Diagnostic Evidence GAuge of Single cells (DEGAS): A transfer learning framework to infer impressions of cellular and patient phenotypes between patients and single cells.

Travis S. Johnson[1,2], Christina Y. Yu[1,2], Zhi Huang[3], Siwen Xu[4], Tongxin Wang[5], Chuangpeng Dong[4], Wei Shao[1], Mohammed Abu Zaid[1], Yijie Wang[3], Christopher Bartlett[6], Yan Zhang[2,7], Yunlong Liu[8], Jie Zhang[8*], Kun Huang[1,8,9*]

[1]Department of Medicine, Indiana University School of Medicine
[2]Department of Biomedical Informatics, College of Medicine, The Ohio State University
[3]School of Electrical and Computer Engineering, Purdue University
[4]School of Informatics and Computing, Indiana University
[5]Department of Computer Science, Indiana University
[6]Nationwide Children's Hospital
[7]The Ohio State University Comprehensive Cancer Center (OSUCCC - James)
[8]Department of Medical and Molecular Genetics, Indiana University School of Medicine
[9]Regenstrief Institute
* To whom correspondence should be addressed (jizhan@iu.edu or kunhuang@iu.edu)

Abstract

With the rapid advance of single cell sequencing techniques, single cell molecular data are quickly accumulated. However, there lacks a sound approach to properly integrate single cell data with the existing large amount of patient-level disease data. To address such need, we proposed DEGAS (Diagnostic Evidence GAuge of Single cells), a novel deep transfer-learning framework which allows for cellular and clinical information, including cell types, disease risk, and patient subtypes, to be cross-mapped between single cell and patient data, provided they share at least one common type of molecular data. We call such transferrable information "impressions", which are generated by the deep learning models learned in the DEGAS framework. Using eight datasets from a wide range of diseases including Glioblastoma Multiforme (GBM), Alzheimer's Disease (AD), and Multiple Myeloma (MM), we demonstrate the feasibility and broad applications of DEGAS in cross-mapping clinical and cellular information across disparate single cell and patient level transcriptomic datasets. Specifically, we correctly mapped clinically known GBM patient subtypes onto single cell data. We also identified previously known neuron loss from AD brains, then mapped the "impression" of AD risk to single cell data. Furthermore, we discovered novel differences in excitatory and inhibitory neuron loss in AD data. From the exploratory MM data, we identified differences in the malignancy of different CD138+ cellular subtypes based on "impressions" of relapse information transferred from MM patients. Through this work, we demonstrated that DEGAS is a powerful framework to cross-infer cellular and patient-level characteristics, which not only unites single cell and patient level transcriptomic data by identifying their latent links using the deep learning approach, but can also prioritize both patient subtypes and cellular subtypes for precision medicine.

Introduction

Large data consortia containing a variety of omics data are widely available for many disease types, which allow researchers to identify multi-level omic perturbations that are associated with disease status and clinical outcomes. Unfortunately, most of such consortia do not contain assays specifically addressing tissue heterogeneity at the cellular level. On the other hand, databases and portals have quickly accumulated with single cell RNA sequencing datasets (scRNA-seq), such as Hemberg lab [1], scRNASeqDB [2], SCPortalen [3], Allen Institute Cell Types Database, and the NCBI Gene Expression Omnibus (GEO) [3]. However, these single cell databases all lack enough patient clinical information to assess how the heterogeneous cell types affect clinical outcomes at the patient level.

In order to transfer the molecular heterogeneity information we learn from scRNA-seq data and apply it to patient-level analysis, there is an urgent need in methodology development to integrate both data types and identify hidden links between the two. However, such integration faces a lot of challenges as different data modalities and difference data sources can have different characteristics, such as quantity, quality, distribution and resolution of the data [4]. For instance, it is common to find studies with a large number patient samples of bulk tissue RNA-seq, whereas studies with scRNA-seq data usually contains a small number of patient samples. Also, most scRNA-seq experiments generate a large number of cells per sample, making the scaling of such data to multiple tissue samples computationally difficult [4]. On top of this, a large patient sample size is often required for statistical studies such as outcome prediction and survival analysis. If traditional methods were used, the resulting scRNA-seq data could end up with cell numbers in the scale of millions. To address such challenges as sample size and computational cost, in this study, we establish a transfer learning framework DEGAS (Diagnostic Evidence GAuge of Single cells) to integrate studies of scRNA-seq and bulk tissue RNA-seq data with the goal to identify the hidden links between the two. Through cross-mapping, DEGAS identifies disease associated cell subtypes while at the same time dissecting patient bulk tissue data into corresponding cell types. The DEGAS framework in its simplest form can be broken into three tasks: 1) correctly labeling cells with a cellular subtype using multitask learning, 2) correctly assigning proportional hazards or clinical labels to patients using multitask learning, and more importantly, 3) generating a subspace for cross-mapping where the patients and cells are comparable using domain adaptation.

Taking transcriptomic data as an example, the rationale behind the DEGAS framework is that since scRNA-seq data and patient-level transcriptomic data share the same set of genes (feature space), there must exist a natural connection between the two data types that can be leveraged to further identify the associations between patients and cells and even cross-map the traits from one data type to the other. Viewing this association as a graph (**Fig. 1**), we can connect the outcomes in patients to the groups of cells, *i.e.*, subtypes, via the common feature space (gene set) between the two. The expression patterns of the genes should also carry at least part of the same biological patterns such as molecular pathways, signaling cascades, and metabolic processes, making the information/knowledge learned from such portion of gene expression patterns transferable between patients and cells. Our assumption is that information

learned from these shared gene expression patterns are simultaneously predictive of both patient outcomes and cellular subtypes.

To determine the link between single cell data and the associated disease state is not new – previous methods mainly utilize unsupervised learning and focused primarily on the number of differentially expressed genes (DEGs) in a given cell type corresponding to some clinical outcome [5, 6]. For example, Gawel *et al.* used enrichment of the cell cluster specific DEGs and multicellular disease models (MCDMs) to visualize the cell type prioritization [7]. Alternatively, Augur did not rely primarily on DEGs, since it decreased the biological resolution to cell type level [8]. Instead, they trained classifiers on each cell type with respect to the disease state of the tissue from which those cells were sampled. The accuracy of the classifier in each cell type was used to prioritize the cell types in relation to the disease state of interest [8]. Both of these methods rely on either prior knowledge to calculate enrichment of DEGs or clinical measurements for each subject from which single cells were extracted.

However, in our DEGAS workflow, we try to incorporate the patient level outcome information with cell type from disparate datasets to perform “cell type prioritization” on scRNA-seq data of a disease that can be attributed to disease-related biological perturbations. This is a novel neural network approach for cell type prioritization, and we hope to achieve the goals that i) train a model simultaneously on both single cell data with cellular information as a label, and patient data with patient information as a label where at least one set of labels, patient or single cell, is available for training; ii) the model is established in such a way that the data distributions between the single cells and patients are reconciled. Multitask learning, a type of transfer learning, is precisely designed to achieve these two goals. Used extensively in computer vision, multitask learning learns a low dimensional representation of the input data to optimally address multiple tasks. Examples of such application in medical science include predicting benign versus malignant tumor samples, as well as subclassification in breast cancer histology images [9, 10]. Recently a new method called SAUCIE has been proposed to denoise the scRNA-seq data, cluster cells, and cluster patients using a multitask learning approach [11]. We further extend this line of research to include datasets with patient outcomes that can be trained simultaneously so that the outcomes can be transferred or cross-mapped between single cells and patients, that is, transfer single cell knowledge learned from deep learning models to patients and patient knowledge to single cells. The major advantage of such transfer learning framework is that, the single cell patients and clinical bulk expression patients, from which the outcomes are being learned, can come from different cohorts. This flexibility not only presents an ingenious way to integrate molecular omic data analysis in different levels, but also virtually merges them into the same cohort, which makes studying a broad variety of heterogeneous diseases possible.

To perform DEGAS, first we obtain the common molecular information, in our case, transcriptomic data, from single cell level (scRNA-seq data) and patient level (bulk expression data); secondly, we apply deep learning models to further learn the connections between the single cell and patient-level gene expression patterns, with the goal to simultaneously minimize a) cell type classification error; b) patient outcome prediction error; and c) the data variance

between the two gene expressions in the hidden layer. Finally, the patient-level clinical data such as survival and clinical subtypes from the patient expression data will be cross-mapped to the populations of single cells in the scRNA-seq data. Similarly, the cell type information from single cells can be cross-mapped to the patient samples. As an analogy to visual perception, we call these transferrable features “impressions” since information from gene expression of disparate data types and studies can be extracted and the characteristics from one data type can be mapped to the other.

DEGAS is developed as a generalizable transfer learning framework/model that can be applied to any disease data as long as the data contain: 1) clinical information for a cohort of patients and/or 2) a separate clustering analysis result on sets of cells from scRNA-seq experiments of the same disease. Either cell type labels or patient outcome labels, but not both, can also be omitted to produce a DEGAS model which only generates “impressions” in one direction. Using this transfer learning framework, these two types of data can be leveraged to construct a DEGAS model and infer cross-mappable results (impressions). We show the general model and workflow in **Fig. 1**. More detailed information is described in the Methods section.

To demonstrate the feasibility and effectiveness of the DEGAS framework, we first tested DEGAS on glioblastoma (GBM) transcriptomic data, which has ground-truth labels of cancer subtypes in both bulk tissue cohorts and singe cell samples. Then we applied it to Alzheimer’s disease (AD) studies in which neuron loss is known. Finally, as an exploratory tool, we applied DEGAS to study multiple myeloma (MM) transcriptomic data, where the disease subtypes and high-risk factors associated with single cells are largely unknown. MM stems from the proliferation of aberrant clonal plasma cells in the bone marrow that secrete monoclonal immunoglobulin protein. It is the second most common blood cancer in the United States and 32,110 new cases will be diagnosed with 12,960 estimated deaths in the United States in 2019 [12]. We combined our newly generated MM scRNA-seq data from six local samples and bulk tissue data from the Multiple Myeloma Research Foundation, then applied DEGAS to infer clinical impressions for plasma cell subtypes and successfully identified a patient subgroup with high risk of relapse.

Results

DEGAS clinical impression framework

In this study, we applied DEGAS to integrate and analyze scRNA-seq and clinical data from three different diseases: GBM, AD and MM. Since the ground truth of GBM subtypes are known, and the neuron loss in AD brains are also known, the GBM and AD datasets primarily served as validation to demonstrate the feasibility and universality of the DEGAS transfer learning approach with high accuracies. We then further expand our study to MM data, which serves as the discovery dataset, since the plasma cell subtypes and high-risk factors are largely unknown for MM. In the MM study, we applied DEGAS on patient data from the Multiple Myeloma Research Foundation (MMRF) and scRNA-seq data from patients at Indiana University School of Medicine (IUSM). Our aim was to identify the cell subtypes as well as a

high-risk subgroup of patients using the impressions of relapse risk on the single cells and impressions of cellular subtype in MM patients. We then applied the results to two separate MM validation datasets, one of which contained plasma cells from normal bone marrow (NHIP), MM precursor conditions: monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM), and MM. We wanted to observe if DEGAS assignment of relapse risk to cell subtypes were higher for more malignant conditions. An additional external validation dataset of patient level expression data with overall survival was used to evaluate whether the patient stratification learned by DEGAS was robust enough to be generalized to an external survival dataset.

