Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

StemSC: A Cross-dataset Human Stemness Index for Single-cell Samples

Hailong Zheng
Harbin Medical University
Jiajing Xie
Fujian Medical University
Kai Song
Harbin Medical University
Jing Yang
Harbin Medical University
Huiting Xiao
Harbin Medical University
Jiashuai Zhang
Harbin Medical University
Keru Li
Harbin Medical University
Rongqiang Yuan
Harbin Medical University
Yuting Zhao
Harbin Medical University
Yunyan Gu
Harbin Medical University
Wenyuan Zhao (zhaowenyuan@ems.hrbmu.edu.cn)
Harbin Medical University https://orcid.org/0000-0002-6477-9434

Research Article

Keywords: stemness, single-cell analysis, cross-dataset, cell Dedifferentiation, tumor microenvironment

Posted Date: June 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-564395/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Stem Cell Research & Therapy on March 21st, 2022. See the published version at https://doi.org/10.1186/s13287-022-02803-5.

Abstract

Background: Stemness is defined as the potential of cells for self-renewal and differentiation. Many transcriptome-based methods for stemness evaluation have been proposed. However, all these stemness indexes showed low correlations with differentiation time and the limitation to identify the high-stemness cells across datasets.

Methods: Here, we constructed a stemness index for single-cell samples (StemSC) based on relative expression orderings (REO) of gene pairs. Firstly, we identified the stemness-related genes by selecting the genes significantly related to differentiation time in all five datasets. Then, we used 13 RNA-seq datasets from both the bulk and single-cell ESC samples to construct the reference REOs. Finally, the StemSC value of a given sample was calculated as the percentage of gene pairs with the same REOs as the ESC samples.

Results: We validated the StemSC by its higher correlations with differentiation time in eight normal datasets and its higher correlations with tumor dedifferentiation in three colorectal cancer datasets and four glioma datasets. By using the StemSC, we can recognize the tissue-specific stem genes and automatically construct the cell differentiation trajectories. StemSC also could provide the same threshold to identify high-stemness cells across datasets. Results showed that the tumor cells with high-stemness had fewer interactions with anti-tumor immune cells. Besides, the immunotherapy-treated patients with high-stemness had worse survival than those with low-stemness.

Conclusions: All above results showed StemSC is a better stemness index to calculate the stemness across datasets, which can help researchers explore the effect of stemness on other biological processes.

Background

Stemness is defined as the self-renewal and differentiation potential of cells [1]. Cancer progression is accompanied by the acquisition of this feature [2]. Besides, quantification of stemness is very helpful to reconstruct cellular differentiation trajectories and explore the role of stemness in tumor tissues [3]. As reported, by using the stemness index, pervasive negative associations were found between cancer stemness and anticancer immunity [4].

With the increasing number of RNA-seq data, many transcriptome-based methods for stemness evaluation have been proposed, such as mRNAsi [5]. However, this method was only trained from bulk data, which limited its performance in single-cell data [3]. Gunsagar S. Gulati has proposed a more suitable method for single cells, named CytoTRACE [3], which showed better performance than all the currently known methods. However, the average correlation between this stemness index and differentiation time was only about 0.6. What's more, all these transcriptome-based methods were vulnerable to batch effect and can't provide individual scores [6], which made these methods unable to identify high-stemness cells across datasets. Therefore, it is necessary to construct a stemness index that

is highly correlated with differentiation time and could be evaluated for single-cell samples across datasets.

In our previous studies, we have identified many signatures based on relative expression orderings (REOs) of gene pairs [7–9], which were not sensitive to batch effects and can be robustly applied to independent validation sets. Based on these unique advantages, we constructed an REO-based stemness index based on bulk samples [6], which showed a high correlation with differentiation time. However, the lack of single-cell samples in its training sets limited its performance in the collected single-cell samples of this study.

In this study, we constructed an REO-based stemness index for single-cell samples, which we called StemSC. Then, we validated StemSC by its higher correlations with differentiation time and tumor dedifferentiation than CytoTRACE in eight independent datasets and the merged datasets. Finally, by using StemSC, we constructed cell differentiation trajectories automatically and explored the influence of the tumor stemness on the tumor microenvironment.

Methods

Data and preprocessing

In this study, we downloaded the gene expression data of 11 human embryonic stem cell (ESC) datasets to reveal the high stability of REOs (Table S1). We also downloaded six datasets with differentiation time (Table S2) for identifying stemness-related genes and 13 datasets (Table S3) for the development of StemSC. Five independent datasets with differentiation time were downloaded to validate the performance of StemSC (Table S4). Three colorectal cancer datasets and four glioma datasets were also downloaded to validate the performance of StemSC in tumor cells (Table S4). We excluded the samples of distant metastatic tumor to focus on the corresponding cancer type. Especially for the glioma dataset GSE117891, we excluded the cells from the normal tissues because these cells were limited to few patients.

For the RNA-seq expression data of both the bulk and single-cell samples, we directly downloaded the processed RPKM, TPM, or count data from the Gene Expression Omnibus [10] (GEO, https://www.ncbi.nlm.nih.gov/geo/), Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra/), and Progenitor Cell Biology Consortium [11] (PCBC, https://www.synapse.org/#!Synapse:syn1773109/wiki/54962). Count data was turned to the RPKM data with the corresponding reference genomes of the datasets. Due to the filtered genes in GSE57872, we downloaded raw single-cell RNA sequencing data from SRA accession SRP042161 [12]. To retrieve the transcriptomic profiles of GSE57872, we built a reference transcriptome based on the GENCODE v19 annotation [13] and mapped the paired-end 25bp reads to the reference transcriptome by using HISAT2 [14] (version 2.1.0, with parameters -q -p 1–5 0–3 0 -k 5 –min-intronlen 20 –max-intronlen 500000 –phred33). The RPKM data of GSE57872 was calculated by using featureCount [15] (with parameters -t exon -g gene_id –primary). Then, for each RNA-seq expression dataset, we mapped the Ensembl gene ID

or gene symbol to the Entrez gene ID by using the reference downloaded from HUGO Gene Nomenclature Committee (HGNC, https://www.genenames.org/download/custom/). Dataset PRJNA482620 was preprocessed in the same way. Especially, we removed the low-quality cells with less than 2000 detected genes for single-cell data.

Consistency evaluation of REOs between datasets

In this study, we calculated the REOs by using the overlapping genes among the datasets with more than three samples. Pairwise comparisons were performed for the expression level of the above genes for each sample. For each gene pair (G_i, G_j), we retained the gene pair with certain REO ($G_i > G_j$ or $G_i < G_j$) in all samples of the dataset, which we called stable REO. The consistency of stable REOs between two datasets was calculated as s/n, where n was the number of shared gene pairs between the stable REOs of two datasets and s was the number of shared gene pairs with the same REOs. The significance of consistency was determined by the cumulative binomial distribution model as follows:

$$P = 1 - \sum_{i=0}^{s-1} \binom{n}{i} (P_0)^i (1 - P_0)^{n-i}$$

Where P_0 is the probability that gene pairs showed same REO pattern ($G_i > G_j$ or $G_i < G_j$) in two datasets by chance ($P_0 = 0.5$).

