Evaluation of colorectal cancer subtypes and cell lines using deep learning

Jonathan Ronen[1], **Sikander Hayat**[2], **and Altuna Akalin**[1,*]

[1]Max Delbrueck Center for Molecular Medicine, Berlin Institute for Medical Systems Biology, Bioinformatics Platform, Berlin, 13125, Germany
[2]Bayer AG, Department of Bioinformatics, 13353 Berlin, Germany
[*]corresponding author: altuna.akalin@mdc-berlin.de

ABSTRACT

Colorectal cancer (CRC) is a common cancer with a high mortality rate and a rising incidence rate in the developed world. The disease shows variable drug response and outcome. Molecular profiling techniques have been used to better understand the variability between tumours as well as cancer models such as cell lines. Drug discovery programs use cell lines as a proxy for human cancers to characterize their molecular makeup and drug response, identify relevant indications and discover biomarkers. In order to maximize the translatability and the clinical relevance of in vitro studies, selection of optimal cancer models is imperative. We have developed a deep learning based method to measure the similarity between CRC tumors and other tumors or disease models such as cancer cell lines. Our method efficiently leverages multi-omics data sets containing copy number alterations, gene expression and point mutations, and learns latent factors that describe the data in lower dimension. These latent factors represent the patterns across gene expression, copy number, and mutational profiles which are clinically relevant and explain the variability of molecular profiles across tumours and cell lines. Using these, we propose a refined colorectal cancer sample classification and provide best-matching cell lines in terms of multi-omics for the different subtypes. These findings are relevant for patient stratification and selection of cell lines for early stage drug discovery pipelines, biomarker discovery, and target identification.

Introduction

Colorectal cancer (CRC) accounts for 10% of cancer related deaths[1], amounting to aproximately 1.4 million new colorectal cases and 693,900 deaths reported in 2012[2]. CRC is not a homogenous disease and can be classified into different subtypes based on molecular and morphological alterations[3]. The disease occurs when normal epithelium cells transform to cancer cells via acquired genetic and epigenetic alterations. Mutations in the Wnt signalling pathway are thought to initiate the transformation to cancer[4,5]. This is followed by deregulation of other signalling pathways such as MAPK, TGF-β, and PI3K–AKT via acquired mutations[5,6]. Since the original description of the molecular pathogenesis of CRC, multiple additional pathways, mutations, and epigenetic changes have been implicated in the formation of CRC[7]. Based on integrative analysis of genomic aberrations observed in the TCGA samples, CRC tumors can be divided into hyper-mutated ($\approx 16\%$) and non-hyper-mutated ($\approx 84\%$) cancers. The hyper-mutated cancers have microsatellite instability (MSI), resulting from defective mismatch repair or DNA polymerase proof-reading mutations[3]. The microsatellite stable (MSS) cancers are characterized by chromosomal instability (CIN), with high occurance of DNA copy number alterations, and mutations in the APC, TP53, KRAS and BRAF genes[3]. In addition, most CRC tumors have aberrantly methylated genes, a subset of which could play a functional role in CRC[7]. A further subset of CRC tumors display a CpG island methylator phenotype (CIMP), where some tumor suppressor genes could be inactivated epigenetically[8]. The diversity of molecular disease mechanisms creates distinct molecular subtypes in CRC with different survival rates and responses to therapy. The different subtype classification schemes based on gene expression developed over the years were recently studied and summarized as the consensus molecular subtypes for CRC[9]. This classification scheme designates four main subtypes called “consensus molecular subtypes” (CMS) with distinguishing features. The CMS1 subtype is defined by hypermutation, microsatellite instability and strong immune activation; CMS2 is defined by chromosomal instability, WNT and MYC signaling activation; CMS3 is defined by metabolic dysregulation; and CMS4 is defined by growth factor β activation, stromal invasion, and angiogenesis[9]. However, 13% of the samples have mixed features and cannot be reliably assigned to consensus subtype based on gene expression alone. These are considered to be subtypes with mixed signatures or samples with intra-tumor heterogeneity. Although purely based on gene expression, follow up analysis of the consensus molecular subtypes revealed distinct copy-number profiles, mutation frequencies and methylation profiles[9].

Thus, in this study, we propose a multi-omics method that can incorporate gene expression, copy number, and mutation data to identify CRC subtypes. In addition, our method is able to match cell lines to each subtype or collection of tumors, as well as matching new tumors to the ones already in the dataset. In the future, it can also be used to assign best-match xenografts or organoid models to the study of each subtype. This would be highly useful, as these have been shown to predict drug response in patients[10]. By leveraging multi-omics datasets, CRC samples that can not be associated to a CMS subtype might also be assigned to an appropriate subtype. In addition, multi-omics signatures, incorporating gene expression programs, point mutations and copy number alterations, are a direct output of the method, and will not need to be generated post-hoc, e.g. by examining mutation rates in groups defined by gene expression profiles, as in the CMS. With these goals in mind, we used deep-learning on multi-omics data in order to refine CRC subtypes in an unsupervised manner, and match them to cell lines.

Among unsupervised methods, latent factor analysis in the form of matrix factorization has been the workhorse for not only multi-omics data analysis but also high-dimensional data analysis in general[11]. This is understandable since these methods reduce the dimensionality and also provide latent factors that summarize the signals from data sets. These signals can be later interpreted in the biological context, and dimensionality reduction helps with subsequent distance metrics and clustering[12]. The latent factor analysis methods mostly revolve around concatenating the data sets to a single matrix and applying a well-known matrix factorization algorithm sometimes with weighting the individual data sets. Multifactor analysis (MFA)[13] and iCluster+[14] are examples of such methods. These types of algorithms sometimes impose orthogonality of factors, i.e. that the factors explain disjoint underlying processes, as in PCA. Orthogonality might be conceptually appealing, but is not a biological necessity. Orthogonal latent factors may be the best for summarizing or reconstruction of a dataset, and still be biologically difficult to interpret. On the other end of the spectrum, there are deep learning based methods which work as dimension reduction methods that can deal with non-linearity and they can generalize well on different problems. In addition, they can be sparse, where each latent factor only depends on a few of the input genes. Sparsity in the relationship between input genes and latent factors simplifies the task of biological interpretation of the model, that is, implicating known biological processes as underlying the values of the latent factors. The disentangled variational autoencoders (β-VAE)[15] have the previously mentioned characteristics and as a deep learning framework, they have flexible architectures that can work with most problem sets. A similar architecture has been shown to be able to stratify cancers by their tissue type based on gene expression profiles[16], and other autoencoder architectures have been used to integrate multi-modal data in robotics[17], as well as protein function prediction[18].

We use a multi-modal, stacked β-VAE architecture for extracting latent factors that are important for defining subtypes and predicting survival of the patients. We call the method, "**M**ulti-omics **A**utoencoder **I**ntegration", or, **maui**. We compare maui's performance to state-of-the-art multi-omics analysis methods, and show that maui outperforms the traditional methods. In addition, we are able to map CRC cell lines to the latent factor space defined by maui. This allows us to check the quality of CRC cell lines, using patterns recognized by the autoencoder as significant in colorectal cancers to rank in-vitro models suitability to study tumors. We then assign suitable cancer cell lines for drug target studies aimed at different colorectal cancer subtypes.

Results

Refining CRC subtypes using multi-omics data

As the colorectal tumor samples in the TCGA data-set (n=519, See Materials and Methods) have been extensively studied and a state-of-the-art subtyping scheme exists for these as the "Consensus Molecular Subtypes" or CMS[9], we wanted to validate the relevance of the latent factors we extract using maui, by checking how well they can recapitulate these subtypes. Table 1 summarizes the tumors' subtype information.

CMS label	Description	# Samples
CMS1	MSI Immune	61
CMS2	Canonical	175
CMS3	Metabolic	60
CMS4	Mesenchymal	123
Total with CMS label		**419**
Without CMS label		100
Total		**519**

Table 1. Summary of TCGA tumors' CMS labels

We extracted latent factors from the multi-omics data using maui, as well as other published methods for multi-omics integration by dimensionality reduction: MOFA[19], and iCluster+[14]. In order to quantify the relashionship between latent factor representation and CMS subtype, and compare the methods, we trained regularized Support Vector Machines using 10-fold cross-validation, at each time predicting the CMS out-of-sample from the latent factor representations. We then computed

'Receiver Operating Characteristics' for each CMS prediction, as well as the mean ROC to measure the performance of classification schemes (See Methods). maui marginally outperforms MOFA on CMS2 prediction, and both maui and MOFA dramatically out-perform iCluster+, showing that latent factors learned using both maui and MOFA allow a supervised learning algorithm to recapitulate the CMS sub-types nearly perfectly (Figure 1A-B).

While the previous analysis is instructive about which latent factors allow a supervised learning task to separate the CMS, maui has an unfair advantage over MOFA, as we ran it with 80 latent factors, whereas MOFA was only run with 20. In order to assess which of the methods is best able to capture the CMS labels, without regard to the number of latent factors, we repeated the previous exercise—predicting the CMS from the latent factors—using an unsupervised learning algorithm. We clustered the samples with a well-defined CMS [i] using k-means on the latent factors and, for a set of K's, computed the Adjusted Mutual Information (AMI) of the clustering with the CMS subtype. We did this with latent factors learned by maui, MOFA, and iCluster+. k-means clustering on the latent factors is able to capture more of the CMS labels with K values of 4–6, but only using maui (Figure 1C). This clustering analysis shows that maui factors are more conductive to CMS prediction, in a fair comparison, as k-means clustering is based on distances, whose computation does not benefit from higher dimensionality (unlike supervised learning) — in fact, the opposite is true[12].

Latent factors inferred by maui are reasonably predictive, using k-means, of the CMS sub-type, especially using K's 4–6 (Figure 1C). In order to pick the best clustering result to focus on, we computed the log-rank statistic for significance of differential survival rates between clusters. $K = 6$ results in the most statistically significant survival difference($P < 0.001$, Figure 1D). Note that the CMS sub-types on their own are not indicative of survival rates in the TCGA data, and that $K = 4$ and $K = 5$, also produce clusters with significant differential survival (Figure S2). Notably, $K = 6$ is preferable to $K = 4$ and $K = 5$, as it is able to tease out a cluster with particularly poor prognosis (cluster 3), which consists mostly of a subset of tumors with the CMS2 designation (Figure S2).

A closer examination of the resulting clusters reveals how closely the maui clusters resemble the CMS sub-types, and where they diverge. Figures 2A-C show that CMS1 is captured by cluster 2, CMS2 is split between clusters 3 and 5, CMS 3 is captured by cluster 0, CMS4 overlaps with cluster 4, and cluster 1 is mixed. Figure S3 leads to a similar conclusion, based on a set of molecular phenotypes introduced in[9]: CMS1 and cluster 2 show the hypermutated (Figure S3A), CIMP (S3B), and microsatellite unstable tumors (Figure S3C). They also have similar mutation rates among TP53, APC, KRAS and BRAF, a set of commonly mutated genes in colorectal cancers.