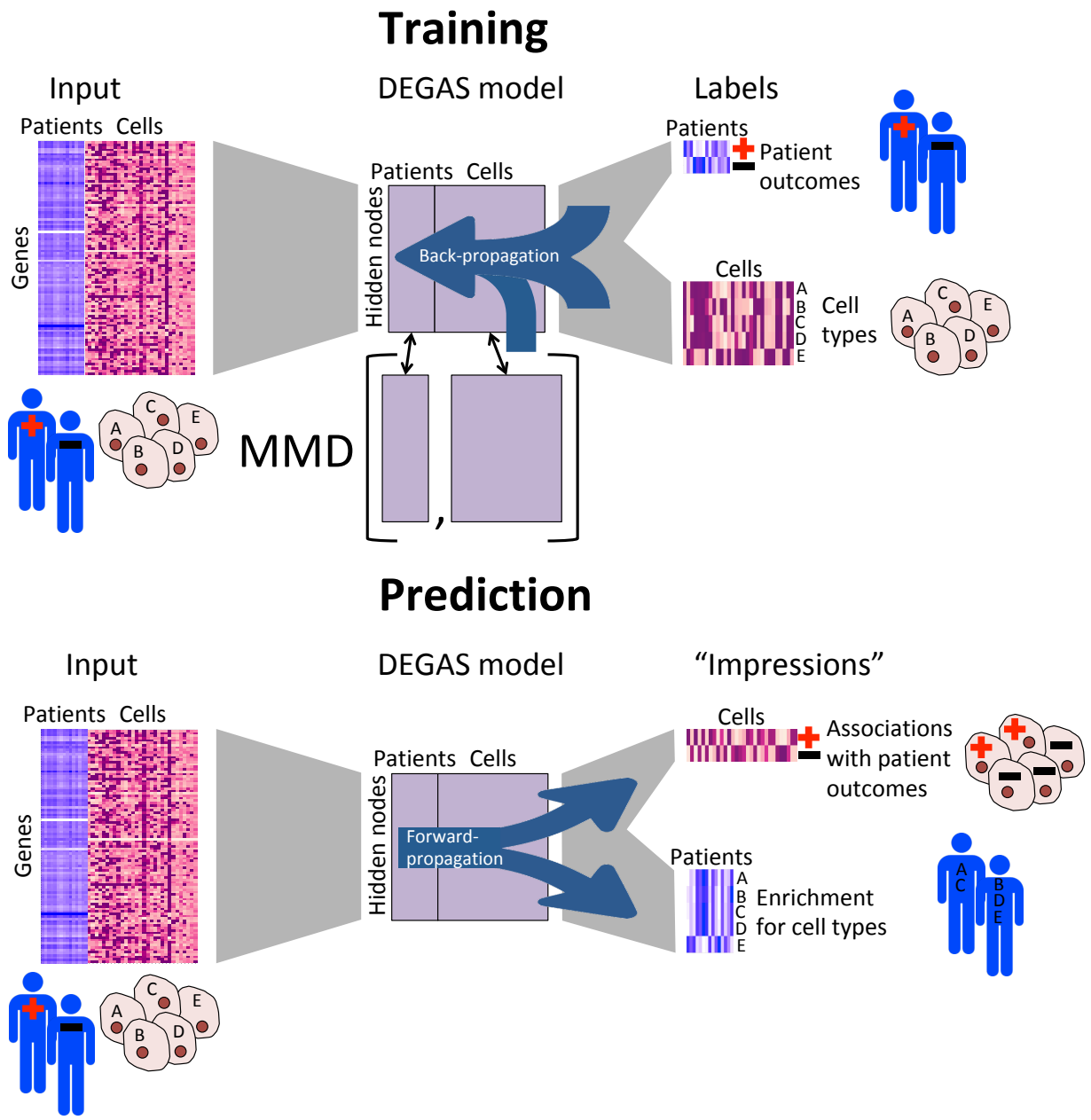

Fig. 1 A workflow diagram of the DEGAS framework. The scRNA-seq and patient expression data are preprocessed into expression matrices. Next a DEGAS model is trained using both single cell and patient outcomes via a multitask

learning neural network while the data distribution differences are reduced between patients and single cells at the final hidden layer using maximum mean discrepancy (MMD), which attempts to match the data distributions from two sources. As the output, this model can be used to infer clinical outcome impressions in single cells and cellular composition impression in patients.

DEGAS correctly mapped single cells to corresponding GBM subtypes

In Patel *et al*. [13], researchers assigned four major GBM tumor subtypes (Proneural, Mesenchymal, Classical, and Neural) to the single cell scRNA-seq data obtained from five GBM tumors. Of the five tumor samples, four had been labeled in the original publication with a single subtype based on the major proportion of cells assigned to each GBM subtype. For GBM bulk tumor tissue expression data, we obtained RNA-seq data for 111 GBM patients from The Cancer Genome Atlas (TCGA), in which the same GBM subtypes were also known. As the simplest form of validation, we used these two datasets as input for the DEGAS model to test if it could reidentify the same GBM subtypes for both single cells and for TCGA cohort. Indeed, DEGAS reidentified the same labels for all four tumors by overlaying GBM subtypes association on each single cell (**Fig. 2A-D**). For the fifth tumor sample, MGH31, it was labeled as a combination of multiple GBM subtypes in the original study and DEGAS also identified mixed cell types for this sample (**Fig. 2E**). In **Fig. 2 A-D**, the groups of cell subtypes with the highest association score (as judged by the median value) matched the ground-truth label of that tumor sample (indicated with the dash line box). Additionally, these relationships can be visualized by plotting the single cells and overlaying the GBM subtype association. It is clear that MGH28 and MGH29, for instance, have a high association with the mesenchymal GBM subtype (**Fig. 2F**). These DEGAS models also proved to be accurate with high AUCs (0.93-0.98) for predicting each of the GBM subtypes in the TCGA patients during cross-validation (**Fig. S1**).

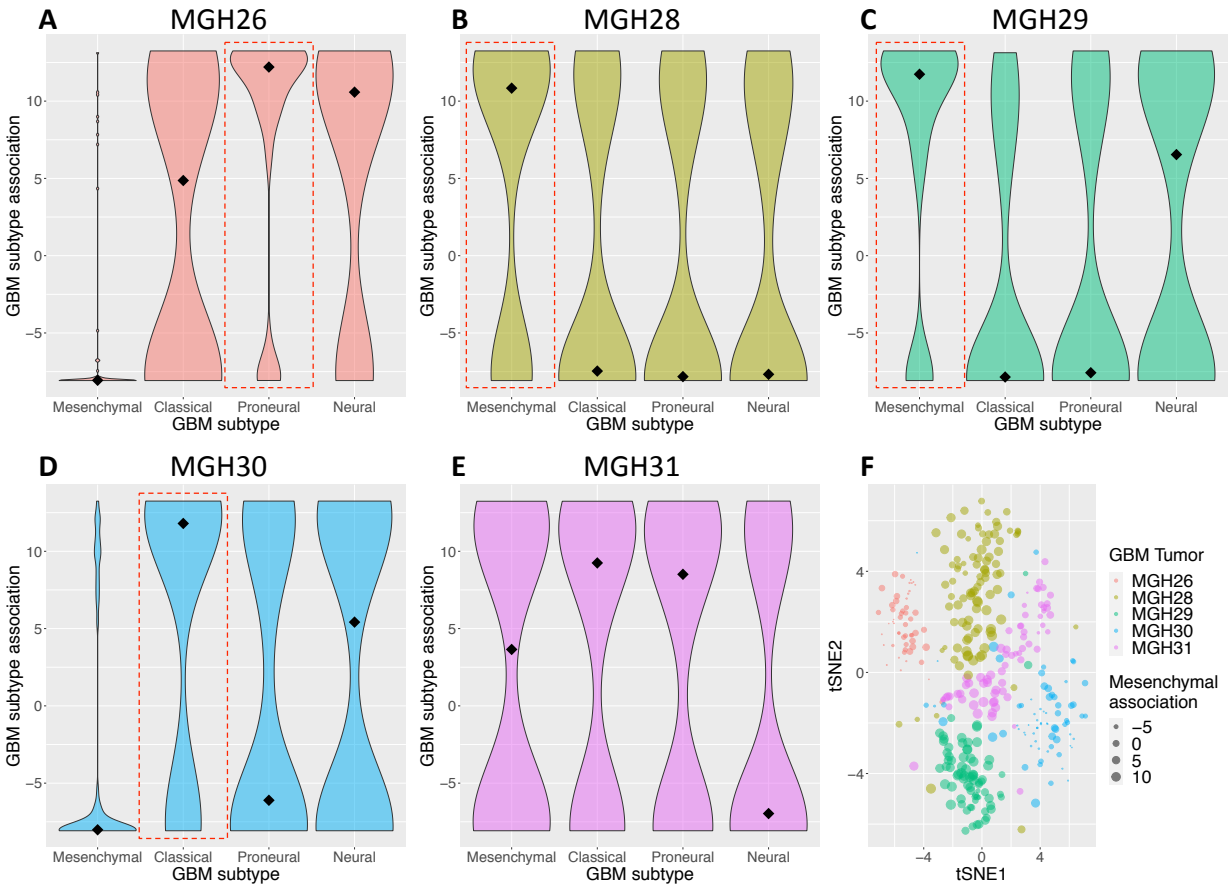

Fig. 2 DEGAS output of the distribution of GBM subtypes in single cells from five GBM tumors. Four of the five tumors had known GBM subtype information from Patel *et al.* (MGH26: Proneural, MGH28: Mesenchymal, MGH29: Mesenchymal, and MGH30: Classical, indicated by dashed red boxes) which were recapitulated by DEGAS. The subtype information for the tumors, MGH26, MGH28, MGH29, and MGH30 were derived from Patel *et al.* where MGH31 did not have a clearly defined subtype in Patel *et al.* The association of cells assigned to each subtype were plotted for each tumor; **A)** MGH26, **B)** MGH28, **C)** MGH29, **D)** MGH30 and **E)** MGH31. Median values are marked by a diamond in each of the violin plots. **F)** The association with the mesenchymal subtype is overlaid on all of the single cells from the five tumors (indicated by the size of the dots). It clearly shows that the single cell from MGH28 and MGH29 have the highest association with the mesenchymal subtype. Note: the tumors in **(F)** are color coded the same as the boxplots in **(A-E)**.