Development of the StemSC

The stemness index for single-cell data, StemSC, was constructed by calculating the similarity of REOs between target cells and ESC samples. Firstly, we identified the stemness-related genes by selecting the genes significantly related to differentiation time in all five datasets (Table S2). Then, due to the lack of single-cell data, we used 13 RNA-seq datasets from both the bulk and single-cell ESC samples to construct the reference REOs (Table S3). We further identified the stable REOs in all above ESC samples as reference REO. Finally, the StemSC value of a given sample was calculated as k/n, where n was the number of the referenced REOs contained in this sample and k was the number of the gene pairs with same REOs as the reference REOs. Additionally, scripts and codes for the StemSC are available for download (https://github.com/Zhao-Wenyuan/StemSC).

Construction of the cellular differentiation trajectories

Here, we provided a method to construct the cellular differentiation trajectories by combining StemSC and Monocle 2 [16]. Firstly, to select the highly variable genes for trajectory inference, we removed the genes detected in less than ten cells and selected the top 5000 genes with the largest product of the coefficient of variation square and mean values. Then, the states and branches were detected using the Monocle v2.16.0 R package. Finally, the root of the differentiation process was automatically identified by choosing the state with the highest mean StemSC values.

Identification of normal cells and the further cell types

Firstly, we downloaded the corresponding normal samples from the GTEx [17] as the reference, unless the original dataset contains peritumoral cells. Next, we used the inferCNV v1.7.1 R package (https://github.com/broadinstitute/inferCNV) to infer the copy number variations (CNV) of all tumor tissue cells by taking these samples as the control. Then, the hierarchical cluster was used to divide these cells into tumor cells with CNV and the normal cells without obvious CNV. We further used SingleR [18] to identify immune cell types for the identified normal cells. In this study, the cell types with less than 50 cells were removed to avoid huge difference in cell numbers.

Cell-cell communication analysis

The intercellular communication analysis was performed by using CellPhoneDB, a python-based tool [19]. We used CellPhoneDB v2.1.1 python package to analyze the potential interaction networks of the low-, high-stemness tumor cells and the major types of immune cells.

Enrichment analysis

The pathway enrichment analyses based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) were conducted and visualized using clusterProfiler v3.16.1 R package [20]. We also used this package to calculate the normalized enrichment score and *p* value for the enrichment of each gene set.

Statistical analysis

In this study, we used Spearman correlation analysis to evaluate the relationship between StemSC and differentiation time. We also used hypergeometric test to assess the significance of enrichment of the differentially expressed genes in the stem signatures and student T-test to assess the differences of the StemSC values between two types of samples. The Benjamini–Hochberg method was utilized to control the false discovery rate (FDR) in the multiple tests. All statistical analyses were carried out with the R 4.0.2 software package (http://www.r-project.org/)

Results

High stability of REOs in single-cell samples

REOs are the features that describe the relative expression orders of gene pairs within a sample. One attribute of their features is that the REOs of RPKM won't change after within-sample normalization, such as TPM or log transformation. Indeed, the REOs of RPKM showed 100% overlap with those after TPM or log transformation in 11 public human ESC datasets (Fig. 1A, Table S1). Another attribute is that the REOs stably showing same pattern in the same type of samples, which we called stable REOs (see Methods), could be retained in other independent bulk datasets, even in paraffin-embedded bulk samples with relative low gene expressions [21]. Thus, we inferred that the stable REOs could also be stably retained in the single-cell samples with low gene expressions. Indeed, we found that the stable REOs

recognized in each of the 11 single-cell datasets exhibited high consistency among datasets, which was similar for bulk datasets (consistency, 0.99 for single-cell datasets and 0.92–0.96 for bulk datasets Fig. 1B). Besides, there was also a high consistency of stable REOs between single-cell datasets and bulk datasets (consistency, 0.97–0.99, Fig. 1B). In addition, we found that increasing the number of datasets can improve the stability of REOs in independent datasets (Fig. 1C). The larger the number of the merged datasets, the more stable the stable REOs are.

Considering that the previous REO-based stemness index did not approach the set value 1 in single-cell ESC samples (Figure S1), it was necessary to add single-cell data to train a signature more suitable for single-cell data. All above results inspired us to add bulk samples to the single-cell samples to build a stemness index that is more suitable for single-cell samples and more robust to batch effects.

Development and validation of StemSC

The development procedure of StemSC is shown in Fig. 2. We collected both bulk and single-cell datasets as training sets because of the shortage of single-cell data and the high consistency of REOs between bulk and single-cell datasets (Fig. 1B). Firstly, to reduce redundant REOs, we identified 437 stemness-related genes by choosing the genes which were significantly associated with differentiation time in all five datasets with different differentiation directions (Spearman, FDR < 0.05, Table S2). Naturally, these genes showed significant enrichment in the pathways related to cell renewal and differentiation, such as cell cycle, Ribosome biogenesis in eukaryotes (Figure S2). Then, we respectively identified 19,937 and 50,827 gene pairs with stable REOs in 92 single-cell and 47 bulk ESC samples from 13 datasets (Table S3). For the 16,848 shared gene pairs of the above two gene pair lists, 99.9% (16,839/16,848) of them had the same REOs, which showed the high consistency of REOs between single-cell and bulk samples again. Finally, the stemness index values of the given human single cells, StemSC, were calculated by the percentage of gene pairs with the same REOs as ESC samples in 16,839 gene pairs (see Methods).

In all five independent validation datasets (Fig. 3A, Table S4), there were strong negative correlations between StemSC values and differentiation time (Spearman correlation; $|r|$, 0.43–0.85, Fig. 3B, D-H), which was greatly higher than that between CytoTRACE and differentiation time (Spearman correlation; $|r|$, 0.14–0.84, Fig. 3B). Further, the robustness of StemSC to batch effect was showed in three following aspects. Firstly, the median StemSC values of ESCs were all centered around the 1 in two independent validation datasets (median StemSC, 0.990 for GSE85066 and 0.984 for GSE109979). Secondly, after combining the two batches of dataset GSE102066, the correlation between StemSC and differentiation time was higher in the merged data than in one of the single batch data (Fig. 3C, I), but not for CytoTRACE. Thirdly, StemSC showed negative correlations with the differentiation time in all validation sets, but CytoTRACE showed a positive correlation in the dataset GSE109979 (Fig. 3B). In addition, to demonstrate that our method can be applied to the cells with different origins, we deleted the ESC samples from the above two datasets with ESCs (GSE85066 and GSE109979). Results showed that there was still a higher negative correlation between the StemSC values and differentiation time than

CytoTRACE (*r*, -0.798 and − 0.583 for StemSC; -0.761 and 0.864 for CytoTRACE). This feature could allow the StemSC to be used in the cancer cells with multiple origins.

The ability of the StemSC to identify stemness- associated genes and construct cellular differentiation trajectories

Given the high correlation between StemSC values and differentiation time, we next explored the ability of the StemSC to identify the stem makers or differentiation factors. Firstly, by ranking all genes according to their correlations with StemSC, we found the enrichment of the stemness-associated genes (the top 100 genes positively correlated with differentiation time) in the positive region and differentiation-associated genes (the top 100 genes negatively correlated with differentiation time) in the negative region in all the validation sets (Fig. 4A). Besides, the majority of the most positive or negative genes showed their role in stemness or differentiation (Fig. 4B, Table S6). For example, *L1TD1*, the most positively correlated gene with stemness, was reported to be a marker for undifferentiated human ESCs [22]. For another example, *CDH2*, the gene trigging the endodermal germ-layer formation [23], showed the most negative correlation with stemness in the endodermal differentiation samples. All above results showed the potential of StemSC to recognize the tissue-specific and stemness-associated genes.

