# Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome

Andrew E. Teschendorff [1,2,3,*] and Tariq Enver [3]

1. CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, 320 Yue Yang Road, Shanghai 200031, China.

2. Department of Women's Cancer, University College London, 74 Huntley Street, London WC1E 6AU, United Kingdom.

3. UCL Cancer Institute, Paul O'Gorman Building, University College London, 72 Huntley Street, London WC1E 6BT, United Kingdom.

*Corresponding author: Andrew E. Teschendorff- a.teschendorff@ucl.ac.uk , andrew@picb.ac.cn

# Abstract

**The ability to quantify differentiation potential of single cells is a task of critical importance. Here we demonstrate, using over 7,000 single-cell RNA-Seq profiles, that differentiation potency of a single cell can be approximated by computing the signaling promiscuity, or entropy, of a cell's transcriptome in the context of an interaction network, without the need for feature selection. We show that signaling entropy provides a more accurate and robust potency estimate than other entropy-based measures, driven in part by a subtle positive correlation between the transcriptome and connectome. Signaling entropy identifies known cell subpopulations of varying potency and drug resistant cancer stem-cell phenotypes, including those derived from circulating tumor cells. It further reveals that expression heterogeneity within single-cell populations is regulated. In summary, signaling entropy allows in-silico estimation of the differentiation potency and plasticity of single-cells and bulk samples, providing a means to identify normal and cancer stem cell phenotypes.**

**Keywords: Single-Cell; RNA-Seq; Stem-Cell; Differentiation; Cancer; Entropy**

**Software Availability:** Signaling Entropy is available as part of the Single Cell Entropy (SCENT) R-package and is freely available from github: https://github.com/aet21/SCENT

One of the most important tasks in single-cell RNA-sequencing studies is the identification and quantification of "intercellular transcriptomic heterogeneity", i.e. variation between the transcriptomes of single cells that is of biological relevance [1-4]. Although some of the observed intercellular transcriptomic variation represents stochastic noise, a substantial component has been shown to be of functional importance [1,5-8]. Very often, this biologically relevant heterogeneity can be attributed to cells occupying states of different potency or plasticity. Thus, quantification of differentiation potency, or more generally functional plasticity, at the single-cell level is of paramount importance. However, currently there is no concrete theoretical and computational model for estimating such plasticity at the single cell level.

Here we make significant progress towards addressing this challenge. We propose a very general model for estimating cellular plasticity. A key feature of this model is the computation of signaling entropy [9], which quantifies the degree of uncertainty, or promiscuity, of a cell's gene expression levels in the context of a cellular interaction network. In effect, signaling entropy uses the transcriptomic profile of a cell to quantify the relative activation levels of its molecular pathways, and more generally that of biological processes, as defined over an a-priori specified protein interaction network. We show that signaling entropy provides an excellent and robust proxy to the differentiation potential of a cell in Waddington's epigenetic landscape [10], and further provides a framework in which to understand the overall differentiation potency and transcriptomic heterogeneity of a cell population in terms of single-cell potencies. Attesting to its general nature and broad applicability, we compute and validate signaling entropy in over 7000 single cells of variable degrees of differentiation potency and phenotypic plasticity, including time-course differentiation data, neoplastic cells and circulating tumor cells (CTCs). This extends entropy concepts that we have previously demonstrated to work on bulk tissue data [9,11-13] to the single-cell level. Based on signaling entropy, we develop a novel algorithm called SCENT (Single Cell Entropy), which can be used to identify and quantify biologically relevant expression heterogeneity in single-cell populations, as well as to reconstruct cell-lineage trajectories from time-course data. In this regard, SCENT differs substantially from other single-cell algorithms like Monocle [14], MPath [15], SCUBA [16], Diffusion Pseudotime [17] or StemID [18], in that it uses single-cell entropy to independently order single cells in pseudo-time (i.e. differentiation potency), without the need for feature selection or clustering.

# Results

### The signaling entropy framework

A pluripotent cell (by definition endowed with the capacity to differentiate into effectively all major cell-lineages) does not express a preference for any particular lineage, thus requiring a similar basal activity of all lineage-specifying transcription factors [9,19]. Viewing a cell's choice to commit to a particular lineage as a probabilistic process, pluripotency can therefore be characterized by a state of high uncertainty, or entropy, because all lineage-choices are equally likely (**Fig.1A**). In contrast, for a differentiated cell, or for a cell committed to a particular lineage, signaling uncertainty/entropy is reduced, as this requires activation of a specific signaling pathway reflecting that lineage choice (**Fig.1A)**. Thus, a measure of global signaling entropy, if computable, could provide us with a relatively good proxy of a cell's overall differentiation potential. Here we propose that differentiation potential can be estimated *in-silico* by integrating a cell's transcriptomic profile with a high quality protein-protein-interaction (PPI) network to define a cell-specific probabilistic signaling process (in effect, a random walk) on the network (**Online Methods**). Mathematically, this random walk is described by a stochastic matrix whose entries reflect the relative interaction probabilities. Underlying the construction of these probabilities is the assumption that two genes, which can interact at the protein level, are more likely to do so if both are highly expressed (**Fig.1A, Online Methods**). Given this stochastic matrix, global signaling entropy is then computed as the entropy rate (abbreviated as SR) of this probabilistic signaling process on the network [20] (**Fig.1B, Online Methods**), and can be thought of as quantifying the overall level of signaling promiscuity of biological processes within the network. In effect, this quantifies the efficiency, or speed, with which signaling can diffuse over the whole network, and therefore measures the number of separate biological processes which are in some sense “active”. Since a committed, or differentiated cell, preferentially activates and deactivates specific processes (pathways) in the network, the expectation is that this would manifest itself as a lower entropy rate since signaling can’t diffuse to the regions of the network describing inactive processes.

### Signaling entropy approximates differentiation potency

To test that signaling entropy correlates with differentiation potency, we first estimated it for 1018 single-cell RNA-seq profiles generated by Chu et al [21], which included pluripotent human embryonic stem cells (hESCs) and hESC-derived progenitor cells representing the 3 main germ-layers (endoderm, mesoderm and ectoderm) (“Chu et al set”, **Supplementary Table 1, Online Methods**). In detail, these were 374 cells from two hESC lines (H1 & H9), 173 neural progenitor cells (NPCs), 138 definite endoderm progenitors (DEPs), 105 endothelial cells representing mesoderm derivatives, as well as 69 trophoblast (TB) cells and 148 human foreskin fibroblasts (HFFs). Confirming our hypothesis, pluripotent hESCs attained the highest signaling entropy values, followed by multipotent cells (NPCs, DEPs), and with less multipotent HFFs, TBs and ECs attaining the lowest values (**Fig.2A**). Differences were highly statistically significant, with DEPs exhibiting significantly lower entropy values than hESCs (Wilcoxon rank sum $P<1e-50$ (**Fig.2A**). Likewise, TBs exhibited lower entropy than hESCs ($P<1e-50$), but higher than HFFs ($P<1e-7$) (**Fig.2A**). Importantly, signaling entropy correlated very strongly with a pluripotency score obtained using a previously published pluripotency gene expression signature [22] (Spearman Correlation = 0.91, $P<1e-500$, **Fig.2B**, **Online Methods**). In all, signaling entropy provided a highly accurate discriminator of pluripotency versus non-pluripotency at the single cell level (AUC=0.96, Wilcoxon test $P<1e-300$, **Fig.2C).** We note that in contrast with pluripotency expression signatures, this strong association with pluripotency was obtained without the need for any feature selection or training.

To further test the general validity and robustness of signaling entropy we computed it for scRNA-Seq profiles of 3256 non-malignant cells derived from the microenvironment of 19 melanomas (Melanoma set, [23], **Supplementary Table 1**). Cells profiled included T-cells, B-cells, natural-killer (NK) cells, macrophages, fully differentiated endothelial cells and cancer-associated fibroblasts (CAFs). For a given cell-type and individual, variation between single cells was substantial and similar to the variation seen between individuals (**Supplementary Fig.1**). Mean entropy values however, were generally stable, showing little inter-individual variation, except for T-cells from 4 out of 15 patients, which exhibited a distinctively different distribution (**Supplementary Fig.1**). In order to assess overall trends, we pooled the single-cell entropy data from all patients together, which confirmed that all lymphocytes (T-cells, B-cells and NK-cells) had similar average signaling entropy values (**Fig.2D**). Intra-tumor macrophages, which are derived from monocytes, exhibited a marginally higher signaling entropy (**Fig.2D**). The highest signaling entropy values were attained by endothelial cells and CAFs (**Fig.2D**), consistent with their known high phenotypic plasticity [24-27]. Importantly, the entropy values for all of these non-malignant differentiated cell-types were distinctively lower compared to those of hESCs and progenitor cells from Chu et al (**Figs.2A & 2D**), consistent with the fact that hESCs and progenitors have much higher differentiation potency. To test this formally, we compared hESCs, mesoderm progenitors, and terminally differentiated cells within the mesoderm lineage (which included all endothelial cells and lymphocytes), which revealed a consistent decrease in signaling entropy between all three potency states (Wilcoxon rank test P<1e-50, **Fig.2E**). Of note, signaling entropy could discriminate progenitor and differentiated cells better than the score derived from the pluripotency gene expression signature [22], attesting to its increased robustness as a general measure of differentiation potency (**Fig.2F**, **Supplementary Fig.2**).

Next, we assessed signaling entropy in the context of a time-course differentiation experiment, whereby hESCs were induced to differentiate into definite endoderm progenitors via the mesoendoderm intermediate [28]. scRNA-Seq for a total of 758 single cells, obtained at 6 timepoints, including origin, 12, 24, 36, 72 and 96 hours post-induction were available (**Online Methods**) [28]. We observed that single cell entropies exhibited a particular large decrease only after 72 hours (**Fig.2G**), consistent with previous knowledge that differentiation into definite endoderm occurs around 3-4 days after induction [28]. To demonstrate the validity of signaling entropy in another species, we next considered a scRNA-Seq data of cells sampled at different embryonic stages in the development of the mouse lung epithelium [29] ("Treutlein set", **Supplementary Table 1, Online Methods**). Signaling entropy decreased continuously until adulthood in line with a gradual increase in differentiation (**Fig.2H**). Moreover, at embryonic day 18, it could discriminate alveolar type cells from a recently discovered bipotent progenitor subgroup [29], albeit with marginal significance due to small cell numbers (**Supplementary Fig.3A**).

To demonstrate the critical importance of the interaction network, we recomputed signaling entropy in the Chu and Treutlein datasets after randomly reshuffling gene expression values over the network (100 and 1000 permutations, respectively). As expected, upon reshuffling, signaling entropy lost its power to discriminate pluripotent from non-pluripotent cells (**Fig.2I**), and did not exhibit a consistent decrease with developmental stage in Treutlein's set (**Supplementary Fig.3B**).

## Robustness to choice of PPI network and NGS platform

Given the importance of the PPI network, it is therefore equally important to verify that signaling entropy is robust to the choice of network. Results were largely unchanged using a different version of a PPI network (**Supplementary Fig.4**). In order to test the robustness of signaling entropy across independent studies, we analyzed scRNA-Seq data for an independent set of single cell hESCs derived from the primary outgrowth of the inner cell mass ("hESC set" [30], **Supplementary Table 1**). Obtained signaling entropy values were most similar to those of single cells derived from the H1 and H9 hESC lines, confirming the robustness of signaling entropy across different studies and next-generation sequencing platforms (**Fig.2J, Supplementary Table 1**).

## Comparison of Signaling Entropy to StemID and SLICE

To further highlight the importance of the PPI network, we decided to compare Signaling Entropy to two other entropy-based potency measures, proposed as part of the StemID [18] and SLICE [31] algorithms, which we note do not use any network information. To provide an objective evaluation, we compared the entropy measures of single cells from well-separated differentiation stages, or by comparing start and end points in time course differentiation experiments, as these cells ought to differ substantially in terms of potency. Adopting this strategy in 4 scRNA-Seq and 1 bulk RNA-Seq dataset, we observed that Signaling Entropy was able to provide high discriminative power in each dataset (**Table 1**). In contrast, we did not find StemID and SLICE to be as accurate or robust (**Table 1**).

## Correlation with potency is independent of cell-cycle phase

A major source of variation in scRNA-Seq data is cell-cycle phase [23,32]. We explored the relation between signaling entropy and cell-cycle phase in a large scRNA-Seq dataset encompassing 3256 non-malignant and 1257 cancer cells derived from the microenvironment of melanomas (Melanoma set, [23], **Supplementary Table 1**). A cycling score for both G1-S and G2-M phases and for each cell was obtained using a validated procedure [23,32,33], and compared to signaling entropy, which revealed a strong yet highly non-linear correlation (**Supplementary Fig.5**). Specifically, we observed that cells with a low signaling entropy were never found in either the G1-S or G2-M phase (**Supplementary Fig.5**). In contrast, cells with high signaling entropy could be found in either a cycling or non-cycling phase. These results are consistent with the view that cycling-cells must increase expression of promiscuous signaling proteins and hence exhibit an increased signaling entropy. Thus, we next asked if signaling entropy correlates with potency when restricting to non-cycling cells. Using the Chu et al dataset, we observed that, although discrimination accuracies were reduced upon correction for cell-cycle phase, signaling entropy could still accurately classify pluripotent from non-pluripotent cell-types (AUC > 0.9, P<1e-5, **Supplementary Fig.6, Supplementary Table 2**). Consistent with this (and now using both cycling and non-cycling cells), the correlation between signaling entropy and potency remained significant when adjusted for cell-cycle scores (**Supplementary Table 2**).