Figures 2C and S3 beg the question of why CMS2 was split into two clusters (3 and 5). In order to investigate whether it is biologically plausible that the CMS group needs to be split into two, we performed a differential expression analysis, identifying marker genes for each cluster. We then ran these lists through a gene set enrichment analysis (See Methods). Figure 2D shows the top pathways associated with each maui cluster. It shows that each cluster is associated with dysregulation in a distinct set of pathways. Specifically, cluster 3 is dominated by TGF-β signaling and leukocyte migration, while cluster 5 is dysregulated in the ErbB, Hippo, and Wnt signaling pathways, demonstrating that these are indeed distinct groups with tumors driven by different oncogenic processes. Further demonstrating this, Figure S1 shows Kaplan-Meier survival curve estimates for clusters 3 and 5, showing that cluster 3 presents a worse prognosis (log-rank $P < 0.001$).

The CMS subtyping scheme, as well as much of the work in the field, is based solely on gene expression profiles. In order to examine whether maui gives better predictions of patient survival with the addition of mutations and copy number data, we also trained a maui model based on gene expression alone. In order to benchmark the survival prediction ability of the maui multi-omics model and the maui-expression-only model, as well as MOFA and iCluster+ models, we first selected, for each model, those latent factors which are individually predictive of survival (See Methods). We then fit a full Cox model using all of the *clinically relevant* factors. For the full Cox models based on clinically relevant latent factors, we computed Harrell's c-Index[21] using 5-fold cross-validation (See Methods). The c-Index is an auc-like measure of prediction accuracy for censored data, with a score of 0.5 being expected by random guessing, and a score of 1.0 being perfect. The maui model based on expression alone achieves a lower score than a maui model with multi-omics data, even when given more genes as input features, indicating that maui is indeed able to find patterns in multi-omics data to which are pertinent to patient survival. maui using multi-omics data also outperforms iCluster+ and MOFA in survival prediction (Figure 1E). The clinically relevant latent factors learned by maui are discussed further in the following section.

One of the advantages of maui over other methods such as iCluster+ and MOFA is that it is able to learn orders of magnitude more latent factors, at a fraction of the computation time (Table 2). In order to compare the utility of maui to that of MOFA and iCluster+, we used a feature-selected data-set with 1,300 genes for most of the analyses presented above (See Methods). We also sought to demonstrate that maui is able to benefit from the computational scalability not offered by MOFA and iCluster+. We repeated the analysis, this time training maui with 6,000 input genes, selecting clinically relevant latent factors, and computing Harrell's c. Figure 1E shows that maui trained using 6,000 genes outperforms maui traiend on 1,300 genes, demonstrating the importance of computational scalability — being more computationally efficient enables learning from more input features,

[i] Some CRC samples do not have a consensus molecular subtype

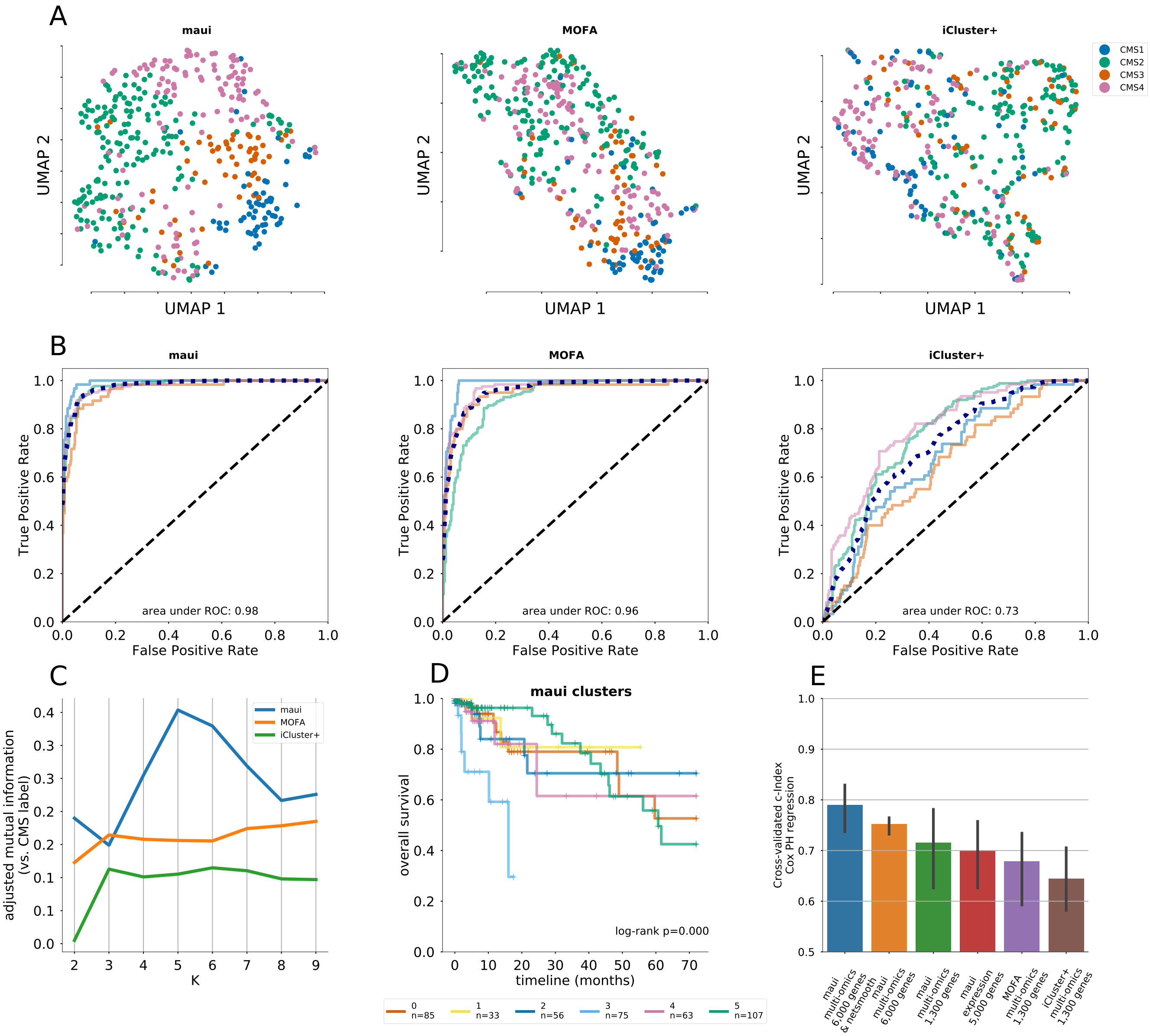


Figure 1. maui, MOFA, iCluster+, and the CMS labels. **A)** UMAP[20] embeddings from latent factors inferred by maui, MOFA, and iCluster+. Each dot represents a tumor, and they are colored by their CMS label. **B)** auROC's for regularized SVM's predicting the CMS label from latent factors (out-of-sample, 10-fold CV). Colored lines are ROCs for each class, with the same colors as in subfigure A. The dotted line is a mean ROC. **C)** The Adjusted Mutual Information (AMI) of clusters obtained from latent factors inferred by maui, MOFA, and iCluster+, using k-means clustering with K ranging from 2 to 9. **D)** Kaplan-Meier estimates and the log-rank statistic for differential survival of different clusters, for the CMS labels and for maui + k-means **E)** Harrel's c-Index for Cox regressions of iCluster+, MOFA, and different flavors of maui.

which brings clinical significance with it.

Finally, we investigated the applicability of using prior information from protein interaction networks for colorectal cancer sub-typing. We and others previously incorporated gene-gene interactions using a method called network-smoothing[22,23]. Network-smoothing is done by allowing binary mutation values to diffuse over a gene network, a process which assigns non-zero "mutation scores" rather than binary mutation values, to genes which either have mutations, or are network-adjacent to mutated genes. We applied the *netSmooth*[23] algorithm (see Methods) to the mutation data prior to passing it to maui and computed Harrell's C index, as above. Figure 1E shows that network smoothing mutations further improves the clinical relevance of latent factors learned when integrating multi-omics data.

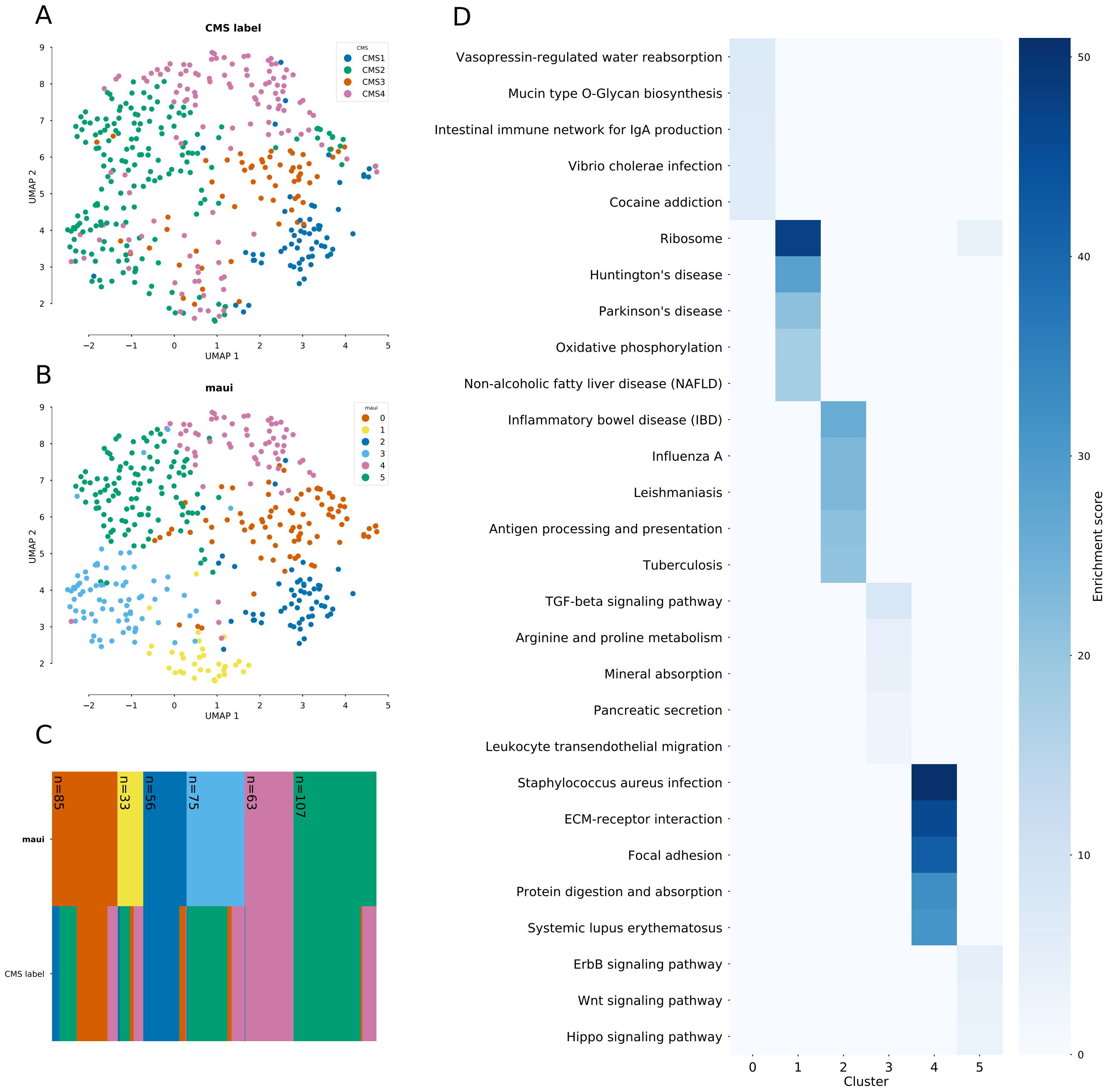


Figure 2. Clustering the tumors using k-means on the latent factors from maui reproduces the CMS labels closely, with the exception of CMS2 being split into two clusters, 3 and 5. **A)** UMAP embedding of tumors colored by CMS label, **B)** UMAP embedding colored by k-means clusters on maui latent factors, **C)** Cluster diagram shows the correspondence between maui clusters and the CMS sub-types: the two rows represent the different labeling schemes (maui clusters and CMS sub-types), and each column represents a sample, which is colored according with its assignment in each row. The legend in subfigures A-B applies to the color scheme in C as well. **D)** Pathways that are enriched in differentially expressed genes for each maui cluster. Clusters show a disjoint set of dysregulated pathways, underlining the different oncogenic processes which underly each group. Cluster 3 and 5 (which together make up the bulk of CMS2) are dominated by dysregulation of TGF-β signaling, and ErbB/Wnt/Hippo signaling, respectively.