Cell lineage trajectory can be determined by using transcriptome-based branch detection tools, such as Monocle 2 [16]. However, users need to enter the starting point of the biological processes. For example, when applied to the dataset GSE109979, Monocle 2 constructed seven possible cell trajectories with different roots (Fig. 4C). On the contrary, our method could identify the correct root by choosing the state with the largest average of StemSC values (Fig. 4C), which was similar to time-based lineage trajectory (Fig. 4D). However, CytoTRACE chose the wrong root in this dataset. Similarly, StemSC recognized the correct roots in the four remaining validation sets (Figure S3). The above results showed a better method to automatically construct cellular differentiation trajectories by combining StemSC and the branch detection tools Monocle 2.

Applicability of StemSC on tumor tissue cells

To validate the applicability of StemSC in tumor cells, we took the single-cell data of colorectal cancer and glioma as examples (Table S4). For three single-cell datasets of colorectal cancer (Table S4), we found that the 30 intestinal stem markers [24] were significantly enriched in the gene list ranked by their correlations with the StemSC values (Fig. 5A-C, Table S7) and the StemSC values were significantly correlated with the sum of expression values of these 30 markers (spearman, $p < 0.05$, Fig. 5D-F). Besides, many of the stem markers showed significantly positive correlations with the StemSC values (Fig. 5G-I). Especially, histological grade reflects the dedifferentiation of cancer tissue. For the dataset GSE81861 with grade information, significantly higher StemSC values were found not only in the 290 tumor tissue cells than the 170 normal tissue cells but also in the 230 low differentiation cells (grade 2) than the 60 high differentiation cells (grade 1) (student T-test, $p = 6E\text{-}9$, Fig. 5J, K), which was more significant than the values calculated by CytoTRACE (student T-test, $p = 0.007$, Fig. 5L).

Similarly, for four single-cell datasets of glioma (Table S4), we firstly derived stem markers from dataset GSE57872 by choosing the top 200 genes with the largest log fold change (FC) values between Glioblastoma stem-like cells and the differentiated cells (edgeR, FDR < 0.05, Table S7). Then, we found these genes were not only enriched in the StemSC-ranked gene list (Fig. 6A-D) but also correlated with StemSC values ($p < 0.05$, Fig. 6E-L) in both the dataset GSE57872 and the other three independent datasets. Besides, for the dataset GSE117891 with grade information, significantly higher StemSC values were not only in the cells with label Grade IV than those with label Grade III – IV but also in the tumoral cells than the peritumoral cells (student T-test, $p < 0.05$, Fig. 6M, N). Particularly, there were significantly lower StemSC values of the 563 differentiated tumor cells than those of the 134 cancer stem cells in the dataset GSE57872 (student T-test, $p < 0.05$, Fig. 6O). Besides, the mixed StemSC values between differentiated tumor cells and cancer stem cells reminded us the existence of high-stemness cells in the differentiated tumor cells. Thus, we divided the 563 differentiated tumor cells into low- and high-stemness cells by the median StemSC value (0.866) of the cancer stem cells. Enrichment analysis showed that the stemness markers were significantly enriched in the gene sets ranked by the FCs between high- and low-stemness cells (Fig. 6P), which revealed that the StemSC threshold 0.866 could be used to distinguish the high-stemness cells in glioma. All above results supported the applicability of StemSC on tumor tissue cells.

The effect of stemness on tumor immune microenvironment

Cancer progression is accompanied by the acquisition of stemness, which greatly affects the immune response of the tumor cells. As reported, the resistance to immune-mediated destruction was shown to be an intrinsic property of cancer stem cells [25]. Similarly, at the bulk tissue level, pervasive negative associations were found between cancer stemness and anticancer immunity [4]. However, at the single-cell level, the effect of stemness on the interaction between tumor cells and tumor microenvironment remains incompletely understood.

Here, we further divided the cells into different kinds of immune cells, low-stemness, and high-stemness tumor cells in the above four glioma datasets except dataset GSE57872, which only has tumor cells. For dataset GSE117891, we used inferCNV (see Methods) to infer the copy number variations (CNV) of all the 4623 tumor tissue cells. Further, we divided these cells into 2724 tumor cells and the 1899 normal cells without obvious CNV by using the hierarchical cluster (Fig. 7A, see Methods). For the identified normal cells, we used SingleR [18] to identify the four immune cell types with more than 50 cells, which was further confirmed by the corresponding cell markers (Fig. 7B). For the identified tumor cells, we used the above threshold 0.866 to classify these cells into 2202 low-stemness and 522 high-stemness cells, which was confirmed by the enrichment of the 200 stemness markers in the FC-ranked gene sets between high- and low-stemness cells (Fig. 7C). Further, cell-cell communication analysis (see Methods) showed that, contrary to the low-stemness cells, the high-stemness cells had fewer connections with each other and fewer interactions with the four types of immune cells (Fig. 7F). A similar result could be found in the two additional glioma datasets (Fig. 7D, E, G, H), which implied the resistance of high-stemness cells to the immunotherapy. To validate this result, we further collected the bulk RNA-seq samples of 13

immunotherapy-treated patients as the mean values of single-cell samples (dataset PRJNA482620). And we found that the median StemSC values of non-responders were significantly higher than those of responders (student T-test, $p < 0.05$, Fig. 7I). Besides, we used the median values to divide the 13 samples into high- and low-stemness groups. Survival analysis showed that the high-stemness group had marginally significantly worse overall survival than the low-stemness group (HR = 6.20; C-index = 0.63; p = 8.18E-2, Fig. 7J), which validated the negative effect of stemness on immunotherapy.

Discussion

Stemness, which describes the differentiation potential of cells, can be a potential index to construct the cellular differentiation trajectories [1, 3]. Here, we developed an REO-based stemness index called StemSC which could be well applied to single-cell samples across datasets. In addition to the high correlation with differentiation time, StemSC can be combined with Monocle 2 to construct the cellular differentiation trajectories automatically. Besides, the insensitivity of StemSC to batch effects made it possible to identify the high-stemness cells in independent datasets. Finally, we found that the interaction between immune cells and tumor cells decreased with the increase of stemness.

REO is the feature that transforms the continuous expression values into discrete values. Thus, using REOs instead of the absolute values of gene expression can effectively avoid overfitting and outliers in the training process. Besides, the larger the numbers of training datasets, the more stable the REO-based signatures are. However, due to the insufficiency of the datasets with differentiation time information and the single-cell samples of ESCs, we chose the limited datasets for training. Similarly, it is hard to explore the ability of StemSC to identify the cancer stem cells individually since the lack of cancer stem cells with experimental validation.

In the future work, it is hopeful to recognize the cancer stem cells individually when there are enough single-cell samples of cancer stem cells. Besides, the development of StemSC can help other researchers to study the cell differentiation trajectories. When applying StemSC to tumor tissue cells, we found that the high-stemness cells had fewer connections with immune cells. The deeper investigations of this phenomenon may reveal new mechanisms of immune cell regulation and provide a new direction for immunotherapy.