## Correlation of expression with degree partly drives potency

In order to gain further biological insight into signaling entropy, we derived an approximation for signaling entropy in terms of the 3-way correlation between the transcriptome, connectome and local signaling entropies (**Online Methods**). This approximation implies that if, on average, network hubs are more highly expressed than low-degree nodes and if they exhibit an increase in their local signaling entropy, then this should generally lead to a more efficient distribution of signaling over the network, and hence to an increased global signaling entropy [12]. We thus posited that in cells with a demand for high phenotypic plasticity (e.g. pluripotent cells), hubs tend to be overexpressed and exhibit increased signaling promiscuity. Using scRNA-Seq data from Chu et al [21], we were able to confirm a weak (Pearson correlation of ~0.2) but significant (P<1e-50) positive correlation of differential gene expression (between hESCs and multipotent cells) with connectivity (**Supplementary Fig.7A**). Importantly, the differential local signaling entropy between hESCs and multipotent cells correlated more strongly with connectivity (Pearson correlation of ~0.64, P<1e-100, **Supplementary Fig.7A**), thus confirming the notion that the increased SR in pluripotent cells is also driven by a more distributed signaling (i.e. increased local entropy) at network hubs. To demonstrate that the Pearson correlation between transcriptome and connectome can be used to approximate signaling entropy (SR), we computed it for all 1018 single-cells in Chu et al, obtaining an excellent agreement with SR ($R^2 = 0.96$, **Supplementary Fig.7B**), and hence also with potency (**Supplementary Fig.7C**). However, we stress that this Pearson correlation approximation is not a substitute for SR, since the definition of SR includes the local signaling entropies (**Fig.1B**), from which important biological information can be extracted. To demonstrate this, we ranked genes in the network according to their differential local signaling entropy (**Online Methods**) and performed Gene Set Enrichment Analysis [34] on the genes exhibiting the most significant increases in local entropy between pluripotent (hESCs) and multipotent cells. Top-ranked enriched biological terms included, besides stemness, genes implicated in mRNA splicing and encoding mitochondrial ribosomal proteins (**Supplementary Table 3, Supplementary Data 1**). This is consistent with recent studies demonstrating that mitochondrial activity influences the global transcription and splicing rate of cells [35-37], and that variations in such activity may influence stemness and differentiation [38-42]. Finally, we also point out that signaling entropy and its Pearson correlation approximation are not equivalent, as there exist networks where both measures yield very different answers (**Online Methods**). For instance, in networks where hubs are not connected to each other (unlike our PPI networks where hubs are generally connected to each other), a positive correlation could lead to a lower signaling entropy (**Supplementary Fig.7D**).

## Quantifying single-cell expression heterogeneity with SCENT

Given that signaling entropy correlates with differentiation potency, we used it to develop the SCENT algorithm (**Fig.1C**). Briefly, SCENT uses the estimated single-cell entropies to infer the distribution of discrete potency states across the cell population (**Fig.1C, Online Methods**). Thus, SCENT can be used to quantify expression heterogeneity at the level of potency. In addition, SCENT can be used to directly order single cells in pseudo-time [14] to facilitate reconstruction of lineage trajectories. A key feature of SCENT is the assignment of each cell to a unique potency state and co-expression cluster, which results in the identification of potency-clusters (which we call "landmarks"), through which lineage trajectories are then inferred (**Online Methods**).

We first tested SCENT on the scRNA-Seq data from Chu et al, which profiled pluripotent and multipotent cells (**Supplementary Table 1**). SCENT correctly predicted a parsimonious 2-state model, with a high potency pluripotent state and a lower potency non-pluripotent progenitor-like state (**Fig.3A**). Interestingly, a small fraction (approximately 4%) of hESCs were deemed to be non-pluripotent cells (**Fig.3B**), consistent with previous observations that pluripotent cell populations contain cells that are already primed for differentiation into specific lineages [5,6]. Supporting this, these non-pluripotent "hESCs" exhibited lower cycling-scores and higher expression levels of neural (*HES1/SOX2*) and mesoderm (*PECAM1*) stem-cell markers, compared to the pluripotent hESCs (**Supplementary Fig.8**). Whereas all HFFs and ECs were deemed non-pluripotent, definite endoderm progenitors (DEPs), TBs and NPCs exhibited mixed proportions, with NPCs exhibiting approximately equal numbers of pluripotent and non-pluripotent cells (**Fig.3B**). Correspondingly, the Shannon index, which quantifies the level of heterogeneity in potency, was highest for the NPC population (**Fig.3C**). In total, SCENT predicted 6 co-expression clusters, which combined with the two potency states, resulted in a total of 7 landmark clusters (**Fig.3D**). These landmarks correlated very strongly with cell-type, with only NPCs being distributed across two landmarks of different potency (**Fig.3E**). SCENT correctly inferred a lineage trajectory between the high potency NPC subpopulation and its lower potency counterpart, as well as a trajectory between hESCs and DEPs (**Fig.3F**). The other cell-types exhibited lower entropies (**Fig.2B & Fig.3F**), and correspondingly did not exhibit a direct trajectory to hESCs, suggesting several intermediate states which were not sampled in this experiment.

To ascertain the biological significance of the two NPC subpopulations (**Fig.3B,E,F**), we first verified that the NPCs deemed pluripotent did indeed have a higher pluripotency score (**Supplementary Fig.9A**), as assessed using the independent pluripotency gene expression signature from Palmer et al [22]. We further reasoned that well-known transcription factors marking neural stem/progenitor cells, such as *HES1*, would be expressed at a much lower level in the NPCs deemed pluripotent compared to the non-pluripotent ones, since the latter are more likely to represent *bona-fide* NPCs. Confirming this, NPCs with low *HES1* expression exhibited higher differentiation potential than NPCs with high *HES1* expression (Wilcoxon rank sum test P<0.0001, **Fig.3G**). Similar results were evident for other neural progenitor/stem cell markers such as *PAX6* and *SOX2* (**Supplementary Fig.9B**). Of note, NPCs expressing the lowest levels of *PAX6, HES1* or *SOX2* were generally always classified by SCENT into a pluripotent-like state (**Fig.3G, Supplementary Fig.9B**). Thus, these results indicate that SCENT provides a biologically meaningful characterization of intercellular transcriptomic heterogeneity.

## SCENT reconstructs lineage trajectories in differentiation

We next tested SCENT in the context of a differentiation experiment of human myoblasts [14], involving skeletal muscle myoblasts which were first expanded under high mitogen conditions and later induced to differentiate by switching to a low serum medium (Trapnell et al set, **Supplementary Table 1**). A total of 96 cells were profiled with RNA-Seq at differentiation induction, as well as at 24h and 48h after medium switch, with a remaining 84 cells profiled at 72h. As expected, signaling entropy was highest in the myoblasts, with a switch to lower entropy occurring at 24h (**Fig.4A**). No further decrease in entropy was observed between 24 and 72h, indicating that commitment of cells to become differentiated skeletal muscle cells already happens early in the differentiation process. Over the whole timecourse, SCENT predicted a total of 3 potency states, with a distribution consistent with the time of sampling (**Fig.4B**). Cells sampled at differentiation induction were made up primarily of two potency states (**Fig.4C,** PS1 & PS2), which differed in terms of *CDK1* expression, consistent with one subset (PS1) defining a highly proliferative subpopulation and with the rest (PS2) representing cells that have exited the cell-cycle (**Supplementary Fig.10**). In total, SCENT predicted 4 landmarks, with one landmark defining undifferentiated (t=0) myoblasts of high potency (**Fig.4D**). Another landmark of lower potency contained cells at all time points, with cells expressing markers of mesenchymal cells (e.g *PDFGRA* and *FN1/LTBP2*) (**Fig.4D**). Cells from this landmark which were present at differentiation induction exhibited intermediate potency expressing low levels of *CDK1* (**Supplementary Fig.10 & Fig.4D**), suggesting that these are "contaminating" interstitial mesenchymal cells that were already present at the start of the time course, in line with previous observations [14,15]. Importantly, SCENT correctly predicts that the potency of all these mesenchymal cells in this landmark does not change during the time-course, consistent with the fact that these cells are not primed to differentiate into skeletal muscle cells, but which nevertheless aid the differentiation process [14,15]. Another landmark of intermediate potency predicted by SCENT defined a trajectory made up of cells expressing high levels of myogenic markers (*MYOG & IGF2*) from 24h onwards (**Fig.4D**). Thus, this landmark corresponds to cells that are effectively committed to becoming fully mature skeletal muscle cells. The final landmark consisted of cells exhibiting the lowest level of potency and emerged only at 48h, becoming most prominent at 72h (**Fig.4D**). As with the previous landmark, cells in this group also expressed myogenic markers, and likely represent a terminally differentiated and more mature state of skeletal muscle cells. In summary, SCENT inferred lineage trajectories that are highly consistent with known biology and with those obtained by previous algorithms such as Monocle [14] and MPath [15]. However, in contrast to Monocle and MPath, SCENT inferred these reconstructions without the explicit need of knowing the time-point at which samples were collected.

## SCENT detects drug resistant cancer stem cell phenotypes

Cancer cells are known to be less differentiated and to acquire a more plastic phenotype compared to non-malignant cells. Hence their signaling entropy should be higher than that of non-malignant cell-types. We confirmed this using scRNA-Seq data from 12 melanomas (Melanoma-set [23], **Supplementary Table 1**), for which sufficient normal and cancer cells had been profiled (**Fig.5A, Supplementary Fig.11**). Although there was some variation in the signaling entropy of cancer cells between individuals, this variation was relatively small in comparison to the difference in entropy between cancer and normal cells. Combining data across all 12 patients, demonstrated a dramatic increase in the signaling entropy of single cancer cells compared to non-malignant ones (Wilcoxon rank sum test P<1e-500, **Fig.5B**).

Since signaling entropy is increased in cancer and correlates with stemness, it could, in principle, be used to identify putative cancer stem cells (CSC) or drug resistant cells. To test this, we first computed and compared signaling entropy values for 38 acute myeloid leukemia (AML) bulk samples from 19 AML patients, consisting of 19 diagnostic/relapse pairs [43]. Confirming that signaling entropy marks drug resistant cell populations, we observed a higher entropy in the relapsed samples (paired Wilcox test P=0.004, **Fig.5C**). For one relapsed sample, scRNA-Seq for 96 single AML cells was available (AML set, **Supplementary Table 1**). We posited that comparing the signaling entropy values of these 96 cells would allow us to identify a CSC-like subset responsible for relapse. Since in AML there are well accepted CSC markers (*CD34, CD96*), we tested whether expression of these markers in high entropy AML single cells is higher than in low entropy AML single cells (**Fig.5D**). Both *CD34* and *CD96* were more highly expressed in the high entropy AML single cells (Wilcox test P=0.008 and 0.032, respectively, **Fig.5D**).

We next computed signaling entropies for 73 circulating tumor cells (CTCs) derived from 11 castration resistant prostate cancer patients (CTC-PrCa set, **Supplementary Table 1**), of which 5 patients exhibited progression under treatment with enzalutamide (an androgen receptor (AR) inhibitor) (n=36 CTCs), with the other 6 patients not having received treatment (n=37 CTCs) [44]. Although of marginal significance, signaling entropy was higher in the CTCs from patients exhibiting resistance (Wilcox test P=0.047, **Fig.5E**). Among putative prostate cancer stem cell markers (e.g. *CD44, CD133, KLF4* and *ALDH7A1*) [44], we observed a positive association of signaling entropy with *ALDH7A1* expression, suggesting that *ADLH7A1* (and not other markers such as CD44) may mark specific prostate CSCs which are resistant to enzalutamide treatment (**Fig.5F**).

## Regulation of single-cell expression heterogeneity

It has been proposed that expression heterogeneity of cell populations is regulated in the sense that the transcriptomes of individual cells within the population differ in a manner which optimizes an objective function, such as pluripotency or homeostasis [3]. To test whether signaling entropy can predict such regulated expression heterogeneity, we compared the distribution of single-cell entropies to the signaling entropy of the bulk population. Specifically, we devised a "measure of regulated heterogeneity" (MRH), which measures the likelihood that the signaling entropy of the cell population could have been observed from picking a single cell at random from that population (**Online Methods, Fig.6A**). We first estimated MRH for the data from Chu et al, for which matched bulk and scRNA-Seq data is available. We first note that although for bulk samples entropy differences between cell-types were smaller, that they were nevertheless consistent with the trends seen at the single-cell level (**Supplementary Fig.12 & Fig.2C**). The MRH for each of the six cell-types (hESCs, NPCs, DEPs, TBs, HFFs, ECs) in Chu et al, revealed evidence of regulated heterogeneity, with the entropy values of bulk samples being significantly higher than that of single-cells (**Fig.6B**). As a negative control, the signaling entropy of the average expression over bulk samples did not exhibit regulated heterogeneity (Normal deviation test P=0.30, **Fig.6B**), as required since bulk samples are not linked in space or time and represent non-interacting cell populations.