DEGAS identifies AD patients with increased microglia and reduced neuron populations

Aside from GBM, AD also has well documented characteristics that can be used as a test bed for DEGAS. Specifically, there is a well-documented reduction in neurons [14-16] and increase in microglia [17-20] in AD patients. AD scRNA-seq data are from the Allen Institute for Brain Science and bulk RNA-seq data are from Mount Sinai Brain Bank (MSBB)[21]. The DEGAS models were trained using either a single neuron cell type or two types (excitatory, and inhibitory neurons), oligodendrocyte, astrocyte, oligodendrocyte progenitor cell (OPC), and microglia. The brain samples were split into groups based on AD diagnosis status (AD^+ or AD^-),

From DEGAS results, we confirmed that in the single cell level, the AD association score was reduced in neuron single cells as previously described [22], as shown by the small point sizes of

neuron cells (**Fig. 3A**) (**Table 1-2**). At the patient level, the neuron enrichment score was also reduced in AD patient brain samples, as shown by the smaller point sizes for AD patients compared to normal patients (**Fig. 3B**). One important finding among AD brain samples is that the decreased neuron pattern was only found on excitatory neurons but not in inhibitory neurons. This supports the similar findings in previous AD studies [22, 23]. However, another study discovered inhibitory neuron loss in AD [24]. In fact, with DEGAS results, although the inhibitory neuron percentage appeared to increase in AD (**Table 3-4**), this could be due to the much greater loss of excitatory neurons. Specifically, the 1.90 mean increase (from -1.14 in AD^- to 0.7 in AD^+) in inhibitory neuron enrichment score (**Table 3**) is much smaller than the 2.84 mean reduction (from 1.70 in AD^- to -1.14 in AD^+ samples) in excitatory neuron enrichment score (**Table 3**) in MSBB samples. Since the enrichment score only reflects the relative proportion of each cell type, the results suggest that the putative gain of inhibitory neuron might actually be a relatively smaller loss in relation to the excitatory neuron loss. Indeed, when inhibitory and excitatory neurons were combined into a single neuron group, the mean reduction in enrichment are actually larger (3.74, **Table 1**) compared to using the excitatory neuron enrichment alone (2.84, **Table 3**). This means that combining the two neuron groups together actually resulted in even greater neuron loss being detected, which indicates that the apparent increase in inhibitory neurons was just an artifact of a differential amount of loss between excitatory and inhibitory neurons. Opposite to the neuron change, we observed a significant increase in the microglia enrichment in AD patients and a significant increase in AD association in microglia single cells (**Fig. 3C**, **Table 2,4**). The patients assigned with increased microglia enrichment score and decreased excitatory neuron enrichment also follow worse outcomes from AD diagnostic scores in multiple brain regions (**Fig. 3D**). Specifically, we evaluated the signature in relation to clinical dementia rating (CDR) [25], Braak and Braak stage (BBS) [26], and β-amyloid plaque mean (Plaque_mean) [27]. The CDR measure evaluates dementia based on interviews with caregivers about a patient behavior and quality of life [25]. The BBS measure evaluates the localization of neurofibrillary tangles in regions of the brain [26]. The Plaque_mean measure is performed during an autopsy to study the presence and size of β-amyloid plaques in the brain of a decreased patient [27] (**Fig. S8**). During 10-fold cross-validation, DEGAS models achieved high cell type prediction AUCs (0.97-1.00) for single cells and high AD diagnosis AUC (0.76) in MSBB patients when inhibitory and excitatory neurons were separated by cell type labels. Similarly, when inhibitory and excitatory neurons were combined into a single neuron cell type, DEGAS models also achieved high cell type prediction AUC (0.90-1.00) for single cells and high AD AUC (0.76) in MSBB patients in 10-fold cross-validation.

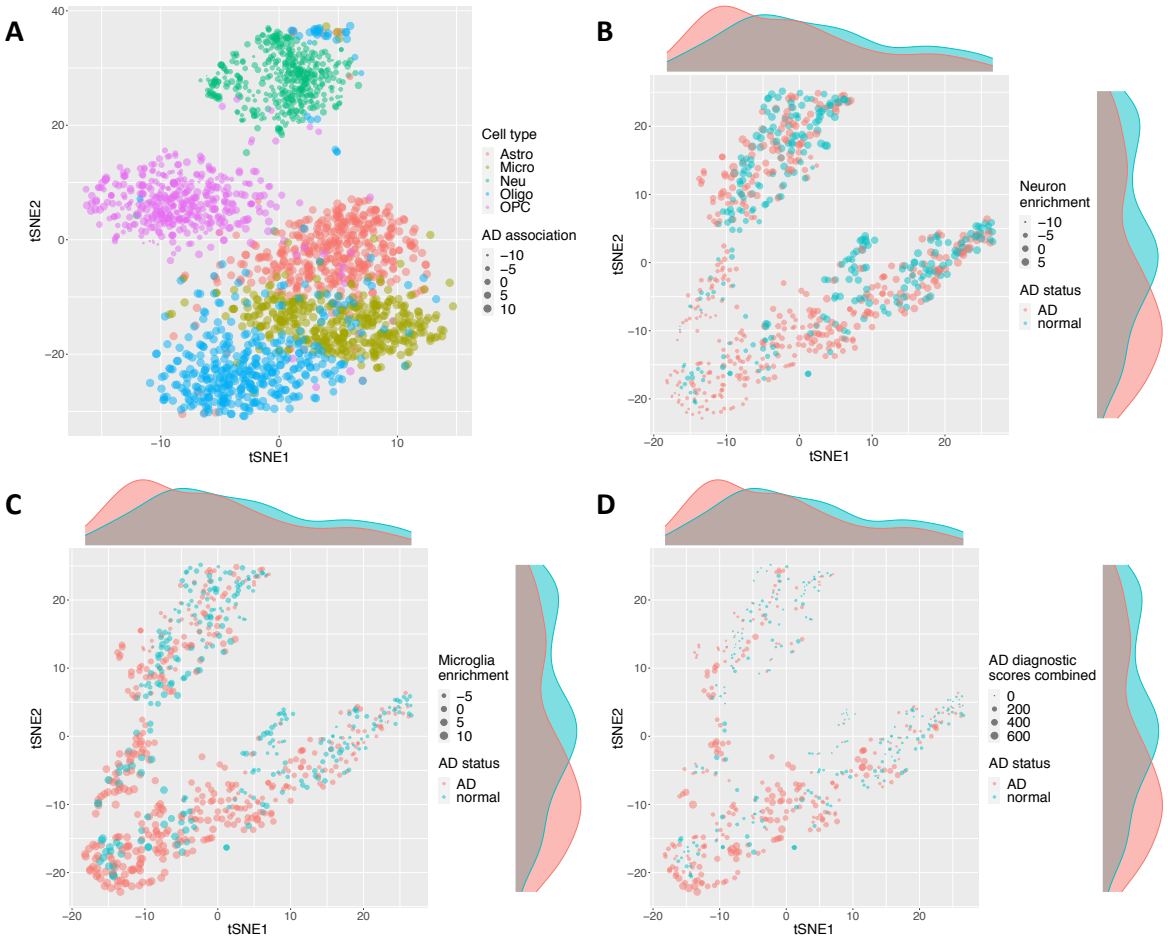

Fig. 3 DEGAS output of relationship between brain cell type and AD diagnosis based on AD scRNA-seq data and AD bulk tissue RNA-seq data. **A)** DEGAS output an AD association score for each single cell. The AD association score is indicated by the dot size and is overlaid on 2000 Allen Institute single cells, 400 from each cell type, showing the negative AD association (smaller neuron dots) in neuron cells. **B)** The neuron enrichment overlaid on MSBB patients showing that the neuron enrichment positively correlated with normal brains. **C)** The microglia enrichment overlaid on MSBB patients showing that the microglia enrichment follows the density of AD patients. **D)** The AD diagnostic test scores combined (CDR, BBS, β-amyloid plaque mean) overlaid on MSBB patients showing that AD diagnostic scores align with AD patients. The blue and red density kernels on the x- and y-axes (B-D) describe the distribution of of AD samples in the t-SNE plots.

Table 1 Comparison of cell type enrichment scores in patients between disease status groups. The DEGAS models were trained using a combined neuron cell type, oligodendrocyte, astrocyte, OPC, and microglia. The patient samples were split into groups based on AD diagnosis status (AD^+ or AD^-) and the cell type enrichment scores were compared between groups using t-tests. See **Table S1** for all cell types.

Cell type	AD+ mean	AD- mean	t-statistics	p-value
Neuron	-1.74	2.00	-7.54	1.83E-13
Microglia	1.43	-2.75	8.62	<2.2E-16

Table 2 Comparison of AD association scores in single cells between cell types. The DEGAS models were trained using a combined neuron cell type, oligodendrocyte, astrocyte, OPC, and microglia. The single cells were split into groups based on their cell type then compared to all other cell types (Cell-type$^+$ or Cell-type$^-$) and the AD association was compared between groups using t-tests. See **Table S2** for all cell types.

Cell type	Cell-type$^+$ mean	Cell-type$^-$ mean	t-statistics	p-value
Neuron	-1.10	10.30	131.34	<2.2E-16
Microglia	11.53	-0.17	-126.38	<2.2E-16

Table 3 Comparison of cell type enrichment score in patients between disease status groups. The DEGAS models were trained using two neuron cell types, inhibitory and excitatory neurons, plus the other four cell types. The patient samples were split into groups based on AD diagnosis status (AD$^+$ or AD$^-$) and the cell type enrichment was compared between groups using t-tests.
See **Table S3** for all cell types.

Cell type	AD+ mean	AD- mean	t-statistics	p-value
Inhibitory	0.76	-1.14	2.60	9.64E-3
Excitatory	-1.14	1.70	-3.89	1.11E-4
Microglia	1.05	-1.58	3.60	3.40E-4

Table 4 Comparison of AD association score in single cells between cell types. The DEGAS models were trained using two neuron cell types, inhibitory and excitatory neurons, plus the other four cell types. The single cells were split into groups based on their cell type then compared to all other cell types (Cell-type$^+$ or Cell-type$^-$) and the AD association was compared between groups using t-tests. See **Table S4** for all cell types.

Cell type	Cell-type$^+$ mean	Cell-type$^-$ mean	t-statistics	p-value
Inhibitory	8.59	-2.68	-136.90	<2.2E-16
Excitatory	-4.64	9.15	191.90	<2.2E-16
Microglia	11.08	-0.13	-78.62	<2.2E-16

DEGAS identifies plasma cell subtypes in IUSM CD138+ scRNA-seq of MM

In the MM study, unlike the previous two datasets, there were no predefined cell type labels, but DEGAS was still capable of analyzing such data. We first used Seurat [28], a commonly used scRNA-seq data analysis tool, to merge and cluster all the CD138+ bone marrow cells from four MM patients whose samples were collected at the IUSM. Seurat generated 11 clusters of cells for the IUSM scRNA-seq study. Since many of these clusters were not distinct from one another (**Fig. 4**), we merged neighboring clusters and reduced the 11 clusters to 5. We further validated these subtype clusters by individually clustering each patient with Seurat and using another scRNA-seq normalization tool Batch Effect ReMoval Using Deep Autoencoders (BERMUDA) [29] on all four patients. We found that the individual clustering experiments closely mirrored the Seurat-CCA clusters (**Fig. S10A-D**, **Table S6**) and that subtype 2 was consistent across MM patients using BERMUDA (**Fig. S10E**). These 5 clusters were the cell subtypes used as the subtype labels in the DEGAS framework. For bulk tissue data from MMRF, the clinical outcome of relapse free survival for 647 patients was used as the input to DEGAS.