Conclusion

We constructed a REO-based stemness index for the single-cell samples, StemSC, which showed high correlations with the differentiation time of embryonic stem cells and high correlations with tumor dedifferentiation. In addition to its ability to construct cellular trajectories, StemSC could also be used to recognized the high-stemness cells across datasets and reveal that the high-stemness tumor cells had fewer connections with anti-tumor immune cells.

List Of Abbreviations

StemSC, stemness index for single-cell samples; REOs, relative expression orderings; ESC, embryonic stem cell; CNV, copy number variations; FDR, false discovery rate; KEGG, Kyoto Encyclopedia of Genes and Genomes

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are available in the in the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) and Progenitor Cell Biology Consortium (PCBC, https://www.synapse.org/#!Synapse:syn1773109/wiki/54962).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China [grant numbers: 61673143; 81572935].

Author contributions

HZ and JX involved in conception and design. KS provided the study material. HZ and JX analyzed and interpreted the data. HZ wrote the manuscript. WZ and YG provided administrative support and financial support. All the authors finally approved the manuscript.

Acknowledgements

Hailong Zheng and Jiajing Xie contributed equally to this work.

References

1. J. Wu, J.C. Izpisua Belmonte, Stem Cells: A Renaissance in Human Biology Research, Cell, 165 (2016) 1572-1585.doi:10.1016/j.cell.2016.05.043
2. S.S. Thorgeirsson, Stemness and reprogramming in liver cancer, Hepatology, 63 (2016) 1068-1070.doi:10.1002/hep.28362

3. G.S. Gulati, S.S. Sikandar, D.J. Wesche, A. Manjunath, A. Bharadwaj, M.J. Berger, F. Ilagan, A.H. Kuo, R.W. Hsieh, S. Cai, M. Zabala, F.A. Scheeren, N.A. Lobo, D. Qian, F.B. Yu, F.M. Dirbas, M.F. Clarke, A.M. Newman, Single-cell transcriptional diversity is a hallmark of developmental potential, Science, 367 (2020) 405-411.doi:10.1126/science.aax0249
4. A. Miranda, P.T. Hamilton, A.W. Zhang, S. Pattnaik, E. Becht, A. Mezheyeuski, J. Bruun, P. Micke, A. de Reynies, B.H. Nelson, Cancer stemness, intratumoral heterogeneity, and immune response across cancers, Proc Natl Acad Sci U S A, 116 (2019) 9020-9029.doi:10.1073/pnas.1818210116
5. T.M. Malta, A. Sokolov, A.J. Gentles, T. Burzykowski, L. Poisson, J.N. Weinstein, B. Kaminska, J. Huelsken, L. Omberg, O. Gevaert, A. Colaprico, P. Czerwinska, S. Mazurek, L. Mishra, H. Heyn, A. Krasnitz, A.K. Godwin, A.J. Lazar, N. Cancer Genome Atlas Research, J.M. Stuart, K.A. Hoadley, P.W. Laird, H. Noushmehr, M. Wiznerowicz, Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation, Cell, 173 (2018) 338-354 e315.doi:10.1016/j.cell.2018.03.034
6. H. Zheng, K. Song, Y. Fu, T. You, J. Yang, W. Guo, K. Wang, L. Jin, Y. Gu, L. Qi, W. Zhao, An absolute human stemness index associated with oncogenic dedifferentiation, Brief Bioinform, (2020).doi:10.1093/bib/bbz174
7. L. Ao, Y. Guo, X. Song, Q. Guan, W. Zheng, J. Zhang, H. Huang, Y. Zou, Z. Guo, X. Wang, Evaluating hepatocellular carcinoma cell lines for tumour samples using within-sample relative expression orderings of genes, Liver Int, 37 (2017) 1688-1696.doi:10.1111/liv.13467
8. R. Chen, J. He, Y. Wang, Y. Guo, J. Zhang, L. Peng, D. Wang, Q. Lin, J. Zhang, Z. Guo, L. Li, Qualitative transcriptional signatures for evaluating the maturity degree of pluripotent stem cell-derived cardiomyocytes, Stem Cell Res Ther, 10 (2019) 113.doi:10.1186/s13287-019-1205-1
9. H. Zheng, K. Song, Y. Fu, T. You, J. Yang, W. Guo, K. Wang, L. Jin, Y. Gu, L. Qi, W. Zhao, Z. Guo, A qualitative transcriptional signature for determining the grade of colorectal adenocarcinoma, Cancer Gene Ther, 27 (2020) 680-690.doi:10.1038/s41417-019-0139-1
10. T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L. Robertson, N. Serova, S. Davis, A. Soboleva, NCBI GEO: archive for functional genomics data sets–update, Nucleic Acids Res, 41 (2013) D991-995.doi:10.1093/nar/gks1193
11. K. Daily, S.J. Ho Sui, L.M. Schriml, P.J. Dexheimer, N. Salomonis, R. Schroll, S. Bush, M. Keddache, C. Mayhew, S. Lotia, T.M. Perumal, K. Dang, L. Pantano, A.R. Pico, E. Grassman, D. Nordling, W. Hide, A.K. Hatzopoulos, P. Malik, J.A. Cancelas, C. Lutzko, B.J. Aronow, L. Omberg, Molecular, phenotypic, and sample-associated data to describe pluripotent stem cell lines and derivatives, Sci Data, 4 (2017) 170030.doi:10.1038/sdata.2017.30
12. A.P. Patel, I. Tirosh, J.J. Trombetta, A.K. Shalek, S.M. Gillespie, H. Wakimoto, D.P. Cahill, B.V. Nahed, W.T. Curry, R.L. Martuza, D.N. Louis, O. Rozenblatt-Rosen, M.L. Suva, A. Regev, B.E. Bernstein, Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma, Science, 344 (2014) 1396-1401.doi:10.1126/science.1254257

13. J. Harrow, A. Frankish, J.M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B.L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J.M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, T.J. Hubbard, GENCODE: the reference human genome annotation for The ENCODE Project, Genome Res, 22 (2012) 1760-1774.doi:10.1101/gr.135350.111
14. D. Kim, J.M. Paggi, C. Park, C. Bennett, S.L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype, Nat Biotechnol, 37 (2019) 907-915.doi:10.1038/s41587-019-0201-4
15. Y. Liao, G.K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics, 30 (2014) 923-930.doi:10.1093/bioinformatics/btt656
16. X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H.A. Pliner, C. Trapnell, Reversed graph embedding resolves complex single-cell trajectories, Nat Methods, 14 (2017) 979-982.doi:10.1038/nmeth.4402
17. G.T. Consortium, D.A. Laboratory, G. Coordinating Center -Analysis Working, G. Statistical Methods groups-Analysis Working, G.g. Enhancing, N.I.H.C. Fund, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, N. Biospecimen Collection Source Site, R. Biospecimen Collection Source Site, V. Biospecimen Core Resource, B. Brain Bank Repository-University of Miami Brain Endowment, M. Leidos Biomedical-Project, E. Study, I. Genome Browser Data, E.B.I. Visualization, I. Genome Browser Data, U.o.C.S.C. Visualization-Ucsc Genomics Institute, a. Lead, D.A. Laboratory, C. Coordinating, N.I.H.p. management, c. Biospecimen, Pathology, Q.T.L.m.w.g. e, A. Battle, C.D. Brown, B.E. Engelhardt, S.B. Montgomery, Genetic effects on gene expression across human tissues, Nature, 550 (2017) 204-213.doi:10.1038/nature24277
18. D. Aran, A.P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, S. Chak, R.P. Naikawadi, P.J. Wolters, A.R. Abate, A.J. Butte, M. Bhattacharya, Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage, Nat Immunol, 20 (2019) 163-172.doi:10.1038/s41590-018-0276-y
19. M. Efremova, M. Vento-Tormo, S.A. Teichmann, R. Vento-Tormo, CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes, Nat Protoc, 15 (2020) 1484-1506.doi:10.1038/s41596-020-0292-x
20. G. Yu, L.G. Wang, Y. Han, Q.Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, OMICS, 16 (2012) 284-287.doi:10.1089/omi.2011.0118
21. R. Chen, Q. Guan, J. Cheng, J. He, H. Liu, H. Cai, G. Hong, J. Zhang, N. Li, L. Ao, Z. Guo, Robust transcriptional tumor signatures applicable to both formalin-fixed paraffin-embedded and fresh-frozen samples, Oncotarget, 8 (2017) 6652-6662.doi:10.18632/oncotarget.14257
22. R.C. Wong, A. Ibrahim, H. Fong, N. Thompson, L.F. Lock, P.J. Donovan, L1TD1 is a marker for undifferentiated human embryonic stem cells, PLoS One, 6 (2011)