We note that for the previous analysis, matched bulk RNA-Seq data is not absolutely required since bulk samples can be approximated by averaging the expression profiles of individual cells in the population. We verified this, although, as expected, the entropy values for the true bulk samples were always marginally higher, in line with the fact that single cell assays only capture a subpopulation of the bulk sample (**Fig.6C**). We also verified that MRH results were not driven by the larger number of dropouts in scRNA-Seq data. Specifically, we simulated bulk samples by aggregating single cells representing the same cell-type and then resampling transcript counts matching to the average number of transcripts seen in single cells (**Online Methods**). We observed that signaling entropy of the simulated bulk did not alter appreciably upon downsampling and that results were unchanged **(Supplementary Fig.13).**

Next, we repeated the MRH analysis for T-cells and B-cells found in melanomas (Melanoma-set, **Supplementary Table 1**), for which sufficient numbers of single cells had been profiled. In all cases, signaling entropies of the bulk were much higher than expected based on the distribution of single-cell entropies (**Supplementary Fig.14**). Evidence for regulated expression heterogeneity was also seen among the melanoma cancer cells from each of 12 patients (Combined Fisher test P<1e-6, **Supplementary Fig.15**). We also analysed RNA-Seq data for 96 single cancer cells from a relapsed patient with acute myeloid leukemia (AML) (AML set [43], **Supplementary Table 1**). The signaling entropy for the AML cell population was 0.88, significantly larger than the maximal value over the 96 cells (SR=0.82, Normal deviation test P<0.001, **Fig.6D**). Again, as a negative control we analysed all 19 bulk AML samples at relapse and diagnosis, treating bulk samples from independent AML patients as if they were single cells from a common population. Estimating the signaling entropy of the average expression profile over all 19 bulk samples did not reveal a value significantly higher than that of the individual bulk samples (Normal deviation test P=0.32, **Fig.6D**).

## Discussion

Although Waddington proposed his famous epigenetic landscape of cellular differentiation many decades ago [10], it has proved challenging to construct a robust molecular correlate of a cell's elevation in this landscape. Here we have made significant progress, demonstrating that the differentiation potency and phenotypic plasticity of single cells, be they normal or malignant, can be estimated *in-silico* from their RNA-Seq profile using signaling entropy. As we have seen, signaling entropy can accurately discriminate pluripotent from multipotent and differentiated cells, without the need for feature selection or training, outperforming a pluripotency gene expression signature and providing a more general measure of differentiation potency.

Importantly, signaling entropy should not be confused with other transcriptional entropy measures, which are estimated over populations of single cells [45,46]. For instance, the "transcriptional entropy" of Richard et al [45] is estimated for single genes across single cells, and therefore reflects the amount of intercellular heterogeneity in the expression of a given gene. Our signaling entropy measure is estimated for a single-cell across genes in the context of a large gene network, which therefore incorporates systems-level information and is genome-wide (**Fig.1A-B**). While the signaling entropy of single-cells will influence the amount of transcriptional heterogeneity and entropy as defined by Richard et al, the precise relation between the two entropies is non-trivial. Indeed, we have here shown how we can assign single-cells into potency states, from which a Shannon Index (SI) over the whole cell population (i.e. using the distribution of potency states over single cells) can then be estimated (**Fig.1C**). This Shannon Index is more analogous to the transcriptional entropy of Richard et al. Indeed, we have shown how this Shannon Index is higher in a population of neural progenitor cells (NPCs) than in a population of hESCs (**Fig.3C**). Thus, the Shannon Index has nothing to do with potency as such, i.e. it does not measure the average differentiation potency of single cells in a cell population. In contrast, our signaling entropy does measure potency of single cells in a cell population. Thus, there is no requirement for our single-cell signaling entropy measure to exhibit a peak before a critical cell-fate transition occurs [45,46]. In contrast, the Shannon Index of a cell population derived from signaling entropy may exhibit the expected hallmarks of criticality. It will be interesting in future to test this with upcoming high resolution timecourse and genome-wide scRNA-Seq data.

The ability of signaling entropy to independently order single cells according to differentiation potency is a central component of the SCENT algorithm, which, as shown here, can help quantify and identify biologically relevant intercellular expression heterogeneity and cell subpopulations. Indeed, key findings which strongly support the validity of SCENT are the following: (i) using SCENT we were able to correctly predict that a hESC population contains a small fraction of cells of lower potency which are primed for differentiation, (ii) SCENT inferred that an assayed neural progenitor cell population was made up two distinct subsets, correctly predicting that only the lower potency subset represents bona-fide NPCs (as determined by expression of known neural stem cell markers), (iii) in a time course differentiation experiment of human myoblasts, SCENT correctly identified a contaminating interstitial mesenchymal cell population, whose potency did not change appreciably during the experiment. We note that this particular insight is not readily obtainable using other algorithms such as Monocle or MPath [14,15]. Thus, the ability of SCENT to assign single cells and cell subpopulations to specific potency states thus adds novel insight and functionality over what can be achieved with other existing algorithms. Alternatively, signaling entropy could be combined with existing algorithms like Monocle [14] or DPT [17,47] to empower their inference, since signaling entropy provides a more unbiased, independent, approach to ordering single cells in pseudo-time, i.e. it constitutes an approach which does not need prior knowledge such as the time point or markers of specific cell-types.

In a proof of principle analysis, we further demonstrated the ability of SCENT to identify putative drug resistant cancer stem cells, encompassing two different cancer-types (AML and prostate cancer), including CTCs. The ability to quantify stemness in cancer cell populations, either in tissue or in circulation, is a task of enormous importance. As shown here, as well as in our previous work on bulk cancer tissue [9,11,13], signaling entropy is, so far, the only single sample measure to have been conclusively demonstrated to robustly correlate with stemness in both normal and cancer cells. Indeed, a recent study by Gruen et al [18] explored a very different measure of transcriptome entropy, but which was not demonstrated to correlate well with differentiation potency or cancer. Likewise, signaling entropy is a more general measure of stemness/plasticity outperforming existing pluripotency expression signatures, as shown here and previously [11].

Importantly, signaling entropy also provides a computational framework in which to understand differentiation potency at the macroscopic (cell population) level from the corresponding potencies of single cells. As shown here, signaling entropy of cell populations, be they normal or malignant cells, exhibit synergy, with the entropy of the bulk being substantially higher than the entropy values of single cells. While no existing assay can measure all single cells in a population, we nevertheless demonstrated that our result is non-trivial, since mixing up bulk samples (to serve as a negative control) did not reveal such synergy. We also showed that these results were not confounded by the larger number of dropouts in scRNA-Seq data. Biologically, increased potency of a cell population as a result of synergistic cell-cell interactions, supports the view that features such as pluripotency are best understood at the cellular population level [3].

Finally, it is important to discuss the technical and biological properties of signaling entropy that underlie its robustness as a measure of differentiation potency. First of all, gene expression values enter the computation of signaling entropy only as gene ratios. Taking ratios of gene expression values and introducing a regularization term to offset dropouts, makes the resulting inference much less sensitive to the sequencing depth, absolute scale and normalization procedure of scRNA-Seq data. Second, signaling entropy is estimated over a fairly large number of genes (8000-10000), making it naturally robust to single gene dropouts. Third, its biological robustness stems in part from differentiation potency being encoded by a subtle positive correlation between the transcriptome and connectome, similar to our previous observations in the context of cancer [12]. Since there is no reason to expect that technical dropouts in scRNA-Seq should correlate with the connectivity of the corresponding protein in a PPI network, such technical effects are expected to average out. Finally, it is worth emphasizing in this context that Signaling Entropy provided a more accurate and robust measure of differentiation potency than other transcriptomic entropy-based measures (those used in StemID and SLICE) which do not use network information.

To conclude, signaling entropy and the SCENT algorithm provide a computational framework to advance our understanding of single-cell biology. We envisage that SCENT will be of great value for quantifying biologically relevant intercellular heterogeneity and for identifying putative normal and cancer stem-cells from scRNA-Seq data.

## Online Methods

### Single cell and bulk RNA-Seq data sets

The main datasets analysed here, the NGS platform used and their public accession numbers are listed in **Supplementary Table 1**. Below is a more detailed description of the samples in each data set:

*Chu et al Set:* This RNA-Seq dataset derives from Chu et al [28]. This set consisted of 4 experiments. Experiment-1 generated scRNA-Seq data for 1018 single cells, composed of 374 hESCs (212 single-cells from H1 and 162 from H9 cell line), 173 neural progenitor cells

(NPCs), 138 definite endoderm progenitors (DEPs), 105 mesoderm derived endothelial cells (ECs), 69 trophoblast cells (TBs), 159 human foreskin fibroblasts (HFFs). Experiment-2 is a time-course differentiation of single-cells, specifically of hESCs induced to differentiate into the definite endoderm, via a mesoendoderm intermediate. Timepoints assayed were before induction (t=0h, n=92), 12 hours after induction (12h, n=102), 24h (n=66), 36h (n=172), 72h (n=138) and 96h (n=188). Experiment-3 matches experiment-1 and consists of RNA-Seq data from 19 bulk samples: 7 representing hESCs, 2 representing NPCs, 2 TBs, 3 HFFs, 3 ECs and 2 DEPs. Experiment-4 consists of 15 RNA-Seq profiles from bulk samples, profiled as part of the time-course differentiation experiment (Experiment-2), with 3 samples per time-point (12h, 24h, 36h, 72h, 96h).

*Melanoma Set:* This scRNA-Seq dataset derives from Tirosh et al [23], and consists of 4645 single-cells derived from the tumor microenvironment of 19 melanoma patients. Of these, 3256 are non-malignant cells, encompassing T-cells (n=2068), B-cells (n=515), Natural Killer cells (n=52), Macrophages (n=126), Endothelial Cells (EndC, n=65) and cancer-associated fibroblasts (CAFs, n=61). The rest of single cells profiled were malignant melanoma cells (n=1257).

*AML Set:* This set derives from Li et al [43]. A total of 96 single cells from a relapsed acute myeloid leukemia (AML) patient (patient ID=130) were profiled. In addition, 38 paired bulk AML samples were profiled from 19 patients (all experiencing relapse), with 19 samples obtained at diagnosis and with the other matched 19 samples obtained at relapse.

*hESC Set:* This set derives from Yan et al [30]. It consists of 124 single cell profiles, of which 90 are from different stages of embryonic development, with 34 cells representing hESCs. These 34 hESCs were derived from the inner cell mass, with 8 cells profiled at primary outgrowth and 26 profiled at passage-10. The 90 single cells from the pre-implantation embryo were distributed as follows: Oocyte (n=3), Zygote (n=3), 2-cell embryo (n=6), 4-cell embryo (n=12), 8-cell embryo (n=20), morulae (n=16), late blastocyst (n=30).

*Trapnell et al set:* This scRNA-Seq set derives from Trapnell et al [14]. It consists of a timecourse differentiation experiment of human myoblasts, which profiled a total of 372 single cells: 96 cells at t=0 (time at which differentiation was induced), 96 at t=24h after induction, another 96 at t=48h after induction, and 84 cells at 72h post-induction.

*CTC-PrCa set:* This scRNA-Seq dataset derives from Miyamoto et al [44].We focused on a subset of 73 single-cells from castration resistant prostate cancers, of which 36 derived from patients who developed resistance to enzulatamide treatment, with the remaining 37 derived from treatment-naïve patients.

*Treutlein set:* This scRNA-Seq dataset derives from Treutlein et al [29]. There are a total of 201 single cells assayed at 4 different stages in the developing mouse epithelium, including embryonic day 14, 16, 18 and adulthood. At E18, a subset of single cells were characterized into alveolar type-1 and type-2 cells (AT1 & AT2), as well as a putative bipotent (BP) subgroup.