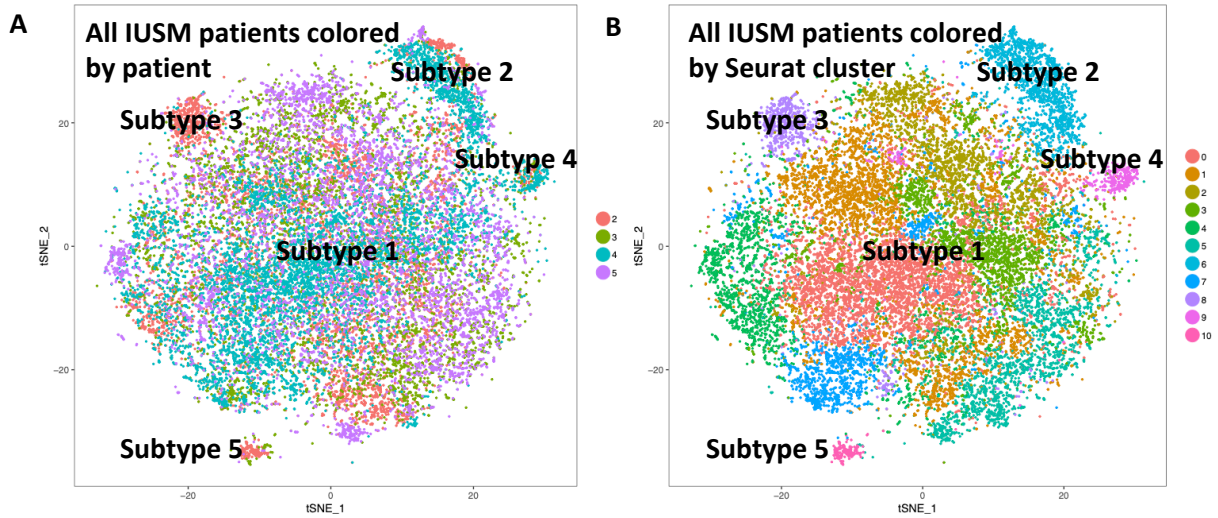

Fig. 4 Clusters generated from Seurat-CCA colored by **A)** IUSM scRNA-seq patient, and **B)** cluster, *i.e.*, subtype.

DEGAS patient stratification and cell type classification on MM

After a DEGAS model was trained on IUSM patient scRNA-seq data with subtype labels and MMRF patients with bulk tissue data and relapse free survival information, performance metrics were calculated via cross-validation. When relapse was treated as a binary outcome, DEGAS was able to achieve a median relapse prediction AUC of 0.68 when simultaneously integrating the single cell data with the patient samples and predicting the subtypes on the single cells. When predicting cellular subtype label in single cells, DEGAS was able to achieve an AUC between 0.92-0.99 for all five of the CD138+ cellular subtypes we identified in the IUSM scRNA-seq data (**Fig. 5A**). Aside from classifying the single cells correctly, DEGAS was able to stratify patients into high and low risk groups based on median relapse risk (p-value = 1.67E-9, **Fig. 5B**). We then applied the trained model on an external patient transcriptomic dataset from Zhan *et al*. [30] for overall survival validation. We demonstrated that the Cox proportional hazards portion of the DEGAS model was robust across datasets, and the impression extracted from the DEGAS framework was capable to stratify patients into low and high-risk groups in the Zhan *et al.* dataset (p-value=1.37E-2, **Fig. 5C**).

DEGAS identifies CD138+ cellular subtypes that are associated with patient relapse

The DEGAS model for the MM study transfers clinical information (impressions) to single cells (*i.e.*, single cells were directly assigned a relapse association), as well as transfers cellular/molecular features (impressions) to patients (*i.e.*, patients are assigned subtype enrichment). We found that subtype 1 and subtype 2 cells were the most important for prognosis. Specifically, subtype 1 was associated with a longer time to relapse and subtype 2 with a shorter time to relapse — IUSM single cells that were subtype 1 had much lower association with relapse than single cells that were subtype 2 (**Fig. 5D**, p-value < 2.2E-16). The MMRF patients who relapsed had significantly lower subtype 1 enrichment (**Fig. 5E**, p-value = 6.2E-9) and higher subtype 2 enrichment (**Fig. 5F**, p-value = 2.90E-3). On an external validation scRNA-seq dataset from Ledergor *et al.*, we found a steady decrease in subtype 1 enrichment

from NHIP (normal control) to the near-MM stage SMM (**Fig. 5H**, p-value = 7.80E-3) and increase in subtype 2 enrichment from NHIP to near-MM stage SMM (**Fig. 5I**, p-value=1.15E-2) and MM (**Fig. 5I**, Kruskal-Wallis p-value = 6.40E-3). The relapse association predicted by DEGAS framework also increased from NHIP to SMM (**Fig. 5G**, p-value = 2.20E-2) and MM (**Fig. 5G**, Kruskal-Wallis p-value = 3.50E-2), which agrees with the order of precursor conditions for MM (NHIP (no disease) ➔ MGUS ➔ SMM ➔ MM).

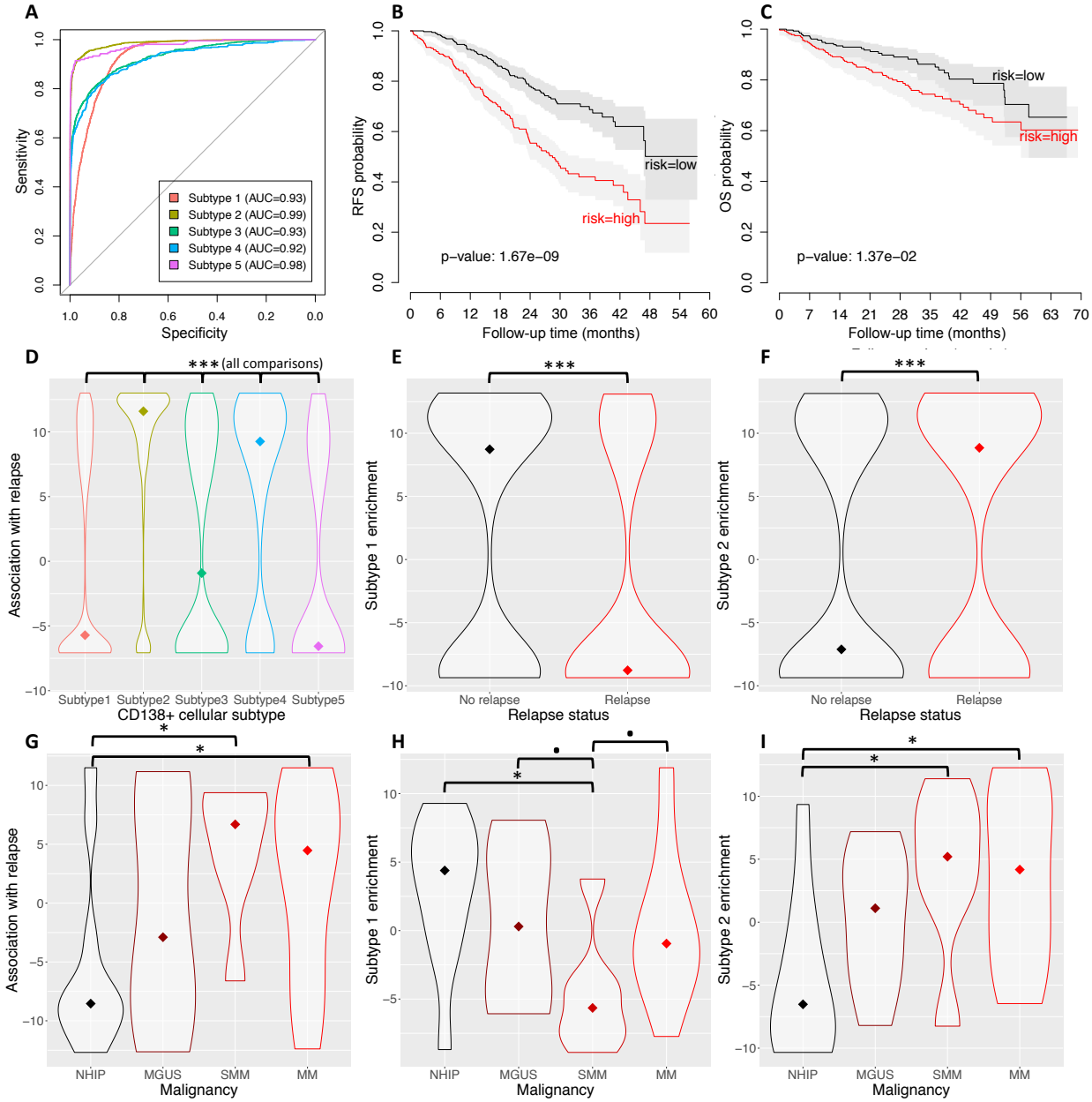

Fig. 5 Association between subtypes and relapse risk. **A)** The ROC curves for predicting the correct subtype **B)** Kaplan Meier curves from cross-validation for the MMRF patients stratified by median proportional hazard. **C)** Kaplan Meier curves of overall survival of Zhan *et al.* external dataset. MM patients stratified by median proportional hazard. **D)** Proportional hazard prediction for NHIP, MGUS, SMM, and MM in the IUSM dataset. **E)** Subtype 1 enrichment for relapsed vs. non-relapsed MMRF patients. **F)** Subtype 2 enrichment for relapsed vs. non-relapsed MMRF patients. **G)** Proportional hazard prediction for NHIP, MGUS, SMM, and MM in the external dataset GSE117156. **H)** Subtype 1

enrichment for NHIP, MGUS, SMM, and MM in the external dataset GSE117156. **I)** Subtype 2 enrichment for NHIP, MGUS, SMM, and MM in the external dataset GSE117156. NHIP: normal hip bone marrow, MGUS: monoclonal gammopathy of undetermined significance, SMM: smoldering multiple myeloma, MM: multiple myeloma. Significance values: • (0.1), * (0.05), ** (0.01), *** (0.001). All median values in violin plots are marked with a diamond. All plots were generated using the default parameters for the DEGAS package described in the section of Methods: Transfer learning using DEGAS.