CRC latent factors are associated with processes related to tumour progression and development

maui is able to infer many latent factors from multi-omics data. This creates an opportunity to select the most interesting latent factors and treat them as potential biomarkers. In order to demonstrate this, we fit Cox Proportional Hazards models[24], fitting

Method	Notes	# Factors	Runtime
iCluster+	Bayesian, MCMC	10	∼11hrs
MOFA	Bayesian, variational	20	20mins
maui	*Multilevel Bayesian, Stochastic Gradient Descent*	*100*	*3 mins*

Table 2. Summary of methods

one regression model for each factor, as above, selecting clinically relevant latent factors. Figure 3B shows the 95% confidence interval of coefficients for these latent factors, showing that high values for some of these latent factors are predictive of a poor prognosis ($\beta > 0$), while others are predictive of more favorable outcomes ($\beta < 0$). In general, these latent factors can be used as biomarkers with a significant prognostic value.

Beyond the potential to use these latent factors values as biomarkers in order to prognosticate, it is important that we are able to interpret what these potential biomarkers represent. maui is very powerful because it can learn highly non-linear relationships in both the generative model $x \sim p(x|z)$ and the inference model $z \sim p(z|x)$. This comes at a certain cost of making interpretation of the factors less straightforward than in a linear matrix factorization approach like PCA or MOFA. In those algorithms, $x = Wz$, giving us direct access to the loading matrix, W, which allows easy interpretation of the drivers of variation in each latent factor. In order to associate neural latent factors with input features, we correlated input features with neuron activations (see Methods). Figure 3A shows the significant correlations ($P_{adj} < 0.01$, see Methods) of each input feature with each latent factor. For instance, this shows that while most latent factors are active in the gene expression domain, most are not significantly affected by mutation data. By correlating latent factors with input features in this way, we can overcome the difficulties presented by the nonlinear relationships between factors and input features, and use the associations in order to find biologically relevant interpretations for neural latent factors.

When we associated survival-related latent factors with gene ids we observed enrichment of pathways known to play a role in CRC such as Wnt signalling and other APC mediated processes (Figure 3C). In addition, one of the the most significantly survival associated factors are enriched in Neuronal growth factor (NGF) signalling associated genes. NGF signalling, which controls neurogenesis, is associated with aggressive colorectal tumours[25,26]. Survival-relevant latent factors also implicate Platelet-Derived Growth Factor (PDGF) signalling which is also associated with stromal invasion and poor prognosis for colorectal cancer patients[27,28]. Thus, in addition to using latent factors as potential biomarkers for prognosis, we can also point at the underlying biological processes that are uncovered by maui, potentially driving future drug-target studies.

Quality assessment of CRC cell lines as models for tumors

Cancer cell lines often differ significantly in their molecular profiles from tumors, as they face a selection pressure that is different from the natural tumor micro-environment, and acquire genomic alterations necessary to adapt to cell culture[29]. As such, not all colorectal-derived cancer cell lines may be appropriate as models for tumors. Further, as cancer cell lines are maintained over time the risk of contamination and mis-labeling rises. For instance, cell lines which were originally annotated as colorectal cell lines have been shown to be derived from other cancers altogether[30]. Therefore, it would be beneficial if we could tell good models from ones which have diverged too much from tumors in their molecular makeup, or ones which have been mis-labeled or contaminated. To that effect, we examined 54 cancer cell lines originating in colon cancer from the Cancer Cell Line Encyclopedia (CCLE). We used maui to infer latent factor values for the cell lines, so that they could be characterized in the same latent space as the tumors. In the latent space, for each cell line, we compiled a list of nearest neighbors, and counted how many of its nearest neighbors are cell lines, as opposed to tumors. The underlying hypothesis is that cell lines are ill-suited to model tumors if they are more similar to other cell lines than to tumors. This similarity we investigate in the space defined by the latent factors from maui. About half of the colorectal cell lines cell lines belong to a "cell line cluster", meaning a majority of their neighbors are other cell lines (Figure 4A). We singled out cell lines where this proportion is above half, and found among them a mis-labeled cell line: *COLO741*, which has been shown to derive from melanoma and not colorectal cancer[ii]. This finding indicates that this method of flagging cell lines as poor models for tumors by the number of other cell lines in their neighborhood has merit.

In order to further examine this method's utility, in lieu of knowledge of other mis-labeled or otherwise inappropriate colon-derived cancer cell lines, we artificially contaminated the data set by adding to the data random sample of 60 non-colon cell lines, under the assumption that these are poorly suited to the study of colorectal cancer tumors[iii]. We then repeated the exercise of counting nearest-neighbors-that-are-cell-lines, with the additional cell lines. The introduction of these true positives

[ii] In more recent versions of the CCLE annotations, this has been fixed.

[iii] The identities of these "known contaminant" cell lines are irrelevant, as we show later that the method works on 100 such random draws.

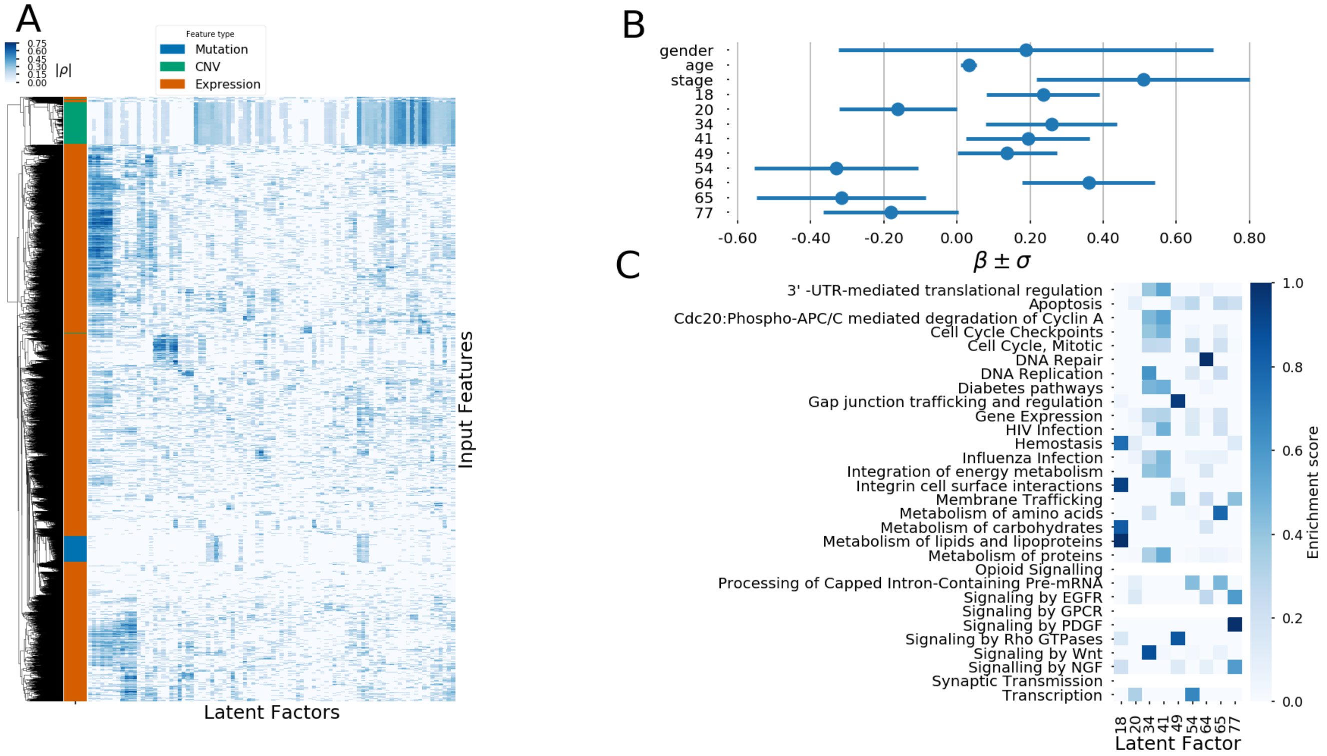


Figure 3. Interpretation of maui latent factors. **A)** A heatmap showing the absolute correlation coefficients of the different input features with the latent factors. Only input features with a mean absolute correlation above a threshold are shown in the heatmap. The row annotation shows the type of input feature, i.e. expression value, mutation, or copy number. **B)** The coefficients in a Cox Proportional Hazards regression, for those latent factors which have significant ($P < 0.05$) effects after controlling for gender, age, and tumor stage. Coefficients also shown for those covariates. **C)** Pathway enrichment scores for genes associated with the latent factors which carry prognostic value (have significant effects in Cox regression).

shows that more cell lines belong to a "cell line cluster" (Figure 4B), and nearly all non-colon derived cell lines have all 5 of their nearest neighbors be other cell lines, while the same is not true for colon-derived cell lines (Figure 4C). Hence, we designated cell lines whose 5 nearest neighbors are other cell lines, as less suitable for the study of tumors (rejected), as these appear to be more similar to other cell lines (even cell lines of other tissues) than to colorectal tumors. Cell lines that have at least one tumor among their 5 nearest neighbors, we retain as suitable models. UMAP embedding of the latent factor space of tumors (with CMS labels, n=419), colorectal cancer cell lines (n=54), and non-colorectal (artificial contamination, n=60) cancer cell lines shows that most contamination (60 ± 1) cell lines are rejected, as well as some of the colon cell lines, and that the non-rejected cell lines are spread among all clusters (Figure 4D). We repeated the analysis with 100 more random draws of 60 additional contaminants, and, rejected any cell line whose 5 nearest neighbors are cell lines. This method consistently rejects almost all known contaminants, as well as about half of the colorectal cell lines (Figure 4E). Note that it is not necessarily a mistake to reject these cell lines; being correctly labeled as originating in colon cancer does not guarantee a cell line to be a good genomic model for a tumor. Moreover, being more similar to non-colon-derived cancer cell lines than to colorectal tumors, is certainly an indication that a particular cell line might not be suitable as a model for colorectal cancers, and the evidence suggests this method successfully rejects almost all known contaminants, indicating that the rejected colon cell lines are likely also poor models. Among the colorectal cell lines, CL40, SW1417, and CW2 are the best candidates, deemed most suitable as models for CRC tumors (Figure 5). On the same scale, cell line COLO320 was one of the lowest ranking cell lines. COLO320 lack mutations in major driver genes in CRC such as BRAF, KRAS, PIK3CA and PTEN, and it has actually a neuroendocrine origin[31,32]. Therefore, COLO320 is possibly a poor model for CRC.

