Published in final edited form as:

Cell. 2018 April 05; 173(2): 338–354.e15. doi:10.1016/j.cell.2018.03.034.

*Correspondance: hnoushm1@hfhs.org (HN), maciej.wiznerowicz@iimo.pl (MW).

[23] These authors contributed equally

[24] Lead Contact

Secondary author list:

Samantha J. Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L. Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C. Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok-Shing Ng, Arvind Rao, Michael Ryan, Jing Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips, Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz, Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau, D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J.M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson, Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J. Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, HarshaVardhan Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbrot, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandoth, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M. Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lichtenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C.S. Silveira, Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann, David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavobilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chandrajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliante, Marc Goodman, Beth Y. Karlan, Curt H. Hagedorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc-Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Hongzheng Zhang, Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric Baudin, Glenn Bubley, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalkanis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang, Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason, Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Akeredolu, Serghei Stepa, Dave Hoon, Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein, Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Castle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas Flotte, Nasra Giama, Thai Ho, Michael Kendrick, Jean-Pierre Kocher, Karla Kopp, Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Smallridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson, Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonzalez, Carmen Behrens, Jolanta Bondaruk, Russell Broaddus, Bogdan Czerniak, Bita Esmaeli, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Alexander J. Lazar, Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta, David Rice, Anil Sood, Pheroze Tamboli, Timothy Thompson, Patricia Troncoso, Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Valerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Graham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein,

Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation

Tathiane M. Malta[1,2,23], **Artem Sokolov**[3,23], **Andrew J. Gentles**[4], **Tomasz Burzykowski**[5], **Laila Poisson**[1], **John N. Weinstein**[6], **Bożena Kamińska**[7], **Joerg Huelsken**[8], **Larsson Omberg**[9], **Olivier Gevaert**[4], **Antonio Colaprico**[10,11], **Patrycja Czerwińska**[12], **Sylwia Mazurek**[12,13], **Lopa Mishra**[14], **Holger Heyn**[15], **Alex Krasnitz**[16], **Andrew K. Godwin**[17], **Alexander J. Lazar**[6], **The Cancer Genome Atlas Research Network**, **Joshua M. Stuart**[18], **Katherine A. Hoadley**[19], **Peter W. Laird**[20], **Houtan Noushmehr**[1,2,23,*], and **Maciej Wiznerowicz**[12,21,22,23,24,*]

Anuradha Gopalan, Douglas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jennifer Smith, Howard Zaren, Joong-Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne-Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl-Friedrich Becker, Klaus-Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne-Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jr., Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, and Armaz Mariamidze

AUTHOR CONTRIBUTIONS

The TCGA Research Network contributed collectively to this study. The contributions of other authors are as follows: epigenetic-derived stemness index, T.M.M., P.W.L., and H.N.; mRNA expression-derived stemness index, A.S.; methodology, T.M.M., A.S., and H.N.; integrative analysis, T.M.M., A.S., and H.N.; clinical analysis, T.B. and L.P; drug analysis, A.C.; data interpretation, T.M.M., A.S., J.H., B.K., H.H., A.J.G., A.C., L.M., P.W.L., H.N., and M.W.; data curation, L.O.; writing, T.M.M., A.S., J.W., B.K., J.H., A.C., J.N.W., H.N., and M.W.; visualization, T.M.M., A.S., H.N., and M.W.; overall concept and coordination, T.M.M., A.S., P.W.L., H.N., and M.W.

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and five tables and can be found with this article online at: *TBD*