e19355.doi:10.1371/journal.pone.0019355

23. F.A. Giger, N.B. David, Endodermal germ-layer formation through active actin-driven migration triggered by N-cadherin, Proc Natl Acad Sci U S A, 114 (2017) 10143-10148.doi:10.1073/pnas.1708116114
24. P. Dalerba, T. Kalisky, D. Sahoo, P.S. Rajendran, M.E. Rothenberg, A.A. Leyrat, S. Sim, J. Okamoto, D.M. Johnston, D. Qian, M. Zabala, J. Bueno, N.F. Neff, J. Wang, A.A. Shelton, B. Visser, S. Hisamori, Y. Shimono, M. van de Wetering, H. Clevers, M.F. Clarke, S.R. Quake, Single-cell dissection of transcriptional heterogeneity in human colon tumors, Nat Biotechnol, 29 (2011) 1120-1127.doi:10.1038/nbt.2038
25. V.S. Bruttel, J. Wischhusen, Cancer stem cell immunology: key to understanding tumorigenesis and tumor immune escape?, Front Immunol, 5 (2014) 360.doi:10.3389/fimmu.2014.00360

Figures

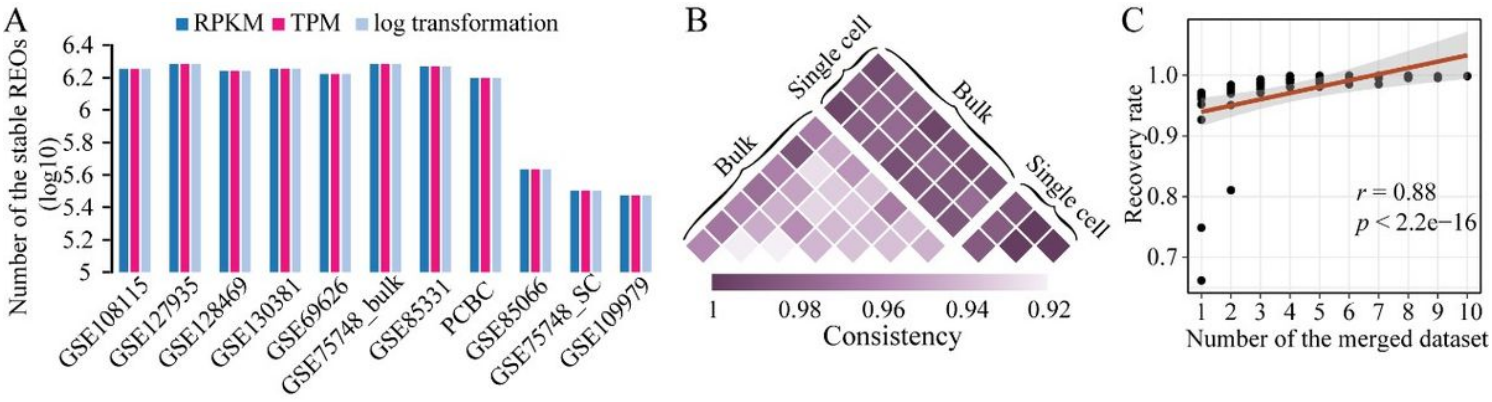

Figure 1

Stability of REOs in both bulk and single-cell ESC samples. (A) The number of stable REOs identified from RPKM, TPM, and log transformation data. (B) The consistency of stable REOs among 11 ESC datasets. (C) The correlation between the number of merged datasets for identifying stable REOs and recovery rate of REOs in the remaining datasets.