A complete subtyping scheme for CRC and appropriate cell lines for the study of each subtype

The Consensus Molecular Subtyping (CMS) scheme[9] is incomplete as it leaves many tumors without a CMS label. By repeating the clustering analysis, and including also tumors that don't have a CMS designation, as well as cancer cell lines, we were able to use maui to assign subtypes to the remaining non-CMS tumors. By also including the cancer cell lines which were

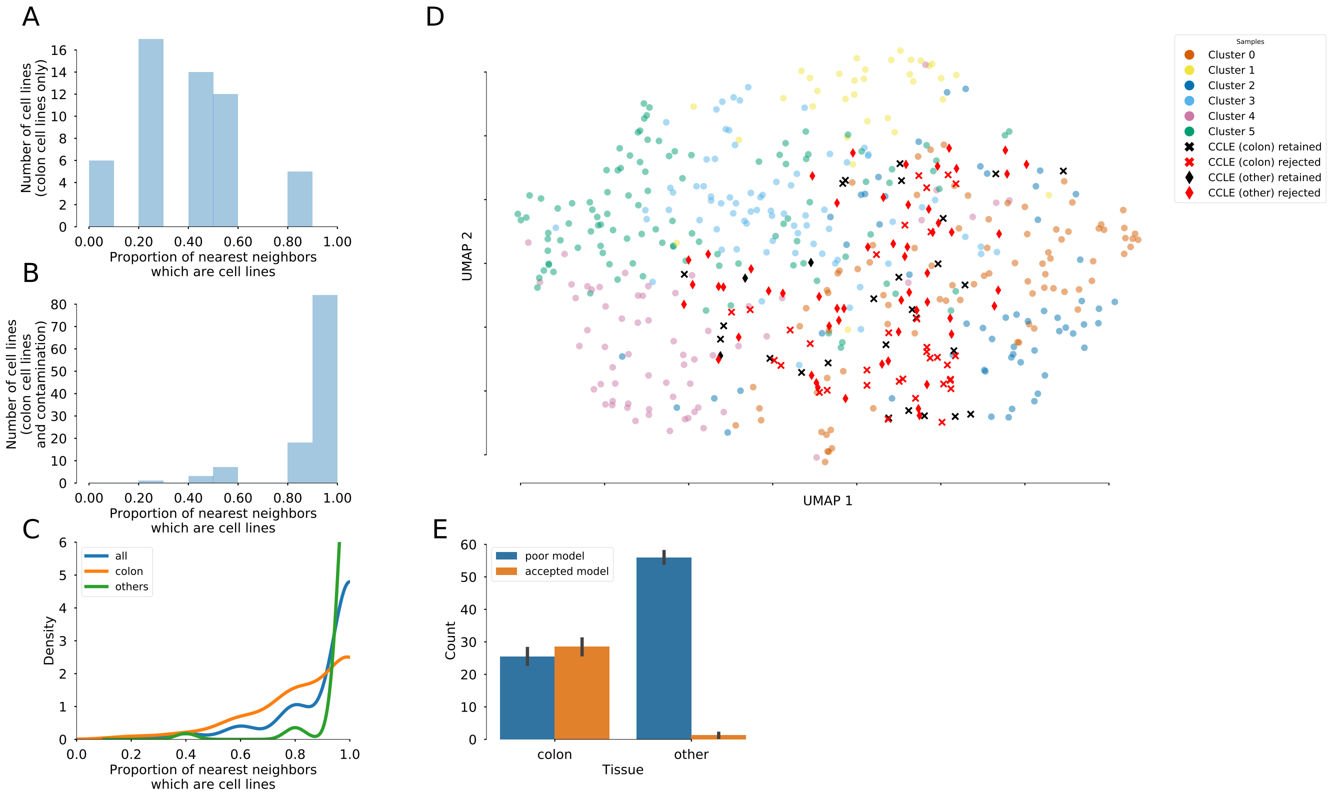
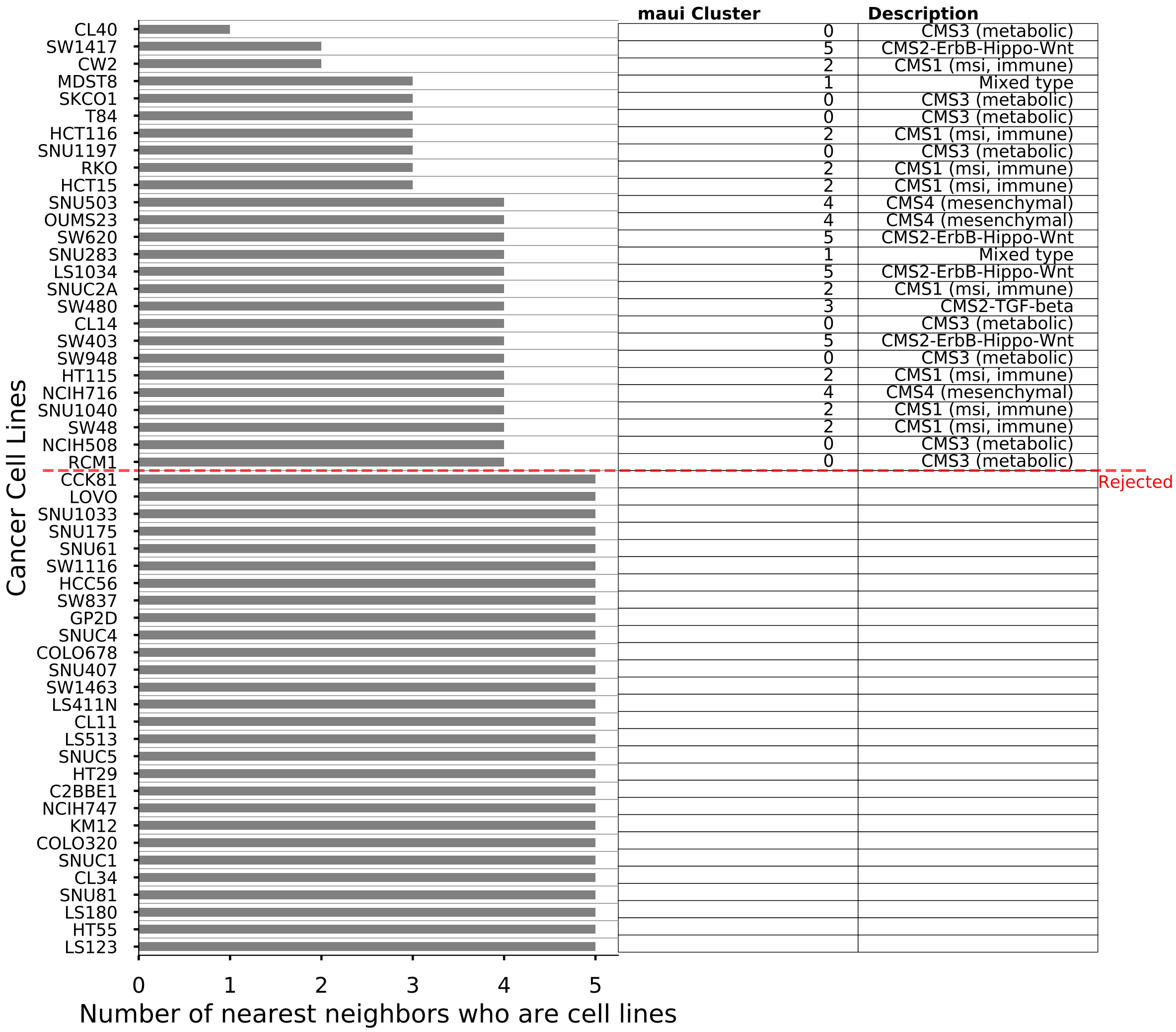


Figure 4. For each cell line, we compiled a list of 5 nearest neighbors in latent factor space, and counted the proportion of those nearest neighbors who are cell lines (as opposed to tumors). Cell lines whose 5 nearest neighbors are all other cell lines, are marked as bad models for tumors, as they are more similar to cell lines than to tumors. **A)** histogram of the proportion of nearest neighbors of cell lines which are also cell lines, colorectal cancers only, **B)**, histogram of the proportion of nearest neighbors of cell lines which are also cell lines, colorectal cancers and non-colorectal cell lines **C)** KDEs of the proportion of cell-line neighbors among all cell lines (colorectal and non-colorectal), broken down by tissue, **D)** UMAP embedding of tumors and cell lines. Crosses are colon-derived cell lines, diamonds are artificial contamination (non-colon derived cancer cell lines). Red cell lines are rejected, black ones are retained as good models. **E)** The proportions of colon and non-colon cell lines which are rejected because their proportion of nearest-neighbor-cell-lines is above the threshold. Nearly all non-colon cell lines are consistently rejected, as well as about half of the colon cell lines.

deemed to be suitable models (see above) in this clustering analysis, we present a novel subtyping scheme for CRC, which covers the whole TCGA cohort (including tumors without a CMS designation), as well as an association of CRC cell lines with these subtypes. We did this using a maui model trained on 6,000 input features, as it is more clinically relevant than the one using 1,300 features (Figure 1E), and produces largely the same cluster assignments as the 1,300 gene model presented above (Figure S4). The non-CMS samples are distributed roughly according to cluster size, as is to be expected for samples that lack a consensus definition (Figure 6A-B), and all clusters have at least one associated cell line (Figure 6C). Table 3 lists the matching cell lines to each cluster, and Figure 5 shows how many of the cell line's nearest neighbors are cell lines, and which cluster it is a best match for. We hope that it can be a useful resource for future drug discovery studies in colorectal cancers.