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

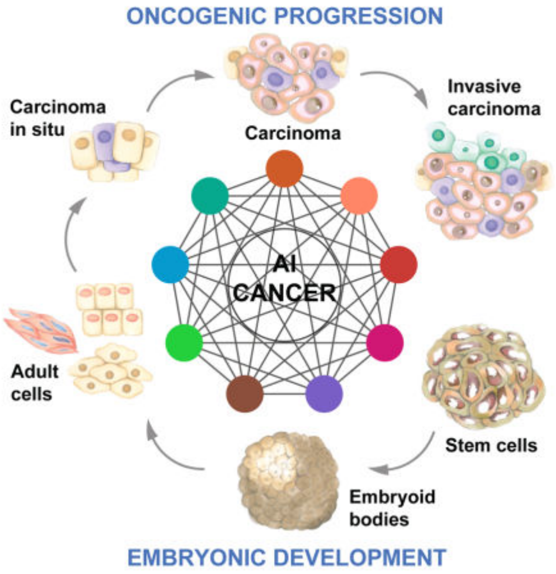
[1]Henry Ford Health System; Detroit, MI, 48202; USA [2]University of São Paulo; Ribeirão Preto, SP, 14049-900; Brazil [3]Harvard Medical School; Boston, MA, 02115; USA [4]Stanford University; Palo Alto, CA, 94305; USA [5]Hasselt University; Diepenbeek, 3590; Belgium [6]The University of Texas MD Anderson Cancer Center; Houston, Texas, 77030; USA [7]Nencki Institute of Experimental Biology of PAS; Warsaw, 02093; Poland [8]Swiss Federal Institute of Technology Lausanne (EPFL); Lausanne, CH-1015; Switzerland [9]Sage Bionetworks; Seattle, WA, 98109; USA [10]Université Libre de Bruxelles; Bruxelles, 1050; Belgium [11]Interuniversity Institute of Bioinformatics in Brussels (IB)2; Bruxelles, 1050; Belgium [12]Poznań University of Medical Sciences; Poznań, 61701; Poland [13]Postgraduate School of Molecular Medicine, Medical University of Warsaw; Warsaw, 02109; Poland [14]George Washington University; Washington DC, 20052; USA [15]Centre for Genomic Regulation (CNAG-CRG); Barcelona, 08003; Spain [16]Cold Spring Harbor Laboratory; Cold Spring Harbor, NY, 11724; USA [17]University of Kansas Medical Center; Kansas City, KS, 66160; USA [18]University of California Santa Cruz; Santa Cruz, CA, 95064; USA [19]University of North Carolina; Chapel Hill, NC, 27599; USA [20]Van Andel Research Institute; Grand Rapids, Michigan, 49503; USA [21]Greater Poland Cancer Center; Poznań, 61866; Poland [22]International Institute for Molecular Oncology; Poznań, 60203; Poland

SUMMARY

Cancer progression involves the gradual loss of a differentiated phenotype and acquisition of progenitor and stem cell-like features. Here, we provide novel stemness indices for assessing the degree of oncogenic dedifferentiation. We used an innovative one-class logistic regression machine learning algorithm (OCLR) to extract transcriptomic and epigenetic feature sets derived from non-transformed pluripotent stem cells and their differentiated progeny. Using OCLR, we were able to identify previously undiscovered biological mechanisms associated with the dedifferentiated oncogenic state. Analyses of the tumor microenvironment revealed unanticipated correlation of cancer stemness with immune checkpoint expression and infiltrating immune system cells. We found that the dedifferentiated oncogenic phenotype was generally most prominent in metastatic tumors. Application of our stemness indices to single cell data revealed patterns of intra-tumor molecular heterogeneity. Finally, the indices allowed for the identification of novel targets and possible targeted therapies aimed at tumor differentiation.

ETOC

Stemness features extracted from transcriptomic and epigenetic data from TCGA tumors reveals new drug targets for anti-cancer therapies



INTRODUCTION

Stemness, defined as the potential for self-renewal and differentiation from the cell-of-origin, was originally attributed to normal stem cells that possess the ability to give rise to all cell types in the adult organism. Cancer progression involves gradual loss of differentiated phenotype and acquisition of progenitor-like, stem cell-like features. Undifferentiated primary tumors are more likely to result in cancer cell spread to distant organs, causing disease progression and poor prognosis, particularly because metastases are usually resistant to available therapies (Friedmann-Morvinski and Verma, 2014; Ge et al., 2017; Shibue and Weinberg, 2017; Visvader and Lindeman, 2012).