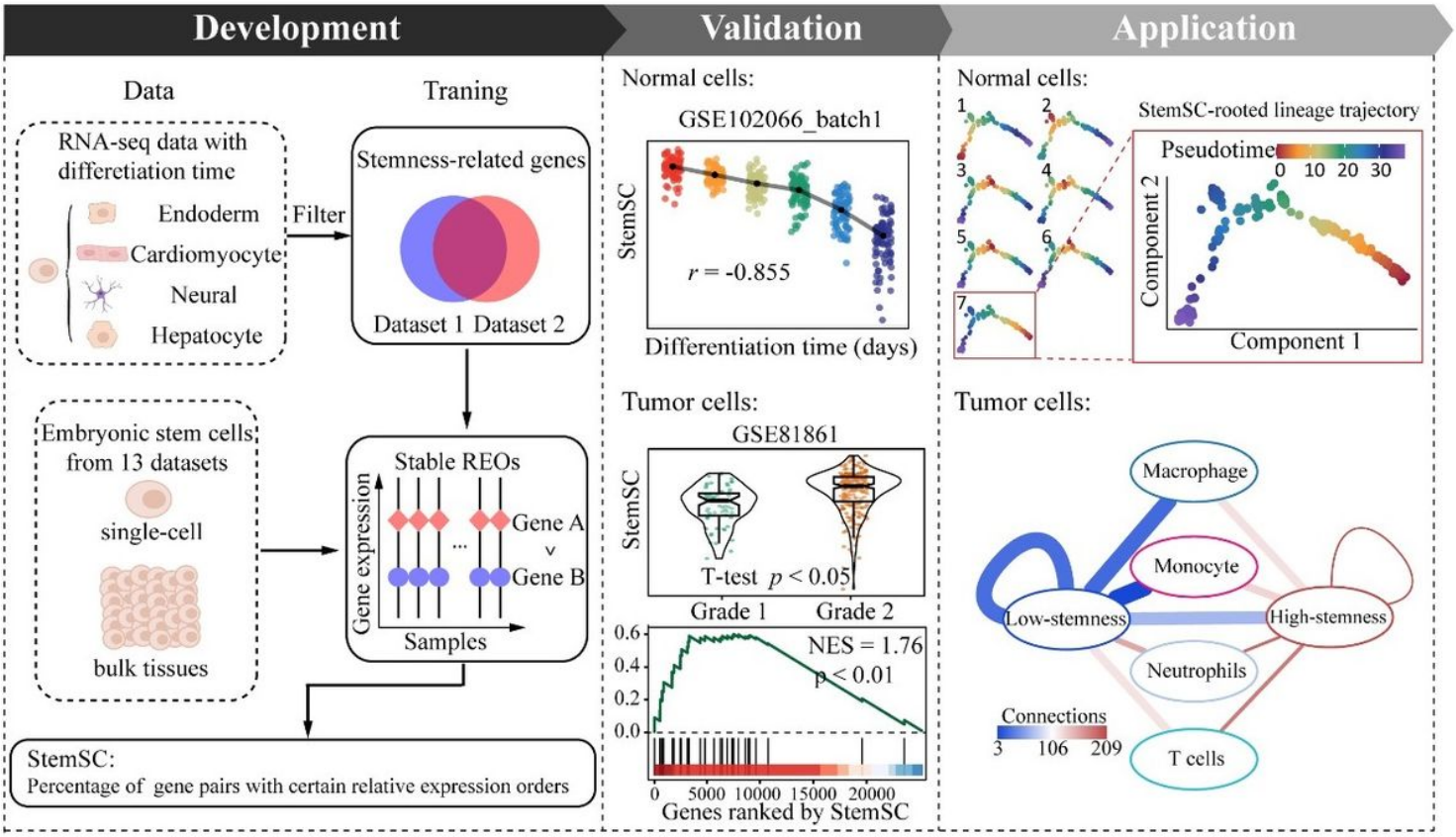

Figure 2

The overall methodology of StemSC.

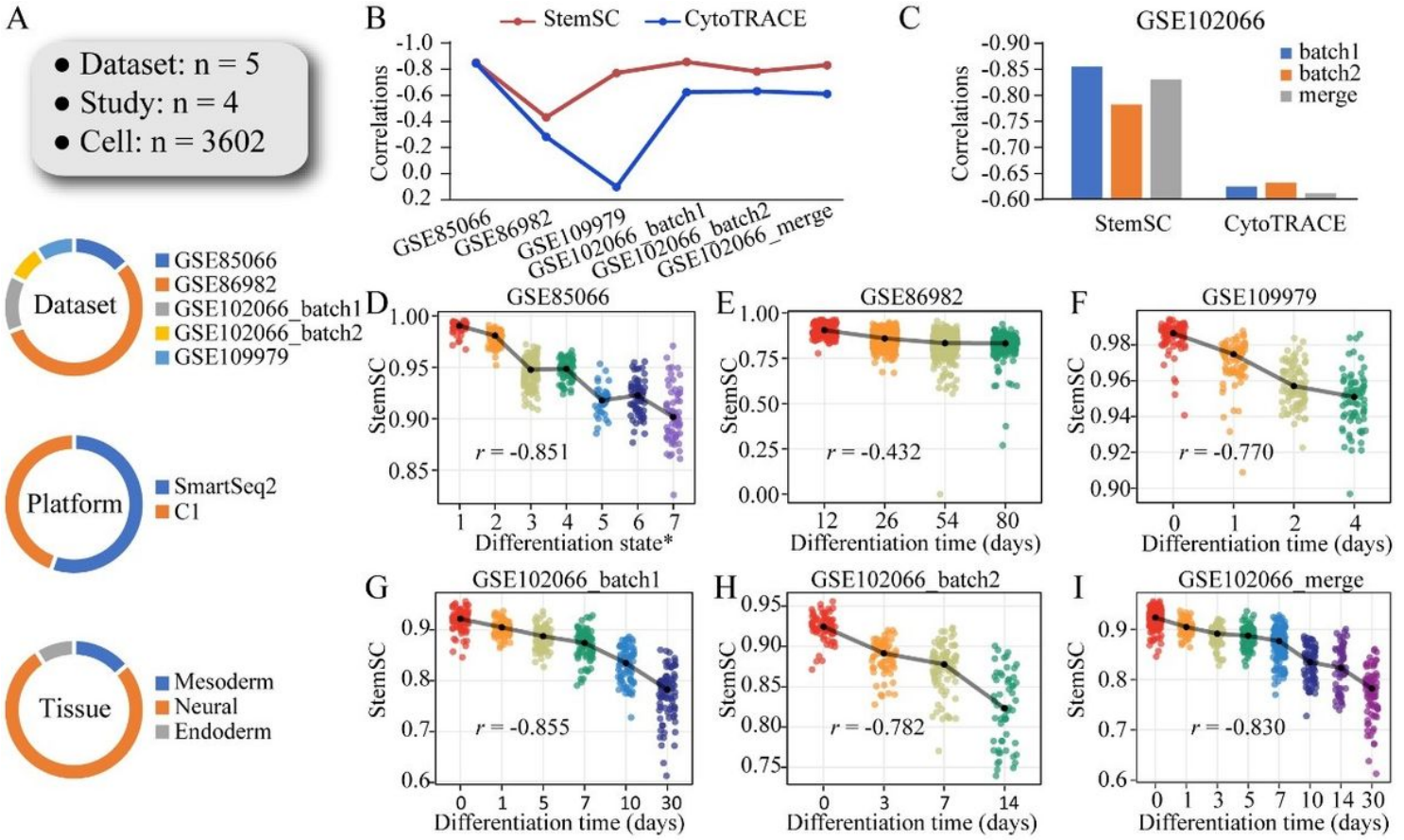

Figure 3

Validation of the StemSC in the single-cell datasets with differentiation time. (A) The general information of validation sets. (B) The correlations between differentiation time and stemness index (StemSC and CytoTRACE) in all validation sets. (C) The changes of correlations between differentiation time and stemness index (StemSC and CytoTRACE) after combining the two batches of GSE102066. (D-I) The high correlations differentiation time and StemSC in each validation set. *Differentiation state of dataset GSE85066 was provided in Table S5