MM distinct prognostic subtypes have distinct co-expression signatures

In order to understand the molecular level differences between subtypes 1 and 2, which behave oppositely in terms of the association to tumor malignancy, we performed gene co-expression analysis on the scRNA-seq data of the two cell subtypes using WGCNA [31]. When we compared the gene co-expression patterns between subtypes 1 and 2, we found that the blue and turquoise modules were shared by the two subtypes and subtype 2 additionally included two unique modules (**Fig. 6A, B**). We performed cell type enrichment analysis on each of the two unique modules, and the results were summarized in **Table 5**. Similarly, we also calculated the differentially expressed genes for subtype 2 (**Supplementary File 2**). The most interesting observation was the down-regulation of CD45 as well as the up-regulation of CD71, CD138, and CD38 gene expressions, which constitutes a CD45-/CD71+/CD138+/CD38+ signature of subtype 2 compared to all other subtypes in the CD138+ fraction. Furthermore, the increased cell type enrichment for erythroblast and progenitor ontology terms (**Table 5**) may indicate a progenitor cell like phenotype for subtype 2.

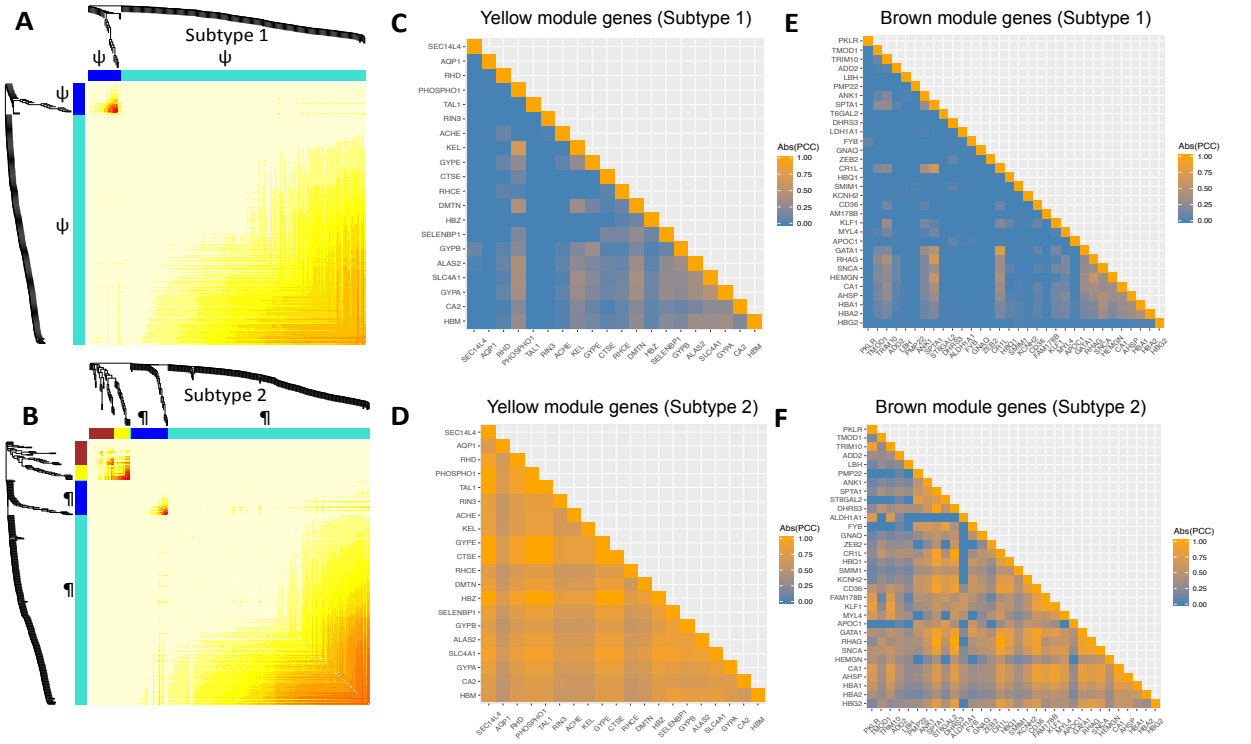

Fig. 6 Gene co-expression network analysis. **A)** Subtype 1 WGCNA output. **B)** Subtype 2 WGCNA output. **C)** Subtype 1 yellow module Pearson Correlation Coefficient (PCC) matrix. **D)** Subtype 2 yellow module PCC matrix. **E)** Subtype 1 brown module PCC matrix. **F)** Subtype 2 brown module PCC matrix. ψ indicates subtype 1 clusters in **Table S7** and ¶ indicates subtype 2 clusters in **Table S7**.

Table 5 Cell type enrichment terms for co-expression modules in subtype 1 and subtype 2.

Subtype	Module	Database	Cell type	p-value	q-value
Subtype 2	yellow	Human Gene Atlas	CD71+_EarlyErythroid	4.12E-22	3.46E-20
Subtype 2	yellow	ARCHS4 Tissues	ERYTHROBLAST	2.71E-17	2.92E-15
Subtype 2	yellow	Jenson TISSUES	K-562_cell	9.61E-9	1.77E-5
Subtype 2	brown	Human Gene Atlas	CD71+_EarlyErythroid	1.28E-15	1.08E-13
Subtype 2	brown	ARCHS4 Tissues	ERYTHROBLAST	1.26E-16	1.36E-14
Subtype 2	brown	Jenson TISSUES	ERYTHROID_CELL	1.79E-7	3.30E-4
Subtype 1 and Subtype 2	blue	GO Biological Process	regulation of B cell activation	4.32E-17	1.10E-13
Subtype 1 and Subtype 2	blue	ARCHS4 Tissues	PLASMA CELL	5.30E-8	2.90E-6
Subtype 1 and Subtype 2	blue	ARCHS4 Tissues	CD19+ B CELLS	7.17E-7	1.55E-5
Subtype 1 and Subtype 2	turquoise	Human Gene Atlas	Lymphoma_burkitts	3.26E-8	2.74E-6
Subtype 1 and Subtype 2	turquoise	Human Gene Atlas	721_B_lymphoblasts	4.18E-4	8.78E-3
Subtype 1 and Subtype 2	turquoise	ARCHS4 Tissues	CD34+ CELL	5.45E-83	5.89E-81

DEGAS is robust to hyper-parameter choice in GBM

To measure the robustness of DEGAS, we also analyzed how the hyper-parameter choices influence the DEGAS results, using a set of 100 randomly generated hyper-parameters and performing 10-fold cross-validation on each set of those 100 sets of hyper-parameters on the GBM datasets. The hyper-parameters that we evaluated were: the number of training steps, batch size for single cells, batch size for patients, number of hidden layer nodes, drop-out retention rate (the percentage of nodes randomly retained at the hidden layer), patient loss weight, MMD loss weight, and L_2 regularization weight. The default parameters used for all previous experiments are listed in the Methods section of Transfer learning using DEGAS. For more information on the range of hyper-parameters that were randomly sampled, please see the Methods section of Evaluation of DEGAS robustness to hyper-parameters in GBM. We discovered that most hyper-parameters did not significantly affect the AUC predicting GBM subtype in TCGA GBM patients with the exception of the drop-out retention rate and number of hidden layer nodes (**Fig. S11, Table 8**). Similarly, most hyper-parameters did not significantly affect the correct assignment of GBM subtype to GBM scRNA-seq tumor, except for the number of hidden layer nodes and the drop-out retention rate (**Fig. S12**, **Table S9**), where an initial phase of increased AUC was seen with the parameter values increasing (for the number of hidden layer nodes < 20 and the drop-out retention rate < 20%). In this case, we suggest users

to keep default settings for these two parameters as described in the Methods. Other than these two parameters, the rest of the hyper-parameters do not affect the high accuracy of the tumor subtype classification. Furthermore, we found that regardless of the hyper-parameters used in our hyper-parameter comparison experiment (including suboptimal ones), the correct GBM subtype was assigned to the corresponding GBM scRNA-seq tumor 62% of the time (**Fig. S13**).

Discussion

In this work, we constructed a transfer learning framework DEGAS to integrate scRNA-seq and patient expression data in order to infer the transferrable "impressions" of patient characteristics in single cells and cellular characteristics in patients. By transfer learning, we trained a model using both scRNA-seq and patient bulk tissue gene expression, then reduced the data differences between the two types in the final hidden layer of our model via domain adaptation while simultaneously predicting cell characteristics using only scRNA-seq data and predicting patient characteristics using only patient expression data. We validated the feasibility and broad applications of our DEGAS framework on datasets from two diseases, GBM and AD, which contained ground truth tumor subtype labels and ground truth cell type-disease associations, respectively. Within the GBM single cell patient cohort, each GBM tumor, from which scRNA-seq data was generated, had a GBM subtype label from Patel *et al*. [13]. The DEGAS results showed that the majority of cells in each tumor were labeled with the same GBM subtype as previously defined in [13]. Specifically, we correctly mapped Proneural, Mesenchymal, Classical, and Neural GBM subtypes to single cells in four GBM tumor samples. This experiment also shows the broad applicability of the model since the single cells had no labels and the patient samples had multiclass labels. DEGAS allows for different categories of output labels to be combined, which may include but are not limited to: classification labels, Cox proportional hazard, and no labels. This allows for a wide variety of applications to adopt the DEGAS framework so that impressions are not limited to only one type of outcome. It also can be applied to other types of molecular data besides transcriptomic data, provided the same feature space (such as genes, or loci) are shared between single-cell data and the patient-level data.

The DEGAS analysis on Alzheimer's disease data further validated our model by correctly identifying the decreased neuron and increased microglia proportions in AD patients. Aside from these known characteristics of AD change, we also discovered that excitatory neuron loss was much greater than inhibitory neuron loss. In fact, the proportion of interneurons appeared to increase slightly though this is most likely attributed to a large decrease in the number of excitatory neurons in relation to a much smaller decrease in the number of inhibitory neurons. Despite the increased efforts put into AD research, there still lacks a general agreement about the change of neuronal subtypes and their roles in AD progression [22-24], making our finding a valuable contribution to this field.

To explore disease with unknown cell types and risk information, we applied DEGAS to MM data. The models were able to assign relapse free survival metrics to subtype populations of

CD138+ cells identified by cell type clustering methods Seurat [28] and BERMUDA [29]. Among the identified subtypes of cells, subtype 2 was the most consistent between IUSM patients visualized by BERMUDA (**Fig. S10E**). Furthermore, we found that the subtype 2 population appeared to have a gradient of cells moving away from the main subtype 1 group, possibly associated with a certain degree of differentiation (**Fig. S10**). Based on cell type enrichment analysis [32] on the gene co-expression modules unique to subtype 2 (*i.e.*, the yellow and brown modules in **Fig. 6**), we found an erythroblast/stem-like signature that may be related to convergent evolution in tumors, since evidence of convergent evolution has already been noted through copy number alterations in circulating tumor cells in multiple cancer types [33]. Another possible explanation could be that subtype 2 is some form of malignant progenitor cell.

Upon further examination, we found evidence that subtype 2 may represent a population of erythroblast-like cells in MM. Tumor-inducible erythroblast-like cells have already been reported in hepatocellular carcinoma (HCC) [34]. These CD71+ erythroblast-like cells have been shown in mice to be derived from embryonic stem cells. It would be unusual to find erythroid cells in the CD138+ fraction but the Vk*MYC mouse model of MM has been shown to contain populations of CD71+/CD138+ erythroid progenitor cells [35]. In addition, there is also evidence of CD45 low/CD138+/CD38+ tumor intiating cells found in patient derived xenograft models of MM [36]. The cell type enrichment analysis from the subtype 2 specific gene co-expression modules indicates CD71+ enrichment in these two gene modules. When we performed differential gene expression analysis between subtype 2 and all other cell subtypes, subtype 2 shows significantly higher SDC1 (also known as CD138) expression levels, increased TFRC (CD71) expression, insignificantly decreased PTPRC (CD45) expression, and significantly decreased PTPRCAP (CD45 associated protein) expression. Since all of the IUSM MM cells in our study had already been FACS sorted for CD138+, it is possible we have identified a population of CD45-/CD71+/CD138+/CD38+ cells in MM with a similar phenotype to the erythroblast-like cells found previously in HCC [34].

DEGAS discovered for the first time this special population of cells (CD45-/CD71+/CD138+ cells) in MM, along with its association with worse prognosis, which would be impossible using the scRNA-seq data alone. The erythroblast-like cells in subtype 2 are the most stable cluster across datasets (**Fig. S10**). We speculate this inter-tumor population stability could be attributed to convergent evolution in MM toward a more erythroblast-like phenotype or a stable pool of malignant progenitor cells that might help the cells to escape therapy. Consequently, this subtype could be targeted using precision immunotherapies that are not restricted to a single patient since the erythroblast-like phenotype, *i.e.*, subtype 2, is present in multiple (3/4) patients.

Based on the validated results in a variety of disease data analyses, we find that DEGAS has broad applications in virtually all diseases with available patient-level and single cell level omic data, as well as clinical data. The tensorflow [37] machine learning code is integrated with a simple R package interface (https://github.com/tsteelejohnson91/DEGAS) which will facilitate researchers to manipulate scRNA-seq and bulk expression data on their own.

Conclusion

DEGAS is a powerful transfer learning tool in integrating different levels of omic data and identifying the latent molecular relationships between populations of cells and clinical outcomes, which we refer to as impressions. We validated the DEGAS framework on GBM and AD by showing DEGAS models were capable of accurately predicting patient characteristics in single cells and cellular characteristics in patients. We then leveraged this transfer learning approach on MM data and identified CD138+ subtype populations, possibly clones or progenitor cells, in MM that were significantly associated with disease relapse. These subtypes contain unique RNA profiles and gene correlations that can be both leveraged as a prognostic biomarker and possibly targeted directly to reduce the risk of relapse. We believe that DEGAS can be a powerful solution to overcome the challenge of integrating patient single cell data with bulk tissue data so that researchers can identify populations of cells associated with an outcome of interest, while as the same time identify patients with certain cell type composition. Furthermore, DEGAS can accommodate flexible data types. This makes it a very general framework that can be applied in multiple different diseases and data types to identify cellular populations that are associated with prognosis or treatment response, or to identify specific patient groups with certain cell subtypes for personalized treatment.

Online Methods

Datasets

In this study we analyzed data from three different diseases, GBM, AD, and MM, to validate the DEGAS framework and apply it for novel discoveries. GBM and AD were primarily used as validation datasets since the ground truth is known. For GBM data, we used scRNA-seq from 5 tumors from Patel *et al*. [13]. and microarrays for the GBM TCGA cohort [38]. For AD data, we used human scRNA-seq from Allen Institute Cell Types Database (https://celltypes.brain-map.org/) and AD patient RNA-seq from the Mount Sinai/JJ Peters VA Medical Center Brain Bank (MSBB) study [21]. We further expanded our inquiry into MM, which served as a discovery dataset. Since the plasma cell subtypes are less understood in relation to clinical outcomes, we aimed to identify subtypes of plasma cells associated with worse prognosis.

For MM analysis, we utilized 647 $CD138^+$-enriched bone marrow patient samples from the Multiple Myeloma Research Foundation (MMRF). These data were generated as part of the Multiple Myeloma Research Foundation Personalized Medicine Initiatives (https://research.themmrf.org). The samples consisted of tumor tissue RNA-seq data and corresponding clinical variables including relapse-free survival time and survival status. Relapse-free survival was defined as the time taken for a patient to relapse after treatment of the initial tumor or the time of death if relapse was not reached. The demographic information of the MMRF patients are shown in **Table 6**. The scRNA-Seq data used in this study were generated at Indiana University School of Medicine (IUSM) and consist of $CD138^+$ plasma cells purified from bone marrow from four MM patients. The low number of patients was a good test case considering most scRNA-seq experiments have few patients. The single cells were

sequenced using 10x Genomics and Illumina NovaSeq6000 sequencer. CellRanger 2.1.0 (http://support.10xgenomics.com/) was utilized to process the raw sequence data. Briefly, CellRanger used bcl2fastq (https://support.illumina.com/) to demultiplex raw base sequence calls generated from the sequencer into sample-specific FASTQ files. The FASTQ files were then aligned to the human reference genome GRCh38 with RNAseq aligner STAR. The aligned reads were traced back to individual cells and the gene expression level of individual genes were quantified based on the number of UMIs (unique molecular indices) detected in each cell. The filtered gene-cell barcode matrices generated by CellRanger were used for further analysis. Additionally a second publicly available scRNA-seq dataset was used for validation, which consisted of NHIP (normal control), MGUS, SMM, and MM patients [39]. A second bulk tissue dataset was used for validating the proportional hazards modeling. This dataset consisted of bulk expression profiling by microarray of CD138+ plasma cells with overall survival information for 559 MM patients [30]. The detailed dataset information are shown in **Table 7**.

Table 6. Summary of the clinical features in each patient cohorts used in training. * Final age category is >90 years.

Glioblastoma Multiforme TCGA	
Feature	**Details**
Sex	74 Male, 37 Female
Age (years)	Range: 14-83, Mean: 56, Median: 58
Clinical GBM subtype	34 Classical, 33 Mesenchymal, 9 Neural, 35 Proneural
Alzheimer's Disease MSBB	
Feature	**Details**
Sex	90 Male, 131 Female
Age (years)	Range: 61-90+, Mean* > 82, Median = 84
AD diagnosis	135 AD, 86 Control
Multiple Myeloma MMRF	
Feature	**Details**
Sex	387 Male, 260 Female
Age (years)	Range: 27-93, Mean: 64, Median: 64
Relapse-free survival time (days)	Range: 13-1753, Mean: 665.4, Median: 629 200 patients relapsed

Table 7. Overview of datasets used in the analysis

Study	Sample size	Data type	Outcome
Patel *et al*., 2014	532 cells (5 patients)	scRNA-seq (SMART-seq)	None
TCGA GBM	111 patients	Microarray	GBM subtype
Allen Institute	47,396 cells (11 patients)	scRNA-seq (SMART-seq)	Brain cell types
MSBB	682 samples (221 patients)	RNA-seq	AD diagnosis
MMRF	647 patients	RNA-seq	Relapse-Free Survival
IUSM	22,968 cells (4 patients)	scRNA-seq (10x Genomics)	Subtype cluster (Subtype 1-5)
Ledergor *et al*., 2019	13,440 cells (35 patients)	scRNA-seq (MARS-seq)	Malignancy (NHIP, MGUS, SMM, MM)
Zhan *et al*., 2006	559 patients	Microarray	Overall Survival

Transfer learning using DEGAS

Cox proportional hazards, patient classification, cell type classification, and maximum mean discrepancy (MMD), a technique used to match distributions across different sets of data [40], were combined to create a multitask transfer learning framework. We called this framework Diagnostic Evidence GAuge of Single cells (DEGAS). DEGAS makes it possible to combine bulk expression from patients with clinical information and single cells with cell subtype information. Since most scRNA-seq datasets do not contain many patients, it is difficult to derive cellular associations with clinical outcomes. DEGAS circumvents this problem using multitask learning and domain adaptation techniques from transfer learning. As a result, each cell type can be given clinical attributes and each patient can be given cellular attributes.

The first step was to find a set of gene expression features that were both informative of cell type and of patient recurrence. The intersection of high variance genes found in the scRNA-seq and patient expression data are used for further analysis. Defining this gene set is up to the user but Seurat-CCA, LASSO selection, and even statistical tests in R can be used to define the gene set. Since these features are the same between patients and single cells, the patients and cells share the same input layer. This makes it possible to predict proportional hazard and cell type regardless of the input sample type (patient or single cell).

As an example of the DEGAS framework, we used a single layer network model for simplicity. However, the following equations can be extrapolated to multiple layers and architectures which are already included in our software. First, a hidden layer was used to transform the genes into a lower dimension using a sigmoid activation function (**Eq. 1**). Where X represents an input

expression matrix, θ_{Hidden} represents the hidden layer weights, and b_{Hidden} represents the hidden layer bias.

$$f_{Hidden}(X) = sigmoid(X^{\mathrm{T}}\theta_{Hidden} + b_{Hidden}) \quad \textbf{Eq. 1}$$

Next, output layers were added for both the patient output and for the single cell output. For the single cells, there could be classification output or no output. Similarly, patients could have Cox proportional hazard output, classification output, or no output. The Cox proportional hazards estimates consisted of a linear transformation to a single output followed by a sigmoid activation function (**Eq. 2**). The classification output consisted of a transformation to the same number of outputs as the number of labels, *i.e.*, patient subtypes, cellular subtypes, using a softmax activation function (**Eq. 3**). Variable X represents an input expression matrix, θ_{Cox} represents the Cox proportional hazard layer weights [41], θ_{Class} represents the classification layer weights, b_{Cox} represents the Cox proportional hazard layer bias, and b_{Class} represents the classification layer bias.

$$f_{Cox}(X) = sigmoid(f_{Hidden}(X)^{\mathrm{T}}\theta_{Cox} + b_{Cox}) \quad \textbf{Eq. 2}$$

$$f_{Class}(X) = softmax(f_{Hidden}(X)^{\mathrm{T}}\theta_{Class} + b_{Class}) \quad \textbf{Eq. 3}$$

To train the DEGAS model, we need to compute three types of loss functions for the Cox proportional hazards output, classification output, and MMD [40] respectively. The Cox proportional hazards loss [41] was calculated only for the patient expression data (X_{Pat}) using the followup period (C), and event status (t) (**Eq. 4**). Similarly, the patient classification loss was only calculated for the patient data (X_{Pat}) using the patient labels (Y_{Pat}). Alternatively, the cellular classification loss was only calculated for the single cell expression data (X_{Cell}) and true subtype label (Y_{Cell}) (**Eq. 5**). However, the MMD loss was calculated between the patient expression data (X_{Pat}) and the single cell expression data (X_{Cell}) (**Eq. 6**).

$$Loss_{Cox} = \sum_{C(i)=1}\left(f_{Cox}(X_{Pat})_i - \sum_{t_j \geq t_i}(\exp\,(f_{Cox}(X_{Pat})_j)\right) \quad \textbf{Eq. 4}$$

$$Loss_{Class} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum(Y_{type,i} - f_{Class}(X_{type})_i)\right) \; where\ type\ \epsilon\ \{Pat, Cell\} \quad \textbf{Eq. 5}$$

$$Loss_{MMD} = MMD(X_{Cell}, X_{Pat}) \quad \textbf{Eq. 6}$$

The overall loss function was additionally weighted using the hyper-parameters, λ_0 (single cell loss function), λ_1 (patient loss function), λ_2 (MMD loss), and λ_3 (regularization loss), so that the importance of each loss term and regularization term could be adjusted (**Eq. 7**). To address more diverse datasets, we also allow for two classification outputs (**Eq. 8**), a single classification output without patient outcome (**Eq. 9**), a single classification output without cell type label (**Eq. 10**), or a single Cox output without cell type label (**Eq. 11**).

$$Loss_{ClassCox} = \lambda_0 Loss_{Class} + \lambda_1 Loss_{Cox} + \lambda_2 Loss_{MMD} + \lambda_3\|\theta\|_2^2 \quad \textbf{Eq. 7}$$

$$Loss_{ClassClass} = \lambda_0 Loss_{Class} + \lambda_1 Loss_{Class} + \lambda_2 Loss_{MMD} + \lambda_3\|\theta\|_2^2 \quad \textbf{Eq. 8}$$

$$Loss_{ClassBlank} = \lambda_0 Loss_{Class} + \lambda_2 Loss_{MMD} + \lambda_3\|\theta\|_2^2 \quad \textbf{Eq. 9}$$

$$Loss_{BlankClass} = \lambda_1 Loss_{Class} + \lambda_2 Loss_{MMD} + \lambda_3\|\theta\|_2^2 \quad \textbf{Eq. 10}$$

$$Loss_{BlankCox} = \lambda_1 Loss_{Cox} + \lambda_2 Loss_{MMD} + \lambda_3\|\theta\|_2^2 \quad \textbf{Eq. 11}$$

In summary, a common hidden layer was used to merge the single cells and patients. Next, an output layer was added to predict the proportional hazards or classes of the patient samples [41]. The loss function for the proportional hazards prediction or patient classification was back-propagated across both layers for each patient. The single cells also had an output layer consisting of a softmax output to predict the cellular subtype of each cell. Error was back-

propagated across both layers from the label output for each cell. Finally, a subspace was learned that can model both the single cells and the patients. To perform this task, we utilized the MMD method [40] to reduce the differences between patients and cells in a low dimensional representation. All of the single cell patients were combined into a single group such that the MMD loss was minimized between patients and single cells from multiple patients. Because there are many different combinations of these outputs, *i.e.*, single cell output followed by patient output, we include ClassCox, ClassClass, ClassBlank, BlankClass, and BlankCox based on equations (7)-(11) in the current version but intend to provide more options in the future.

To keep the analyses consistent, the same set of hyper-parameters were used in all of the experiments in this study, except for the robustness to hyper-parameters experiment, where they were intentionally altered to test the influences on the output results. These are considered the default hyper-parameters in the DEGAS package but can be changed. They are: training steps 2000, single cell batch size 200, patient batch size 50, hidden layer nodes 50, drop-out retention rate 50%, patient loss weight (λ_1) 3, MMD loss weight (λ_2) 3, L_2 regularization weight (λ_3) 3.

Validating DEGAS using GBM data

The scRNA-seq data from the Patel *et al*. study [13] were downloaded from NCBI Gene Expression Omnibus (GSE57872). The single cell expression values were previously normalized to TPM containing 5,948 genes with mean($\log_2$(TMP))>4.5 retained in the data table. The top 20% variance genes were retained for training. These values were then converted to z-scores then standardized to a range of [0,1] for each sample. The TCGA GBM microarray expression data was downloaded from Firebrowse (http://firebrowse.org/). Microarray data were used since it contains more patient samples for training with GBM subtype information than RNA-seq data. Likewise, the top 20% variance genes were retained for training and these expression values were converted to z-scores then standardized to a range of [0,1] for each sample. The GBM subtype labels for the TCGA patients were downloaded from Verhaak *et al*. [42]. The intersection of genes between single cells and patients were used for the final model training. Since subtype labels were only available for the GBM patient samples, we trained a BlankClass DEGAS model (**Eq. 10**). This model minimizes the MMD loss between single cells and patients while minimizing the classification loss only in GBM patients. We split the dataset into 10 groups and performed 10-fold cross-validation by leaving out a single patient group during training. After cross-validation, we normalized the GBM subtype output using quantile normalization (quantile normalized output represented by x in **Eq. 12**) and increased the variance of these quantile normalized outputs (**Eq. 12**) which we call association scores for patient outcome in single cells or enrichment scores for cell type classification in patients. These association scores were overlaid on the GBM single cells and now referred to as GBM subtype association scores because GBM subtype from patients is overlaid on single cells. We plotted these association scores stratified by GBM subtype for each tumor individually (**Fig. 2A-E**). We then compared the proportions of these cell types to the previously defined GBM types from the original publication marked red dashed boxes in **Fig. 2**. We also visualized the GBM subtypes association in single cells by calculating a low dimensional representation using tSNE and overlaying the kNN smoothed GBM subtype associations (**Fig. 2F,S2-S5**). To make the scatter

plots of cells and patients more informative, kNN smoothing was used by averaging each point's GBM subtype association value with its five nearest neighbors in tSNE. The model performance was shown with the receiver operating curve (ROC) and area-under-curve (AUC) for each of the GBM subtype labels in the TCGA patients from cross-validation (**Fig. S1**).

$$\begin{matrix} association\ score, \\ enrichment\ score \end{matrix} = \begin{cases} log_2(1E4 \cdot x), x > 0 \\ x, x = 0 \\ -log_2(-1E4 \cdot x), x < 0 \end{cases} \quad \textbf{Eq. 12}$$

Validating DEGAS and exploration using AD data

For AD datasets, we were primarily interested in identifying known relationships between cell types and AD diagnosis. For these reasons, we downloaded all of the adult Human scRNA-seq data from the Allen Brain Institute. Only inhibitory neurons, excitatory neurons, oligodendrocytes, astrocytes, microglia, and oligodendrocyte progenitor cells (OPCs) were retained in the analysis due to the extremely low sample sizes in the remaining cell types. In some analysis, the inhibitory and excitatory neuron groups were merged into a single neuron group. These data were then log_2 transformed, converted to sample-wise z-scores, and then standardized to [0,1] by each sample. Genes were only retained as features if they had 70% non-zero values. Of the remaining genes the top 30% variance genes were retained for training to keep the feature set larger. The labels for the single cells consisted of these major cell types listed above. The AD brain data was downloaded from Mount Sinai/JJ Peters VA Medical Center Brain Bank (https://www.synapse.org/#!Synapse:syn3157743). Each of the RNA-seq samples were either from an AD patient or a normal brain sample. The binary outcomes of AD case or normal were used as the label for the model. As in the other experiments, the RNA-seq values were log_2 transformed, converted to sample-wise z-scores, and standardized to [0,1] for each sample. The top 50% variance genes were retained for training to keep the feature set larger. The intersection of the patient genes and single cell genes were using to train the final model. Using the cell type classification for each single cell and the AD/normal classification for each MSBB patient we were able to train a ClassClass DEGAS model (**Eq. 8**). The performance was evaluated using 10-fold cross-validation by leaving out each group during training once. As in the GBM experiments (**Eq. 12**) we quantile normalized and increased the variance of the cell type output in patients and the AD diagnosis output in single cells resulting in the cell type enrichment and AD associations respectively. Student's t-tests were performed on the cell type enrichment between the patients with different disease status (AD vs. Normal) (**Table 1, 3, S1, S3**) or on AD association score between different cells with each cell type (**Table 2, 4, S2, S4**). In addition, patients were plotted overlaid with kNN smoothed cell type enrichment and single cells were plotted overlaid with kNN smoothed AD association (**Fig. 3**). Furthermore, to evaluate DEGAS performance, ROC and AUC were computed for the single cells during cross-validation for each cell type in the single cell data. Similarly, AD diagnosis ROC and AUC were computed from the MSBB patient RNA-seq (**Fig. S6**). We also repeated the same above analysis without merging the inhibitory and excitatory groups into a single neuron group so that the single cell labels were six types instead of five, *i.e.*, inhibitory neurons, excitatory neurons, oligodendrocytes, astrocytes, microglia, and OPCs.

Preprocessing of IUSM MM single cell data

The scRNA-seq data generated at IUSM were first combined into a dataset using Seurat-CCA [28]. This initial dataset integration allowed conserved subtypes of cells to be identified across datasets. All four patient dataset counts were loaded into a Seurat object. They were normalized, scaled, biased cells removed, and high variance genes identified following the Seurat online vignette. Using the union of high variance genes, multi-canonical correlation analysis was run across all four datasets, the subspaces were aligned across patients, the aligned single cells were plotted with t-SNE [43], and clusters of cells were identified. The raw expression values for the high variance genes identified by Seurat were $\log_2$ transformed, converted to z-scores, and then scaled to [0,1].

Furthermore, each IUSM scRNA-seq patient was individually clustered using Seurat to check the replicability of the clusters and were plotted with UMAP [44]. We used Rand, Fowlkes and Mallows's index (FM), and Jaccard index (JI) to measure the cluster consistency between single patient clustering experiments and the merged all-patient clustering results (**Fig. S10**, **Table S6**). The four single patient clustering results, one for each IUSM scRNA-seq patient, were used as input into BERMUDA [29] to visualize and evaluate the original Seurat clustering (**Fig. S10**).

Preprocessing of MMRF patient data

MMRF patients with bulk tissue RNA-seq and clinical data were used in MM analysis. We used relapse-free survival (RFS) with the time to first relapse or death. TPM values for the MMRF patient gene expression data and the RFS survival data were used as the input for DEGAS, these values were $\log_2$ transformed, converted to z-scores, and scaled to [0,1]. The union of the features identified by Seurat in the single cell data and the features selected in the MMRF patient data were used as the final feature set. The features retained in the MMRF data were identified by fitting an elastic-net Cox model [45] to the TPM values based on the RFS.

Evaluate DEGAS performance on MM datasets

AUC was calculated for each of the output labels for the single cells and for patient labels if a classification output was used for the patient data (**Fig. 5A**). Cox proportional hazard output was used for patients, a log-rank test was calculated for each patient so that the hazard ratio and p-value could be evaluated based on patient stratification by median proportional hazard (**Fig 5B**). Additionally, the same models were used to predict risk in the GSE2658 dataset which had information on OS. The output for each GSE2658 sample averaged across all 10 DEGAS models and stratified by median risk to show the robustness of the cox output across datasets (**Fig. 5C**). It is worth noting that the performance on each of these tasks individually should decrease in DEGAS since multiple tasks are being optimized simultaneously. The benefit and insight of DEGAS come from generating a feature space that combines traits from both tasks and allows information unavailable in one dataset to be transferred to another, *i.e.*, generate “impressions”.

Identifying MM cell types associated with prognosis

Gene expression profile for each of the MMRF patients, as a result of the trained transfer learning model, can be deconvoluted into the proportion of MM cell types identified in the single

cell MM data. In a similar fashion, the single cells from MM patients can be assigned proportional hazards based on the MMRF Cox section of the model. During each step of cross-validation and after training, each MMRF patient gene expression data in the validation set was deconvoluted into MM subtypes. Each single cell in the validation set was assigned relapse risk by feeding those samples through the Cox output layer. In this way, we can infer the association with relapse risk of specific cell types as well as the cell type enrichment contained in each MMRF sample. The raw output, like the GBM and AD experiments, were quantile normalized and the variance increased (**Eq. 12**) into the association scores of AD risk in single cells and enrichment scores of cell types in patients that we used for further analysis. We plotted these relationships and conducted Student's t-tests on the subtype vs. association with relapse in single cells (**Fig. 5D**) and the subtype enrichment vs. relapse status in patients (**Fig. 5E,F**).

Analysis of gene co-expression of prognostic cell types

For each cell type, we performed gene co-expression analysis across all four IUSM patients. The Pearson correlation coefficient (PCC) was calculated for each pair of genes and used as the edge weight for the co-expression network mining. Next, applying the co-expression mining tool WGCNA [31], we identified modules of co-expressing genes for different cell subtypes. For modules of interest, we compared the modules' correlation in each cell subtype. Additionally, we used the gene sets from each of the gene co-expression modules to identify enrichment of cell type using EnrichR [32]. Furthermore, Student's t-tests were calculated cell subtype 1 vs all cell subtypes and cell subtype 2 vs. all cell subtypes using the batch corrected gene expression values from Seurat. These values were stored in (**Supplementary File 1** and **Supplementary File 2**) respectively.

Evaluation of DEGAS robustness to hyper-parameters in GBM

Using the GBM dataset, we evaluated the robustness of DEGAS model outputs to hyper-parameters by repeating 10-fold cross-validation 100 times with randomly generated hyper-parameters following a uniform distribution. The range of hyper-parameters used in training consisted of: training steps 1,000-3,000, single cell batch size 100-300, patient batch size 20-100, hidden features 10-100, drop-out retention rate 0.1-0.9, Patient loss weight (λ_1) 0.2-5, MMD loss weight (λ_2) 0.2-5, L_2 regularization weight (λ_3) 0.2-5. The output for each of the 100 hyper-parameters was quantile normalized.

Using these outputs we performed two tests. One was to evaluate the loss in performance based on changing the hyper-parameters where performance was measured with AUC among the TCGA GBM patients labeled by patient GBM subtype (Mesenchymal, Classical, Proneural, Neural). In this test, we calculated the spearman correlation and plotted the scatter plot between the AUC of each of the four GBM subtype labels and the hyper-parameters used (**Fig. S11**, **Table S8**).

Next, we evaluated whether or not the correct GBM subtype labels (Mesenchymal, Classical, Proneural, Neural) could be recapitulated in the GBM scRNA-seq tumors that had known GBM subtypes (MGH26: Proneural, MGH28: Mesenchymal, MGH29: Mesenchymal, MGH30: Classical). To do this for each tumor (MGH26, MGH28, MGH29, MGH30), the rank of the

correct label was calculated by calculating the mean of each GBM subtype association across all of the cells in that tumor. This resulted in each of the 100 random hyper-parameters having a rank for each GBM subtype for each of the GBM scRNA-seq tumors (4 highest ranked, 1 lowest ranked). Ideally all GBM scRNA-seq tumors would have a rank of 4 indicating the correct GBM subtype was ranked the highest regardless of hyper-parameters (**Fig. S13**). Similarly, we also calculated the Spearman correlation and plotted the scatter plot between correct label rank and the hyper-parameters used (**Fig. S12**, **Table S9**).

Acknowledgements
We thank the Center for Computational Biology and Bioinformatics for the computational resources and work space to complete the research. We also thank the MMRF for the data generated as part of the Multiple Myeloma Research Foundation Personalized Medicine Initiatives (https://research.themmrf.org and www.themmrf.org), the Allen Institute for Brain Science for the data generated as part of their cell types database, and Mount Sinai/JJ Peters VA Medical Center for the data generated as a part of their brain bank.

Author Contributions
TSJ, CYY, JZ, and KH conceived and designed the project. TSJ, CYY, SX performed the analyses. TSJ and ZH designed the software package. TSJ, CYY, SX, CD, MA, YW, CB, YZ, YL, JZ, and KH interpreted the results. ZH, TW, WS, YW, and CB provided technical guidance. TSJ, CYY, JZ, and KH wrote the manuscript. JZ and KH supervised the project.

Funding
National Institutes of Health NLM-NRSA Fellowship F31LM013056 to T.S.J. and The Ohio State University (Columbus, OH) and departmental start-up funding from the Indiana University School of Medicine (Indianapolis, IN) to K.H.

Conflict of Interest
The authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Kiselev, V.Y., A. Yiu, and M. Hemberg, *scmap: projection of single-cell RNA-seq data across data sets.* Nat Methods, 2018. **15**(5): p. 359-362.
2. Cao, Y., et al., *scRNASeqDB: A Database for RNA-Seq Based Gene Expression Profiles in Human Single Cells.* Genes (Basel), 2017. **8**(12).
3. Abugessaisa, I., et al., *SCPortalen: human and mouse single-cell centric database.* Nucleic Acids Res, 2018. **46**(D1): p. D781-D787.
4. Lahnemann, D., et al., *Eleven grand challenges in single-cell data science.* Genome Biol, 2020. **21**(1): p. 31.
5. Mathys, H., et al., *Single-cell transcriptomic analysis of Alzheimer's disease.* Nature, 2019. **570**(7761): p. 332-337.
6. Rossi, M.A., et al., *Obesity remodels activity and transcriptional state of a lateral hypothalamic brake on feeding.* Science, 2019. **364**(6447): p. 1271-1274.
7. Gawel, D.R., et al., *A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases.* Genome Med, 2019. **11**(1): p. 47.
8. Skinnider, M.A., et al., *Cell type prioritization in single-cell data.* bioRxiv, 2019: p. 2019.12.20.884916.
9. Araujo, T., et al., *Classification of breast cancer histology images using Convolutional Neural Networks.* PLoS One, 2017. **12**(6): p. e0177544.
10. Bardou, D., K. Zhang, and S.M. Ahmad, *Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks.* IEEE Access, 2018. **6**: p. 24680-24693.
11. Amodio, M., et al., *Exploring single-cell data with deep multitasking neural networks.* Nat Methods, 2019. **16**(11): p. 1139-1145.
12. Institute, N.C., *Cancer Statistics*, N.C. Institute, Editor. 2019: Cancer.gov.
13. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.
14. de Wilde, M.C., et al., *Meta-analysis of synaptic pathology in Alzheimer's disease reveals selective molecular vesicular machinery vulnerability.* Alzheimers Dement, 2016. **12**(6): p. 633-44.
15. DeKosky, S.T. and S.W. Scheff, *Synapse loss in frontal cortex biopsies in Alzheimer's disease: correlation with cognitive severity.* Ann Neurol, 1990. **27**(5): p. 457-64.
16. Donev, R., et al., *Neuronal death in Alzheimer's disease and therapeutic opportunities.* J Cell Mol Med, 2009. **13**(11-12): p. 4329-48.
17. Akiyama, H., *Inflammatory response in Alzheimer's disease.* Tohoku J Exp Med, 1994. **174**(3): p. 295-303.
18. Glass, C.K., et al., *Mechanisms underlying inflammation in neurodegeneration.* Cell, 2010. **140**(6): p. 918-34.
19. Hemonnot, A.L., et al., *Microglia in Alzheimer Disease: Well-Known Targets and New Opportunities.* Front Aging Neurosci, 2019. **11**: p. 233.
20. Holtman, I.R., et al., *Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: a co-expression meta-analysis.* Acta Neuropathol Commun, 2015. **3**: p. 31.
21. Wang, M., et al., *The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease.* Sci Data, 2018. **5**: p. 180185.
22. Fu, H., et al., *Tau Pathology Induces Excitatory Neuron Loss, Grid Cell Dysfunction, and Spatial Memory Deficits Reminiscent of Early Alzheimer's Disease.* Neuron, 2017. **93**(3): p. 533-541 e5.

23. Hardy, J., et al., *Region-specific loss of glutamate innervation in Alzheimer's disease.* Neurosci Lett, 1987. **73**(1): p. 77-80.
24. Solas, M., E. Puerta, and M.J. Ramirez, *Treatment Options in Alzheimer s Disease: The GABA Story.* Curr Pharm Des, 2015. **21**(34): p. 4960-71.
25. Morris, J.C., *The Clinical Dementia Rating (CDR): current version and scoring rules.* Neurology, 1993. **43**(11): p. 2412-4.
26. Braak, H. and E. Braak, *Neuropathological stageing of Alzheimer-related changes.* Acta Neuropathol, 1991. **82**(4): p. 239-59.
27. Murphy, M.P. and H. LeVine, 3rd, *Alzheimer's disease and the amyloid-beta peptide.* J Alzheimers Dis, 2010. **19**(1): p. 311-23.
28. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nat Biotechnol, 2018. **36**(5): p. 411-420.
29. Wang, T., et al., *BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes.* Genome Biol, 2019. **20**(1): p. 165.
30. Zhan, F., et al., *The molecular classification of multiple myeloma.* Blood, 2006. **108**(6): p. 2020-8.
31. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**(1): p. 559.
32. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W90-7.
33. Gao, Y., et al., *Single-cell sequencing deciphers a convergent evolution of copy number alterations from primary to circulating tumor cells.* Genome Res, 2017. **27**(8): p. 1312-1322.
34. Han, Y., et al., *Tumor-Induced Generation of Splenic Erythroblast-like Ter-Cells Promotes Tumor Progression.* Cell, 2018. **173**(3): p. 634-648 e12.
35. Bordini, J., et al., *Erythroblast apoptosis and microenvironmental iron restriction trigger anemia in the VK*MYC model of multiple myeloma.* Haematologica, 2015. **100**(6): p. 834-841.
36. Kim, D., et al., *CD19-CD45 low/- CD38 high/CD138+ plasma cells enrich for human tumorigenic myeloma cells.* Leukemia, 2012. **26**(12): p. 2530-7.
37. Abadi, M., et al. *Tensorflow: A system for large-scale machine learning*. in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016.
38. Cancer Genome Atlas Research, N., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
39. Ledergor, G., et al., *Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma.* Nat Med, 2018. **24**(12): p. 1867-1876.
40. Gretton, A., et al., *A kernel two-sample test.* Journal of Machine Learning Research, 2012. **13**(Mar): p. 723-773.
41. Ching, T., X. Zhu, and L.X. Garmire, *Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data.* PLoS Comput Biol, 2018. **14**(4): p. e1006076.
42. Verhaak, R.G., et al., *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.* Cancer Cell, 2010. **17**(1): p. 98-110.
43. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(Nov): p. 2579-2605.
44. Becht, E., et al., *Dimensionality reduction for visualizing single-cell data using UMAP.* Nat Biotechnol, 2018.

45. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* J Stat Softw, 2010. **33**(1): p. 1-22.