An increasing number of genomic, epigenomic, transcriptomic, and proteomic signatures have been associated with cancer stemness. Those molecular features are causally connected to particular oncogenic signaling pathways that regulate transcriptional networks that sustain the growth and proliferation of cancer cells (Ben-Porath et al., 2008; Eppert et al., 2011; Kim et al., 2010). Transcriptional and epigenetic dysregulation of cancer cells frequently leads to oncogenic de-differentiation and acquisition of stemness features by altering core signaling pathways that regulate the phenotypes of normal stem cells (Bradner et al., 2017; Young, 2011). Cell-extrinsic mechanisms can also affect maintenance of the undifferentiated state, largely through epigenetic mechanisms. Tumors comprise a complex, diverse, integrated ecosystem of relatively differentiated cancer cells, cancer stem cells, endothelial cells, tumor-associated fibroblasts, and infiltrating immune cells, among other cell types. The microenvironment of a tumor, considered as a pathologically formed “organ” is frequently characterized by hypoxia, as well as by abnormal levels of various cytokines, growth factors, and metabolites (Lyssiotis and Kimmelman, 2017). It provides numerous opportunities for cell-cell signals to modulate the epigenome and expression of stem cell-like programs in cancer cells, frequently independent of their genetic backgrounds (Gingold et al., 2016).

Over the last decade, The Cancer Genome Atlas (TCGA) has illuminated the landscapes of primary tumors by generating comprehensive molecular profiles composed of genomic, epigenomic, transcriptomic, and (post-translational) proteomic characteristics (Hoadley et al., 2014; Tomczak et al., 2015), along with histopathological and clinical annotations. The resulting resource enabled us to analyze cancer stemness quite extensively in almost 12,000 samples of 33 tumor types.

First, we defined signatures to quantify stemness, using publicly available molecular profiles from normal cell types that exhibit various degrees of stemness. By multi-platform analyses of their transcriptome, methylome, and transcription factor binding sites using an innovative one-class logistic regression machine-learning algorithm (OCLR) (Sokolov et al., 2016), we obtained two independent stemness indices. One (mDNAsi) was reflective of epigenetic features, the other (mRNAsi) was reflective of gene expression. We then identified associations between the two stemness indices and novel oncogenic pathways, somatic alterations, and microRNA and transcriptional regulatory networks. Those features correlated with, and perhaps govern, cancer stemness in particular molecular subtypes of TCGA tumors. Importantly, higher values for stemness indices were associated with biological processes active in cancer stem cells and with greater tumor dedifferentiation, as reflected in histopathological grade.

Metastatic tumor cells appeared more dedifferentiated phenotypically, probably contributing to their aggressiveness. We also found tumor heterogeneity at the single-cell level by measuring stemness in transcriptome profiles obtained from individual cancer cells. Using CIBERSORT to profile immune cell types in TCGA tumors, we obtained insight into the interface of the immune system with stemness. Finally, we identified compounds specific to selected molecular targets and mechanisms that may eventually lead to novel treatments that trigger differentiation and exhaust the stemness potential of highly aggressive neoplasms.

RESULTS

DNA methylation- and mRNA expression-based stemness classifiers

We analyzed publicly available non-tumor and tumor datasets for which transcriptomic and epigenomic molecular profiles were available (Figure 1A). We derived stemness indices using an OCLR algorithm trained on stem cells (SC; ESC/iPSC) classes and their differentiated ecto-, meso- and endoderm progenitors. We chose OCLR because it does not penalize misclassification of stem cell-derived progenitors at different stages of differentiation, which still carry some of the undifferentiated features in their molecular profiles (its output was also validated against Random Forests in Figure S1A). OCLR-based transcriptomic and epigenetic signatures were applied to TCGA datasets to calculate the mRNAsi and mDNAsi. Each stemness index (si) ranges from low (zero) to high (one) stemness (Table S1). The tumor samples stratified by the indices were used for the integrative analyses.

mRNA expression-based stemness index—We validated the mRNAsi by applying it to an external dataset composed of both stem cells and somatic differentiated cells (Nazor et al., 2012) (Figure 1B), and by scoring molecular subtypes of breast cancers and gliomas that are characterized by different degrees of oncogenic dedifferentiation associated with pathology and clinical outcome (Figures S1B and S1C). All stem cell samples attained higher si values than samples from differentiated cells. TCGA tumors display various degrees of cancer stemness as revealed by mRNAsi (Figure 1C, left) and mDNAsi (Figure 1C, right). Germ-cell tumors, basal breast cancer, and Ly-Hem cancers displayed highly dedifferentiated phenotypes in comparison to other tumor types.