## The Single-Cell Entropy (SCENT) algorithm

There are five steps to the SCENT algorithm: (1) Estimation of the differentiation potency of single cells via computation of signaling entropy, (2) Inference of the potency state distribution across the single cell population, (3) Quantification of the intercellular heterogeneity of potency states, (4) Inference of single cell landmarks, representing the major potency-coexpression clusters of single cells, (5) Lineage trajectory (or dependency network) reconstruction between landmarks. We now describe each of these steps:

Computation of signaling entropy: The computation of signaling entropy for a given sample proceeds using the same prescription as used in our previous publications [9,11]. Briefly, the normalized genome-wide gene expression profile of a sample (this can be a single cell or a bulk sample) is used to assign weights to the edges of a highly curated protein-protein interaction (PPI) network. The construction of the PPI network itself is described in detail elsewhere [11], and is obtained by integrating various interaction databases which form part of Pathway Commons (www.pathwaycommons.org) [48]. The weighting of the network via the transcriptomic profile of the sample provides the biological context. The weight of an edge between protein $i$ and protein $j$, denoted by $w_{ij}$ , is assumed to be proportional to the normalized expression levels of the coding genes in the sample, i.e. we assume that $w_{ij} \sim x_i x_j$ . We interpret these weights (if normalized) as interaction probabilities. The above construction of the weights is based on the assumption that in a sample with high expression of $i$ and $j$, that the two proteins are more likely to interact than in a sample with low expression of $i$ and/or $j$. Viewing the edges generally as signaling interactions, we can thus define a random walk on the network, assuming we normalize the weights so that the sum of outgoing weights of a given node $i$ is 1. This results in a stochastic matrix, $P$, over the network, with entries

$$p_{ij} = \frac{x_j}{\sum_{k\in N(i)} x_k} = \frac{x_j}{(Ax)_i}$$

where $N(i)$ denotes the neighbors of protein $i$, and where $A$ is the adjacency matrix of the PPI network ($A_{ij}=1$ if $i$ and $j$ are connected, 0 otherwise, and with $A_{ii}$=0). The signaling entropy is then defined as the entropy rate (denoted $Sr$) over the weighted network, i.e.

$$Sr(\vec{x}) = -\sum_{i=1}^{n} \pi_i \sum_{j \in N(i)} p_{ij} \log p_{ij}$$

where $\pi$ is the invariant measure, satisfying $\pi P=\pi$ and the normalization constraint $\pi^T\mathbf{1}=1$. The invariant measure, also known as steady-state probability, represents the relative probability of finding the random walker at a given node in the network (under steady state conditions i.e. long after the walk is initiated). Nodes with high values thus represent nodes that are particularly influential in distributing signaling flux in the network. In the steady-state we can assume detailed balance (conservation of signaling flux, i.e. $\pi_i p_{ij} = \pi_j p_{ji}$ ), and it can be shown [9] that $\pi_i = x_i(Ax)_i/(x^TAx)$. Given a fixed adjacency matrix $A$ (i.e. fixing the topology), it can also be shown [9] that the maximum possible $Sr$ among all compatible stochastic matrices $P$, is the one with $P = \frac{1}{\gamma} v^{-1} \otimes A \otimes v$ where $\otimes$ denotes product of matrix entries and where $v$ is the dominant eigenvector of $A$, i.e. $Av=\lambda v$ with $\lambda$ the largest eigenvalue of $A$. We denote this maximum entropy rate by *maxSr,* and define the normalized entropy rate (with range of values between 0 and 1) as

$$SR(\vec{x}) = \frac{Sr(\vec{x})}{maxSr}$$

Throughout this work, we always display this normalized entropy rate.

Inference of potency states: In this work, we show that signaling entropy (i.e. the entropy rate *SR)* provides a proxy to the differentiation potential of single cells. We can model a cell population as a statistical mechanical model, in which each single cell has access to a number of different potency states. For a large collection of single cells we can estimate their signaling entropies, and infer from this distribution of signaling entropies the number of underlying potency states using a mixture modeling framework. Since *SR* is bounded between 0 and 1, we first conveniently transform the *SR* value of each single cell into their logit-scale, i.e. $y(SR)=log_2(SR/(1-SR))$. Subsequently, we fit a mixture of Gaussians to the $y(SR)$ values of the whole cell population, and use the Bayesian Information Criterion (BIC) (as implemented in the *mclust* R-package) [49] to estimate the optimal number $K$ of potency states, as well as the state-membership probabilities of each individual cell. Thus, for each single cell, this results in its assignment to a specific potency state.

Quantifying intercellular heterogeneity of potency states: For a population of $N$ cells, we can then define a probability distribution $p_k$ over the inferred potency states. For $K$ inferred potency states, one can then define a normalized Shannon Index ($SI$):

$$SI = -\frac{1}{\log K} \sum_{k=1}^{K} p_k \log p_k$$

which measures the amount of heterogeneity in potency within the single-cell population (1=high heterogeneity in potency, 0=no heterogeneity in potency).

Inference of co-expression clusters and landmarks***:*** With each cell assigned to a potency state, we next perform clustering (using the scRNA-seq profiles) of the single cells. We use the Partitioning-Around-Medoids (PAM) algorithm with the average silhouette width to estimate the optimal number of clusters, a combination which was found to be among the most optimal clustering algorithms in applications to omic data [50]. Clustering of the cells is performed over a filtered set of genes that are identified as those driving most variation in the complete dataset, as assessed using SVD. In detail, we perform a SVD on the full z-scored normalized RNA-seq profiles of the cells, selecting the significant components using RMT [51] and picking the top 5% genes with largest absolute weights in each significant component. The final set of genes is obtained by the union of those identified from each significant componente. PAM-clustering (with a Pearson distance correlation metric) of all cells results in the assignment of each cell into a co-expression cluster, with a total number of $n_p$ cell-clusters. Thus, each cell is assigned to a unique potency state and co-expression cluster. Finally, landmarks are identified by selecting potency-state cluster combinations containing at least 1 to 5% of all single cells. Importantly, each of these landmarks has a specific potency state and mean signaling entropy value, allowing ordering of these landmarks according to potency.

Inference of lineage trajectories: For each landmark in step-4, we compute centroids of gene expression using only cells that are contained within that landmark and defined only over the genes used in the PAM-clustering. Partial correlations [52,53] between the centroid landmarks are then estimated to infer trajectories/dependencies between landmarks. Significant positive partial correlations may indicate transitions between landmarks. Since each landmark has a signaling entropy value associated with it, directionality is inferred by comparing their respective potency states.

## A fast Pearson correlation approximation

Under certain assumptions (to be discussed below), there is a useful approximation to signaling entropy, which also provides important biological insight. It entails first using an approximation for the steady-state probability (invariant measure) $\pi$. As before, in the steady-state, we can assume the detailed balance condition (conservation of signaling flux: i.e. $\pi_i p_{ij} = \pi_j p_{ji}$ ), so that the invariant measure satisfies $\pi_i \sim x_i (Ax)_i$ [9]. If we now take a global mean field approximation, that is, if we replace the expression values of the neighbors of gene *i,* with the mean expression value over all genes in the network, it then follows that $\pi_i \sim x_i\ k_i$ , where $k_i$ is the connectivity of gene/protein $i$ in the network. Hence, $SR = \sum_i \pi_i S_i \sim \sum_i x_i k_i S_i$ , which is effectively the 3-way correlation between the transcriptome, connectome and local signaling entropies. If we assume further that the dynamic range of local signaling entropies $S_i = -\sum_{j\in N(i)} p_{ij} \log p_{ij}$ is small (which for realistic PPI networks is often the case [12]), and also assuming that the local entropies correlate positively with node-degree, we obtain that $SR \sim x_i\, k_i$ , i.e the signaling entropy is approximately the Pearson correlation of the cell´s transcriptome and the connectome from the PPI network.

Importantly, we stress that (i) this approximation is an empirical one which works reasonably well for the realistic PPI networks considered here, and (ii) that the signaling entropy and its Pearson correlation approximation are not equivalent, since there exist networks where the two measures give widely different answers. In particular, if a network has scale-free topology, but with the hubs not connected to each other, then a positive correlation between expression and connectivity may not lead to a higher signaling entropy. For instance, if the low-degree nodes ("bottlenecks") linking the hubs have very low expression then signaling flux can't be distributed over the network, leading to a lower entropy rate compared to an expression configuration where all genes have similar expression values (see Supplementary Fig.7). For realistic PPI networks, hubs are generally connected to each other and for these type of networks, the Pearson approximation works well. We note that for a 8,393 node network with 300,916 edges, the computation of *SR* for 100 samples takes approximately 370 seconds on an Intel Xeon CPU E3-1575M 3.00GHz, whereas that of its Pearson correlation approximation only takes 1/10 seconds, thus although the approximation is computationally much faster, the computation of *SR* for 1 sample only takes about 4 seconds.

## Ranking genes according to differential local entropy

Since signaling entropy is obtained as a weighted average over local signaling entropies (i.e. $SR = \sum_i \pi_i S_i$) with the local entropies defined by $S_i = -\sum_{j\in N(i)} p_{ij} \log p_{ij}$ , the latter can be used to identify genes in the network where the signaling flux distribution differs between two phenotypes. Specifically, we use the normalized version of the local signaling entropy, defined by $NS_i = -\frac{1}{log k_i} \sum_{j\in N(i)} p_{ij} \log p_{ij}$ , which is bounded between 0 and 1, thus allowing genes of different connectivity to be compared. Thus, for each gene and each sample, we can compute a local entropy and genes can then be ranked according to the difference in local entropy using an empirical Bayes framework [11,54] to derive moderated t-statistics which reflect the significance in differential local entropy. Adjustment for multiple-testing was performed using the Benjamini-Hochberg procedure.

## Gene Set Enrichment Analysis (GSEA)

We performed GSEA on the top-ranked genes, ranked according to differential local entropy between pluripotent and non-pluripotent cells. Specifically, we focused on the genes exhibiting increased local signaling entropy in pluripotent cells, and focused on a range of thresholds (top 500, 600, 700, 800, 900, 1000) to assess robustness. Enrichment was performed using a one-tailed Fisher's exact test, as implemented by us previously [55]. Enrichment was assessed against the Molecular Signatures Database (http://software.broadinstitute.org/gsea/msigdb) [34].

## Application to mouse scRNA-Seq data

In our application to mouse scRNA-Seq data, we first converted mouse gene Ensembl IDs into their human homologs using the AnnotationTools Bioconductor package [56]. Only those mapping to a unique human homolog were considered. The resulting set of genes were then integrated with our human PPI network.

## Estimation of cell-cycle and TPSC pluripotency scores

To identify single cells in either the G1-S or G2-M phases of the cell-cycle we followed the procedure described in [23]. Briefly, genes whose expression is reflective of G1-S or G2-M phase were obtained from [32,33]. A given normalized scRNA-Seq data matrix is then z-score normalized for all genes present in these signatures. Finally, a cycling score for each phase and each cell is obtained as the average z-scores over all genes present in each signature.

To obtain an independent estimate of pluripotency we used the pluripotency gene expression signature of Palmer et al [22], which we have used extensively before [11]. This signature consists of 118 genes that are overexpressed and 39 genes that are underexpressed in pluripotent cells. The TPSC score for each cell with scRNA-Seq data is obtained as the t-statistic of the gene expression levels between the overexpressed and underexpressed gene categories. Optionally, the scRNA-Seq is z-score normalized beforehand and the t-statistic is obtained by comparing expression z-scores. However, we note that the z-score procedure uses information from all single cells, so the fairest comparison to signaling entropy means we ought to compare expression levels. We note that the TPSC scores obtained from z-scores or expression levels were highly correlated and did not affect any of the conclusions in this manuscript.

## Comparison analysis of bulk and single-cell RNA-Seq data

Since signaling entropy (SR) can be computed for each single-cell, one can compare the predicted entropies of bulk samples (cell population) to those of the single cells making up that population. To test whether the entropy of the bulk deviates markedly from that of single cells, we computed a z-score, by comparing the entropy of the bulk to that of the single cells where the latter distribution is modeled as a Gaussian. This z-score is called the measure of regulated heterogeneity (MRH), since it assesses whether the transcriptomes of single cells differ in a regulated synergistic manner, increasing entropy (potency) well above that of single cells. In the case where matched bulk samples were not available, we simulated bulk samples in two distinct ways. In one approach, we simply averaged the single cell transcriptomes before computing SR. In a second approach, which corrects for the large number of dropouts present in scRNA-Seq data, by first aggregate the transcript counts of all single cells, and then downsample counts so as to match to the average number of transcripts per single-cell. Robustness to the specific downsampling draw was tested by performing 100 Monte-Carlo samplings.

### Other entropy measure proxies for differentiation potency

Briefly, we describe two other entropy-based measures for approximating differentiation potency in a single-cell context, but which do not make use of a PPI network. One measure is part of the StemID algorithm [18]. However, the original StemID algorithm does not estimate differentiation potency of single cells. Instead it provides estimates for single cell clusters, which are inferred by clustering the expression profiles of single cells. Thus, for a given cluster
*k*, StemID computes a potency which is proportional to $\delta E_k$, where

$$\delta E_k \equiv median_{c\in k}(E_c) - min_l(median_{c\in l}(E_c))$$

where $E_c$ is the information entropy of cell *c,* defined by $E_c = -\sum_{g=1}^{N} q_{gc} \log q_{gc}$ (where *N* is the number of genes and where $q_{gc}$ is the normalized number of reads mapping to gene *g* in cell *c*). Thus, in order to objectively compare to our signaling entropy measure, which does not use information of other cells when estimating potency of a given cell*,* we here use $E_c$ as the potency estimate from StemID. Another information entropy based measure is part of the SLICE algorithm, proposed by Guo et al [31]. Briefly, in this approach, genes are first clustered into related GO-terms to define *m* functional gene clusters. For a given cell *c*, relative activity of each functional cluster *k* is estimated from the average expression of genes mapping to that cluster. These activity scores are then normalized so that they can be interpreted as probabilities $q_{kc}$, and subsequently the potency of cell *c* is estimated as the information entropy $H_c = E_B[-\sum_{k=1}^{m} q_{kc} \log q_{kc}$ where the expectation is taken over a number of bootstraps over genes. We compute this information entropy using the R-script provided in Guo et al [31].

