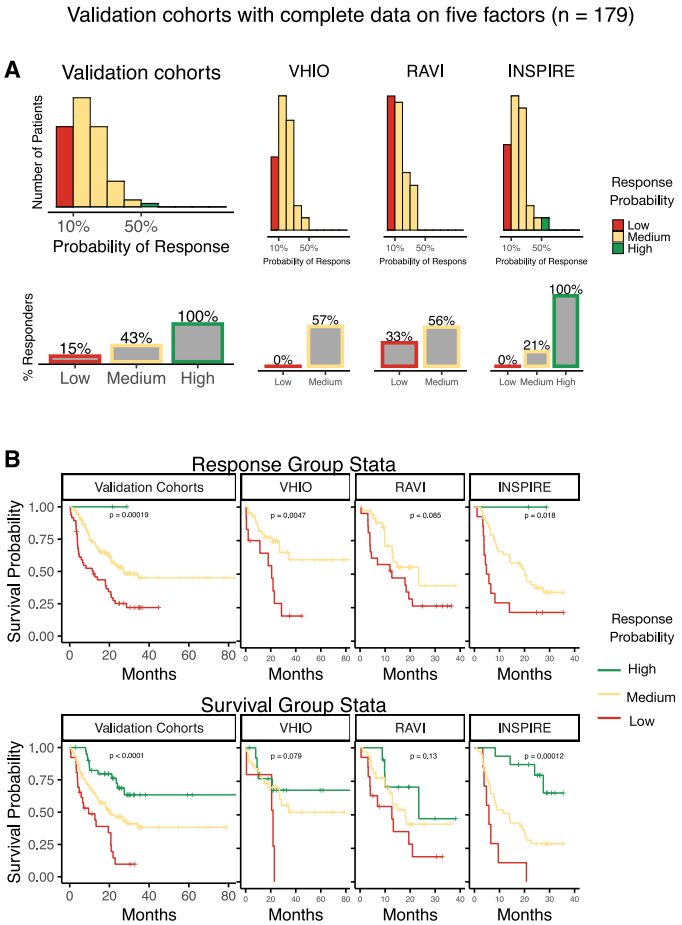
Extended Data Fig. 7 | Relative importance of the five latent factors in the prediction of response or overall survival across patients in the HMF-CPI cohort. The line plots represent the contribution of the values of each latent factor (Scaled feature values) across patients to the predictions cast by the response (BOR) and overall survival (OS) multivariate models. The effects are illustrated through the Shapley Values (Methods and Supplementary Note 1). Thus, in each plot, the line corresponding to each latent factor follows the relative influence of the values of the feature used to measure the latent factor on the predictions obtained through the model across all patients. Lines with positive slope correspond to latent factors that increase either the probability of response or the hazards with the increase in their value. The bar plots below the line plots represent the overall importance of each latent factor (using the standard deviation of the Shapley values) across all predictions of each model in each cohort. **a**) Representation of the relative importance of the latent factors in the prediction of response to CPI across the pan-cancer cohort and each tumor type separately within the HMF-CPI cohort. **b**) Representation of the relative importance of the latent factors in the prediction of overall survival (hazards) to CPI across the pan-cancer cohort and each tumor type separately within the HMF-CPI cohort.