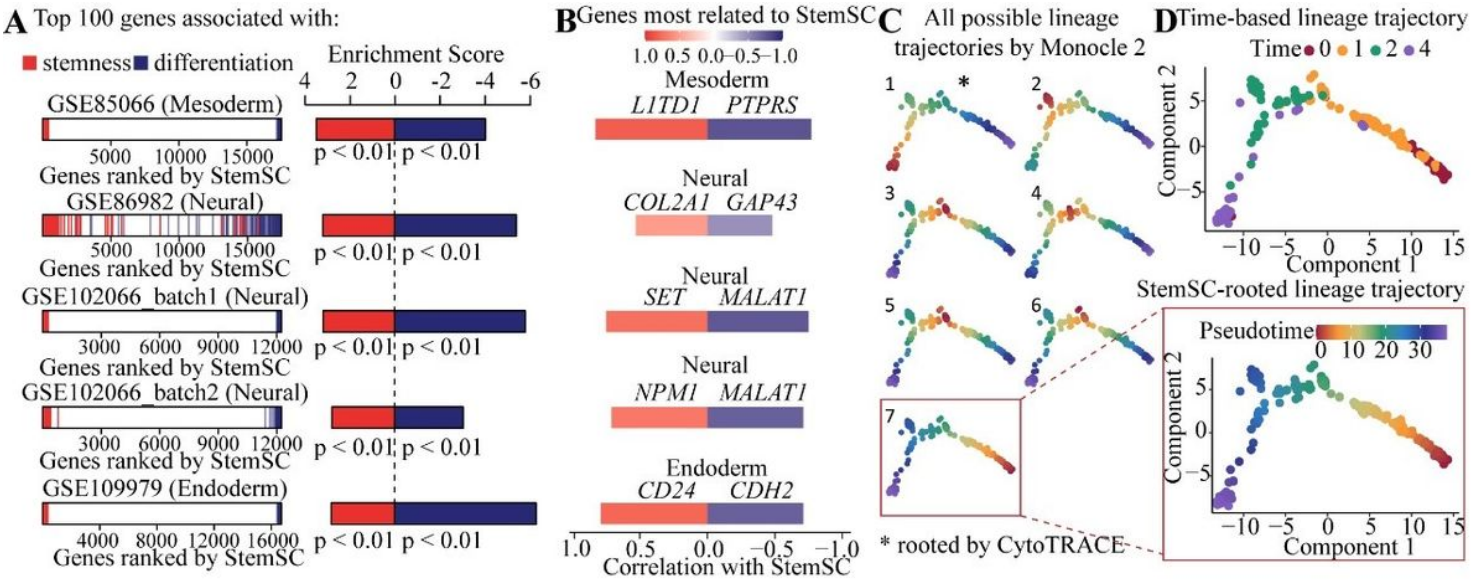

Figure 4

The abilities of StemSC to Identify the stemness-related genes and cellular differentiation trajectories. (A) The enrichment of the top 100 stemness-associated or differentiation-associated genes (the top 100 genes positively or negatively correlated with differentiation time) in the StemSC-ranked gene list. (B) Genes most positively or negatively correlated with StemSC. (C) Construction of lineage trajectory by combining Monocle 2 and StemSC. (D) The time-based lineage trajectory.

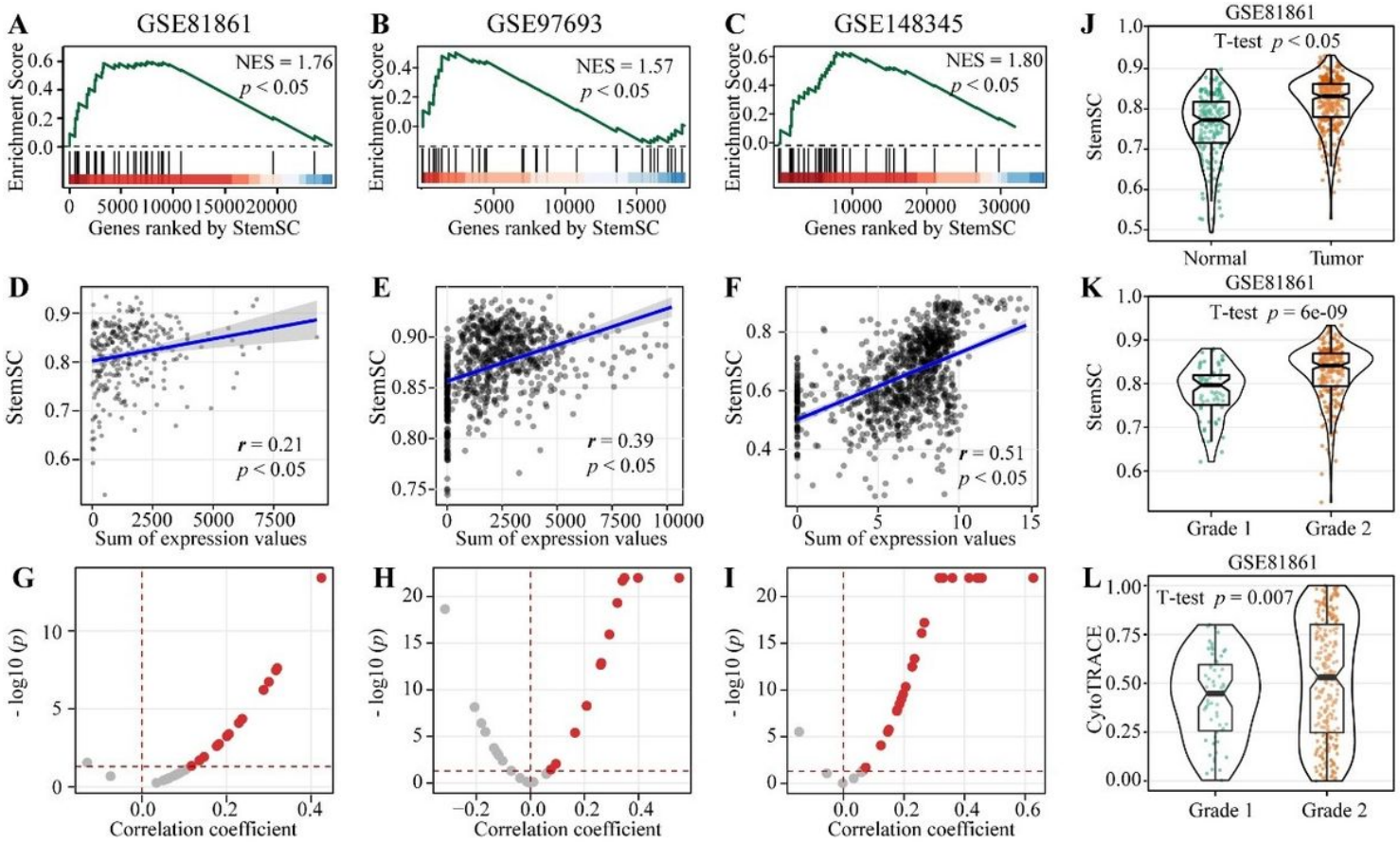

Figure 5

The validation of StemSC in colorectal cancer. (A-C) Enrichment of the 30 intestinal stem cell markers in the StemSC-ranked gene list. (D-F) The correlation between StemSC and the sum of gene expression values of the 30 intestinal stem cell markers. (G-I) The correlations between the StemSC and the gene expression values of 30 intestinal stem cell markers. (J) The significant difference of StemSC between tumor and normal tissue cells. The difference of stemness index between cells with different grades by using StemSC (K) and CytoTRACE (L).

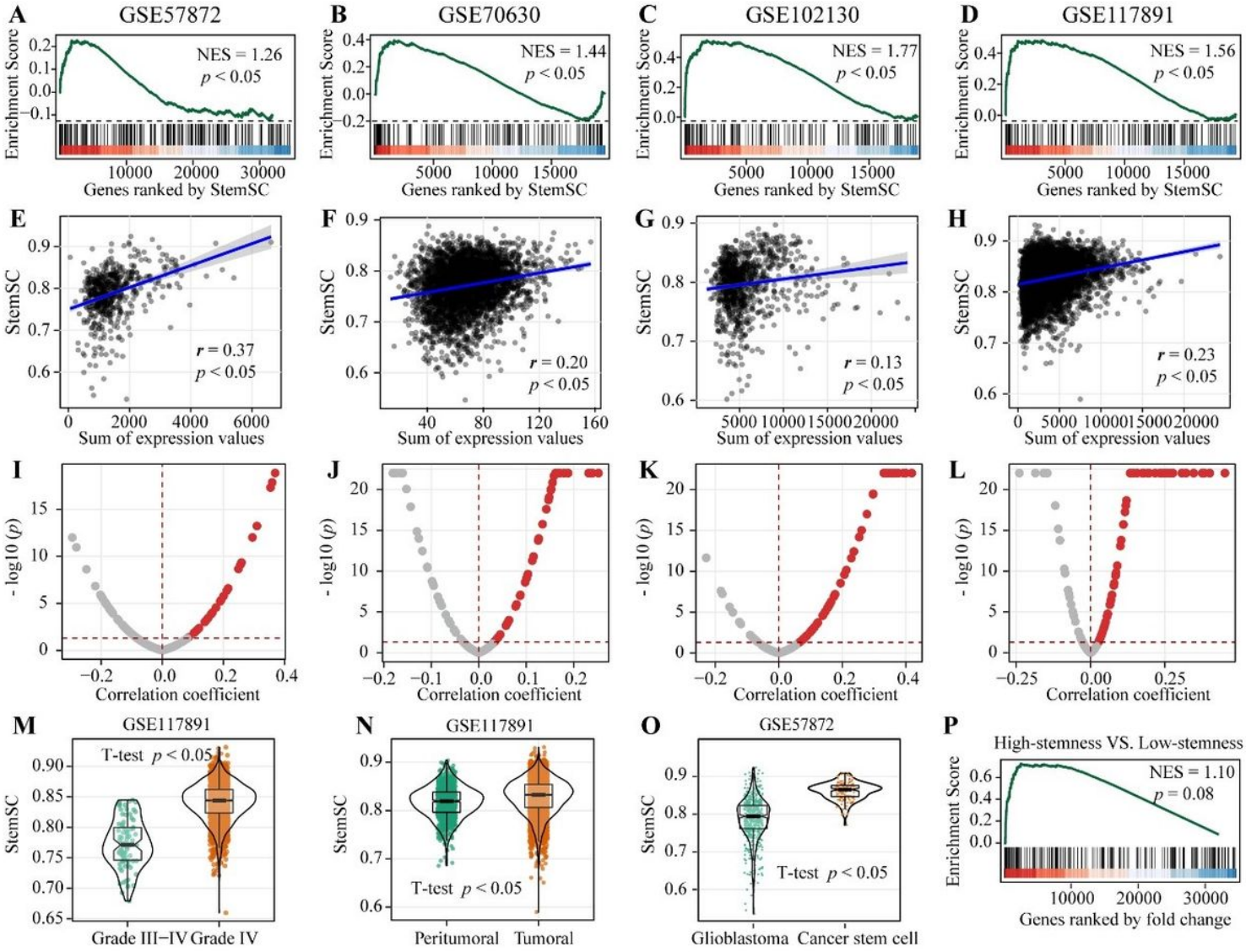

Figure 6

The validation of StemSC in glioma. (A-D) Enrichment of the 200 glioma stem markers in the StemSC-ranked gene list. (E-H) The correlation between StemSC and the sum of gene expression values of the 200 glioma stem markers. (I-L) The correlations between the StemSC and the gene expression values of the 200 glioma stem markers. The significant difference of StemSC between (M) different grades (N) tumor and normal tissue cells (O) cancer stem cells and differentiated cells. (P) Enrichment of the 200 glioma stem markers in the gene sets ranked by the FCs between high- and low-stemness cells.

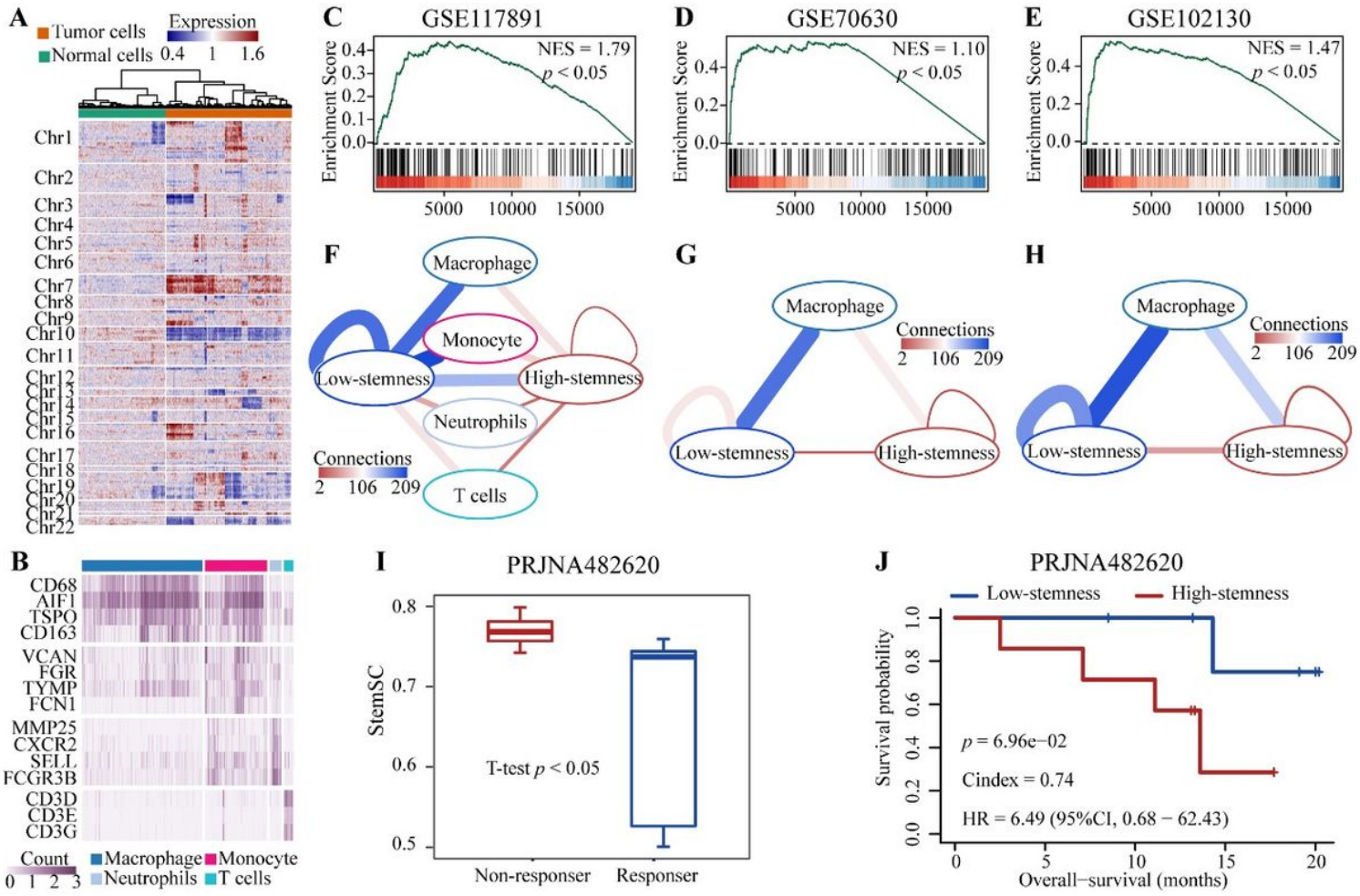

Figure 7

The effect of stemness on tumor immune microenvironment. (A) The hierarchical cluster of the inferred copy number variation in the tumoral tissue cells of dataset GSE117891. (B) The expressions of the corresponding markers for the four types of immune cells in the dataset GSE117891. (C-E) The enrichment of the 200 stemness markers in the gene sets ranked by the log FCs between high- and low-stemness cells. (F-H) Interaction network of immune cells, high- and low-stemness cells. (I) The higher median StemSC values in the non-responders than in the responders. (J) The Kaplan–Meier curves of overall survival in the high- and low-stemness groups.

Supplementary Files

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

- Supplementarymaterials.docx