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

**Data Availability:** All data analyzed in this manuscript is already publicly available from the following GEO (www.ncbi.nlm.nih.gov/geo/) accession numbers: GSE72056, GSE83533, GSE75748, GSE36552, GSE52529, GSE67980, GSE52583. All data is also available on request from the authors.

**Supplementary Material** All Supplementary Tables and Figures can be found in the Supplementary Information document.

**Competing Interests** The authors declare that they have no competing interests.

**Author Contributions** Manuscript was conceived and written by AET. Statistical analyses were performed by AET. TE contributed useful feedback.

## Acknowledgements

This work was supported by NSFC (National Science Foundation of China) grants, grant numbers 31571359 and 31401120 and by a Royal Society Newton Advanced Fellowship (NAF project number: 522438, NAF award number: 164914). The author also wishes to thank Guo-Cheng Yuan for stimulating discussions.

## References

1. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* **33**, 155-60 (2015).
2. Levsky, J.M., Shenoy, S.M., Pezo, R.C. & Singer, R.H. Single-cell gene expression profiling. *Science* **297**, 836-40 (2002).
3. MacArthur, B.D. & Lemischka, I.R. Statistical mechanics of pluripotency. *Cell* **154**, 484-9 (2013).
4. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133-45 (2015).
5. Pina, C. *et al.* Single-Cell Network Analysis Identifies DDIT3 as a Nodal Lineage Regulator in Hematopoiesis. *Cell Rep* **11**, 1503-10 (2015).
6. Pina, C. *et al.* Inferring rules of lineage commitment in haematopoiesis. *Nat Cell Biol* **14**, 287-94 (2012).
7. Kalmar, T. *et al.* Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol* **7**, e1000149 (2009).
8. Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* **450**, 1230-4 (2007).
9. Teschendorff, A.E., Sollich, P. & Kuehn, R. Signalling entropy: A novel network-theoretical framework for systems analysis and interpretation of functional omic data. *Methods* **67**, 282-93 (2014).
10. Waddington, C.R. *Principles of Development and Differentiation*, (Macmillan Company, New York,

1966).

11. Banerji, C.R. *et al.* Cellular network entropy as the energy potential in Waddington's differentiation landscape. *Sci Rep* **3**, 3039 (2013).

12. Teschendorff, A.E., Banerji, C.R., Severini, S., Kuehn, R. & Sollich, P. Increased signaling entropy in cancer requires the scale-free property of protein interaction networks. *Sci Rep* **5**, 9646 (2015).

13. Banerji, C.R., Severini, S., Caldas, C. & Teschendorff, A.E. Intra-tumour signalling entropy determines clinical outcome in breast and lung cancer. *PLoS Comput Biol* **11**, e1004115 (2015).

14. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-6 (2014).

15. Chen, J., Schlitzer, A., Chakarov, S., Ginhoux, F. & Poidinger, M. Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nat Commun* **7**, 11988 (2016).

16. Marco, E. *et al.* Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc Natl Acad Sci U S A* **111**, E5643-50 (2014).

17. Haghverdi, L., Buttner, M., Wolf, F.A., Buettner, F. & Theis, F.J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**, 845-8 (2016).

18. Grun, D. *et al.* De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell* **19**, 266-77 (2016).

19. Lee, T.I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-13 (2006).

20. Gomez-Gardenes, J. & Latora, V. Entropy rate of diffusion processes on complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **78**, 065102 (2008).

21. Chu, L.F. *et al.* Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol* **17**, 173 (2016).

22. Palmer, N.P., Schmid, P.R., Berger, B. & Kohane, I.S. A gene expression profile of stem cell pluripotentiality and differentiation is conserved across diverse solid and hematopoietic cancers. *Genome Biol* **13**, R71 (2012).

23. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-96 (2016).

24. Lacorre, D.A. *et al.* Plasticity of endothelial cells: rapid dedifferentiation of freshly isolated high endothelial venule endothelial cells outside the lymphoid tissue microenvironment. *Blood* **103**, 4164-72 (2004).

25. Oliver, G. & Srinivasan, R.S. Endothelial cell plasticity: how to become and remain a lymphatic endothelial cell. *Development* **137**, 363-72 (2010).

26. Kalluri, R. The biology and function of fibroblasts in cancer. *Nat Rev Cancer* **16**, 582-98 (2016).

27. Chen, W.J. *et al.* Cancer-associated fibroblasts regulate the plasticity of lung cancer stemness via paracrine signalling. *Nat Commun* **5**, 3472 (2014).

28. Chu, L.-F. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definite endoderm. *Genome Biol* **17**(2016).

29. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371-5 (2014).

30. Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* **20**, 1131-9 (2013).

31. Guo, M., Bao, E.L., Wagner, M., Whitsett, J.A. & Xu, Y. SLICE: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res* (2016).

32. Macosko, E.Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-14 (2015).

33. Whitfield, M.L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* **13**, 1977-2000 (2002).

34. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).

35. das Neves, R.P. *et al.* Connecting variability in global transcription rate to mitochondrial variability. *PLoS Biol* **8**, e1000560 (2010).

36. Johnston, I.G. *et al.* Mitochondrial variability as a source of extrinsic cellular noise. *PLoS Comput Biol* **8**, e1002416 (2012).

37. Guantes, R. *et al.* Global variability in gene expression and alternative splicing is modulated by mitochondrial content. *Genome Res* **25**, 633-44 (2015).

38. Schieke, S.M. *et al.* Mitochondrial metabolism modulates differentiation and teratoma formation capacity in mouse embryonic stem cells. *J Biol Chem* **283**, 28506-12 (2008).

39. Wanet, A., Arnould, T., Najimi, M. & Renard, P. Connecting Mitochondria, Metabolism, and Stem Cell Fate. *Stem Cells Dev* **24**, 1957-71 (2015).

40. Sukumar, M. *et al.* Mitochondrial Membrane Potential Identifies Cells with Enhanced Stemness for Cellular Therapy. *Cell Metab* **23**, 63-76 (2016).

41. Hu, C. *et al.* Energy Metabolism Plays a Critical Role in Stem Cell Maintenance and Differentiation. *Int J Mol Sci* **17**, 253 (2016).

42. Folmes, C.D. & Terzic, A. Energy metabolism in the acquisition and maintenance of stemness. *Semin Cell Dev Biol* **52**, 68-75 (2016).

43. Li, S. *et al.* Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat Med* **22**, 792-9 (2016).

44. Miyamoto, D.T. *et al.* RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science* **349**, 1351-6 (2015).

45. Richard, A. *et al.* Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. *PLoS Biol* **14**, e1002585 (2016).

46. Mojtahedi, M. *et al.* Cell Fate Decision as High-Dimensional Critical State Transition. *PLoS Biol* **14**, e2000640 (2016).

47. Angerer, P. *et al.* destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241-3 (2016).

48. Cerami, E.G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* **39**, D685-90 (2011).

49. Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. & Ruzzo, W.L. Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977-87 (2001).

50. Wiwie, C., Baumbach, J. & Rottger, R. Comparing the performance of biomedical clustering methods. *Nat Methods* **12**, 1033-8 (2015).

51. Teschendorff, A.E., Zhuang, J. & Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **27**, 1496-505 (2011).

52. Schafer, J. & Strimmer, K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21**, 754-64 (2005).

53. Barzel, B. & Barabasi, A.L. Network link prediction by global silencing of indirect correlations. *Nat*

*Biotechnol* **31**, 720-5 (2013).

54. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).

55. Teschendorff, A.E. *et al.* An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One* **4**, e8274 (2009).

56. Kuhn, A., Luthi-Carter, R. & Delorenzi, M. Cross-species and cross-platform gene expression studies with the Bioconductor-compliant R package 'annotationTools'. *BMC Bioinformatics* **9**, 26 (2008).

# Tables

| **Dataset** | | **Signaling Entropy** | **SLICE** | **StemID** |
|---|---|---|---|---|
| *scRNA-Seq* | | | | |
| Chu1 (Pl > NonPl) | **P** | 3e-132 | ~1 | 3e-58 |
| | **AUC** | **0.96** | <0.5 | 0.79 |
| Chu2 (0h > 96h) | **P** | 2e-38 | 0.94 | 1e-22 |
| | **AUC** | **0.97** | <0.5 | 0.86 |
| Trapnell (0h>72h) | **P** | 6e-9 | 0.0003 | 2e-10 |
| | **AUC** | **0.74** | 0.65 | **0.75** |
| Treutlein (E14>Adult) | **P** | 5e-27 | 6e-26 | 5e-27 |
| | **AUC** | **1** | **0.998** | **1** |
| *Bulk RNA-Seq* | | | | |
| Chu3 (Pl > NonPl) | **P** | 4e-5 | 0.001 | 0.76 |
| | **AUC** | **0.99** | 0.90 | <0.5 |

***Table-1: Comparison of Signaling Entropy to SLICE and StemID as measures of differentiation potency in scRNA-Seq and bulk RNA-Seq datasets.*** *Table lists one-tailed Wilcoxon rank sum test P-values and associated (one-tailed) AUCs, testing whether entropy is higher in the pluripotent or multipotent cells compared to the less potent cells in various scRNA-Seq and bulk RNA-Seq datasets. In Chu1, the comparison is between pluripotent (hESCs, n=374, Pl) and non-pluripotent (n=644, NonPl) single cells. In Chu2, the comparison is between hESCs (0h, n=92) and definite endoderm progenitors sampled 96h later (n=188). In Trapnell, the comparison is between human myoblasts (0h, n=96) and differentiated skeletal muscle cells (72h, n=84). In Treutlein, the comparison is between early lung progenitors (E14, n=45) and mature alveolar cells (n=46). In Chu3, the comparison is between bulk hESCs (n=7) and non-pluripotent samples (n=12).*

# Figure Legends

**Figure-1: The Single-Cell Entropy (SCENT) algorithm. A) Signaling entropy of single cells as a proxy to their differentiation potential in Waddington's landscape.** Depicted on the left is a population of cells with cells occupying either a pluripotent (magenta), a progenitor (cyan) or a differentiated state (green). The potency state of each cell is determined by a complex function of the transcriptomic profile $\vec{x}$ of the cell. For a given interaction between proteins *i* and *k* in the network, signaling in a given cell occurs with a probability $p_{ik} \sim x_i x_k$, defining a stochastic matrix $P=(p_{ik})$. In a pluripotent state, there is high demand for phenotypic plasticity, and so promiscuous signaling proteins (i.e those of high connectivity) are highly expressed (red colored node) with all major differentiation pathways kept at a similar basal activity level (grey edges). The probability of signaling between protein *i* and *k*, $p_{ik}$, is therefore $1/k_i$ where $k_i$ is the connectivity of protein *i* in the network. Thus the local signaling entropy around node *i* is maximal. In a differentiated state, commitment to a specific lineage (activation of a specific signaling pathway shown by red colored node) means that most $p_{ij}$~0 , except when $j=k$, so that $p_{ik}$~1. Thus, local signaling entropy around node *i* is close to zero. **B) Estimation of signaling entropy.** An overall measure of signaling promiscuity of the cell is given mathematically by the signaling entropy rate (SR), which is a weighted average of local signaling entropies $S_i$ over all the genes/proteins in the network, with weights specified by $\pi$ (the steady-state probability satisfying $\pi P=\pi$). It is proposed that SR provides a proxy to the elevation in Waddington's landscape, quantifying differentiation potential of cells (i.e the number of accessible cell-fates within a given lineage). **C) Quantification of intercellular heterogeneity and reconstruction of lineage trajectories.** Estimation of signaling entropy at the single-cell level across a population of cells, allows the distribution of potency states in the population to be determined through Bayes mixture modelling which infers the optimal number of potency states. From this, the heterogeneity of potency states in a cell population is computed using Shannon's Index. To infer lineage trajectories, SCENT uses a clustering algorithm over dimensionally reduced scRNA-Seq profiles to infer co-expression clusters of cells. Dual assignment of cells to a potency state and co-expression cluster allows the identification of landmarks as bi-clusters in potency-coexpression space. Finally, partial correlations between the expression profiles of the landmarks are used to infer a lineage trajectory network diagram linking cell clusters according to expression similarity, with their height or elevation determined by their potency (signaling entropy).