https://doi.org/10.1038/s41588-024-01899-0



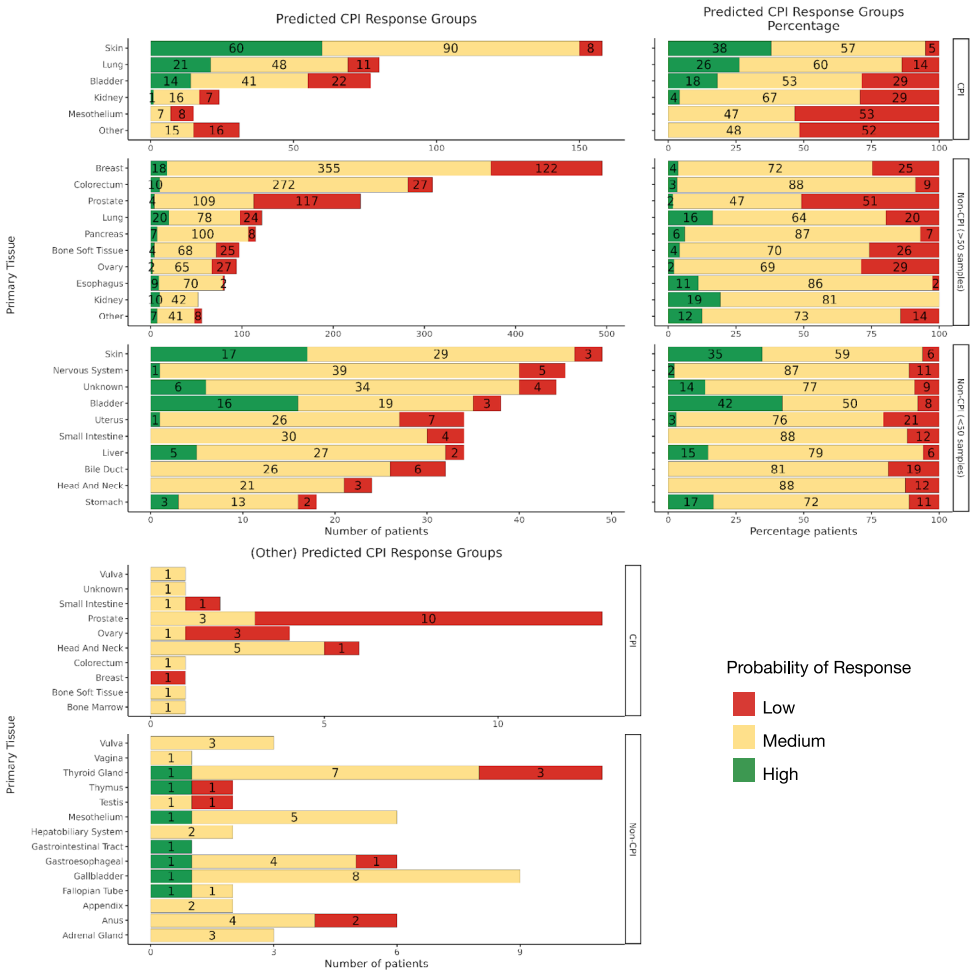
Extended Data Fig. 8 | Comparison of response and survival models using Shapley values. a) Showing a comparison of response and survival hazard estimates. The points are color coded red for low responders (<10% probability response), yellow for medium responders (10-50% probability) and green for high responders(>50%). The estimates we obtained from XGboost models trained on representative biomarkers of the five latent factors across patients in the HMF-CPI cohort to predict CPI response and survival. **b**) Exploring the determinants of the distribution of hazards across patients with low probability of response (scatterplot). The patients in this group have been subdivided into two smaller groups based on their predicted hazard, represented by dots of different shades of red separated by the horizontal line in the value of predicted hazard 1.5. The line plots represent the distribution (quantiles) of Shapley values (see Methods) calculated for these two subgroups of patients for the five latent factors. The two lines appear more separated in the distributions of Shapley values of tumor proliferative potential and TGF-beta activity in the microenvironment. This indicates that it is the values of these two latent factors that contribute the most to the separation between these two groups of patients. **c**) Example of the predicted CPI response and survival of one patient in the HMF-CPI cohort broken down by Shapley values.