Using GSEA, we compared our signature to 16 gene sets that were associated with stemness in cancer and healthy cells in previous studies (Ben-Porath et al., 2008; Ivanova et al., 2006; Kim and Orkin, 2011; Mathur et al., 2008; Palmer et al., 2012; Sato et al., 2003; Venezia et al., 2004; Yan et al., 2011). These sets spanned 2,564 unique genes, with no two sets overlapping by more than 134 genes. In all cases, the published stemness gene sets were significantly enriched in mRNAsi (Figure 2A). We found that "Cancer Hallmark" gene sets were significantly enriched, as were MYC targets which significantly contributed to the positive side of the signature (Hanahan and Weinberg, 2011). This is consistent with MYC being one of the transcription factors that drive pluripotency in ESC (Young, 2011).

Wnt/β-catenin and TGFβ signaling pathways were significantly enriched on the negative side of the stemness signature. This negative enrichment does not imply absence of specific signals in cancer stem cells, but rather that this signaling is lower relative to stem cell-derived progenitors, as captured by the signature weights. This is again consistent with other GSEA results, as both signaling pathways are known mediators of the EMT mechanism (Gonzalez and Medici, 2014). We also computed the correlation of mRNAsi against mRNA expression of published pan-cancer EMT markers (Mak et al., 2016), which revealed significant correlations with for most tumors (Figures S2C). This is consistent with the biology of ESCs, which grow as epithelioid, polygonal cells *in vitro* and epithelial cancer precursors having stem-like properties. Importantly, most TCGA samples are primary tumors of an epithelial phenotype. Most skin melanoma cases come from lymph nodes and this tumor type shows higher expression of vimentin, a key marker of mesenchymal phenotype. mRNAsi is positively correlated with other core stem cell factors: EZH2, OCT4, and SOX2 (Figure 2B and Table S2). Finally Moonlight analysis of the oncogenic signatures from MSigDB further validated our gene expression based index and confirmed engagement of MYC, EZH2, along with E2F3, MTOR, SHH in driving oncogenic dedifferentiation (Figure 2C) (Colaprico et al., 2018).

DNA methylation-based stemness index—We defined the mDNAsi using OCLR by combining: 1) supervised classification between ESC/iPSC and their progenies; 2) stem cell signatures associated with pluripotency-specific genomic enhancer elements based on ChromHMM from Roadmap, and 3) ELMER, which uses DNA methylation to identify enhancer elements and correlates their state with the expression of nearby genes. 219 CpG probes (Figure S2A) were selected in training OCLR using the PCBC datasets. By selecting probes previously defined to be active stemness-specific enhancers, we confirmed the ability of our approach to derive an mDNAsi. Since we focused exclusively on hypomethylated, functionally important CpG probes associated with stem cells, we further explored cis-activated genes.

We scored each TCGA sample using the mDNAsi and used an external dataset to confirm that stem cells had higher mDNAsi than differentiated samples (Figure 1B, left plot). TCGA tumor types show different degrees of inferred dedifferentiated phenotype (Figure 1C, right). Within these, Individual tumor samples show variation for cancer stemness. As anticipated, TCGA samples derived from the primary tumors show higher cancer stemness indices compared to non-tumor samples obtained from adjacent normal tissue-of-origin (Figure S1E, bottom).