**Figure-2: Signaling entropy correlates with differentiation potency of single cells. A)** Violin plots of the signaling entropy (SR) against cell-type (hESC=human embryonic stem cells, NPC=neural progenitor cells, DEP=definite endoderm progenitors, TB=trophoblast cells, HFF=human foreskin fibroblasts, EC=endothelial cells (mesoderm progenitor derivatives)). Number of single cells in each class is indicated. Total number is 1018. Wilcoxon rank sum test P-values between each cell-type (ranked in decreasing order of SR) are given. Diamond shaped data points correspond to the matched bulk samples. **B)** Scatterplot of the signaling entropy (SR, y-axis) against an independent mRNA expression based pluripotency score (TPSC, x-axis) for all 1018 single cells. Cell-type is indicated by color. Spearman Correlation Coefficient (SCC) and associated P-value are given. **C)** Violin plot comparing the signaling entropy (SR) between the hESCs and all other (non-pluripotent) cells. P-value is from a Wilcoxon rank sum test. Inlet figure is the associated ROC curve, which includes the AUC value. **D)** Violin plot of signaling entropy (SR) values for non-malignant single cells found in the microenvironment of melanomas. Number of single cells of each cell-type are given (CAF=cancer associated fibroblasts, EndC=endothelial cells, MacPH=macrophages, T=T-cells, B=B-cells, NK=natural killer cells). Wilcoxon rank sum test P-values between EndC and MacPH, and between MacPH and all lymphocytes are given. **E)** Signaling entropy (SR) as a function of differentiation stage within the mesoderm lineage. Differentiation stages include hESCs (pluripotent), mesoderm progenitors of endothelial cells (multipotent) and differentiated endothelial and white blood cells. Wilcoxon rank sum test P-values between successive stages are given. **F)** ROC curves and AUC values for discriminating the progenitor and differentiated cells within the mesoderm lineage for signaling entropy (SR) and the t-test pluripotency score (TPSC). **G)** Signaling entropy (SR, y-axis) as a function of time in a single-cell time course differentiation experiment, starting from hESCs at time=0h (time of differentiation induction) into definite endoderm (which occurs from 72h onwards). Number of single cells measured at each time point is given. Wilcoxon rank sum test P-values between the first 4 time points and 72h, and between 72h and 98h are given. **H)** Signaling entropy (SR, y-axis) as a function of developmental stage in the differentiation of the distal mouse lung epithelium. Number of single cells measured at each stage is given. Wilcoxon rank sum test P-values between embryonic day 14 (E14) and all other stages are given. **I)** Comparison of the SRs in C) (left panel) to the case where expression values are randomly reshuffled before computation of SR (middle panel). Right panels compare the corresponding ROC curves and AUC values. **J)** As C), but now splitting the hESCs into cells from H1 and H9 lines, and including an additional independent set of 90 single hESCs profiled with a different NGS platform.

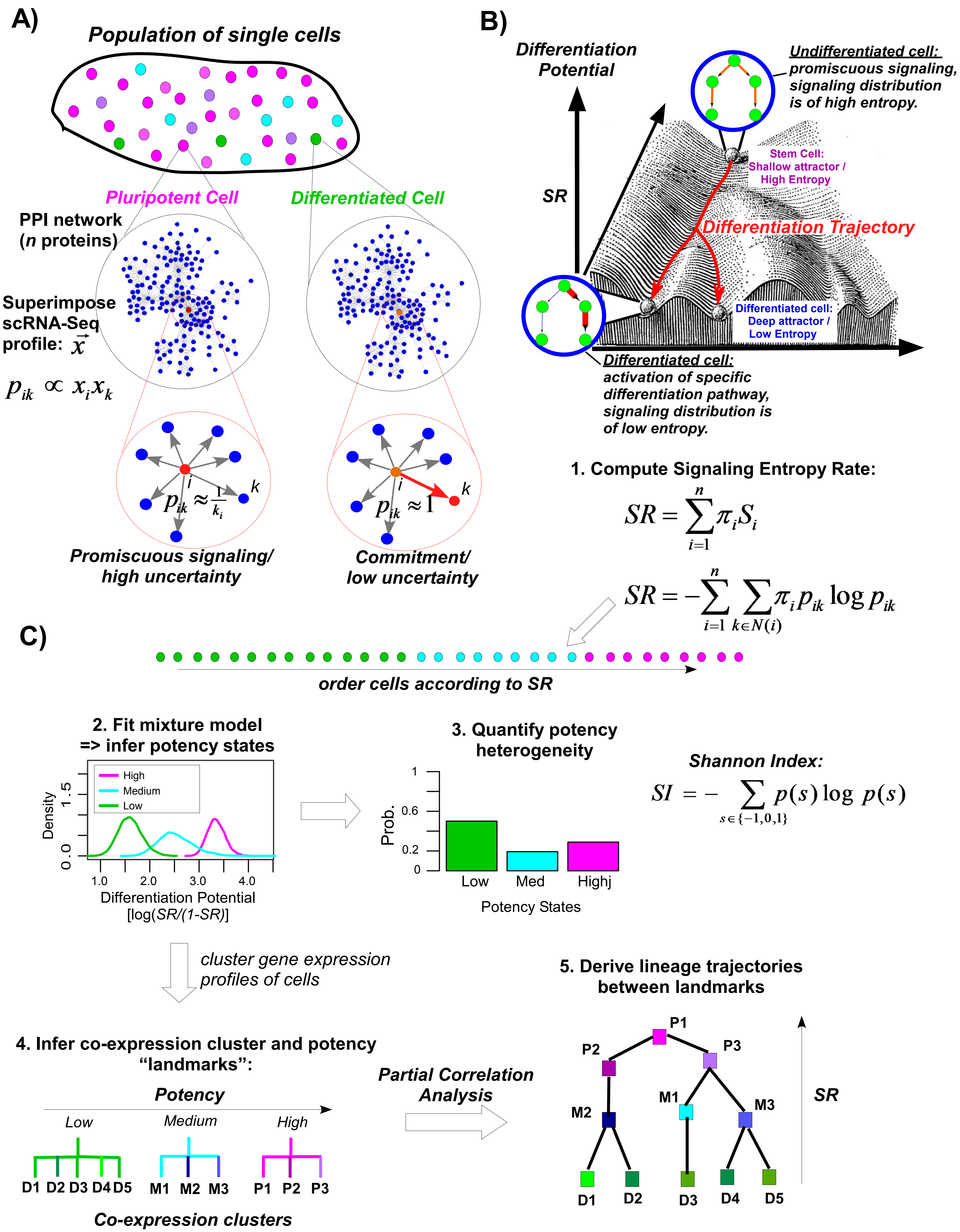
**Figure-3: SCENT identifies single cell subpopulations of biological significance. A)** Fitted Gaussian mixture model to the signaling entropies of 1018 single cells (scRNA-Seq data from Chu et al) using a logit scale for the signaling entropies (x-axis, $\log_2$[SR/(1-SR)]). BIC predicted only 2-states: a high energy/entropy pluripotent state (magenta-PS1) and a lower-energy non-pluripotent state (cyan-PS2). Number of cells categorized into each state is indicated in plot. **B)** Barplot comparing, for each cell-type, the probability that a cell from this cell population is in the pluripotent (prob(Pl)) or non-pluripotent state (probe(NonPl). Cell-types include human embryonic stem cells (hESCs), neural progenitor cells (NPCs), definite endoderm progenitors (DEPs), trophoblast cells (TBs), human foreskin fibroblasts (HFFs) and endothelial cells (ECs). **C)** Barplot of the corresponding Shannon Index for each cell-population type. **D)** Distribution of single cell numbers between inferred potency states and co-expression clusters, as predicted by SCENT. In brown, we indicate "landmark clusters" which contain at least 5% of the total number of single cells. **E)** Distribution of single cell-types among the 7 landmark clusters. **F)** Inferred lineage trajectories between the 7 landmarks which map to cell-types. Border color indicates potency state: magenta=PS1, cyan=PS2. **G) Left panel:** Scatterplot of signaling entropy (SR) vs mRNA expression level of a neural stem/progenitor cell marker, HES1, for all NPCs. NPCs categorized as pluripotent are shown in magenta, NPCs categorized into a non-pluripotent state are shown in cyan. NPCs of high and low HES1 expression (as inferred using a partition-around-medoids algorithm with k=2) are indicated with triangles and squares, respectively. **Right panel:** Corresponding boxplot comparing the differentiation potency (SR) of NPCs with low vs. high HES1 expression. P-value is from a one-tailed Wilcoxon rank sum test.

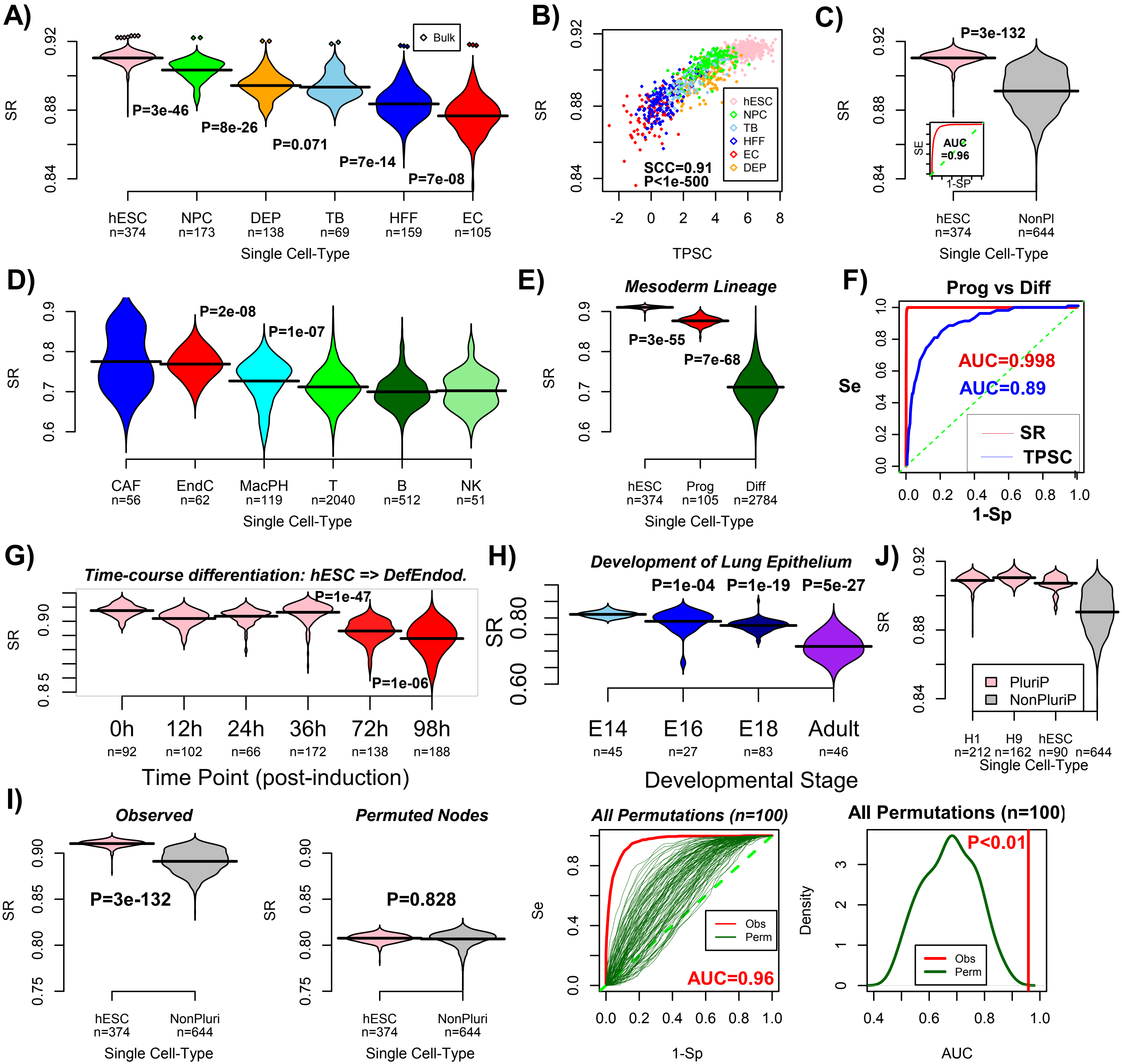
**Figure-4: SCENT dissects distinct lineage trajectories in human myoblast differentiation. A)** Signaling entropy (SR) vs. time point (0h, 24h, 48h, 72h) for a total of 372 single cells, collected during a time course differentiation experiment of human myoblasts (scRNA-Seq from Trapnell et al). Violin plots show the density distribution of SR values at each time point. P-value is from a one-tailed Wilcox rank sum test comparing timepoint 0h to 24h. **B)** SCENT Gaussian Model fit to SR values predicts 3 potency states (PS1, PS2, PS3). **C)** Probability distribution of potency states at each timepoint. **D)** Co-expression heatmap of highly variable genes obtained by SCENT predicting 3 main clusters. Single cells have been ordered, first by cluster, then by potency state and finally by their time of sampling, as indicated. Landmarks are indicated by rectangular boxes, and distribution of single cells across landmarks and timepoints is provided in table. Genes have been clustered using hierarchical clustering. Genes that are markers of the different landmarks have been highlighted. **E)** Inferred lineage trajectories between landmarks. Diagram illustrates an inferred two-phase trajectory, with one trajectory describing myoblasts of high potency (t=0, cyan circle) differentiating into skeletal muscle cells of intermediate potency (t=24 and 48) (blue circles) and a mixture of terminally differentiated and intermediate potency skeletal muscle cells (t=72) (grey and blue circle, respectively). A second trajectory/landmark describes a different cell-type (interstitial mesenchymal cells) whose intermediate potency state does not change during the time-course (blue stars).