https://doi.org/10.1038/s41588-024-01899-0



Extended Data Fig. 9 | Stratification of patients in validation cohorts using multivariate machine learning models. a) The histograms represent the distribution of the probability of response to CPI of patients across three of the validation cohorts (those with complete data on all five latent factors), either combined or separate. The bars are colored red (probability of response below 0.1, low), yellow (probability between 0.1 and 0.5, medium) or green (probability above 0.5, high). The absolute number of patients across the three cohorts in each group (low, medium, high) are shown in the horizontal bar below the combined histogram. The barplots below present the percentage of patients in each of the groups who actually showed response to CPI according to the data of each cohort. **b**) Top panel: Kaplan-Meier curves resulting from the aforementioned stratification of patients across the three cohorts, either combined or separate. Bottom panel: Kaplan-Meier curves resulting from stratifying the patients across the three cohorts based on their predicted probability of survival according to the hybrid models trained on survival data. The p-value for each cohort (annotated in the plot) was calculated via a one-sided logrank test.

https://doi.org/10.1038/s41588-024-01899-0



Extended Data Fig. 10 | Application of multivariate machine learning models to identify patients with high probability to respond to CPI across the entire HMF cohort. Bars represent the number of patients with metastatic tumors from different sites of origin in the HMF cohort who received (top) or did not receive (bottom) CPI as treatment. All patients with an annotation of having received a treatment (other than CPI) for the metastatic tumor and for which an annotation of the organ of origin of the primary tumor was available were included in this group (N = 2,497). The colored segments in the bars at the left represent the absolute number of patients with low (below 0.1), medium (between 0.1 and 0.5) or high (above 0.5) predicted probability of response. These bars have been separated based on the total number of patients of each tumor type, and x-axes representing the relative scales of each plot have been added. To the right side of the plot, the percentage of patients of each tumor type including more than 15 cases are represented as stacked bar plots, to facilitate comparability between tumor types. An important fraction of patients with tumors from the same origin as those in the HMF-CPI cohort (for example, in the lung) present high predicted probability of response to CPI. Interestingly, patients with tumors of other origins, who are not typically considered as candidates for CPI treatment also exhibit high predicted response probability.

nature portfolio

Corresponding author(s): Nuria Lopez-Bigas and Abel Gonzalez-Perez

Last updated by author(s): Jul 15, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	HMF cohort: Data of the Hartwig Medical Foundation (HMF) cohort were obtained from the HMF database (version DR-263_update1). Several other features computed for the tumors of these patients were also obtained directly from the HMF database, while others were computed by us. A through list of these features and their values is presented in Table S3. Validation cohorts INSPIRE: Data for 64 patients in the INSPIRE basket trial (NCT02644369) were obtained from the link shared in Data Accessibility. Lyon: Data for 315 patients treated at several hospitals in Lyon and Paris were obtained from the link shared in Data Accessibility. MARIATHASAN: Data for several cohorts of tumors (including those within clinical trials CheckMate 038 and CheckMate 067 and two cohorts published within other studies) compiled by Campbel et al (2023), totaling 315 patients were obtained from the link shared in Data Accessibility. RAVI: Data for 325 patients within the SU2C-MARK cohort published in an article by Ravi et al. (2023), were obtained from the link shared in Data Accessibility. VHIO: Data for 74 patients were obtained directly from the Cancer Genomics Group, and clinical data were provided by attending oncologists. Details on all the cohorts, and the source of the data obtained for each are described in Supplementary Table 1.
Data analysis	TRanscriptomics expression features consisting on summarization of the expression of gene sets representing different biological processes, or derivation via Cibersort of infiltrating immune cell populations were also calculated. Somatic HLA loss of heterozygosity in the tumors was estimated using the LILAC tool. For the systematic analysis of the association of these features with checkpoint inhibitors response or survival, each feature was tested individually for strength of association using logistic (R "glm" function) or Cox ("coxph" function from the "surival" package in R) regressions.