Discussion

Colorectal cancer (CRC) is a heterogeneous disease, with different sub-types being driven by different kinds of genomic alterations, e.g. hypermutated tumors, tumors showing chromosomal instability, etc. Multi-omics data analysis has the potential to increase the understanding of different subtypes of the disease, and new methods which scale computationally are necessary as the amount of available data increases. Apart from stratifying patients into clinically relevant subgroups, it is necessary to find potential drug targets specific to each subtype. Most drug target discovery studies use cancer models such as cell lines, organoids, or xenografts, and it is thus necessary to match these cancer models to the appropriate subtype in each study, or if a cancer model is inappropriate for the study of any subtype, to be able to flag it as such.



	maui Cluster	Description
CL40	0	CMS3 (metabolic)
SW1417	5	CMS2-ErbB-Hippo-Wnt
CW2	2	CMS1 (msi, immune)
MDST8	1	Mixed type
SKCO1	0	CMS3 (metabolic)
T84	0	CMS3 (metabolic)
HCT116	2	CMS1 (msi, immune)
SNU1197	0	CMS3 (metabolic)
RKO	2	CMS1 (msi, immune)
HCT15	2	CMS1 (msi, immune)
SNU503	4	CMS4 (mesenchymal)
OUMS23	4	CMS4 (mesenchymal)
SW620	5	CMS2-ErbB-Hippo-Wnt
SNU283	1	Mixed type
LS1034	5	CMS2-ErbB-Hippo-Wnt
SNUC2A	2	CMS1 (msi, immune)
SW480	3	CMS2-TGF-beta
CL14	0	CMS3 (metabolic)
SW403	5	CMS2-ErbB-Hippo-Wnt
SW948	0	CMS3 (metabolic)
HT115	2	CMS1 (msi, immune)
NCIH716	4	CMS4 (mesenchymal)
SNU1040	2	CMS1 (msi, immune)
SW48	2	CMS1 (msi, immune)
NCIH508	0	CMS3 (metabolic)
RCM1	0	CMS3 (metabolic)
CCK81		
LOVO		
SNU1033		
SNU175		
SNU61		
SW1116		
HCC56		
SW837		
GP2D		
SNUC4		
COLO678		
SNU407		
SW1463		
LS411N		
CL11		
LS513		
SNUC5		
HT29		
C2BBE1		
NCIH747		
KM12		
COLO320		
SNUC1		
CL34		
SNU81		
LS180		
HT55		
LS123		

Figure 5. For all colon-derived cell lines, we compiled lists of their 5 nearest neighbors. The barplot shows how many of those 5 were other cell lines. Cell lines where all 5 nearest neighbors are other cell lines are rejected, those having at least one nearest neighbor who is a tumor are kept and assigned to clusters, as shown in the table on the right.

We have developed an autoencoder-based method, called **maui**, for integrating data from multi-omics experiments, and demonstrated it using RNA-seq, SNPs and CNVs. The autoencoder infers latent factors which explain the variation across the different data modalities. The latent factors infered by maui capture important biology such as different gene expression programs, gene-based mutational profiles, copy number profiles, and their interactions. We showed that, using maui to learn latent factors in multi-omics data, we get latent factors which are predictive of previously described CRC subtypes (the Consensus Molecular Subtypes, CMS), and outperforms the other methods we benchmarked it against, namely iCluster+ and MOFA. maui also outperforms MOFA and iCluster+ in survival prediction regardless of the CMS subtypes. Performance-wise, maui can extract more latent factors from larger datasets, at a fraction of the computational cost of both iCluster+ and MOFA, making maui better suited to the analysis of the larger datasets we expect to see more of in the future. While maui reproduced the CMS nearly perfectly, it revealed that one of the CMS subtypes, CMS2, is in fact two distinct tumor subtypes, with different survival characteristics, and different underlying gene expression programs, pointing at distinct oncogenic processes.

The latent factors can also be individually associated with the genes they are driven by, as well as by their individual relevance to survival prediction, and when we performed a pathway analysis on the latent factors which are individually most predictive of patient survival, we observed enrichments of pathways which are known to play a role in CRC, such as WNT signaling and other APC-mediated processes, NGF signaling, and PDGF signaling[33]. While the association of latent factors to individual genes is not as straightforward using maui as it is using matrix factorization methods, the relevance of the implicated pathways is promising. We also proposed a way to use the latent factors learned by maui to predict the fitness of cancer cell

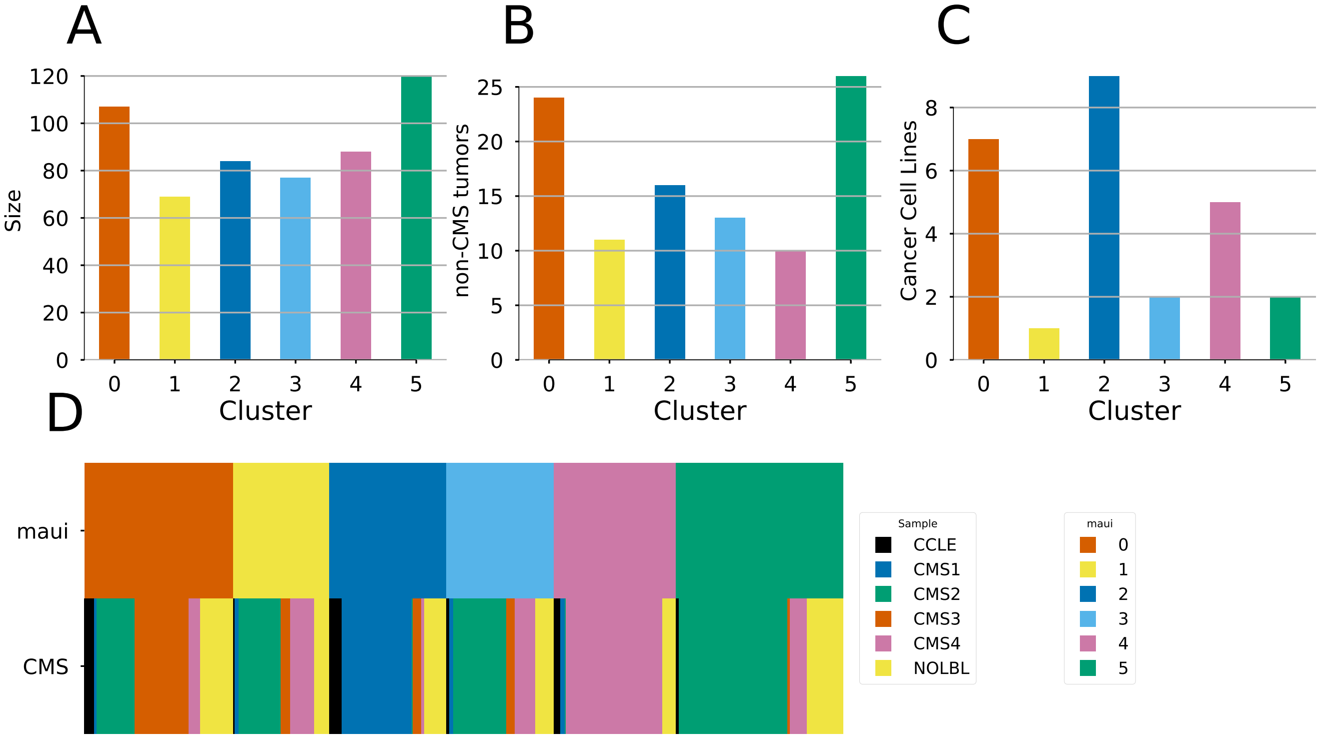


Figure 6. **A)** The sizes (number of samples) of the clusters, **B)** The number of non-CMS tumors assigned to each cluster, **C)** the number of cancer cell lines associated with each cluster **D)** Cluster diagram shows the correspondence between maui clusters and the CMS sub-types: the two rows represent the different labeling schemes (maui clusters and CMS sub-types), and each column represents a sample, which is colored according with its assignment in each row. The *NOLBL* samples without a defined CMS sub-type are distributed among all clusters, as are cancer cell lines (CCLE).

lines as models for CRC generally, as well as for specific subtypes. In order to address the first question, we hypothesized that cell lines which are poor models for the study of CRC tumros will show higher similarity to other cell lines than to CRC tumors. By including non-colorectal cell lines in the sample and computing for each cell line the proportion of samples in its latent-space neighborhood which are also cell lines (as opposed to tumors), a simple threshold of that proportion is enough to correctly predict that 98% of non-colorectal cell lines are poor models for CRC. The threshold method also predicts that approximately 45% of the colorectal cell lines are poor models for CRC, a prediction which still needs to be validated by new experiments. Using this approach, we were able to reject previously known inappropriate cell lines such as COLO 741 and COLO320[30–32]. By including the predicted good model cell lines in the clustering analysis, we were also able to assign CRC subtype specific cell lines, a finding with far reaching potential for subtype specific drug trials. We hope in the future it can be tested whether our approach to predicting fitness of cancer cell lines as models for tumors can be verified, and extended to other cancer models, such as organoids and xenografts. In that way, maui could become an indispensable part of drug discovery pipelines and speed up new therapeutics.

The CRC subtypes we used as a starting point for this study were defined based on gene expression profiles alone. As we wanted to use multi-omics data to refine these subtype definitions, we were limited to a subset of the tumors used in the CMS definition. We used only samples from the TCGA which had measurements for both gene expression, mutations, and copy numbers (n=519), while the CMS study used a larger cohort (n=4,151) and only gene expression profiles. Consequently, it is unclear whether the splitting of the CMS2 subtype into two clusters which we have proposed above would hold when presented with a larger dataset. Only once a larger multi-omics dataset is available will this question be answered.

While the autoencoder architecture of maui is able to do inference in larger data at a fraction of the time compared with matrix factorization methods such as MOFA and iCluster+, the resulting model is more challenging to interpret biologically, i.e. linking genes with latent factors is not as straightforward as in matrix factorization. We have proposed to solve this by using correlations of the input genes and the latent factors, picking the most significant ones heuristically. While we were able to show that such latent factor—gene relationships capture meaningful biology and recapitulate known associations between dysregulation of certain pathways and patient survival, this method is potentially less robust than matrix factorization to these

Cluster	Description	Cell lines
0	CMS3 (metabolic)	SW948, CL14, SNU1197, RCM1, NCIH508, CL40, T84, SKCO1
1	Mixed type	SNU283, MDST8
2	CMS1 (msi, immune)	CW2, HT115, SNU1040, HCT15, SW48, HCT116, RKO, SNUC2A
3	CMS2-TGF-beta	SW480
4	CMS4 (mesenchymal)	OUMS23, SNU503, NCIH716
5	CMS2-ErbB-Hippo-Wnt	SW403, LS1034, SW620, SW1417

Table 3. maui clusters and the cancer cell lines associated with them

associations, and might require more user involvement in the analysis pipeline.

In this study we have developed a deep learning based multi-omics integration method (maui) and shown that it can be used to define clinically relevant subtypes of CRC, as well as predict the fitness of cancer cell lines as models for the study of tumors, and an association of cell lines to CRC subtypes. The latent factors inferred by maui are also interpretable in biological context, and predictive of patient survival, which enables the associations between underlying oncogenic processes, and patient survival. We benchmarked maui against two state-of-the-art methods for multi-omics data integration, and showed that not only is it more effective in defining clinically meaningful subtypes, it also does so with superior computational efficiency. Being orders of magnitude faster will enable maui to be used in studies with larger cohorts and more omics types, as these experiments become more abundant in the future. Further, maui is a general tool for multi-omics integrations, and may be used outside of the cancer context as well, in basic biology studies.

Materials and Methods

Data

We obtained data for tumors from the TCGA-COAD and TCGA-READ project designations of the Genomic Data Commons [iv] using the *TCGAbiolinks* R package[34]. We downloaded the CMS annotations for the TCGA tumors from the Colorectal Cancer Subtyping Consortium (CRCSC) [v]. Table 1 summarizes the subtype information. The gene expression data (mRNA) is HTSeq - FPKM. Mutations were downloaded as MAF files, filtered to include non-synonymous mutations only, and represented as a binary mutation matrix where $m_{ij} = 1$ if and only if gene i carries a non-synonymous mutation in sample j. Copy number alterations are GISTIC calls by gene, represented as a real-valued matrix where c_{ij} is the GISTIC segment mean for the segment containing gene i in sample j.

Cancer Cell Line Encyclopedia data was obtained from the CCLE portal [vi], and is the same data types as the TCGA data, with the exception that transcriptome profiles are RPKM-normalized and not FPKM. We considered 54 cancer cell lines originating from the colon.