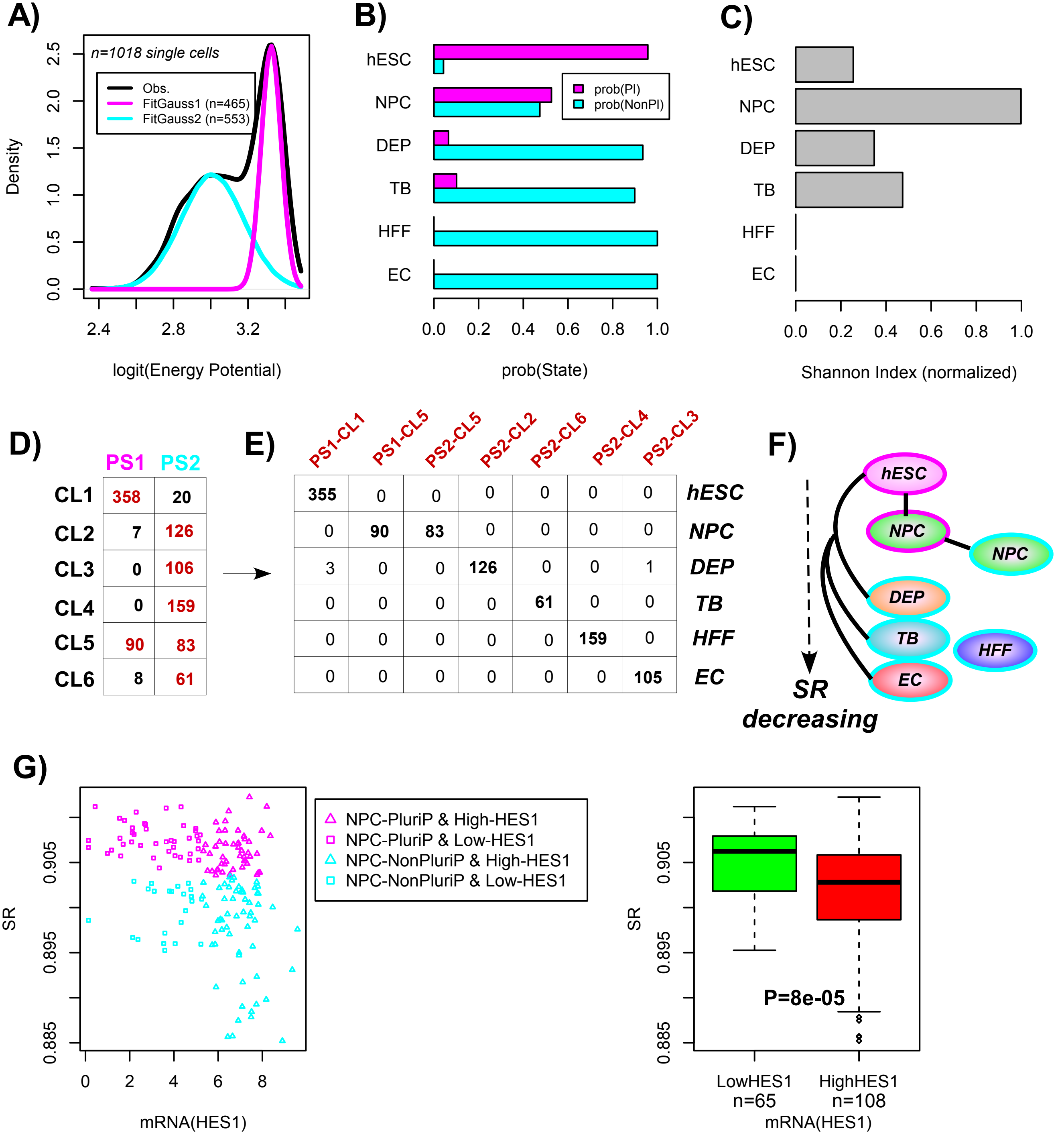
**Figure-5: Increased signaling entropy in cancer cells and identification of drug resistant cancer stem cells. A)** Boxplots of the signaling entropy (SR) for single melanoma cancer cells (C ) compared to non-malignant (NotC) cells for 3 different melanoma patients (patient IDs given above each plot). Numbers of single cells are given below each boxplot. P-value is from a Wilcoxon rank sum test. **B)** As A), but now pooled across all 12 patients. **C)** Comparison of signaling entropy (SR) of 19 diagnostic acute myeloid leukemia bulk samples to relapsed samples from the same patients. Wilcox rank sum test P-value (one-tailed paired) is given. **D)** Sorting of 96 single AML cells from one patient according to signaling entropy and comparison of mRNA expression of AML CSC markers between low and high SR groups. P-values from a one-tailed Wilcox test. **E)** Comparison of signaling entropy (SR) of circulating tumor cells from metastatic prostate cancer patients who did not receive AR inhibitor treatment (UNTR) to those which developed resistance (RESIST). P-value from a one-tailed Wilcox test. **F)** Sorting of 73 single CTCs according to SCENT (signaling entropy, SR) into low and high SR groups. Correlation of gene expression of one putative CSC marker (ALDH7A1) with SR.

**Figure-6: Signaling entropy predicts regulated expression heterogeneity of single-cell populations. A)** Definition of the measure of regulated expression heterogeneity (MRH). The MRH is a z-statistic, obtained by measuring the deviation of the signaling entropy (SR) of the bulk expression profile from the mean of single-cell entropies, taking into account the variability of single-cell entropies in the population. **B)** Barplots of MRH for each cell-type population from Chu et al, representing the degree to which the signaling entropy of the cell population is higher than that of single-cells. P-values are from a one-tailed normal-deviation test. Dashed line indicates the line P=0.05. AvgBulkS compares the signaling entropy of the average expression over all bulk samples to that of the individual bulk samples, indicating that although the RHM is positive (signaling entropy increases), that it is not significantly higher than that of the individual bulk samples. **C)** Scatterplot of the signaling entropy of bulk samples (y-axis), representing 6 cell-types (hESCs, NPCs, DEPs, TBs, HFFs, ECs) against the corresponding signaling entropies of these cell populations obtained by first averaging the expression profiles of single-cells ("Simulated Bulk", x-axis). $R^2$ value and P-value are given with green dashed line representing the fitted regression. Observe how the signaling entropy of bulk samples is always higher than that obtained from first averaging expression of single cells, in line with the fact that the assayed single cells are a subpopulation of the full bulk sample. **D) Left panel:** Comparison of the signaling entropy of an acute myeloid leukemia (AML) bulk sample (red line and point) to the signaling entropies of 96 single AML cells (blue) from that bulk sample. P-value is from a one-tailed normal deviation test. **Right panel:** Comparison of the MRH value for the matched 96 single cells and bulk AML sample (SCs) to the MRH values obtained by comparing the signaling entropy of the average expression over 19 AML bulk samples to the signaling entropies of each individual AML bulk sample. The 19 AML bulk samples come in pairs, obtained at diagnosis (dgn) and relapse (rel), which are shown separately. P-values are from a one-tailed normality deviation test.

## The Single-Cell Entropy (SCENT) algorithm

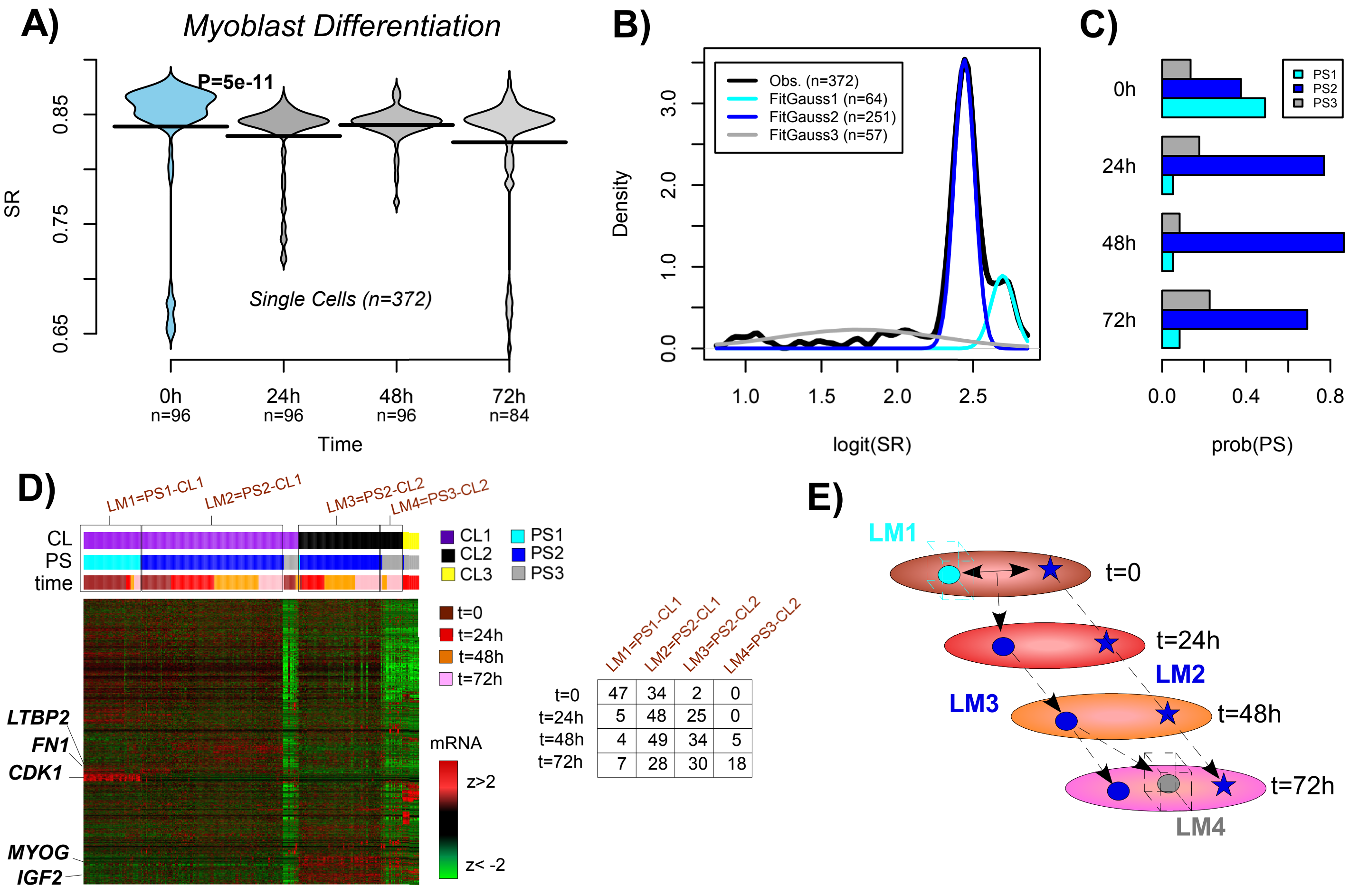



A)
SR
0.92
0.88
0.84
Bulk
P=3e-46
P=8e-26
P=0.071
P=7e-14
P=7e-08
hESC n=374
NPC n=173
DEP n=138
TB n=69
HFF n=159
EC n=105
Single Cell-Type
B)
SR
0.92
0.88
0.84
hESC
NPC
TB
HFF
EC
DEP
SCC=0.91
P<1e-500
-2 0 2 4 6 8
TPSC
C)
SR
0.92
0.88
0.84
P=3e-132
SE
AUC =0.96
1-SP
hESC n=374
NonPl n=644
Single Cell-Type
D)
SR
0.9
0.8
0.7
0.6
P=2e-08
P=1e-07
CAF n=56
EndC n=62
MacPH n=119
T n=2040
B n=512
NK n=51
Single Cell-Type
E)
Mesoderm Lineage
SR
0.9
0.8
0.7
0.6
P=3e-55
P=7e-68
hESC n=374
Prog n=105
Diff n=2784
Single Cell-Type
F)
Prog vs Diff
Se
0.0 0.2 0.4 0.6 0.8 1.0
AUC=0.998
AUC=0.89
SR
TPSC
0.0 0.2 0.4 0.6 0.8 1.0
1-Sp
G)
Time-course differentiation: hESC => DefEndod.
SR
0.90
0.85
P=1e-47
P=1e-06
0h n=92
12h n=102
24h n=66
36h n=172
72h n=138
98h n=188
Time Point (post-induction)
H)
Development of Lung Epithelium
SR
0.80
0.60
P=1e-04
P=1e-19
P=5e-27
E14 n=45
E16 n=27
E18 n=83
Adult n=46
Developmental Stage
J)
SR
0.92
0.88
0.84
PluriP
NonPluriP
H1 n=212
H9 n=162
hESC n=90
n=644
Single Cell-Type
I)
Observed
SR
0.90
0.85
0.80
0.75
P=3e-132
hESC n=374
NonPluri n=644
Single Cell-Type
Permuted Nodes
SR
0.90
0.85
0.80
0.75
P=0.828
hESC n=374
NonPluri n=644
Single Cell-Type
All Permutations (n=100)
Se
0.8
0.4
0.0
Obs
Perm
AUC=0.96
0.0 0.2 0.4 0.6 0.8 1.0
1-Sp
All Permutations (n=100)
Density
3
2
1
0
P<0.01
Obs
Perm
0.4 0.6 0.8 1.0
AUC