Generalized linear models and their native maximum likelihood based tools were used for all estimation, standard error calculation, and hypothesis testing (details in Supp. Note 1).
For somatic mutations by gene, only genes with at least 1 mutation per20 samples were kept for the analyses. For RNA expression, only coding genes with mean andstandard deviation of adjusted TPM values greater than 0.5 were considered. For the driver features, only driver genes mutated in at least 1 in 30 samples were included. Similarly, formutational signatures, only signatures with exposure greater than .02 for at least 1 in 20 samples were included. Second, all features were standardized to have a mean 0 and standard deviation of 1 across the CPI samples. Each feature was tested individually for strength of association to best overall response, progression free survival, and overall survival. Generalized linear models and their native maximum likelihood based tools were used for all estimation, standard error calculation, and hypothesis testing.
We used the Benjamini Yekutieli (BY) multiple testing threshold to control the false discovery rate.
Multivariate models were fitted using the Extreme Gradient Boosting (XGBoost) package in R. The fine tuning of the parameters of the XGBoost are described in Supplementary Note 1. Separate models were trained solely on TMB (the value computed on the HMF-CPI, INSPIREand VHIO cohorts) and PDL1 expression (the continuous value reported in each cohort viawhole transcriptome RNAseq or NanoString). Shapley values for predictions cast by the models on all patients were computed in R, using the predict function applied to the XGBoost output, setting the argument contribution = TRUE.
Clusters were built using the "hclust" package in R with the Ward.D2 algorithm.
Silhouette scores were calculated using the "cluster" package in R.
All analyses were implemented through ad hoc scripts in python (version 3.8.8) and R (version 3.6.1). All software needed to reproduce all analyses described in the paper and generate the figures is provided in https://bitbucket.org/bbglab/immune_biomarkers and https://bitbucket.org/bbglab/hartwig_biomarkers

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Access to the HMF-CPI data can be obtained through request to the Hartwig MedicalFoundation (https://www.hartwigmedicalfoundation.nl/en/data/data-acces-request/).
The validation datasets can be obtained as follows:
INSPIRE, https://github.com/pughlab/inspire-genomics/;
Lyon, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE159067, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161537, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162519, and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162520;
MARIATHASAN, http://research-pub.gene.com/IMvigor210CoreBiologies;
PARKER ICI, https://github.com/ParkerICI/MORRISON-1-public;
RAVI, https://zenodo.org/records/7625517;
VHIO, https://bitbucket.org/bbglab/hartwig_biomarkers

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex information is available for an important number of patients across the HMF and validation cohorts.
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	Details of the Hartwig Medical Foundation, and validation cohorts can be found in Table S1.
Recruitment	Recruitment of patients in the HMF and validation cohorts are described in the original sources. In the case of the VHIO cohort, it was integrated by patients arriving at VHIO hospital who received checkpoint inhibitors therapy.
Ethics oversight	The HMF and validation cohorts are shared within the public domain. The transfer of data concerning the VHIO cohort for this research was approved by ethics committees at the VHIO and IRB, counting with the informed consent of patients.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes correspond to all samples available from each dataset (Supplementary Table 1). These sample sizes provide enough statistical power to detect the features integrating the 5 latent factors as significant after correction for multiple testing. VHIO cohort: 479 metastatic cancer patients who received CPI INSPIRE cohort: 76 cancer patients Lyon cohort: 315 cancer patients MARIATHASAN: 348 cancer patients PARKER ICI: 315 cancer patients RAVI: 362 cancer patients VHIO: 75 cancer patients Details on all cohorts are available in the Methods section, Table S1 and the Supplementary Note
Data exclusions	All patients with CPI therapy in the three cohorts were included in the study. For the training and application of the multivariate models, only patients for which the five latent factors could be estimated were included.
Replication	The five latent factors are validated across 6 independent cohorts. In the pooled validation, all latent factors were succesfully validated. In separate datasets, some latent factors do not reach significance (see main manuscript).
Randomization	Randomization in the selection of training and test sets in the HMF cohort for the construction of multivariate models was applied
Blinding	This is a bioinformatics analysis of data previosly generated; no prior hypothesis existed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA