Cell-attribute aware community detection improves differential abundance testing from single-cell RNA-Seq data

Alok Maity
SINH https://orcid.org/0000-0001-9275-6400
Andrew Teschendorff (andrew@sinh.ac.cn)
CAS-MPG Partner Institute for Computational Biology https://orcid.org/0000-0001-7410-6527

Article

Keywords: network science, single cell RNA-Seq, graph theory, community detection, aging, immune system, cancer risk

Posted Date: November 4th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2199519/v1

Version of Record: A version of this preprint was published at Nature Communications on June 5th, 2023. See the published version at https://doi.org/10.1038/s41467-023-39017-z.

Abstract

Variations of cell-type proportions within tissues could be informative of biological aging and disease risk. Single-cell RNA-sequencing offers the opportunity to detect such differential abundance patterns, yet this task can be statistically challenging due to the noise in single-cell data, inter-sample variability and because differential abundance (DA) patterns are often characterized by small effect sizes. Here we present a novel DA-testing paradigm called ELVAR, which, unlike the popular Louvain clustering method, takes cell attribute information into account when inferring cell-states within the high-dimensional single-cell manifold. We validate ELVAR using both simulated and real single-cell and single-nucleus RNA-Seq data, demonstrating improved inference over the popular Louvain algorithm and competing DA-testing methods. In lung tissue, ELVAR detects a decrease in the naïve Cd4 + T-cell proportion with age, as well as a shift of alveolar macrophages towards an M2 polarization program. In colon tissue, ELVAR predicts increased stem-cell and T-regulatory fractions in polyps preceding adenoma. In summary, leveraging cell attribute information when inferring cell communities can denoise single-cell data and help retrieve more robust cell states for subsequent DA-testing. ELVAR is available as an open-source R-package.

Introduction

Detecting shifts of cell-type proportions in relation to aging, exposures or disease risk factors is an important task to improve our understanding of disease predisposition and disease onset [1]. Single-cell technologies, and single-cell RNA sequencing (scRNA-Seq) [2] in particular, offer the opportunity to detect such differential abundance (DA) patterns, although this task can be statistically challenging [3]. One key challenge is that, almost inevitably, any assessment of DA across biological conditions entails a comparison of cell-type numbers from assays performed in different subjects. Thus, biological inter-subject variability, as well as technical batch effects, can potentially confound naïve DA-analyses [4]. Technical batch effects can be addressed by performing scRNA-Seq assays in a number of different subjects representing the same biological condition, called sample replicates. Using such biological replicates also helps gauge the biological inter-sample variability, which can be substantial. For instance, it is now well recognized that different individuals may age at different rates [5], and that such variations in biological age may be associated with underlying shifts in T-cell proportions [6]. Thus, taking inter-sample variability into account is critically important for DA-testing. Another major challenge is that the differential abundance patterns of most interest will be of small effect size, involving shifts in cellular states of a given cell-type. A concrete example is immuno-senescence, where the relative proportion of naïve to mature CD4 + T-cells in blood decreases with age [6]. While scRNA-Seq technology allows relatively easy discrimination of major cell-types (e.g. fibroblasts from epithelial cells), the discrimination of underlying cell-states within a cell-type is much harder due to the noisy nature of single-cell data. Thus, it is critically important to devise methods that can robustly identify cell-states in the background of such noisy data, to ensure that the subsequent quantification of DA-patterns is reliable.

In response to these challenges, various statistical DA-testing algorithms have been proposed [4, 7–9], with some of the more recent methods (e.g. [4]) taking sample replication into account. Algorithms also

differ in terms of whether they estimate DA from the discrete cell clusters that are usually inferred from scRNA-Seq data [1, 3, 10], or, as advocated by newer methods, from fuzzier representations known as cellular neighborhoods, in recognition of the fact that cells generally cluster by broad cell-types and not by underlying cell-states (or cell subtypes) [4, 7, 8]. We reasoned however that one key underlying problem with clustering-based approaches is that cell clusters are not inferred using cell-state or cell-attribute information into account. Indeed, we posited that if cell clustering algorithms could be generalized to take cell attribute information into account, for instance, cell-state or a cell's biological condition, that this would improve the signal-to-noise ratio, and, in so doing, allow retrieval of more biologically relevant clusters and robust cell-states. In support of our hypothesis, we note that clustering graph nodes by taking node-attributes into account has been a fruitful approach in network science generally [11], hence this novel paradigm ought to be beneficial for tackling the uncertainties and noise associated with single-cell omic data. Moreover, cell attributes themselves often encode information that is part of the intrinsic process that generates cell-states, so not using such information may lead to incomplete or imprecise clusters. A final reason to consider cell-attribute aware clustering is that it can help discern cellular states that are shared across sample replicates, and which are therefore more likely to be biologically relevant.

To test our hypothesis, we here adapt a node-attribute aware community detection algorithm called EVA [12], which is a generalization of the very popular Louvain clustering method [13]. Of note, despite Louvain's great popularity in the single-cell analysis field [3], the Louvain algorithm only takes the network structure into account when inferring cell communities. EVA can be viewed as a direct extension of the Louvain algorithm allowing multiple cell attributes to be incorporated when inferring communities. This novel concept allows clustering of cells, not only by similarity in the high-dimensional state-space, but also by how similar their attributes (e.g. age, disease stage) are. Here we develop a novel R-implementation of EVA and incorporate it into a novel algorithmic pipeline for DA-testing called ELVAR, that we subsequently validate very extensively on both simulated as well as real datasets, demonstrating improved sensitivity over competing methods.

Results

Rationale of ELVAR algorithm

Detecting subtle shifts in cell-state or cell-subtype proportions across different conditions from scRNA-Seq data is particularly challenging when these factors only drive a small fraction of the overall data variance. Indeed, with scRNA-Seq data, cells generally cluster by the main cell-types present in a tissue, for instance, epithelial, immune and fibroblast types. However, more refined clusterings that clearly discriminate more similar cell-types or cell-states from each other (e.g. different CD4 + T-cell subtypes, or differentiated vs undifferentiated epithelial cells) are generally not forthcoming. Likewise, there are biological conditions (e.g. age of a cell) that may only cause a relatively small change in the transcriptome of cells, rendering the identification of cell-states representing these conditions very difficult. These challenges arise because in scRNA-Seq data relevant components of biological variation often carry less variance compared to technical factors, which may prevent the segregation of cells by

cell-state or subtype. We hypothesized that the identification of robust cellular states could be improved if clustering analyses were to include cell-attributes into account when inferring cellular communities. Specifically, following standard dimensional reduction and inference of a cell-cell nearest neighbor graph (Fig. 1a) [14], we posited that more relevant cellular states could be identified if the inference of communities in this graph were to use cell-attribute information, since this would favor clustering solutions where cells within communities are predominantly of one condition or cell-subtype. To test our hypothesis, we make use of an extension of the popular Louvain algorithm called EVA [12], which in contrast to Louvain, aims to maximize an objective function which also depends on a purity index, which measures how homogeneous inferred communities are in relation to some particular cell-attribute (e.g. age or disease stage) (Fig. 1b). For a given cell-attribute value, communities enriched for cells taking on that value can subsequently be identified (Fig. 1c). Of note, whilst these communities may contain cells from multiple sample replicates, ELVAR does not impose this, safeguarding flexibility and power. If cells have additional attributes such as cell-state or cell-subtype, negative binomial regressions (NBRs) can then be used to assess if the proportions of these cell-states change across the main attribute representing distinct biological conditions (i.e. age or disease stage) [4] (Fig. 1d). We call the resulting DA-testing pipeline ELVAR.

Validation of EVA/ELVAR on simulated data

In order to test our R-implementation of EVA, we devised a simulation model based on scRNA-Seq data from the Tabula Muris Senis [15], consisting of 200 scRNA-Seq profiles representing classical monocytes from one particular mouse, with 100 cells defining a perturbed state (P) and the remaining 100 representing an unperturbed normal (N) condition (**Methods**). We simulated differences in gene expression between the two conditions to be subtle, only involving 0.2% of all genes. Dimensional reduction and visualization with t-stochastic neighborhood embedding (t-SNE) did not reveal any clustering structure except for the distinctively non-random distribution of perturbation-state within the main cluster (**SI fig.S1a**). Louvain clustering over the cell-cell nearest neighbor graph in the higher dimensional state manifold revealed a more complex clustering structure with nine clusters that globally correlated with perturbation state **(SI fig.S1a**). This rich clustering structure not only reflects differences in perturbation state, but also other biological heterogeneity within a typical cell population. Applying EVA with parameter values *a* favoring high purity clustering solutions revealed, as required, stronger correlations with perturbation state, as evaluated using the adjusted Rand Index (ARI) or with Chi-Square statistic P-values (**SI fig.S1a-c**). Restricting to EVA solutions with the same number of inferred communities as Louvain (n = 9), also displayed improved ARI and more significant chi-square statistic P-values compared to Louvain (**SI fig.S1d**). We verified that the improvement of EVA over Louvain was independent of the resolution parameter, which also controls number and size of inferred communities (**Methods, SI fig.S2).**

Next, we generalized the simulation model to include cells from different age groups and multiple mouse replicates, and by altering the frequency of the perturbation state between age groups in order to simulate age-related differential abundance (Fig. 2a). We identified an index purity parameter *a* = 0.8 as a

reasonable choice for EVA on this dataset (Fig. 2b, **Methods**). EVA inferred enriched age-related communities from the nearest neighbor cell-cell graph (Fig. 2c), and using negative binomial regressions to account for inter-mouse replicate variation, we correctly inferred the expected increase of the perturbation state fraction with age (Fig. 2d-e). Importantly, statistical significance attained by EVA was stronger compared to both sequential (deterministic) and non-sequential (stochastic) Louvain algorithms (Fig. 2e, **Methods**). Of note, whilst this improvement over Louvain comes at the expense of a higher computational complexity, EVA runtimes are feasible for reasonably sized cell-cell networks (**SI fig.S3**). For instance, for a network with approximately n = 30,000 cells, EVA runtime on a typical professional workstation only takes around 15 minutes, with runtime growing only mildly faster than $O\left(n\log n\right)$ (**SI fig.S3**) [12].

ELVAR predicts age-associated differential abundance of immune-cell subsets in lung

To validate ELVAR in the context of real scRNA-Seq datasets, we first considered the case of aging in Cd4 + T-cells. As described now by several studies, the naïve subset of Cd4 + T-cells decreases with age in blood, contributing to the well-known phenomenon of immuno-senescence [6, 16]. Hence, ELVAR should be able to predict an analogous age-related shift from naïve Cd4 + T-cells to the more mature subtype in tissues with significant amount of immune-cell infiltration, such as lung [17]. To this end, we considered the lung-tissue 10X scRNA-Seq dataset from the Tabula Muris Senis [15], due to ample profiling of Cd4 + T-cells in this tissue across at least 5 age groups, ranging from one-month (1m) to 30 month-old mice (30m). After QC, a total of 537 Cd4 + T-cells from 11 mice remained, their ages being distributed as 143 (1m), 122 (3m), 67 (18m), 107 (21m) and 98 (30m). Cells from any given age-group were derived from at least two mice (**SI table.S1)**, allowing us to take sample variability into account. Of the 537 Cd4 + T-cells, 186 were identified as being in the naïve state due to expression of *Lef1*, a well-known marker for naïve Cd4 + T-cells [16]. We used Seurat to perform feature selection, dimensional reduction and visualization (**Methods**), resulting in two broad clusters that correlated with age and Cd4 + T-cell subtype (Fig. 3a). Thus, this represents an “easy” scenario where ELVAR should be able to predict the expected shift to a more mature Cd4 + T-cell phenotype. Since ELVAR contains a parameter a that determines the relative importance of a cell’s attribute (i.e. age) when inferring clusters, we first ran ELVAR a total of 100 times for each of 9 choices of a ranging from a = 0.1 to a = 0.9, in order to identify at which parameter value, ELVAR departs from the ordinary Louvain algorithm in terms of clustering quality metrics (**Methods**). We observed that the number of inferred communities did not grow appreciably until a reached a value a = 0.8 (Fig. 3b). Importantly, at this value, purity was substantially higher compared to the ordinary Louvain algorithm, whilst the decrease in overall modularity was relatively small (Fig. 3b). Thus, from this analysis, we concluded that a = 0.8 is the optimal parameter value. For each age group and for each of 100 runs at this optimal a = 0.8 value, ELVAR communities enriched for cells from that age-group were identified (**Methods**, Fig. 3c-d). Ignoring sampling variability revealed a significant skew towards lower naïve cell-fractions in older mice (Fig. 3e). Taking mouse replicates into account also revealed a skew independently of mouse-ID (Fig. 3f). To confirm this, we ran negative binomial regressions, which revealed highly significant and robust negative and positive associations for naïve and mature Cd4t cells,

respectively (Fig. 3g). Importantly, the associations of T-cell subfractions with age were significantly stronger as inferred using ELVAR compared to the deterministic and non-deterministic Louvain algorithms (Fig. 3g, **SI fig.S4, Methods**). Thus, the improvement of ELVAR over Louvain is solely due to the incorporation of cell-attribute information and not due to differences in whether a stochastic or deterministic optimization step is implemented. Using the same data from the TMS [15], we were able to replicate these results for 1124 lung alveolar macrophages, of which 464 were annotated as polarization type M1, 214 as M2, with the rest (n = 446) undetermined (**Methods, SI table.S2**). Specifically, by using ELVAR we were able to detect that the relative distribution of M1 to M2 macrophages among enriched communities shifts towards a lower M1/M2 ratio with increased age (**SI fig.S5**). Importantly, ELVAR revealed a stronger age-related decrease in M1 polarization compared to the deterministic and non-deterministic Louvain algorithms (**SI fig.S5-S6**). In summary, these data demonstrate the ability of ELVAR to detect subtle age-related differential abundance patterns in immune-cell subsets of solid tissues, and that the improvement over Louvain is due to the use of cell-attribute information when inferring cellular communities.

ELVAR compares favorably to DAseq and miloR

In order to better understand the advantages ELVAR offers, we compared it to two competing methods, DAseq [7] and miloR [4]. Of note, although DAseq and miloR allow DA to be assessed in relation to one main cell-attribute, they do not explicitly allow assessment of DA of additional cell-attributes (e.g. cell-states) relative to the main one. Thus, in order to compare ELVAR to DAseq and MiloR in their ability to detect DA of cell-states in relation to another attribute such as age, we adapted the DAseq and MiloR algorithms to this particular DA-task (**Methods**). Applying all these methods to the Cd4t-cell and alveolar macrophage scRNA-Seq datasets, with age as the primary cell-attribute and mature/naïve Cd4t and M1/M2 polarization status as the secondary attributes, respectively, revealed much stronger levels of statistical significance for ELVAR compared to DAseq or miloR (Fig. 4a-d). For instance, whilst all 3 methods correctly predicted an age-related increase and decrease of mature and naive Cd4t cells, respectively, DAseq and MiloR only attained marginal levels of significance, in contrast to the much stronger levels of statistical significance obtained with ELVAR (Fig. 4a-b). In the case of alveolar macrophages, none of the methods predicted age-related changes of M1 or M2 cells, except for ELVAR which did predict an age-related decrease of M1 cells, with DAseq and MiloR failing to pass statistical significance (Fig. 4c-d).

We reasoned that the improved sensitivity of ELVAR may be related to its ability to capture larger communities specifically enriched for the youngest and oldest cells. To test this, we computed the fraction of captured cells for each age-group, defined as cells of a given age-group that belong to significantly age-enriched communities (ELVAR) or to significantly age-associated cellular neighborhoods/regions (MiloR/DAseq) (**Methods**). Supporting our hypothesis, we observed that in both the Cd4t and alveolar macrophage data, ELVAR captured significantly more cells from the youngest and oldest age-groups compared to MiloR or DAseq (Fig. 4e-f). It is noteworthy that in the case of Cd4t-cells, for which the reduction of naïve cells with age is a well-known established fact, ELVAR's improvement

over MiloR and DAseq was specially pronounced for the oldest age-groups (30m), whilst there was no improvement for the intermediate group (18m). This supports the view that MiloR and DAseq struggle to capture larger communities of old cells, probably due to these cells displaying higher heterogeneity and therefore less prone to cluster together in local neighborhoods. In contrast, by incorporating age as a cell-attribute when inferring communities, ELVAR is able to go beyond the local neighborhoods to capture larger clusters of old cells, thus improving power and facilitating the detection of age-related shifts in underlying cell-states.

ELVAR predicts increased stem-cell fraction in polyps

To demonstrate broad applicability of ELVAR, we next applied it to a single-nucleus RNA-Seq (snRNA-Seq) dataset of colon cancer progression, encompassing 8 normal samples from healthy individuals, 16 normal samples from unaffected familial adenomatous polyposis (FAP) cases, 41 polyps from predominantly FAP cases and 4 colorectal cancer adenomas, encompassing over 200,000 cells [18]. We asked if ELVAR could detect disease-associated shifts in the epithelial and lymphocyte compartments. We note that in the original study by Becker et al [18] an increase in the epithelial stem-cell and regulatory T-cell fractions was observed, but only when analyzing scATAC-Seq data, and not from the snRNA-Seq data itself which displayed very high (97–99%) sparsity. We reasoned that ELVAR's improved sensitivity would allow detecting these shifts from the snRNA-Seq itself. To ensure robustness, we performed the analysis in two independent ways (**Methods**). In the first approach, we restricted to a subset of samples for which the QC processing and cell-type annotation was already provided in the original study [18] (**SI table.S3-S4**). Dimensional reduction and visualization with UMAP [19] on this subset revealed that enterocytes and lymphocytes displayed gradients correlating weakly with disease stage and cell-state (**SI fig.S7**). Applying ELVAR to the cell-cell similarity graphs, derived separately for enterocytes and lymphocytes, revealed statistically significant progressive increases in the stem-cell and regulatory T-cell fractions, despite the relatively small numbers of samples (Fig. 5a-d,5f-i). Of note, these increases in the stem-cell and T-regulatory cell fractions were not observed, or were not as significant, when applying DAseq and MiloR, respectively, further attesting to the improved sensitivity displayed by ELVAR (Fig. 5e,5j).

In the second approach, we re-analyzed the full snRNA-Seq dataset, performing QC and annotating cells into broad enterocyte, goblet, immune-cell, stromal and endothelial cell categories (**Methods**). Briefly, to annotate, we identified broad cell-types using only normal samples and well-known cell-type specific markers, to subsequently build an mRNA expression reference matrix, which was then used in a robust partial correlation framework [20, 21] to annotate all cells from all disease stages (**Methods**). Only cells that were confidently annotated into one of the broad categories were taken forward for further analysis (**Methods**). Whilst the high sparsity of the snRNA-Seq data precluded reliable annotation of T-regulatory cells, in the case of stem-like cells, we applied a recently validated single-cell method called CancerStemID [22], that first estimates differentiation activity for a number of colon-specific transcription factors (TFs) [23] across all cells, subsequently identifying stem-like cells as those displaying the lowest average differentiation activity (**Methods**). In CancerStemID, differentiation activity of a TF is estimated

using a corresponding pool of direct and indirect targets called a TF-regulon [23, 24], which as shown by us improves robustness [23]. We observed that the average differentiation activity of the colon-specific TFs decreased during cancer progression (Fig. 6a), and identified 38,667 stem-like and 65,432 non-stem cells, with the stem-like cells displaying much lower levels of differentiation activity (Fig. 6a). Next, we used Seurat to build the k-nn cell-cell graph (k = 50), on which we then applied ELVAR (100 runs) with disease stage as the main cell attribute (Fig. 6b). Alluvial plots indicated that inferred communities enriched for disease stages were also predominantly associated with either stem or non-stem cells (Fig. 6c). Using only cells that were part of communities enriched for disease stages, ELVAR confirmed an increase in the stem-cell fraction, which was particularly pronounced at the polyp-stage (Fig. 6d), and which was further validated using NBRs to account for inter-subject variability (Fig. 6e). In summary, all these data demonstrate that ELVAR is able to detect shifts in disease relevant cell-states from the noisy and challenging snRNA-Seq data, thus extending and confirming similar shifts derived from the scATAC-Seq data [18].

Discussion

Recent scRNA-Seq studies proposing DA-testing methods have argued that reliance on hard clustering algorithms (e.g. Louvain) to infer cellular states and their DA-patterns is problematic due to the small effect sizes associated with these states, which means that clusters representing these states are not well-defined, requiring a fuzzier representation of communities known as neighborhoods or regions [4, 7]. Here we have shown that an alternative solution is to use cell-attribute information when clustering cells. By using cell-attribute information in the community inference procedure, one can more readily discern cellular states defined by the attribute itself, thus helping to circumvent the noise and orthogonal sources of variation which would otherwise preclude identification of such states. Furthermore, in the applications to aging and colon cancer progression considered here, ELVAR displayed higher sensitivity than MiloR/DA-Seq, to detect biologically plausible DA-shifts, such as the age-related shift from naïve to mature Cd4 + T-cells in lung tissue (thus mirroring the corresponding known shift in blood [16, 25]), or the increased epithelial stem-cell fraction in polyps, which represent important biological insights. We also contributed a theoretical understanding underpinning this improved sensitivity, as demonstrated by ELVAR's ability to detect larger cellular communities representing immune-cell states in aged cells, which competing methods like MiloR or DAseq could not resolve due to increased intercellular heterogeneity of aged cells.