We considered only tumors (from TCGA) and cancer cell lines (CCLE) which have "complete data", that is, available measurements in all three assays: gene expression, SNVs, and CNVs.

We used gene-wise MAD statistic, computed directly on the raw data described above, in order to select the most informative features. For the comparisons with MOFA and iCluster+, we used the 1,000 genes with the highest MAD for gene expression, 200 for mutations, and 100 for copy number alterations, for a total of 1,300 input features. We down-sampled the features so strongly in order to make a comparison against iCluster+ viable, and with a larger feature space the runtime would become untenable (Table 2).

For the final clustering analysis, we used a larger feature space, with 5,000 gene expression values, 500 mutations and 500 CNVs for a total of 6,000 features, taking advantage of maui's neural network architecture which allows for larger feature spaces to undergo feature selection as part of the training.

We fit the autoencoder using all TCGA samples, both with and without a CMS label (n=519, Table 1) as well as colon-derived cancer cell lines (n=54), for a total training set size of 573. For the analysis that depends on a CMS label being available, the input features were the latent factors, and the samples only those TCGA samples with a well-defined CMS label (n=419, See Table 1).

[iv] https://portal.gdc.cancer.gov
[v] http://sagebionetworks.org/research-projects/colorectal-cancer-subtyping-consortium-crcsc/
[vi] https://portals.broadinstitute.org/ccle

Network-smoothing of multi-omics data

We applied *netSmooth*[23] to the binary mutation matrix prior to feeding it into the neural network of maui. The method uses the protein-protein interactome (PPI) in order to *smooth* noisy molecular assays, in effect incorporating prior data from countless previous experiments, in order to improve the signal-to-noise-ratio. The intuition behind the method is that genes seldom act alone, and genes in close neighborhoods in the PPI are expected to behave similarly. For instance, interacting proteins tend to be co-expressed[35], and somatic mutations or amplifications/deleteions in adjacent genes (in the PPI) may lead to similar dysfunctions.

The algorithm is a simple *Random Walks with Restarts* diffusion process on the PPI, described by the iterative process

$$F_{t+1} = \alpha A F_t + (1-\alpha) F_0,$$

where F is a data matrix (gene expression, mutations, etc.), A is the degree-normalized adjacency matrix of the PPI, and $(1-\alpha)$ is the restart rate. The process is guaranteed to converge, and has a closed-form solution

$$F_\infty = (I - \alpha A)^{-1} F_0.$$

In order to pick the optimal α value , we performed a grid search over a range between 0 and 1, and picked the lowest α value within 1 standard deviation of the highest score on the Harrel's c-Index benchmark (Figure 1E).

Latent factor model for multi-omics data

Starting from different data matrices x_i from different modalities, we call the full multi-omics data set $x = [x_1, x_2, ..., x_m]$.

We define a generative model $x \sim p(x|z)$. Graphically, our model looks like Figure 7a, a Bayesian latent variable model where the variation in the data x is explained by the variation in a smaller set of latent factors, z. In order to infer the latent variables $z \sim p(z|x)$, as $p(z|x)$ is generally intractable, we proceed with a variational Bayes framework, i.e. approximating $p(z|x) \approx q_\theta(z|x)$, where $q_\theta(z|x)$ is a simple class of distribution, and minimizing the Kullback-Leibler divergence $D_{KL}(q_\theta(z|x) \| p(z|x))$. This is equivalent to maximizing the Evidence Lower Bound (ELBO)[36]:

$$ELBO = E_q[log(p_\phi(x|z))] - D_{KL}(q_\theta(z|x) \| p_\phi(z)).$$

We follow[37] and re-parameterize z_i^l as

$$z_i^l = \mu_i + \sigma_i \varepsilon_l$$

where

$$\varepsilon_l \sim \mathcal{N}(0, \mathbf{I}),$$

which allows us to construct the Autoencoder shown in Figure 7b.

The first half of the autoencoder, leading from x to z (the "encoder") is a neural network which will be trained to compute $q_\theta(z|x)$, that is, θ denotes the weights of the encoder network. The second half, the "decoder" network, is a neural network which will be trained to compute $p_\phi(x|z)$, so ϕ denotes the weights of the decoder network. Thanks to the reparametrization of z, the path from x to $\hat{x}$ is differentiable, via backpropagation, in θ and ϕ, and thus we can use gradient descent to optimize a loss function that is differentiable in θ and ϕ.

Setting the loss function of the neural network to the negative ELBO

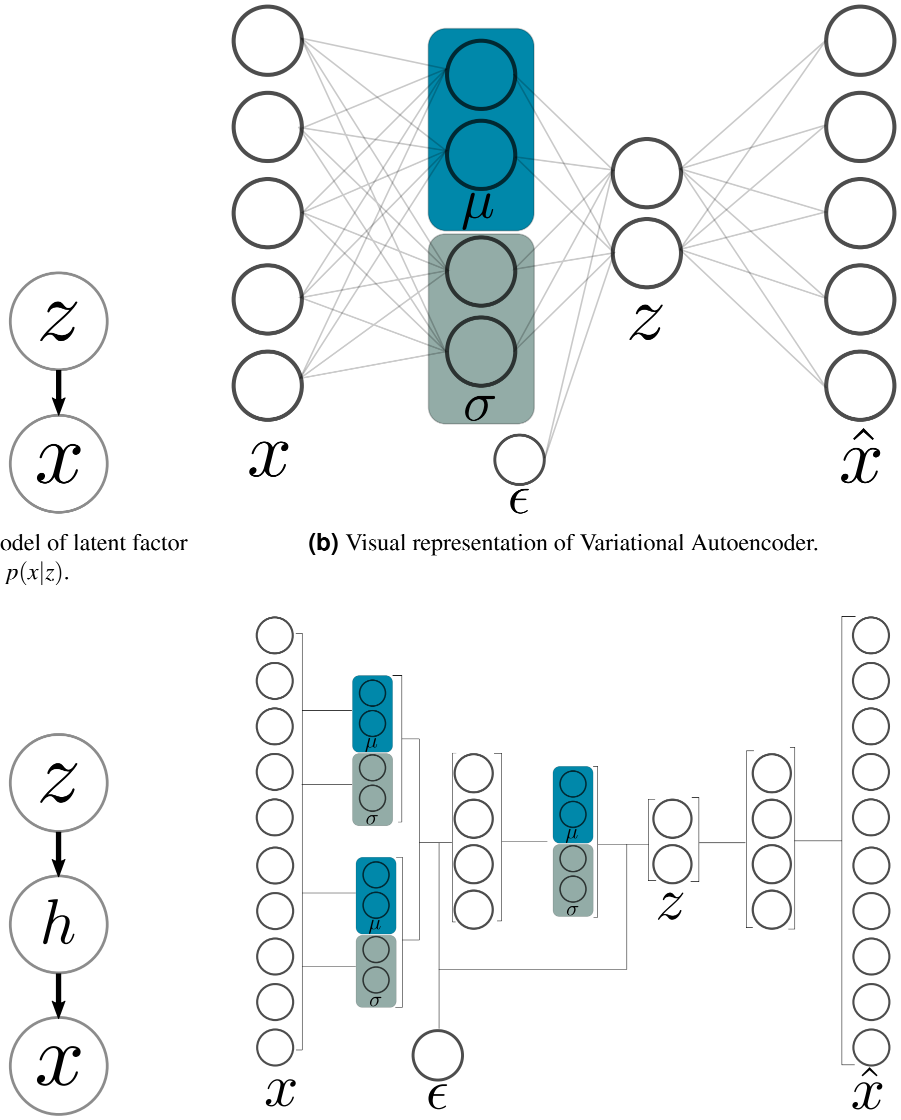
$$l = -E_q[log(p_\phi(x|z))] + D_{KL}(q_\theta(z|x) \| p_\phi(z)),$$

we see that the first term is equivalent to the cross-entropy reconstruction loss, and the second term, the KL-divergence between $q_\theta(z|x)$ and the prior $p_\phi(z)$ can be seen as a regularization term, which will push the z's to their prior distribution.

Stacking autoencoders

The Variational Autoencoder described above is for a one-layer bayesian framework, i.e. Figure 7a. But Autoencoders may be stacked[38] to produce deeper neural network architectures. Deep architectures have more than one layer of nonlinearities, and can thus more compactly capture highly nonlinear functions. We introduce a hidden layer to our Bayesian latent variable model (Figure 7c).

Using the reparametrization trick as above, and specifying the full loss function, inference in the generative model (Figure 7c can be done by backpropagation in the stacked variational autoencoder model (Figure 7d).



(a) Plate model of latent factor model; $x \sim p(x|z)$.

(b) Visual representation of Variational Autoencoder.

(c) Plate model of stacked latent factor model.

(d) Visual representation of Stacked Variational Autoencoder.

Figure 7. Graphical models and neural network schematics of corresponding Autoencoders. a, b: latent variable model. c, d: multilevel latent variable model.

Model regularization

Deep neural networks have many parameters, making them very flexible. This flexibility, however, comes at a cost—deep models are prone to over-fitting: the generation of models which explain the training data well, but generalize poorly to new data. In addition, deep nets are prone to producing complex relationships between many variables. In the case of a latent variable model, that means latent factors that change with the variation of any of a large number of input features, a property which makes the task of interpreting the biological meaning of those latent factors difficult. In technical terms, we wish to enforce sparsity in $q_\theta(z|x)$, so that each latent factor will depend on fewer of the inputs.

In order to address the first issue of potential over-fitting, we use Batch Normalization[39]. When fitting the model, we segment the data into mini-batches, at each iteration computing derivatives and making updates to the model based on that sample. Using Batch Normalization, each feature is scaled and centered in each mini-batch. We feed all of the training examples to the model fitting procedure until the entire training set is exhausted, and then we segment it into new minibatches and repeat the process, for a specified number of epochs. This way, each time a training sample is passed to the model, it will be slightly different, which is roughly equivalent to adding noise, which has been shown to work as a regularizer in Denoising Autoencoders[40] and prevent over-fitting. Further, Batch Normalization addresses another issue - that of Internal Covariate Shift. Internal Covariate Shift happens when the distributions of activations of internal nodes in the neural network changes while training. Reducing Internal Covariate Shift enables us to pick higher learning rates, and thus speeds up inference considerably.

The second mode of regularization, encouraging disentangled representations where latent factors depend only on a few input features, is partially achieved by the KL term in the loss function, as that penalizes distributions of z's which are far from the Gaussian prior. However, it has been shown that the reconstruction loss (the first term in the loss function) is generally much greater than the KL loss[15]. We therefore use their proposed method and add a multiplier to the loss function, allowing us to weigh the relative importance of the terms:

$$l = -E_q[log(p_\phi(x|z))] + \beta D_{KL}(q_\theta(z|x) \| p_\phi(z))$$

In order to ensure the network finds a good representation before it starts regularizing, we use the method proposed by[41], where β is initially 0, and is gradually increased by $\beta = \beta + \kappa$ until its value reaches 1.