Most of our selected probes fell within non-promoter elements, yet the SOX2-OCT4 transcription factor binding motif is one of the most highly enriched signatures within these regions. The SOX2-OCT4 complex is a critical master regulator of pluripotency and stemness, and is highly enriched in tumor samples with high mDNAsi (Figure 2D).

Correlations of mRNAsi and mDNAsi—Since the inputs for mDNAsi and mRNAsi are not necessarily complementary, we explored stratification of glioma samples by the epigenetically regulated-mRNAsi (EREG-mRNAsi), a stemness index generated using a set of stemness-related epigenetically regulated genes. The EREG-mRNAsi, based on both RNA expression and epigenetics, elucidates the discrepancy between mDNAsi and mRNAsi and shows a positive correlation with both indices (Figure S1F). Both mRNAsi and mDNAsi show good correspondence for a majority of tumors (Figures S1F and S2B). We observed major discrepancies in the case of LGG, THCA, and THYM. For gliomas, mDNAsi is correlated positively with tumor pathology and clinical features, while mRNAsi shows a negative correlation. This result could arise from a high frequency of IDH1/2 mutations and resulting DNA hypermethylation.

Stemness index can stratify recognized undifferentiated cancers

We examined BRCA, AML, and gliomas to study if the mRNAsi/mDNAsi predict stemness in poorly differentiated tumors. In BRCA, we found a strong association between the stemness index and known clinical and molecular features (Figure 3A, left). The mRNAsi was highest in the basal subtype, known to exhibit an aggressive phenotype associated with an undifferentiated state. BRCA samples with high mRNAsi were more likely to be ER-negative, and enriched for *FAT3* and *TP53* mutations. We noted that high mRNAsi was associated with higher protein expression of FOXM1, CYCLINB1 and MSH6 as well as higher microRNA-200 family expression (Figure 3A, right). Invasive lobular type of BRCA (ILC), characterized by better prognosis in comparison to invasive ductal carcinoma (IDC), has a lower mRNAsi (Figure 3A, right). We also applied our indices to non-TCGA BRCA samples (Reyngold et al., 2014), and found a similar correlation between mRNAsi and mDNAsi in those samples. Moreover, mRNAsi also stratified BRCA samples with distinct histology in this dataset (Figure S1B). Using datasets with estimated tumor cell type composition provided by the epigenetic deconvolution method (Onuchic et al., 2016) we found that both mRNAsi and mDNAsi were more highly correlated with malignant epithelial cells than with normal epithelial cells suggesting that our indices identify distinct cancerous epithelial cell populations characterised by different features or degrees of stemness (Figure S1D).

We found an association between the mRNAsi, RNA expression subtypes previously defined by TCGA, and the French-American-British (FAB) classification of AML (Figure 3B). The mRNAsi showed the strongest correlation with the stage of myeloid differentiation of the AML samples. FAB subtypes M0 (undifferentiated), M1 (with minimal maturation), and M2 (with maturation) were characterized by high mRNAsi. In contrast, M3 well-matured promyelocytic subtype, which is associated with benign chromosomal abnormalities and favorable clinical outcome had low mRNAsi (Figure 3B, right upper). High mRNAsi was associated with higher expression of miR-181c-3p, miR-22-3p, and miR-30b-3p (Figure 3B, right bottom).

We found a strong association between high mDNAsi, high pathologic grade and recently published molecular subtypes of glioma (Figure 3C). mDNAsi was low in less aggressive gliomas that are characterized by codel and G-CIMP-high features and was highest in highly aggressive GBMs characterized by IDH mutations (G-CIMP-low) and poor clinical outcome. Also, high mDNAsi is strongly associated with more aggressive classical and mesenchymal subtypes of GBM, suggesting that it can stratify tumors with distinct clinical outcomes. We also found that high mDNAsi was associated with mutations in *NF1* and *EGFR* and infrequent mutations in *IDH1, TP53, CIC,* and *ATRX* (Figure 3C, left), with higher expression of ANNEXIN-A1 protein and lower expression of ANNEXIN-A7, and with expression of the miR-200 family (Figure 3C, right bottom).