The improved inference of DA with cell-attribute aware community detection inevitably comes at the expense of increased computational complexity and runtimes. However, runtimes remain very feasible and hence the increased computational complexity does not present a limitation. Indeed, a typical scRNA-Seq study may profile on the order of 200k cells, encompassing on the order of 10 cell-types, hence on the order of 20k cells per cell-type. Since ELVAR is aimed at detecting subtle shifts of underlying cell-states within one of these cell-types (e.g. mature vs naïve Cd4 + T-cell states), the typical cell-cell graphs on which we would apply ELVAR would have on the order of 20k cells. On such a network, ELVAR can

complete the task in approximately 10–15 minutes. On cell-cell networks encompassing ~ 100k cells, one ELVAR run would complete in the order of 80–120 minutes, and multiple runs can be easily parallelized.

Finally, it is important to re-emphasize the need to infer cellular states within the context of the high-dimensional single-cell state manifold. Given two cellular attributes that are defined to a large extent independently of this high-dimensional manifold, one can in principle always perform DA-testing between these two attributes using NBRs, without the need to apply a cell clustering algorithm to the state-manifold. For instance, in the case of age and a cell-state defined by the binary expression of a single marker gene, DA-testing with multiple replicates could be done using NBRs on cell-counts within this two-dimensional attribute space (age, cell-state). However, in the context of scRNA-Seq data, such an approach has been shown to be suboptimal [4], because it assumes, unreasonably so, that all cells sharing common attributes define the same cellular states. Thus, a cell-attribute aware clustering and DA-testing pipeline such as ELVAR strikes an optimal balance, allowing more biologically relevant and robust cell-states to be inferred, whilst simultaneously also removing the many noisy and rogue cells that are not part of these states.

In summary, ELVAR, and the cell-attribute aware clustering algorithm on which it is based, is a useful addition to the arsenal of statistical methods for DA-testing in scRNA-Seq data. Given the richness and complexity of single-cell omic data, including multi-omic data, general network science approaches will continue to find novel successful applications in this area.

Methods

Single-cell RNA-Seq datasets

We here analyzed the following scRNA-Seq datasets:

Tabula Muris Senis (TMS): This mouse scRNA-Seq dataset [15] encompasses many different tissue-types with samples collected at 6 different ages: 1, 3, 18, 21, 24 and 30 months. Data object files were downloaded from figshare https://doi.org/10.6084/m9.figshare.8273102.v2

We used the normalized data as provided in the h5ad files. We focused on the lung-tissue 10X dataset, because it contained one of the largest numbers of immune-cell subtypes with good representation across age-groups including multiple mouse replicates.

Colon cancer development: This is a human snRNA-Seq dataset [18] encompassing colon samples collected from healthy individuals, normal samples from unaffected individuals with FAP, polyps from FAP and non-FAP cases, and colorectal adenomas. We analysed both the processed snRNA-seq data available from GitHub (https://github.com/winstonbecker/scCRC_continuum), as well as the full unprocessed data available from GEO (GSE201348). Processed data were stored as Seurat objects that included donor, disease stage and cell-type annotation information.

The EVA algorithm

Here we describe the recently published algorithm (EVA: Extended LouVain Algorithm) to identify homogeneous communities in a network with node attributes [26]. Let $G = (V, E, A)$ denote a graph where V, E and A are the set of vertices, edges and node (cell) attributes, respectively. Node/cell attributes can be categorical or numerical such that $A(v)$, with $v \in V$, identifies the set of cell attribute values associated with cell v. In our applications, we will mostly consider one cell attribute (typically the cell's age or perturbation state) but below we formulate the model for any number of node attributes. With EVA, the goal is to identify a network partition, i.e. a mutually exclusive set of communities/clusters, that maximizes a topological clustering criterion as well as node label homogeneity within each community. Thus, the measure we wish to maximize consists of two components: the modularity *Q* that measures the extent to which the partitioning captures clusters of high-edge density, and a purity index *P* that measures the homogeneity of the communities in relation to node attributes.

In more detail, the modularity *Q* of a partition quantifies the edge density within communities relative to that expected under an appropriate null distribution [27], and is defined by

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \gamma \frac{k_v k_w}{2m} \right] \delta(c_v, c_w) \quad (1)$$

where *m* is the number of edges, A_{vw} represents the edge weight between nodes v and w, $k_v = \sum_w A_{vw}$ and $k_w = \sum_v A_{vw}$ are the sum of weights of the edges linked to nodes v and w, respectively. The Kronecker $\delta(c_v, c_w)$ function is 1 when nodes v and w are in the same community (c) and 0 otherwise. The resolution parameter γ, which typically takes values in the range $0 < \gamma \leq 1$, controls the number and size of inferred communities with higher resolution values leading to a greater number of smaller communities. For $\gamma = 1$, *Q* can take values between 0 and 0.5 [12].

The purity index is defined by the average node label homogeneity (i.e. purity) over all inferred communities [26]:

$$P = \frac{1}{|c|} \sum_{c \in C} P_c \quad (2)$$

where P_c represents the purity of community c. The purity of a given community (c) is defined as the product of frequencies of the most frequent node attribute values within the$c$

$$P_c = \prod_A \frac{max_a \left(\sum_{v \in c} a(v) \right)}{|c|} \quad (3)$$

where A is the set of node attributes, and $a(v)$ is an indicator function for node v taking value 1 if $a = A(v)$, 0 otherwise. P_c attains maximum purity 1 when all nodes in the community (c) have same attribute value sequence. In the case of just one cell attribute, this corresponds to the case where all cells in the community have the same attribute value. Of note, *P* takes values in the range 0 to 1.

Finally, the EVA algorithm is defined by the optimization of a generalized modularity function *Z*

$$Z = \alpha P + (1 - \alpha) Q \, (4)$$

where α is the purity index parameter taking values also in the range 0 to 1. Of note, for $\alpha = 0$, *Z = Q* and we recover the Louvain algorithm if $\gamma = 1$ [13] (or a modified Louvain algorithm if $\gamma < 1$). At the other extreme ($\alpha = 1$), *Z = P*, and the clustering algorithm only cares about maximizing purity subject to network connectivity constraints. For a given α and γ, we optimize *Z* following the algorithmic implementation of Citraro and Rossetti [12].

The ELVAR algorithm

Building on EVA, we developed an algorithm and R-package called ELVAR for differential abundance (DA) testing in scRNA-Seq data. Specifically, the biological question being addressed by ELVAR is whether the proportion of a given cell-state (or cell subtype) changes in relation to some other factor or cell attribute (e.g. age, disease stage). Given the potentially high technical and biological variability, such DA-testing should ideally be carried out in scenarios where multiple replicates are available [4]. ELVAR is designed for complex scenarios where (i) the source of variation associated with the cellular states or subtypes is relatively small and (ii) where sample replicates are available. The ELVAR algorithm consists of 4 main steps that we now describe:

1. *Construction of the k nearest-neighbor (knn) cell-cell graph*: As input, the EVA algorithm requires a connected nearest neighbor cell-cell graph, as generated for instance using Seurat's *FindNeighbors* function. The number of nearest neighbors *k* should be chosen sensibly in relation to the total number of cells n_{tot}. For instance, we aim for a ratio $n_{tot}/k \sim 50$, so that for a 1000 cell graph, the number of neighbors is 20. To run EVA, the input graph must also have at least one vertex/node (i.e. cell) attribute, which will be used when inferring communities in the graph. In our applications this main cell attribute will be age or disease status. In general, cells also have other attributes besides the one being used in the community-inference process. For instance, these additional attributes could be the sample replicate (individual/mouse) from which the cell derives, or a particular cell-state or subtype. To avoid confusion, we will call the attribute used in the clustering or community inference as the "main attribute", whilst the attribute being interrogated in DA-testing as the secondary attribute.
2. *Selection of purity index parameter a*: Another important input to the EVA algorithm is the value of the purity index parameter *a* (called α in previous section), which controls the relative importance of purity *P* over modularity *Q* when optimizing *Z.* Typically, we recommend running the EVA algorithm 50 to 100 times for different *a* parameter values ranging from 0.1 to 0.9 (the extremes *a = 0* and *a = 1* are not interesting), in order to assess how *P, Q* and *Z* vary as a function of *a.* It is also important to record the number of inferred communities, as this also depends on the value of *a.* Following the recommendation of Citraro et al [12], we choose *a* such that when compared to Louvain (*a = 0*) there is a clear increase in the purity *P* without much degradation to the modularity *Q.* Ideally, the number of inferred communities must also be higher compared to Louvain, although this is not absolutely

necessary. Thus, although for a fixed *a* EVA works by optimizing *Z*, we do not choose the *a* value which maximizes *Z*, because for most real-world networks, *Z* will increase with *a* and will be maximal when *a = 1.* For typical *knn* cell-cell networks, the optimal *a* value is generally in the range 0.7 to 0.9. As far as the resolution parameter is concerned, we generally consider $\gamma = 1$, as this allows direct benchmarking to the original Louvain algorithm which is also defined for $\gamma = 1$.

3. *Inference of enriched communities with EVA*: Having inferred the optimal parameter *a* value, we now rerun EVA with this input parameter *a* value, to infer communities. The next step is to then identify those communities that are enriched for specific main attribute values. This is done for each main attribute value in turn, using a Binomial test with a stringent Bonferroni-adjusted P-value < 0.05/(number of communities * number of attribute values) threshold. Only communities enriched for specific attribute values are taken forward for further analysis. Thus, the purpose of this important step is to remove cells that don't clearly define cell-states associated with the attribute value. Having found all enriched communities for a given attribute value, we then group together all cells from these enriched communities to define a cell-group per attribute value. The cells within a cell-group may derive from distinct sample replicates.
4. *DA testing for secondary attributes*: For each cell group associated with a main attribute value *v*, we next count the number of cells with any given secondary attribute value contributed by each replicate *r*, whilst also recording the total number of cells contributed by that sample replicate. For each secondary attribute value, we then run a negative binomial regression (NBR) of the cell count against the main attribute value (here assumed ordinal e.g. age-group or disease stage) with the sample replicate's total cell count being the normalization factor. Mathematically, we model the cell count n_{st} of secondary attribute value *t* and sample *s*, as a negative binomial (NB)

$$n_{st} \sim NB(\mu_{st}, \phi_t)$$

where μ_{st} is the mean number and ϕ_t is the dispersion parameter. We further assume that $\mu_{st} = \mu_{vrt} = f_{vt} n_{vr}$, where f_{vt} is the fraction of cells of type *t* when they have main attribute value *v*, and where n_{vr} is the total number of cells derived from sample *s* (which has main attribute value *v* and *r* is the replication index). We next assume that the log of f_{vt} is a linear function of *v*, so that the final regression is of the form

$$\log \mu_{vrt} = \alpha_t + \beta_t \log n_{vr} + \gamma_t v_{vr}$$

$$\log \mu_{st} = \alpha_t + \beta_t \log n_s + \gamma_t v_s$$

Thus, given the cell counts n_{st} we simply run a negative binomial regression (NBR) against the two covariates v_s and $\log n_s$, to find out if the cell counts vary significantly with the main attribute value *v.* The covariate $\log n_s$ plays the role of a normalization factor, accounting for the total of number of cells contributed by sample *s.* Wald z-statistics and P-values of association are obtained from this NBR.

Benchmarking against Louvain is done by direct comparison of these statistics. Because the original Louvain algorithm is deterministic, whilst ELVAR is not (in ELVAR optimization is performed in a non-

sequential random manner), for benchmarking we perform 100 distinct ELVAR runs, comparing the distribution of Wald-test statistics to the Louvain-derived one with a one-sided Wilcoxon rank sum test. Of note, further below we also describe how we benchmark ELVAR to a non-sequential randomized version of Louvain, where the Louvain output may also differ between runs.

ELVAR pseudocode

Below is an outline of pseudocode for running ELVAR. We assume the scRNA-Seq data is encoded in a Seurat object *seu.o* with all the required meta-information, including a main cell attribute to be used in the clustering (clustering attribute "CA"), and a secondary cell attribute for DA-testing that we call "SA". The first step is to normalize the data and to build the k-nearest neighbor cell-cell graph:

Step-1 (Normalization and construction of cell-cell graph):

- *seu.o <- FindVariableFeatures(seu.o,selection.method="vst");*
- *seu.o <- ScaleData(seu.o,features = rownames(seu.o));*
- *seu.o <- RunPCA(seu.o);*
- *Elbowplot(seu.o);* ### to determine number of significant PCs: topPC

The choice of *k* in specifying the number of nearest neighbors should be chosen sensibly. Typically the ratio of number of cells/k should be around 50, so assuming 1000 cells, *k* should be around 20:

- *seu.o <- FindNeighbors(seu.o,dims = 1:topPC,k.param = 20);*
- *adj.m <- as.matrix(seu.o@graphs$RNA_nn); diag(adj.m) <- 0;*
- *gr.o <- graph.adjacency(adj.m,mode="undirected");*
- *vertexN.v <- names(V(gr.o));*
- *vertex_attr(gr.o,name="CA") <- seu.o@meta.data$CA;*
- *is.connected(gr.o); ### check graph is connected (if not, incrementally increase k).*

Step-2 (Estimate optimal purity index parameter a):

- *aOPT <- FindOptimalPurityIndex(gr.o, nRuns = 100);*

Step-3 (Inference of enriched communities with EVA):

- *eva.o <- Eva_partitions(gr.o,alpha = aOPT,Vattr.name="CA");*
- *comm.o <- ProcessEVA(eva.o,seu.o);*

Step-4 (Do DA-testing of a secondary attribute with negative binomial regressions):

- *nbr.o <- DoDA(eva.o,seu.o,comm.o,DAattr="SA");*

The object *nbr.o* is typically the output of the *glm.nb* function from the *MASS* R-package, and statistics of association between the SA (e.g. Cd4t activation status) and CA (e.g. age) attributes can be extracted using *summary(nbr.o)$coeff* .

Louvain with random non-sequential node selection (LouvainRND)

Because EVA is a natural extension of Louvain, it is natural to benchmark it against the ordinary Louvain algorithm as implemented by Blondel et al [13], which is also implemented in the *igraph* R package. The ordinary Louvain algorithm gives a deterministic network partition output, i.e. every run of the Louvain algorithm results in the same partition. This is because during an optimization run nodes are visited and assessed for local moving sequentially. On the other hand, EVA builds upon algorithmic details implemented in the Leiden algorithm [28], which results in potentially different partitions every time EVA is run. Specifically, as with the Leiden algorithm, during an optimization run in EVA, we consider random (non-sequential) selection of nodes, which can therefore result in a different partition every time EVA is run. Thus, when benchmarking EVA against Louvain, it is important to account for these implementation differences. We do this by also benchmarking EVA against a modified version of Louvain (called LouvainRND), where during an optimization run, nodes are visited randomly, which may also result in different partitions for different runs.

Simulation Models

The analysis benchmarking EVA against Louvain, was performed in the context of two simulation models. In the first case, we selected 200 classical monocyte cells from the TMS lung tissue scRNA-Seq 10X dataset, ensuring all cells derive from the same mouse (mouseID = 19) and thus from the same age. For 100 of these cells, we then modified their scRNA-Seq profiles, simulating a “perturbed” cell-state, as follows. We randomly selected 50 genes among all genes not expressed in any of the 200 cells. For each perturbed cell, we then randomly subselected 20 genes from these 50, whose values were then altered in the cell, by randomly drawing 20 non-zero expression values from the distribution of non-zero expression values of the whole data matrix. Thus, this procedure generates a weak but significant co-expression structure among the 100 perturbed cells. Seurat was then applied to the 20138 gene x 200 cell scRNA-Seq data matrix, with VST feature selection followed by PCA. Top-8 PCs were selected to build the k-nearest neighbor cell-cell graph using k = 6. Louvain clustering algorithm as implemented in igraph was used to infer communities. EVA was run on the same cell-cell graph using a cell’s perturbation state as cell-attribute. Since the EVA result depends on initialization, we performed a total of 100 runs for each of nine choices of purity index parameter a (a = 0.1, 0.2, …, 0.8, 0.9). The final value of a was chosen heuristically as the value at which purity increased compared to the Louvain solution (a = 0) without compromising modularity too much. The quality of the EVA and Louvain clustering was assessed using the Adjusted Rand Index against the cell’s perturbation state, as well as using Chi-Square statistics.

For the second simulation model, we selected all classical monocytes from mice with mouse-IDs 0 and 1 representing 1 month old mice (201 & 284 cells), mouse IDs 2 and 3 representing 3 month old mice (51 &

60 cells), mouse IDs 13 and 14 representing 21 month old mice (104 & 138 cells) and mouse IDs 21 and 22 representing 30 month old mice (94 & 61 cells). For young mice (1 and 3 month old mice), 25% of cells were perturbed using the same procedure described previously. For old mice (21 and 30 months), the frequency of perturbed cells was increased to 50%. Thus, this model simulates an age-related increased in a perturbed state. The cell-cell graph was derived as before, this time using the top 10 PCs and k = 20, i.e k was increased in line with the larger number of cells (n = 993). Louvain and EVA were run on this cell-cell graph, in this case using age as the cell-attribute information. As before, EVA was initially run a 100 times for each of nine choices of *a* parameter, in order to select an optimal *a* based on overall purity and modularity values. Using the optimal *a* value, EVA was then compared to Louvain in its ability to predict the increased frequency of the perturbed cell-state in the older mice.

Application of ELVAR to detecting shifts in lung tissue Cd4 + T-cell subtypes

As part of the 10X lung-tissue TMS set, a total of 551 Cd4 + T-cells were profiled to allow testing of a shift in naïve to mature subtypes with age. We removed cells from 4 mice each contributing less than 10 cells, leaving a total of 537 cells from 11 mice representing five age-groups: 143 (1m), 122 (3m), 67 (18m), 107 (21m) and 98 (30m). Cells expressing Lef1, a well-known marker of naïve Cd4 + T-cells [16], were defined as naïve (n = 186), the rest as mature (n = 351). ELVAR was then applied to determine if the naïve/mature proportions change with age. Cell-cell graph was constructed using Seurat with VST and 8 top PCs and k = 10. EVA was run a total of 100 times with a = 0.8 (the optimal value in this dataset).

Application of ELVAR to M1/M2 polarization analysis in lung alveolar macrophages

As part of the 10X lung-tissue TMS set, lung alveolar macrophages were abundantly profiled (n = 1261) to allow testing of a shift in M1/M2 macrophage polarization with age. We removed cells from mice displaying batch effects, leaving a total of 1124 cells from 15 mice representing five age-groups: 517 (1m), 184 (3m), 193 (18m), 91 (21m) and 139 (30m). In order to annotate these 1124 lung alveolar macrophages into M1/M2 subtypes, we first identified 5 robust murine M1 (Cd80, Cd86, Fpr2, Tlr2, Cd40) and 5 robust M2 markers (Egr2, Myc, Arg1, Mrc1, Cd163) from the literature [29]. In an initial annotation, we declared cells as M1 if they co-expressed at least 2 of the 5 M1 markers, and similarly for M2. Cells annotated to both M1 and M2 subtypes were re-assigned an undetermined (UD) category alongside all other cells not annotated to either M1 and M2, resulting in 308 M1, 195 M2 and 621 UD-cells. We reasoned that UD-cells clustering predominantly with either M1 or M2 cells could be re-assigned to M1/M2 subtypes. To this end, we developed an iterative algorithm that reassigns the status of UD-cells to either M1 and M2, depending on their relative proportions among the neighbors of a given UD-cell. In more detail, we used the cell-cell graph as inferred using Seurat, and a multinomial test with $P < 0.05$ threshold to identify the UD-cells whose polarization status could be reassigned to either M1 or M2 status. We also required the absolute difference between the proportion of M1 and M2 neighbors of a given UD-cell to be larger than 0.2. This procedure was iterated 20 times, but numbers already converged

after 7 iterations, resulting in 464 M1, 214 M2 and 446 UD-cells. ELVAR was then applied to determine if the M1/M2 proportions change with age. Cell-cell graph was constructed using Seurat with VST and 9 top PCs and k = 20. EVA was run a total of 100 times with a = 0.7 (the optimal value in this dataset).

Application of ELVAR to colon cancer progression

We applied ELVAR to explore if the fractions of epithelial stem-cells and T-regulatory cells changes with disease progression. The analysis was performed in two ways. In the first approach, we downloaded Seurat objects from https://github.com/winstonbecker/scCRC_continuum which contain QC-processed data and cell-type annotations for a subset of samples. In the case of epithelial cells, the analysis was performed on a subset of the data consisting of stem-cells, TA2 & TA1 transit amplifying progenitors, enterocyte progenitors, immature enterocytes and differentiated enterocytes cell states. We randomly picked 1000 cells from each cell state (thus a total of 6000 cells) in order to reduce the computational runtime because two cell states contained ≥ 30k cells. Next, we removed cells from 2 donors each contributing less than 10 cells and cells from 1 donor displaying a batch effect, leaving a total of 5810 cells from 11 donors representing four disease stages: 1153 (Normal), 1672 (Unaffected), 2911 (Polyp) and 74 (Adenocarcinoma). The distribution of cell-states was 843 stem-cells, 971 TA2, 1000 TA1, 999 enterocyte progenitors, 998 immature enterocytes and 999 enterocytes number of cells. A cell-cell graph was constructed using Seurat with variance stabilization for feature selection, and selecting the 15 top PCs with k = 20. ELVAR was run a total of 100 times with a = 0.8 and disease stage as the main cell-attribute for community detection. For the analysis of T-regulatory (Tregs) cells we focused on the subset of data consisting of Tregs, NK, Naïve T, CD4 + and CD8 + cells. We removed cells from 3 donors each contributing less than 10 cells and cells from 1 donor displaying a batch effect, resulting in a total of 6171 cells from 8 donors representing four disease stage groups: 381 (Normal), 3721 (Unaffected), 1900 (Polyp) and 169 (Adenocarcinoma). The distribution of cells across cell-types was: 472 Tregs, 245 NK, 1042 Naïve T, 3063 CD4 + and 1349 CD8 + number of cells. The cell-cell similarity graph was constructed using Seurat with variance stabilization for feature selection, selecting the 15 top PCs and number of nearest neighbors k = 20. ELVAR was run a total of 100 times with a = 0.8 with disease stage as the main clustering attribute.

In the second approach, we downloaded the raw count snRNA-Seq matrices from GEO (GSE201348). Data was normalized using Seurat with variance stabilization for feature selection, leaving a total of 380,527 cells. Because the cell-type annotation for the full dataset was not provided, we applied dimensional reduction, clustering, UMAP visualization and well-known marker genes from Becker et al [18] to the cells from the normal samples only, to annotate well separated cell clusters into enterocyte, goblet, immune-cell, stromal and endothelial cell categories. We then used Wilcoxon tests and marker-specificity scores [20, 21] to build an mRNA expression reference matrix for these 5 broad cell categories. With this mRNA expression reference matrix in place, we then used our robust partial correlation framework [21] to estimate cell-type probabilities for all cells from all disease stages. Using a probability threshold of > 0.7, we were thus able to confidently annotate 1866 endothelial cells, 104009 enterocyte cells, 78421 goblet cells, 24973 immune-cells and 7941 stromal cells. Because of the very high sparsity

of the snRNA-Seq data, in order to confidently identify stem-like cells among the 104,009 enterocytes, we applied our validated CancerStemID algorithm [22, 23] which approximates stemness of single-cells from the estimated differentiation activities of tissue-specific transcription factors. In this instance, we used a set of 56 colon-specific TFs and their associated regulons, already validated by us previously [23]. The regulons were applied to the snRNA-Seq data, to estimate transcription factor differentiation activity (TFA) for each of the 56 TFs across each of the 104,009 enterocyte cells. We then declared stem-cell like cells as those displaying average TFA levels over the 56 TFs less than a threshold given by the 5% quantile of the average TFA distribution defined over the normal cells only. For ELVAR analysis, we only retained samples contributing at least 50 cells. For all other samples, all cells up to a maximum of 500 randomly selected cells were chosen, resulting in a total of 31,385 cells, drawn from 69 samples (8 normal, 16 unaffected FAPs, 41 polyps and 4 CRCs), encompassing 3761 normal, 7443 unaffected, 18558 polyp and 1623 CRC cells. The cell-cell k = 50 nearest neighbor graph was constructed using Seurat, and ELVAR run a 100 times with $a = 0.8$, using disease stage as main cell attribute.

Comparison of ELVAR to DAseq and MiloR

We compare the three algorithms in their ability to detect proportions of cell-states as a function of a main cell-attribute. In our context this main cell-attribute could be age or disease-status, and cell-states could refer to Cd4t activation status, or differentiation stage. What the three algorithms have in common is the inference of groups of cells that display differential abundance relative to the main cell-attribute. The methods differ in how these groups of cells are inferred. In ELVAR, we use the main cell-attribute information when inferring cellular communities from the nearest neighbor cell-cell graph, subsequently identifying those that display enrichment for any specific attribute value (e.g. age-group). In contrast, DAseq [7] and MiloR [4] infer local regions, or potentially overlapping cellular neighborhoods, displaying significant DA of age-groups. Thus, one way to compare all three algorithms is by selecting the cells that appear in these significant communities/regions/neighborhoods, and subsequently running negative binomial regressions of cell-state counts vs age-group, taking biological replicates into account and normalizing for the total number of cells that each replicate sample contributes to each age-group, as described earlier for ELVAR.

To understand the difference in performance between methods, we developed the following metric. Methods may display different sensitivity to detect DA of a secondary attribute relative to the main one because the significantly associated cell groups derived from each method (i.e. age-group enriched communities in the case of ELVAR, age-associated neighborhoods/regions in the case of MiloR/DAseq) may capture different numbers of cells. To make this clear, consider a scenario where one of the methods (call it “X”) can’t detect a cell group with sufficient numbers of old cells, say it detects a cell-group with at most 10 old cells, with 6 of these belonging to one cell-state “A”, with the remaining 4 belonging to another cell-state “B”. In contrast, another method “Y” does infer a large enough cell-group consisting of old-cells, say 30 cells with 15 belong to state “A” and 15 belonging to state “B”. As far as young cells are concerned, all methods are able to infer a cell-group with a considerable number of cells, say 50 young cells, with 40 belonging to state “A” and 10 belonging to state “B”. Because method "X” was not able to

identify a cell-group with sufficient numbers of old-cells, it lacks power to detect the relative decrease of state “A” with age (two-tailed Fisher-test P = 0.22), whilst method “Y” has the power to detect it (two-tailed Fisher-test, P = 0.006). Whilst this hypothetical example ignores the variation due to sample replicates or variations due to replicate cell numbers, it clearly illustrates that the fraction of captured cells (*fCaptCells*) per main attribute value will strongly influence a method’s power to detect DA of an underlying cell-state with respect to this main cell attribute. Mathematically, we define the fraction of captured cells per attribute value *a* and from method *m* by:

$$fCaptCells_{ma} = \frac{|(\#Cells with attribute value = a) \cap (\#Cells in groups from method m)|}{(\#Cells with attribute value = a)}$$

Specifically, a method that attains higher $fCaptCells_{ma}$ across the whole range of attribute values *a*, including the extremes if the attribute is ordinal, will display higher power.

Declarations

Data Availability: The snRNA-Seq dataset of colon cancer progression is publicly available from GEO (www.ncbi.nlm.nih.gov/geo/) under accession number GSE201348. The TMS scRNA-Seq data is available from *https://doi.org/10.6084/m9.figshare.8273102.v2*.

Code Availability: ELVAR is freely available as an R-package from *https://github.com/aet21/ELVAR*

Acknowledgements: This work was supported by NSFC (National Science Foundation of China) grants, grant numbers 31970632 and 32170652. We would like to thank the TMS consortium for making their data open-access and to everyone who supports open-access data.

References

1. Ramachandran P, Dobie R, Wilson-Kanamori JR, Dora EF, Henderson BEP, Luu NT, Portman JR, Matchett KP, Brice M, Marwick JA, et al: **Resolving the fibrotic niche of human liver cirrhosis at single-cell level**. Nature 2019, **575**:512–518.
2. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al: **mRNA-Seq whole-transcriptome analysis of a single cell**. Nat Methods 2009, **6**:377–382.
3. Kiselev VY, Andrews TS, Hemberg M: **Challenges in unsupervised clustering of single-cell RNA-seq data**. Nat Rev Genet 2019, **20**:273–282.
4. Dann E, Henderson NC, Teichmann SA, Morgan MD, Marioni JC: **Differential abundance testing on single-cell data using k-nearest neighbor graphs**. Nat Biotechnol 2021.
5. Horvath S, Raj K: **DNA methylation-based biomarkers and the epigenetic clock theory of ageing**. Nat Rev Genet 2018, **19**:371–384.
6. Jonkman TH, Dekkers KF, Slieker RC, Grant CD, Ikram MA, van Greevenbroek MMJ, Franke L, Veldink JH, Boomsma DI, Slagboom PE, et al: **Functional genomics analysis identifies T and NK cell**

activation as a driver of epigenetic clock progression. Genome Biol 2022, **23**:24.

7. Zhao J, Jaffe A, Li H, Lindenbaum O, Sefik E, Jackson R, Cheng X, Flavell RA, Kluger Y: **Detection of differentially abundant cell subpopulations in scRNA-seq data**. Proc Natl Acad Sci U S A 2021, **118**.
8. Burkhardt DB, Stanley JS, 3rd, Tong A, Perdigoto AL, Gigante SA, Herold KC, Wolf G, Giraldez AJ, van Dijk D, Krishnaswamy S: **Quantifying the effect of experimental perturbations at single-cell resolution**. Nat Biotechnol 2021, **39**:619–629.
9. Lun ATL, Richard AC, Marioni JC: **Testing for differential abundance in mass cytometry data**. Nat Methods 2017, **14**:707–709.
10. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M: **SC3: consensus clustering of single-cell RNA-seq data**. Nat Methods 2017, **14**:483–486.
11. Yang J, McAuley J, Leskovec J: **Community Detection in Networks with Node Attributes**. In *IEEE 13th International Conference on Data Mining*. IEEE; 2013
12. Citraro S, Rossetti G: **Identifying and exploiting homogeneous communities in labeled networks**. Applied Network Science 2020, **5**:55.
13. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E: **Fast unfolding of communities in large networks**. J Stat Mech: Theory and Experiment 2008, **10**:P10008.
14. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R: **Integrating single-cell transcriptomic data across different conditions, technologies, and species**. Nat Biotechnol 2018, **36**:411–420.
15. Tabula Muris C: **A single-cell transcriptomic atlas characterizes ageing tissues in the mouse**. Nature 2020, **583**:590–595.
16. Elyahu Y, Hekselman I, Eizenberg-Magar I, Berner O, Strominger I, Schiller M, Mittal K, Nemirovsky A, Eremenko E, Vital A, et al: **Aging promotes reorganization of the CD4 T cell landscape toward extreme regulatory and effector phenotypes**. Sci Adv 2019, **5**:eaaw8330.
17. Zheng SC, Webster AP, Dong D, Feber A, Graham DG, Sullivan R, Jevons S, Lovat LB, Beck S, Widschwendter M, Teschendorff AE: **A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix**. Epigenomics 2018, **10**:925–940.
18. Becker WR, Nevins SA, Chen DC, Chiu R, Horning AM, Guha TK, Laquindanum R, Mills M, Chaib H, Ladabaum U, et al: **Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer**. Nat Genet 2022, **54**:985–995.
19. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell EW: **Dimensionality reduction for visualizing single-cell data using UMAP**. Nat Biotechnol 2018.
20. Zhu T, Liu J, Beck S, Pan S, Capper D, Lechner M, Thirlwell C, Breeze CE, Teschendorff AE: **A pan-tissue DNA methylation atlas enables in silico decomposition of human tissue methylomes at cell-type resolution**. Nat Methods 2022, **19**:296–306.
21. Teschendorff AE, Zhu T, Breeze CE, Beck S: **EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data**. Genome Biol 2020, **21**:221.

22. Liu T, Zhao X, Lin Y, Luo Q, Zhang S, Xi Y, Chen Y, Lin L, Fan W, Yang J, et al: **Computational identification of preneoplastic cells displaying high stemness and risk of cancer progression**. Cancer Res 2022.
23. Teschendorff AE, Wang N: **Improved detection of tumor suppressor events in single-cell RNA-Seq data**. NPJ Genom Med 2020, **5**:43.
24. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context**. BMC Bioinformatics 2006, **7 Suppl 1**:S7.
25. Maity AK, Hu X, Zhu T, Teschendorff AE: **Inference of age-associated transcription factor regulatory activity changes in single cells**. Nat Aging 2022, **2**:548–561.
26. Citraro S, Rossetti G: **Identifying and exploiting homogeneous communities in labeled networks**. Applied Network Science 2020, **5**:55.
27. Newman ME: **Analysis of weighted networks**. Phys Rev E Stat Nonlin Soft Matter Phys 2004, **70**:056131.
28. Traag VA, Waltman L, van Eck NJ: **From Louvain to Leiden: guaranteeing well-connected communities**. Sci Rep 2019, **9**:5233.
29. Jablonski KA, Amici SA, Webb LM, Ruiz-Rosado Jde D, Popovich PG, Partida-Sanchez S, Guerau-de-Arellano M: **Novel Markers to Delineate Murine M1 and M2 Macrophages**. PLoS One 2015, **10**:e0145342.

Figures

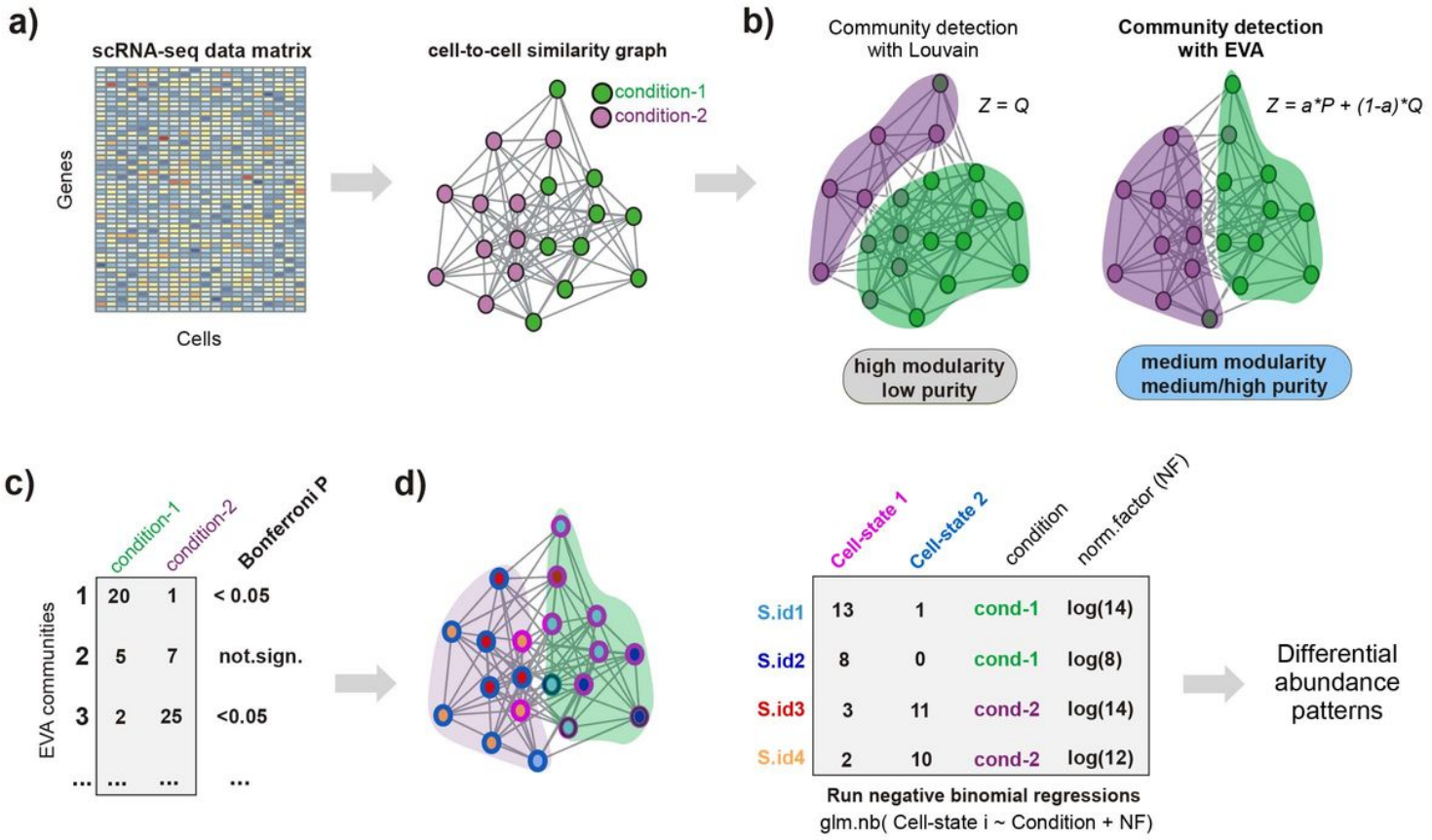

Figure 1

Flowchart of ELVAR algorithm. a) Given a scRNA-Seq data matrix with cells derived from various conditions (e.g. age-groups), one first derives a cell-cell similarity graph using standard pipelines like Seurat. Cells may also differ in terms of cell-state and the sample replicate it is derived from. **b)** To infer communities from this cell-cell graph, we use an extended Louvain algorithm (EVA) which, unlike the standard Louvain algorithm, takes cell attribute information into account when deriving the communities. In this case, the cell-attribute could be the condition it is derived from, in which case the inferred communities will be more enriched for cells of the same condition, as shown. Compared to the standard Louvain algorithm, which aims to maximize the overall modularity Q of the communities, EVA aims to maximize a weighted sum of Q and the overall purity P (a measure of how pure the communities are in relation to the conditions). The a parameter controls the relative importance of Q and P when maximizing the objective function Z. **c)** EVA communities that are significantly enriched for cells from a particular condition are selected for further downstream analysis, thus removing noisy cellular neighborhoods. **d)** For a given condition, cells from all communities enriched for that condition are merged and the distribution of underlying cell-states from each sample replicate are computed. Finally, negative binomial regressions are used to infer if given cell-state fractions vary significantly with condition, whilst taking sampling variability into account.

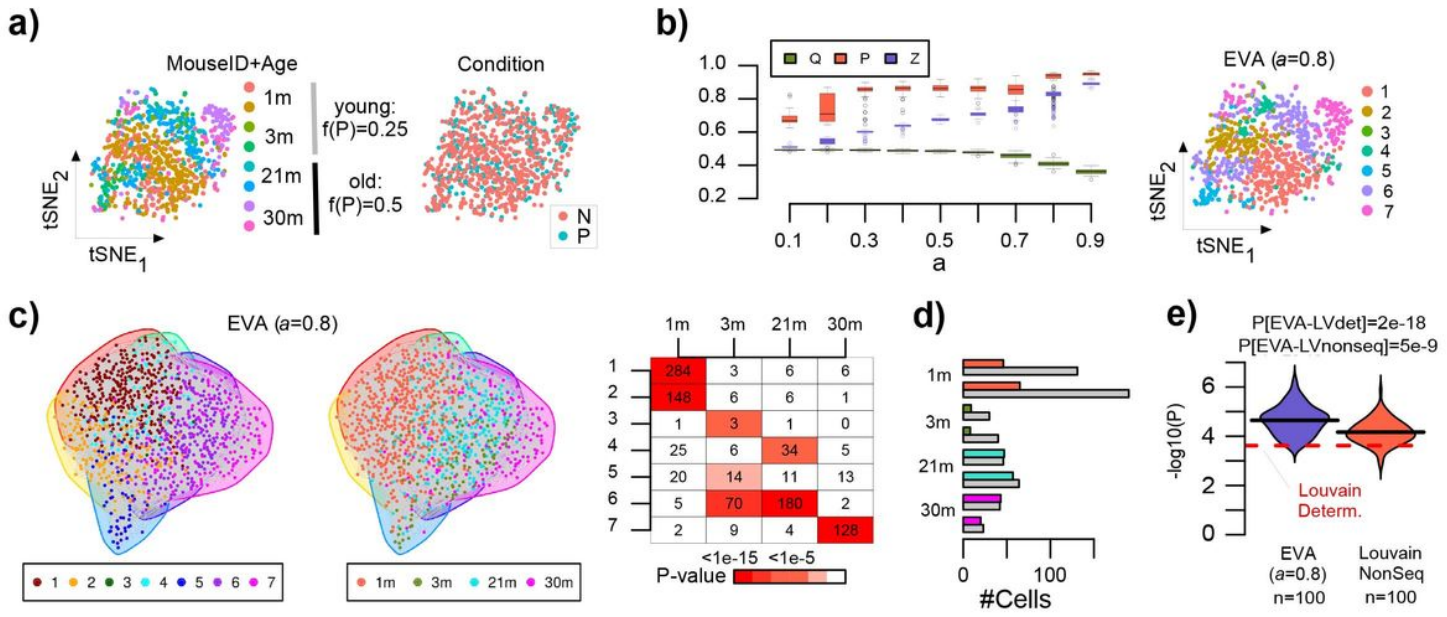

Figure 2

Benchmarking EVA against Louvain. a) tSNE visualization of a simulated scRNA-Seq dataset consisting of 993 mouse c cells drawn from 8 mice encompassing 4 age-groups (1 month, 3 months, 21 months and 30 months) with two mouse replicates per age-group. In the right panel, cells are annotated by perturbation state, where the frequency of cells being in the perturbed (P) state increases from 0.25 in young mice (1 & 3m) to 0.5 in old mice (21 & 30m). **b)** Boxplots displaying the modularity (Q), the purity (P) and generalized modularity (Z) as a function of purity index parameter *a* for EVA. Each boxplot contains the values of 100 distinct EVA runs**.** Right panel is the same tSNE plot as in a) but with cells annotated by the clusters of one particular EVA run. **c)** Nearest neighbor cell-cell graph on which the EVA algorithm is run. Left panel: cells annotated by clusters inferred in one particular EVA run. Middle panel: cells annotated by age-group. Right panel: confusion matrix between the communities inferred with EVA (same run) and age-groups, with the number of cells and Binomial test P-value of enrichment shown. **d)** Barplot displays the number of normal (N, grey bars) and perturbed (P, colored bars) cells from each mouse and age-group, using only cells from EVA communities enriched for specific age-groups (same run as in c)). **e)** Violin plot compares the statistical significance (y-axis, $-\log_{10}P$) of P-values from a negative binomial regression of perturbed cell number against age-group as derived from ELVAR (100 runs) against the corresponding statistical significance value derived from using deterministic and non-sequential (non-deterministic, 100 runs) Louvain. P-values shown are from a one-sided Wilcoxon rank sum test comparing the 100 ELVAR values to the deterministic Louvain one, or to the 100 values from the non-sequential Louvain.