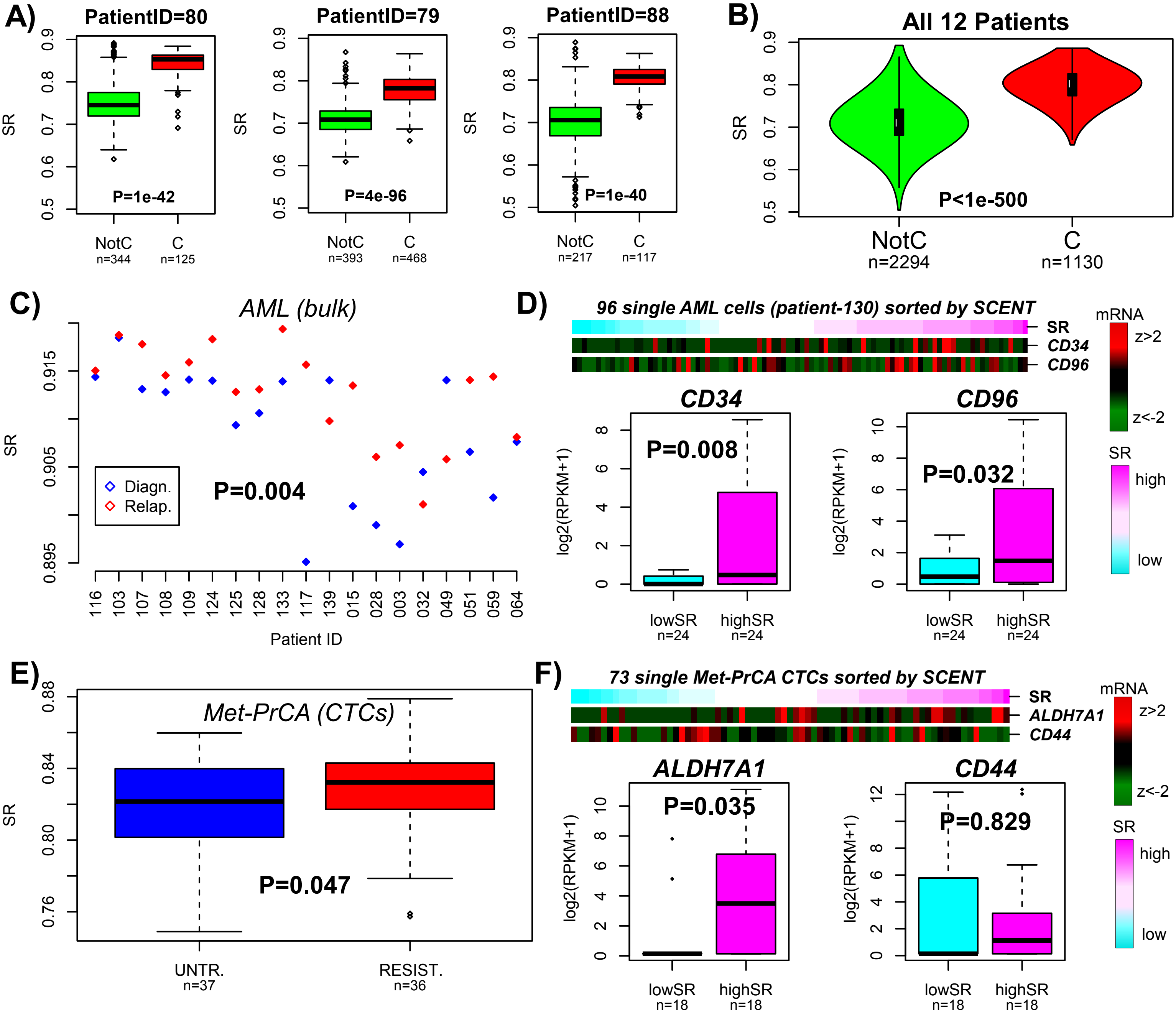


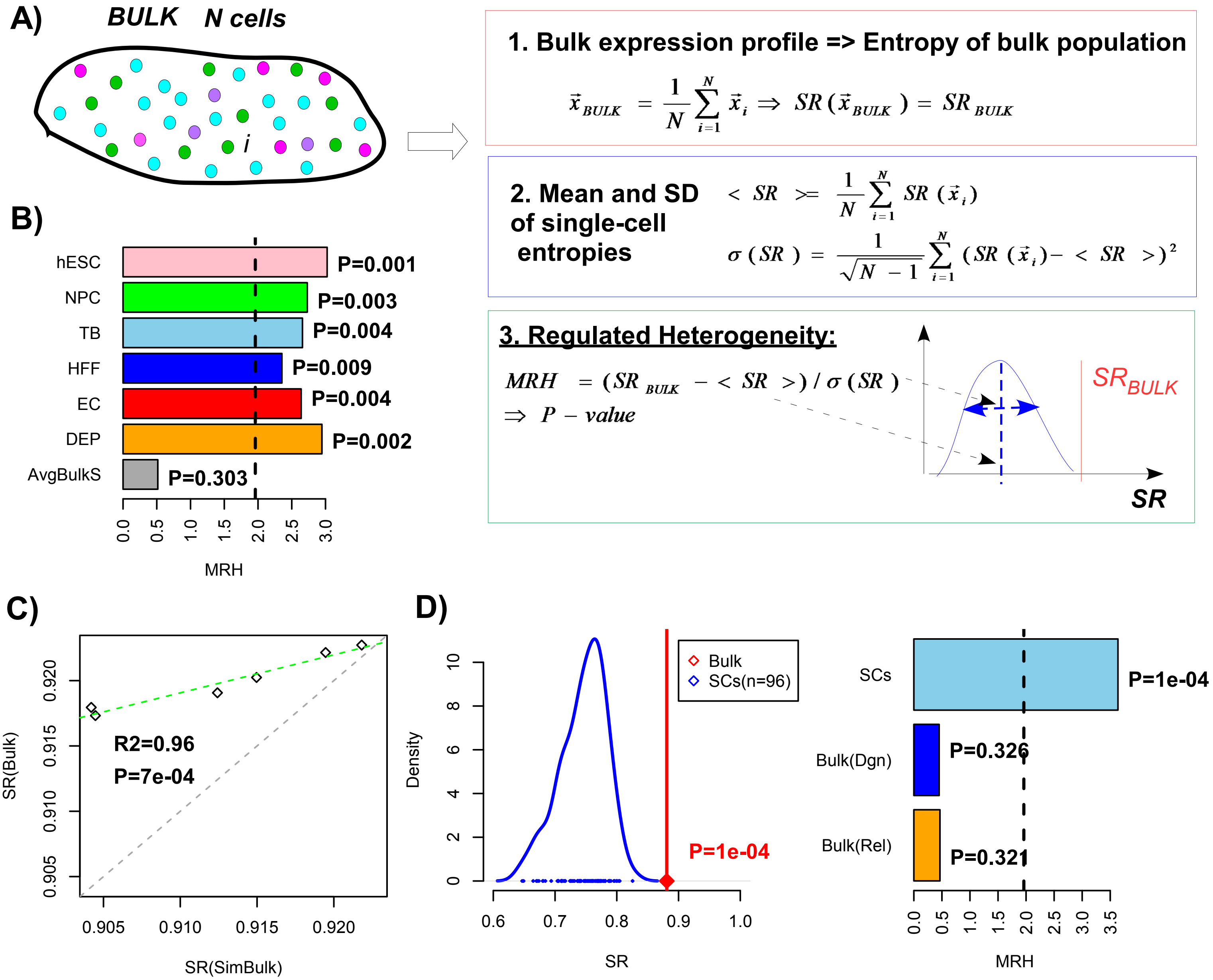
|  | PS1 | PS2 |
|---|---|---|
| CL1 | 358 | 20 |
| CL2 | 7 | 126 |
| CL3 | 0 | 106 |
| CL4 | 0 | 159 |
| CL5 | 90 | 83 |
| CL6 | 8 | 61 |

| PS1-CL1 | PS1-CL5 | PS2-CL5 | PS2-CL2 | PS2-CL6 | PS2-CL4 | PS2-CL3 |  |
|---|---|---|---|---|---|---|---|
| 355 | 0 | 0 | 0 | 0 | 0 | 0 | hESC |
| 0 | 90 | 83 | 0 | 0 | 0 | 0 | NPC |
| 3 | 0 | 0 | 126 | 0 | 0 | 1 | DEP |
| 0 | 0 | 0 | 0 | 61 | 0 | 0 | TB |
| 0 | 0 | 0 | 0 | 0 | 159 | 0 | HFF |
| 0 | 0 | 0 | 0 | 0 | 0 | 105 | EC |



|  | LM1=PS1-CL1 | LM2=PS2-CL1 | LM3=PS2-CL2 | LM4=PS3-CL2 |
|---|---|---|---|---|
| t=0 | 47 | 34 | 2 | 0 |
| t=24h | 5 | 48 | 25 | 0 |
| t=48h | 4 | 49 | 34 | 5 |
| t=72h | 7 | 28 | 30 | 18 |


A)
PatientID=80
SR
0.5
0.6
0.7
0.8
0.9
P=1e-42
NotC
n=344
C
n=125
PatientID=79
SR
0.5
0.6
0.7
0.8
0.9
P=4e-96
NotC
n=393
C
n=468
PatientID=88
SR
0.5
0.6
0.7
0.8
0.9
P=1e-40
NotC
n=217
C
n=117
B)
All 12 Patients
SR
0.5
0.6
0.7
0.8
0.9
P<1e-500
NotC
n=2294
C
n=1130
C)
AML (bulk)
SR
0.895
0.905
0.915
Diagn.
Relap.
P=0.004
116
103
107
108
109
124
125
128
133
117
139
015
028
003
032
049
051
059
064
Patient ID
D)
96 single AML cells (patient-130) sorted by SCENT
SR
CD34
CD96
mRNA
z>2
z<-2
SR
high
low
CD34
log2(RPKM+1)
0
2
4
6
8
P=0.008
lowSR
n=24
highSR
n=24
CD96
log2(RPKM+1)
0
2
4
6
8
10
P=0.032
lowSR
n=24
highSR
n=24
E)
Met-PrCA (CTCs)
SR
0.76
0.80
0.84
0.88
P=0.047
UNTR.
n=37
RESIST.
n=36
F)
73 single Met-PrCA CTCs sorted by SCENT
SR
ALDH7A1
CD44
mRNA
z>2
z<-2
SR
high
low
ALDH7A1
log2(RPKM+1)
0
2
4
6
8
10
P=0.035
lowSR
n=18
highSR
n=18
CD44
log2(RPKM+1)
0
2
4
6
8
12
P=0.829
lowSR
n=18
highSR
n=18


A)
BULK
N cells
i
1. Bulk expression profile => Entropy of bulk population
$\vec{x}_{BULK} = \frac{1}{N}\sum_{i=1}^{N}\vec{x}_i \Rightarrow SR(\vec{x}_{BULK}) = SR_{BULK}$
2. Mean and SD of single-cell entropies
$<SR> = \frac{1}{N}\sum_{i=1}^{N} SR(\vec{x}_i)$
$\sigma(SR) = \frac{1}{\sqrt{N-1}}\sum_{i=1}^{N}(SR(\vec{x}_i) - <SR>)^2$
3. Regulated Heterogeneity:
$MRH = (SR_{BULK} - <SR>)/\sigma(SR)$
$\Rightarrow P - value$
$SR_{BULK}$
SR
B)
hESC
NPC
TB
HFF
EC
DEP
AvgBulkS
P=0.001
P=0.003
P=0.004
P=0.009
P=0.004
P=0.002
P=0.303
0.0
0.5
1.0
1.5
2.0
2.5
3.0
MRH
C)
SR(Bulk)
0.905
0.910
0.915
0.920
R2=0.96
P=7e-04
0.905
0.910
0.915
0.920
SR(SimBulk)
D)
Density
0
2
4
6
8
10
◇ Bulk
◇ SCs(n=96)
P=1e-04
0.6
0.7
0.8
0.9
1.0
SR
SCs
Bulk(Dgn)
Bulk(Rel)
P=1e-04
P=0.326
P=0.321
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
MRH