Model selection

The stacked VAE presented above is a class of models which are parameterized by the number of hidden units (the dimensionality of h), N_{hidden}, and the number of latent factors, N_{latent}. In order to pick the optimal model, we searched the space spanend by (N_{hidden}, N_{latent}) and computed a compound benchmark score at each point. The compound benchmark score is the average of the scores of: the AUC in the supervised CMS prediction task, the AMI in the unsupervised CMS sub-type prediction task, the $-log_{10}p$ of the multivariate log-rank test for differential survival statistics, and the c-index[21] from the Cox proportional hazards model. The results are presented in Figure 8, where the optimal benchmark score is achieved at $N_{hidden} = 1500$ and $N_{latent} = 80$. Note that the general trend of the score profile is towards better scores at higher N_{latent}, but the optimum we found breaks with that trend, indicating that this level of model complexity achieves a compromise between the most latent factors, and "the curse of dimensionality" which applies to larger latent spaces, and hurts some of the metrics.

Model implementation

We implemented the model using Keras (v2.1.5) using a Tensorflow (v1.6.0) backend. We used Rectified Linear Units for all activations except for the last layer which is Sigmoids, for all features. We trained our network for 600 epochs using mini-batches of size 100 and $\kappa = .01$. We used the Adam optimizer[42].

Association of latent factors with genomic features

The Stacked Variational Autoencoder model described above computes latent factors $z = f(x)$ where $f(x)$ is a nonlinear function which may not necessarily be well-approximated by a linear $z \approx Wx$, as in models such as MOFA or iCluster+. The architecture and depth of the neural network also makes it nontrivial to associate the input genomic features (gene expression, mutations, etc.) with the different latent factors. However, in order to make biological sense of the latent factors, it is necessary to make that association. In order to do that, we computed Pearson's ρ for each latent factor with each input feature, and call a latent factor associated with an input feature if $p < 0.001$.

Gene set enrichment

In order to find out if the genes associated with latent factors (Figure 3), or with clusters (Figure 2) belong to known pathways, we used *Enrichr*[43,44] via the python package *gseapy* [vii]. We used pathways (gene sets) defined by KEGG[45–47].

[vii] version 0.9.4, available from PyPI https://pypi.org/project/gseapy

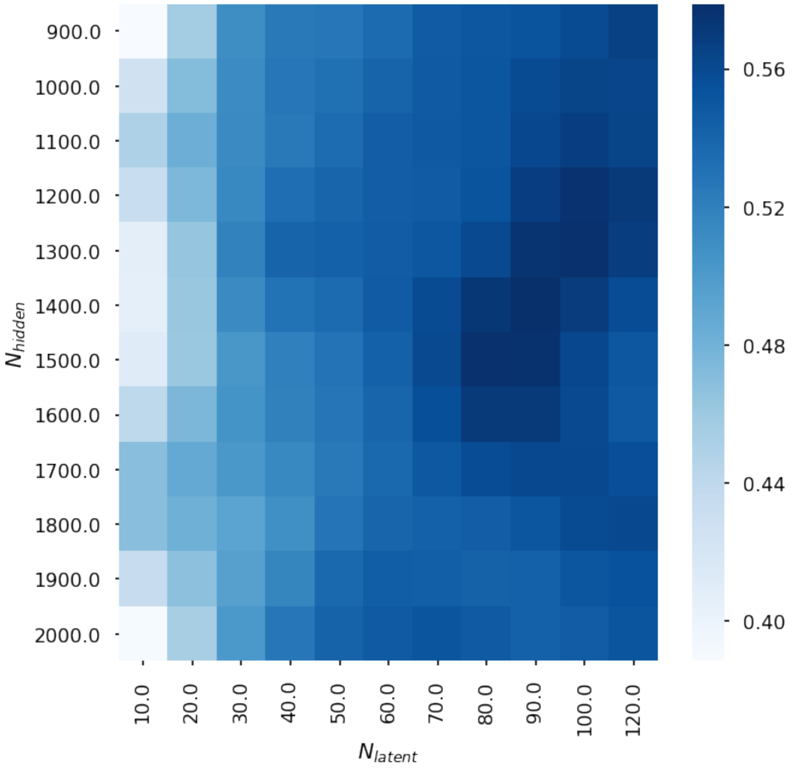


Figure 8. The composite benchmark score in the space defined by N_{hidden}, the number of hidden units, and N_{latent}, the number of latent factors in a model. The optimal parameters are $N_{hidden} = 1500$ and $N_{latent} = 80$

Survival analysis

We rely on overall survival data from the TCGA annotations for all survival analyses.

In order to assess the prognostic value of latent factors inferred by our deep learning approach, we fit a Cox Proportional Hazards model[24],

$$ln \frac{h(t)}{h_o(t)} = \sum_i \beta_i x_i,$$

where the left hand side is the logarithm of the hazard ratio, and x's are co-variates. We assess the predictive value of each latent factor separately, while controlling for the patient's age, gender, and tumor stage at diagnosis. We compute confidence intervals for the coefficient β associated with the latent factor, and pick the latent factors with FDR-correction and $\alpha = 0.95$.

In order to compare the prognostic value of different models, we compute the c-index[48–50] and use 5-fold cross-validation[51].

The log-rank statistics reported in Figure 1D and Figure S2 are multivariate log-rank test, under a null hypothesis that all groups have the same survival function, with an alternative hypothesis that *at least one group* has a different survival function.

All survival analysis was done using the python package *lifelines* [viii].

Comparing models' survival-predictive value

In order to compare maui to MOFA and iCluster+ (as well as to a gene expression only-based maui model), we used Harrell's C[21] in a Cox Proportional Hazards[24] regression model. The c-Index was computed for Cox models based only on clinically relevant factors, which we select using individual, unregularized Cox models, one per factor, while controlling for patient age, sex, and tumor stage. In those individual factor models, we used Efron's method to compmute confidence intervals, and only kep the latent factors with statistically significant (adjusted P-value < 0.05) nonzero coefficients in the individual Cox models. Having selected clinically relevant latent factors from each model (maui, MOFA, iCluster+, maui-expression, maui-netsmooth), we fit a full Cox regression using those, and ran a cross validated out-of-sample c-Index calculation using regularized Cox PH regression, searching for the optimal result among the regularizers $1, 10, 100, 1000, 10000$. The results reported in Figure 1F are the best regularized model for each of the methods.

[viii] https://lifelines.readthedocs.io/en/latest/

Quality assessment of CRC cell lines for modeling tumors

In order to assess the fitness of different cancer cell lines as models for tumors, we computed the pairwise euclidean distance between each of the samples (TCGA and CCLE), in the space of the latent factors derived from maui. Then, we computed, for each cell line, the proportion of its 5 nearest neighbors which are also cell lines, the working hypothesis being that cell lines that form "cell line clusters" are more cell-line like than tumor like, and likely less fit as models for tumors. The choice of $K = 5$ for the number of nearest neighbors is immaterial, as the method is largely insensitive to the choice of K among a wide range (Figure S5).

References

1. Bowel cancer statistics, cancer research uk. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer. Accessed: 2018-10-31.

2. Torre, L. A., Siegel, R. L., Ward, E. M. & Jemal, A. Global cancer incidence and mortality rates and trends–an update. *Cancer Epidemiol. Biomarkers & Prev.* **25**, 16–27, DOI: 10.1158/1055-9965.epi-15-0578 (2015).

3. Müller, M. F., Ibrahim, A. E. K. & Arends, M. J. Molecular pathological classification of colorectal cancer. *Virchows Arch.* **469**, 125–134, DOI: 10.1007/s00428-016-1956-3 (2016).

4. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

5. Fearon, E. R. Molecular genetics of colorectal cancer. *Annu. Rev Pathol* **6**, 479–507 (2011).

6. Parsons, D. W. *et al.* Colorectal cancer: mutations in a signalling pathway. *Nature* **436**, 792 (2005).

7. Lao, V. V. & Grady, W. M. Epigenetics and colorectal cancer. *Nat. Rev. Gastroenterol. & Hepatol.* **8**, 686–700, DOI: 10.1038/nrgastro.2011.173 (2011).

8. Toyota, M. *et al.* CpG island methylator phenotype in colorectal cancer. *Proc. Natl. Acad. Sci.* **96**, 8681–8686, DOI: 10.1073/pnas.96.15.8681 (1999).

9. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015). [PubMed Central:PMC4636487] [DOI:10.1038/nm.3967] [PubMed:26457759].

10. Vlachogiannis, G. *et al.* Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. *Science* **359**, 920–926, DOI: 10.1126/science.aao2774 (2018).

11. Tini, G., Marchetti, L., Priami, C. & Scott-Boyer, M.-P. Multi-omics integration—a comparison of unsupervised clustering methodologies. *Briefings Bioinforma.* DOI: 10.1093/bib/bbx167 (2017).

12. Trunk, G. V. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis Mach. Intell.* **PAMI-1**, 306–307, DOI: 10.1109/tpami.1979.4766926 (1979).

13. de Tayrac, M., Le, S., Aubry, M., Mosser, J. & Husson, F. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics* **10**, 32 (2009).

14. Mo, Q. *et al.* Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 4245–4250 (2013).

15. Li, Y. *et al.* Disentangled variational auto-encoder for semi-supervised learning (2017). arXiv:1709.05047.

16. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *bioRxiv* DOI: 10.1101/174474 (2017). https://www.biorxiv.org/content/early/2017/10/02/174474.full.pdf.

17. Cadena, C., Dick, A. & Reid, I. D. Multi-modal auto-encoders as joint estimators for robotics scene understanding. In *Robotics: Science and Systems XII*, DOI: 10.15607/rss.2016.xii.041 (Robotics: Science and Systems Foundation).

18. Gligorijević, V., Barot, M. & Bonneau, R. deepNF: deep network fusion for protein function prediction. *Bioinformatics* DOI: 10.1093/bioinformatics/bty440 (2018).

19. Argelaguet, R. *et al.* Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124, DOI: 10.15252/msb.20178124 (2018).

20. McInnes, L. & Healy, J. Umap: Uniform manifold approximation and projection for dimension reduction (2018). arXiv:1802.03426.

21. Pencina, M. J. & DAgostino, R. B. OverallC as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat. Medicine* **23**, 2109–2123, DOI: 10.1002/sim.1802 (2004).

22. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115, DOI: 10.1038/nmeth.2651 (2013).

23. Ronen, J. & Akalin, A. netSmooth: Network-smoothing based imputation for single cell RNA-seq. *F1000Res* **7**, 8 (2018).

24. Breslow, N. E. Analysis of survival data under the proportional hazards model. *Int. Stat. Rev. / Revue Int. de Stat.* **43**, 45–57 (1975).

25. Jobling, P. *et al.* Nerve-Cancer Cell Cross-talk: A Novel Promoter of Tumor Progression. *Cancer Res.* **75**, 1777–1781 (2015).

26. Liebig, C. *et al.* Perineural invasion is an independent predictor of outcome in colorectal cancer. *J. Clin. Oncol.* **27**, 5131–5137 (2009).

27. Kitadai, Y. *et al.* Expression of activated platelet-derived growth factor receptor in stromal cells of human colon carcinomas is associated with metastatic potential. *Int. J. Cancer* **119**, 2567–2574 (2006).

28. Steller, E. J. *et al.* PDGFRB promotes liver metastasis formation of mesenchymal-like colorectal tumor cells. *Neoplasia* **15**, 204–217 (2013).

29. Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330, DOI: 10.1038/s41586-018-0409-3 (2018).