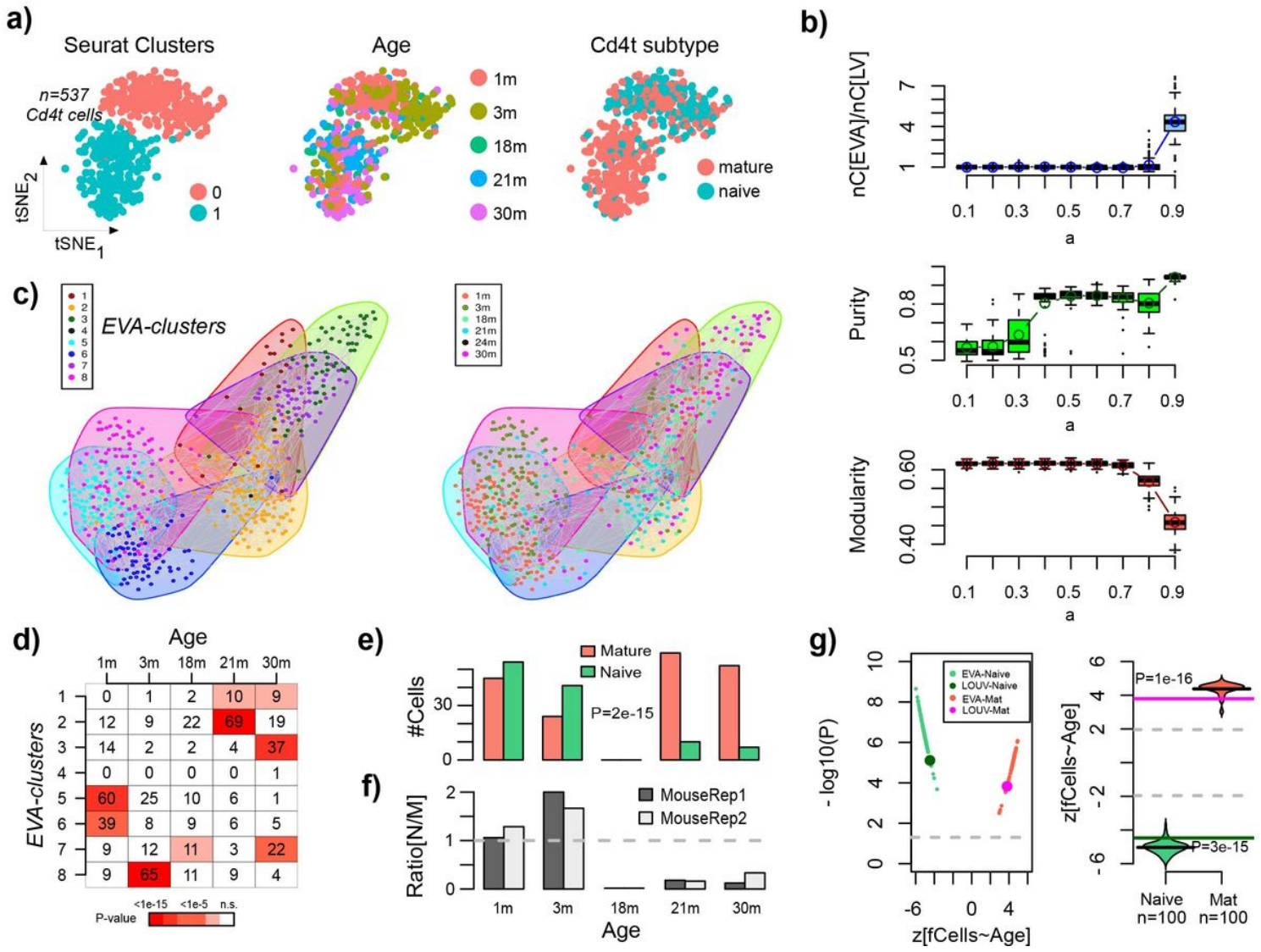

Figure 3

Validation of ELVAR in Cd4t cells from lung tissue. a) tSNE visualization of 537 Cd4+ T-cells with cells annotated by inferred Seurat cluster (left), by age-group (middle) and by Cd4+ T-cell subset (naïve vs mature) (right). **b)** Top panel: boxplots display the number of communities inferred using EVA against the purity parameter *a* , normalized relative to the number of communities inferred with the Louvain algorithm (*a=0*). Each boxplot represents the distribution over 100 different runs. Middle panel: As top-panel but now with y-axis displaying the purity of the clusterings. Lower panel: As top-panel but now with y-axis displaying the modularity of the clusterings. **c)** Left panel: Cell-cell nearest neighbor graph inferred using Seurat, with cell colors indicating the inferred EVA communities from one typical run. Right panel: as left-panel, but with cells now colored by age-group. **d)** Matrix entries give the number of cells per EVA-cluster and age-group, with color indicating the P-value of enrichment, for one particular run. For a given age-group, only cells from enriched clusters are taken forward using a Bonferroni-adjusted threshold (typically around 0.001). **e)** Barplot displaying the number of mature and naïve Cd4t cells per age-group only using enriched clusters from d). P-value is from a two-tailed Fisher-test. **f)** Barplots display for each mouse-replicate the ratio of naïve to mature Cd4t cells. **g)** Left panel: scatterplot of z-statistics (x-axis) vs.

statistical significance (-$\log_{10}$P-value) (y-axis) derived from a negative binomial regression analysis. There are 100 datapoints, one for each run, with the statistics derived from the naïve and mature cell-fractions displayed in different colors as indicated. The darker bigger datapoints are the result from the ordinary deterministic Louvain (LV) algorithm. Right panel: violin plots displaying the z-statistics (100 runs) derived from the negative binomial regression for the case of naïve and mature cell-type fractions, with solid horizontal lines indicating the z-statistics from the ordinary deterministic Louvain algorithm. P-values derive from a one-tailed Wilcoxon rank sum test comparing the distributions in the violins to the Louvain-values.

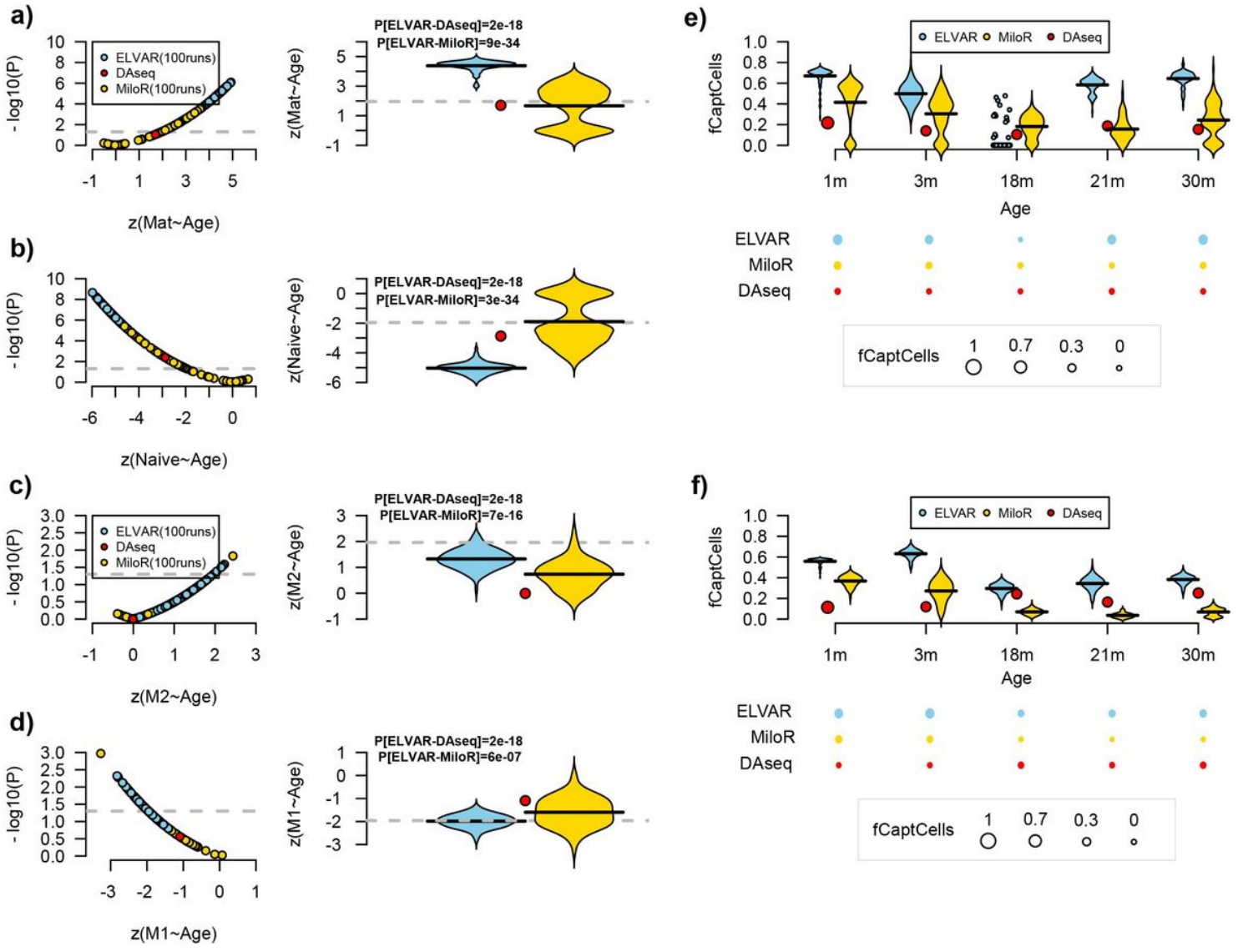

Figure 4

ELVAR compares favorably to DAseq and MiloR. a) Left: Scatterplot of significance levels (y-axis, -$\log_{10}$P) vs z-statistics (x-axis) derived from negative binomial regressions of the mature Cd4t cell fraction vs age, as estimated from the lung-tissue Cd4t scRNA-Seq dataset from the Tabula Muris Senis. For ELVAR and MiloR, results are displayed for 100 different runs. Right: Violin plots displaying the corresponding z-statistics. P-values derive from a one-tailed Wilcoxon rank sum test comparing ELVAR derived z-statistics

to those from DAseq and MiloR, respectively **b)** As a), but for the naïve Cd4t cell fraction. **c-d)** As a-b), but for the lung-tissue alveolar macrophage scRNA-Seq dataset from the Tabula Muris Senis, and considering M2 and M1 macrophages, respectively. **e)** Upper panel: Violin plots displaying the fraction of captured Cd4t cells (fCaptCells, y-axis) of each age-group (x-axis) and by each method. By captured cells we mean those found in significantly age-group enriched communities (ELVAR) or significantly age-associated neighborhoods/regions (MiloR/DAseq). For ELVAR and MiloR we performed 100 runs, as results differ between runs. Lower panel: balloon plot displaying the average fraction of each method and age-group. **f)** As e) but for the alveolar macrophages.

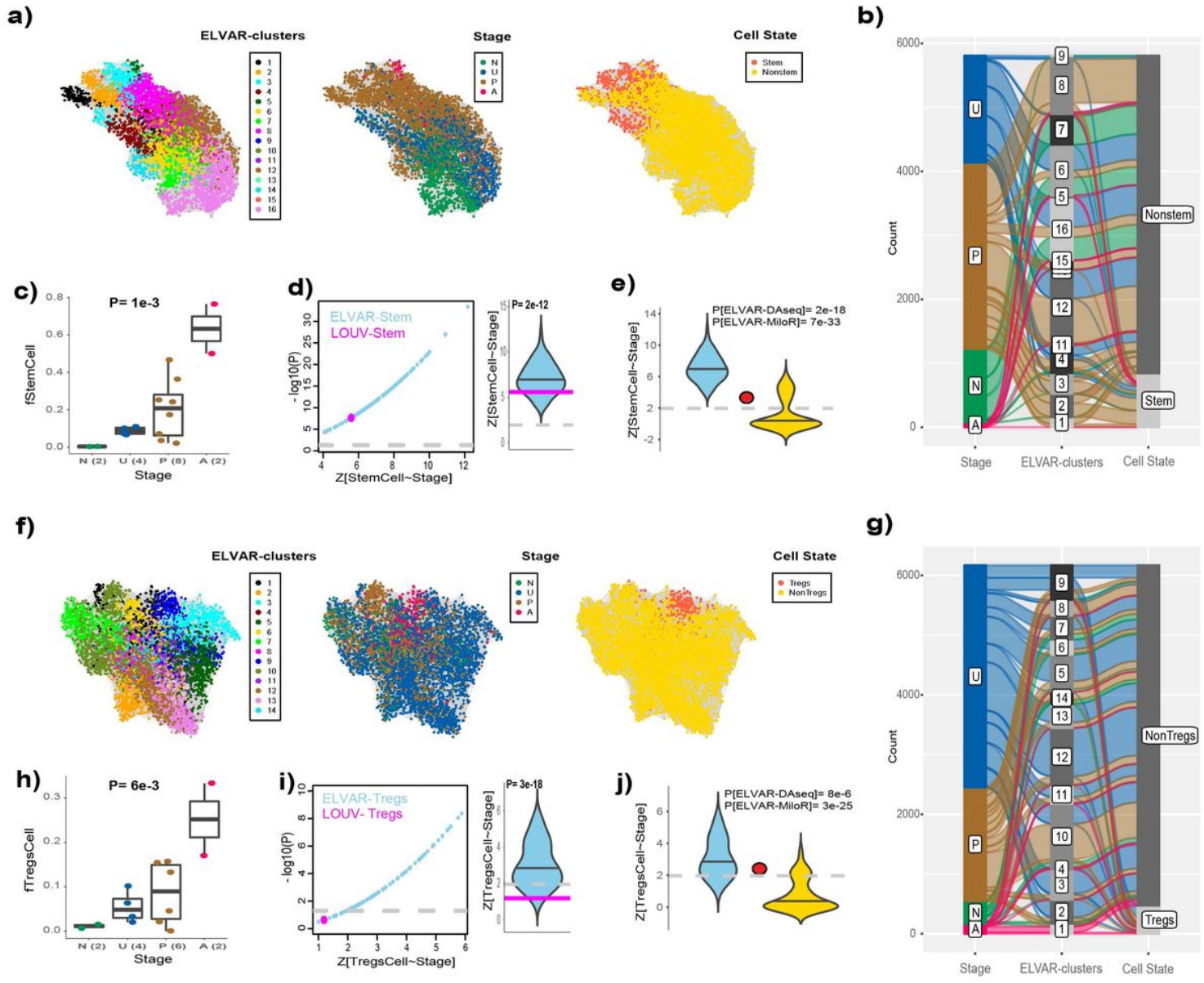

Figure 5

ELVAR predicts increased stem-cell and T-regulatory cell fractions in polyps. a) Left panel: The cell-cell similarity graph inferred using Seurat on scRNA-Seq data with epithelial enterocyte lineage cells annotated by community membership, as inferred using ELVAR. Middle and Right panels depict the same

graph but with cells annotated by disease stage (N=normal, U=unaffected, P=polyp, A=adenoma) and cell-state. Data is shown for one representative ELVAR run. **b)** Alluvial plot displaying composition of ELVAR communities according to disease stage and cell-state. **c)** Boxplot displaying the stem-cell fraction as a function of disease stage, considering only cells that are part of significantly enriched ELVAR-clusters. P-value derives from a linear regression. **d)** Left panel: Scatterplot of z-statistics (x-axis) vs. statistical significance (-log10 P-value) (y-axis) derived from a negative binomial regression analysis of stem-cell counts vs. disease stage including a normalization factor. There are 100 datapoints, one for each ELVAR run. The magenta colored datapoint is the result from the deterministic Louvain (LV) algorithm. Right panel: violin plot displaying the same z-statistics (100 ELVAR runs) with solid horizontal line indicating the z-statistic from the deterministic Louvain algorithm. P-value derives from a one-tailed Wilcoxon rank sum test comparing the ELVAR distribution to the Louvain-value. **e)** Violin plots comparing the ELVAR (skyblue) and MiloR (gold) z-statistics (100 runs each) of stem-cell counts against disease stage. The red datapoint is the z-statistic derived from the DA-seq method. P-values derive from a one-tailed Wilcoxon rank sum test comparing the ELVAR distribution to the DA-seq value and ELVAR to MiloR distributions. **f,g,h,i-j)** As a-e), but now for the lymphocytes and T-regulatory cell fraction.

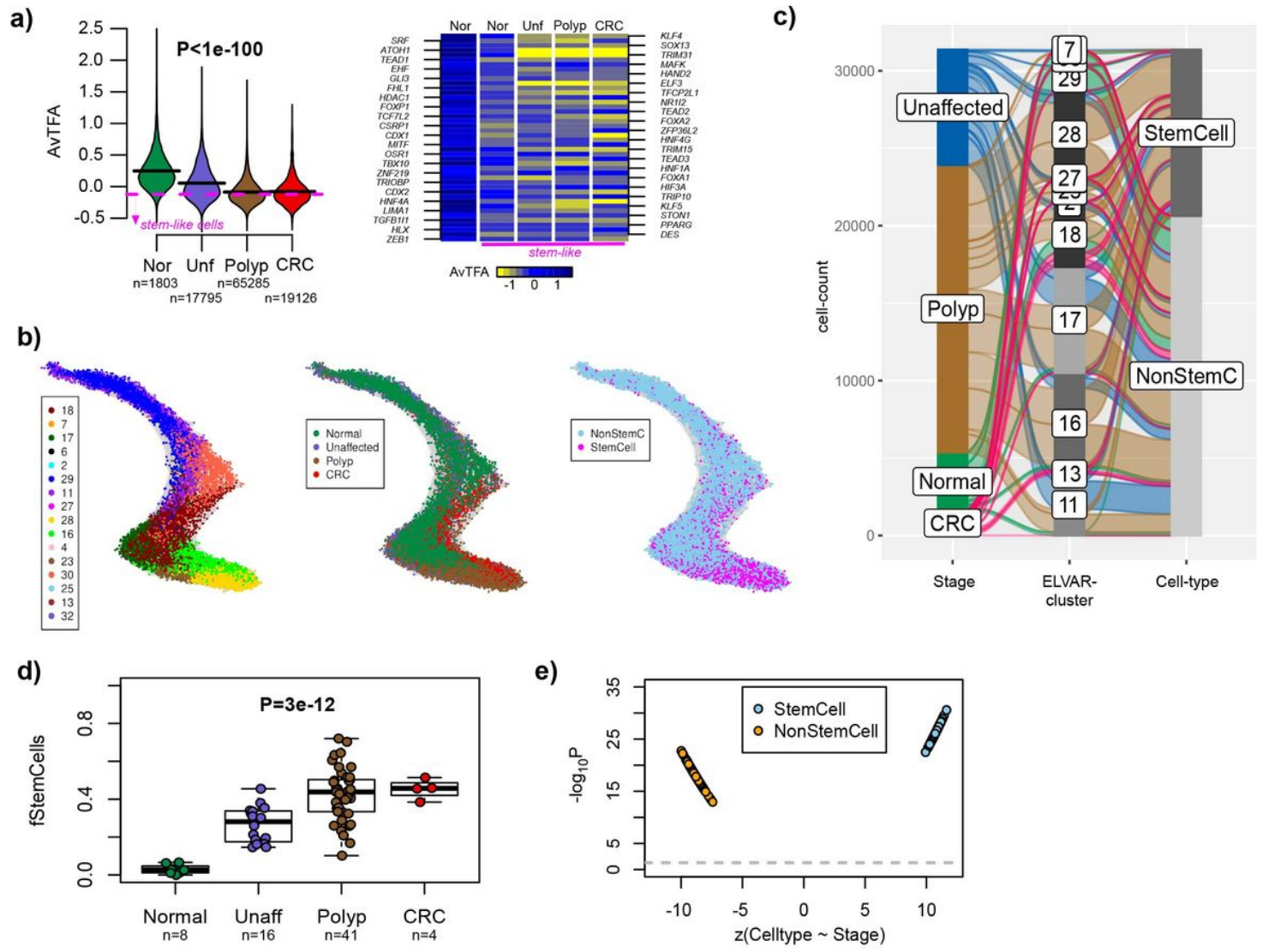

Figure 6

ELVAR validation of increased stem cell fraction in polyps. a) Left panel: Violin plots comparing the average differentiation activity (AvTFA), as estimated over 56 colon-specific TFs, across the four disease stages (Nor=normal, Unf=unaffected FAP cases, Polyp=predominantly FAP cases with polyps, CRC=colorectal adenocarcinoma (predominantly non-FAP)). P-value is two-sided from a linear regression. Pink dashed line indicates the 95% lower quantile of AvTFA values. Number of cells in each disease stage is indicated. Right panel: Heatmap of TFA values for a subset of 44 colon-specific TFs that display lower TFA in unaffected + polyp-carrying FAP cases compared to normal, with TFA values averaged over all normal cells, normal stem-like cells, unaffected stem-like cells, polyp stem-like cells and CRC stem-like cells. The stem-like cells were defined as in a). **b)** Cell-cell nearest neighbor graph (k=50) with cells annotated by inferred EVA/ELVAR community, disease stage and cell-state (stem-like vs non-stem cell). Data is shown for one EVA run. **c)** Alluvial plot for the same EVA run, displaying the composition of communities according to disease stage and cell-state. **d)** Boxplot comparing the fraction of stem-like cells in each disease stage, with the fraction computed for each independent sample using only cells from enriched communities. P-value is two-sided from a linear regression. **e)** Corresponding negative binomial regression analysis, displaying the level of statistical significance (y-axis) to the z-statistics of cell counts vs disease stage (x-axis). For each cell-state, there are 100 values representing 100 distinct ELVAR runs. Grey dashed line indicates the $-\log_{10}(0.05)$ significance level.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuppTablesELVAR.xls
- SuppInfoELVAR.pdf