30. Medico, E. *et al.* The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nat Commun* **6**, 7002 (2015).

31. Ahmed, D. *et al.* Epigenetic and genetic features of 24 colon cancer cell lines. *Oncogenesis* **2**, e71 (2013).

32. Berg, K. C. G. *et al.* Multi-omics of 34 colorectal cancer cell lines - a resource for biomedical studies. *Mol. Cancer* **16**, 116 (2017).

33. Kuipers, E. J. *et al.* Colorectal cancer. *Nat Rev Dis Primers* **1**, 15065 (2015).

34. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).

35. Bhardwaj, N. & Lu, H. Correlation between gene expression profiles and protein–protein interactions within and across genomes. *Bioinformatics* **21**, 2730–2738, DOI: 10.1093/bioinformatics/bti398 (2005). /oup/backfile/content_public/journal/bioinformatics/21/11/10.1093/bioinformatics/bti398/2/bti398.pdf.

36. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. DOI: 10.1080/01621459.2017.1285773 (2016). arXiv:1601.00670.

37. Kingma, D. P. & Welling, M. Auto-encoding variational bayes (2013). arXiv:1312.6114.

38. Bengio, Y., Lamblin, P., Popovici, D. & Larochelle, H. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, 153–160 (2007).

39. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift (2015). arXiv:1502.03167.

40. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **11**, 3371–3408 (2010).

41. Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K. & Winther, O. Ladder variational autoencoders (2016). arXiv:1602.02282.

42. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization (2014). arXiv:1412.6980.

43. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinforma.* **14**, 128 (2013).

44. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–97 (2016).

45. Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30, DOI: 10.1093/nar/28.1.27 (2000).

46. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462, DOI: 10.1093/nar/gkv1070 (2015).

47. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361, DOI: 10.1093/nar/gkw1092 (2016).

48. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546 (1982).

49. Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. & Rosati, R. A. Regression modelling strategies for improved prognostic prediction. *Stat Med* **3**, 143–152 (1984).

50. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* **15**, 361–387 (1996).

51. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics (Springer New York Inc., New York, NY, USA, 2001).

Acknowledgements

The work shown here is based upon data generated by the TCGA Research Network [ix] and the Cancer Cell Line Encyclopedia [x].

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing interests** (mandatory statement).

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

[ix] http://cancergenome.nih.gov/

[x] https://portals.broadinstitute.org/ccle

Supplementary material

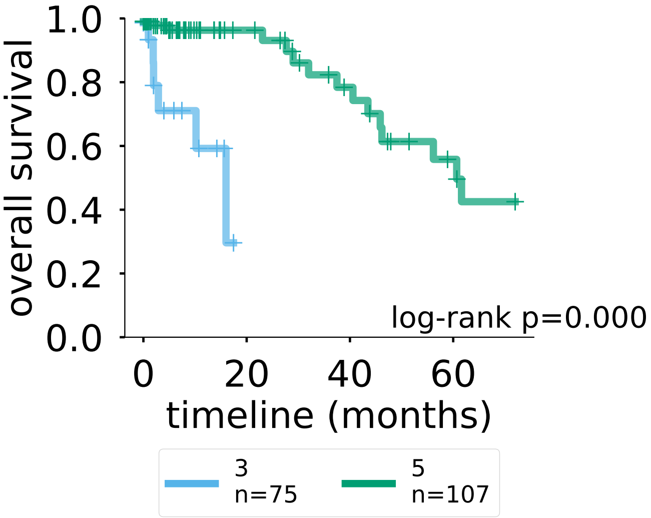


Figure S1. Kaplan-Meier curves for maui clusters 3 and 5. Cluster 3 appears to be more aggressive tumors with a worse prognosis ($P < 0.001$).

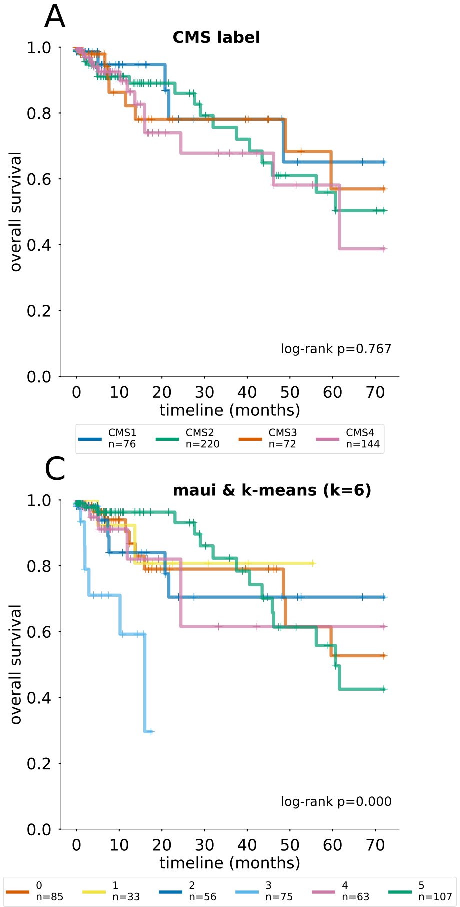


Figure S2. Kaplan Meier curves and log-rank tests for differential survival statistics for the CMS sub-types, as well as maui clusters using k-means with different K's.

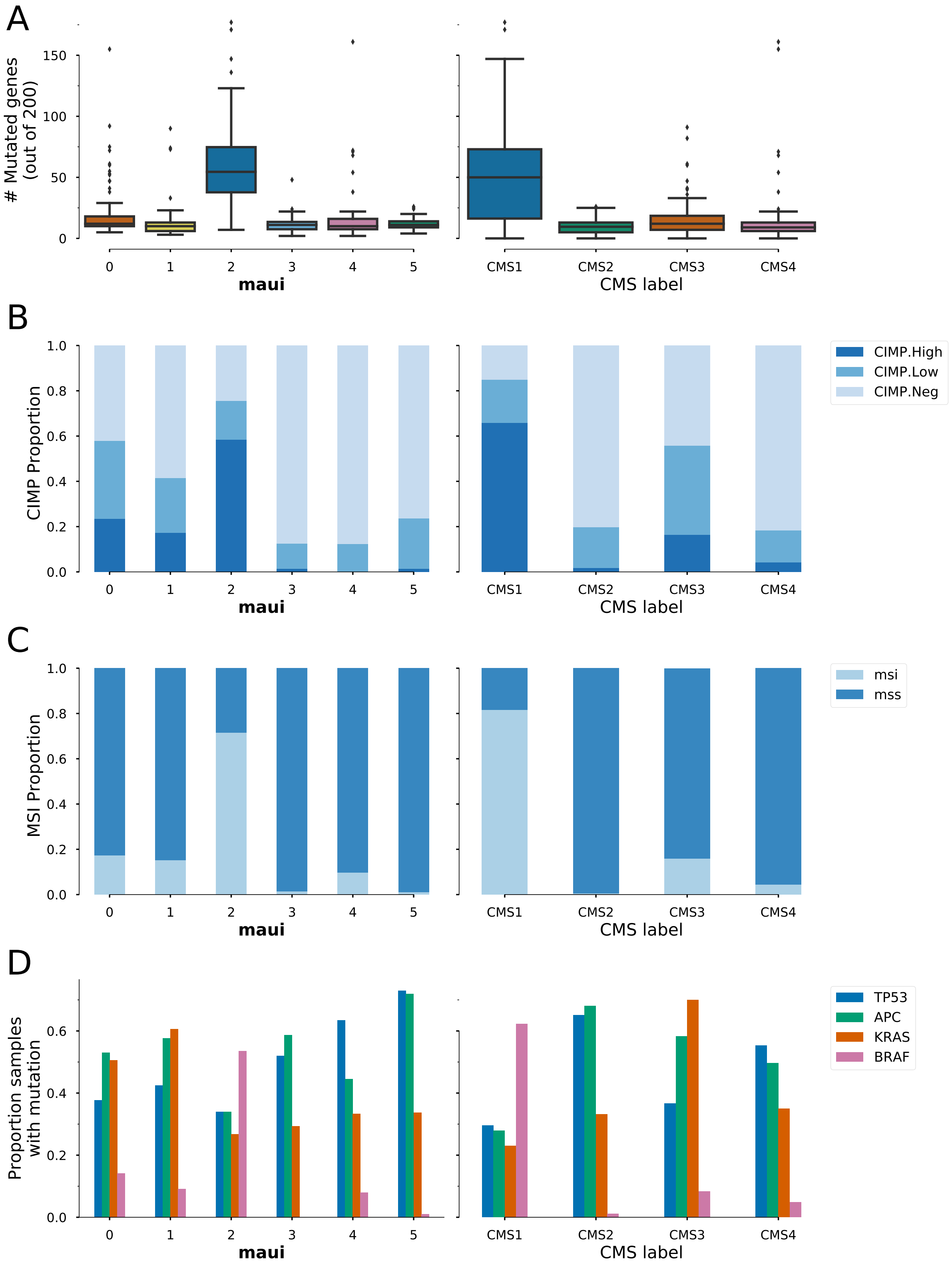


Figure S3. Molecular markers and their distribution in CMS sub-types (left column) and maui clusters (right column). **A)** utational load, **B)** CIMP phenotype, **C)** Microsatellite instability, and **D)** the prevalence of mutations in a key set of colorectal cancer genes.

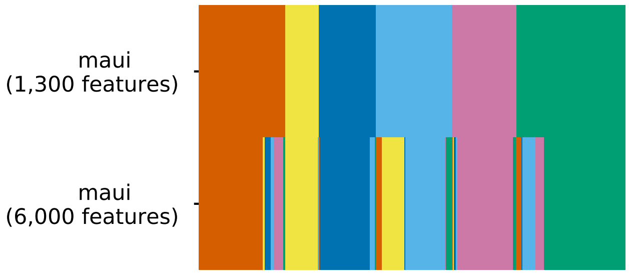


Figure S4. Correspondence of maui clusters when training using 1,300 genes and 6,000 genes. Each column is a sample, and they are colored by their cluster assignment. Clusters are mostly the same when using more input features, with some refinements taking place.

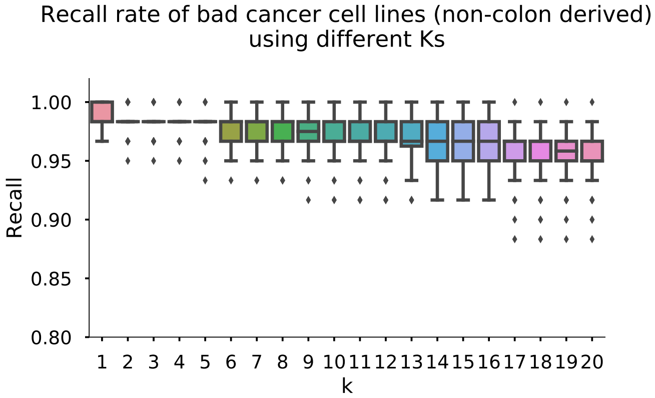


Figure S5. We repeated the exercise of Figure 4E, that is, adding non-colon cell lines to the mix, and calculating the proportion of each cell line's K nearest neighbors, that are also cell lines (as opposed to tumors). Setting the threshold at 0.95, the method correctly identifies most non-colon cell lines as bad models for colorectal tumors. The recall rate is $\frac{\text{predicted bad models among non-colon cell lines}}{\text{number of non-colon cell lines}}$, and is largely insensitive to the choice of K.