We obtained similar results on non-TCGA glioma samples for which both mRNA expression and DNA methylation data were available (Turcan et al., 2012) (Figure S1C). The negative correlation between mDNAsi and mRNAsi was restricted to LGG samples, specifically the IDH mutant subtypes (G-CIMP high and codel). IDH1 mutations are known to reduce cell differentiation, and high values of the mRNAsi in a subset of IDH mutant gliomas might capture this phenomenon (Lu et al., 2012).

Pan-cancer stemness landscape

Next, we tested the ability of our indices to identify previously unexplored features of cancer stemness across all TCGA tumors. First, we performed an enrichment analysis by sorting all TCGA samples by stemness index for each tumor type and looking for associations with mutations, molecular and clinical features. The most salient associations of mRNAsi and mDNAsi are presented in Figure 4, while the following results of the comprehensive analyses are shown in the supplementary material: associations with mutations (Figure S3), associations with miRNA expression and protein abundance (Figure S4), associations with the tumor grading and clinical outcome (Figure S5).

Correlations of mRNAsi and mDNAsi with mutations in genes, miRNA and expression of proteins—We found a strong association between mDNAsi and known molecular subtypes, somatic mutations in *SETD2* and *TP53* genes, and with tobacco smoking status in LUAD (Figures 4A and S3). Current smokers and recently reformed smokers have higher mDNAsi than non-smokers or long-term reformed smokers. This suggests that the stemness of LUAD tumors might be activated in response to environmental stimuli such as smoking, and might influence te aggressiveness of the tumor. We also found an association between mDNAsi and higher protein expression of CYCLINB1 and FOXM1, which is a pro-stemness transcription factor upstream of CYCLINB1 (Figure 4A, lower plots). FOXM1 has been associated with dedifferentiation in pancreatic cancer cells (Bao et al., 2011) as well as tumor proliferation in the kidney (Xue et al., 2012) and ovarian (Wen et al., 2014) cancers. Our result suggests that it could be a driver of dedifferentiation and proliferation in breast and lung cancers. Stemness of LUAD tumors is also associated with lower expression of ANNEXIN-A1 (Figure 4A). ANNEXIN-A1 has been indicated as a differentiation marker in pancreatic (Bai et al., 2004) and urothelial cancers (Kang et al., 2012), therefore we suspect that the relationship between ANNEXIN and FOXM1 expression and tumor differentiation may extend to other tumor types (Figure S4C).

Analyses of HNSC samples revealed that high indices are correlated with *NSD1* mutation, E-cadherin protein expression, miR-200-3p, and previously identified classical molecular subtypes (Figure 4B). *NSD1* mutation was recently linked in HNSC tumors to blockade of cellular differentiation and promotion of oncogenesis (Papillon-Cavanagh et al., 2017). Interestingly, miR-200 family members have been implicated in cancer initiation and metastasis, as well as self-renewal of healthy stem cells (Gregory et al., 2008; Tellez et al., 2011). HNSC tumors with high mDNAsi have reduced programmed death ligand 1 (PD-L1) protein level (Figure 4B).

In LIHC samples, we found an association between mRNAsi and high pathological grade (Figure 4C). Negative associations between mRNAsi and the probability of OS or PFS were detected (Figures 4E and S5C). In contrast to the majority of tumor types, LIHC samples with high mRNAsi have low expression of miR-200 family members (Figure 4C). The miR-200 family is known to be associated with progression of hepatocellular carcinoma (Tsai et al., 2017; Wong et al., 2015), and the miR-200b-ZEB1 circuit has been suggested as a master regulator of stemness in these cancers (Tsai et al., 2017). We found associations of mRNAsi with higher CYCLINB1 and ACC1 and with lower PD-L1 and ANNEXIN A1 protein expression in LIHC (Figure 4C). ACC1 was associated with pathomorphological markers of LIHC aggressiveness (vascular invasion and poor differentiation) and its upregulation was correlated with poor OS and disease recurrence in hepatocellular carcinoma patients (Wang et al., 2016). LIHC samples with high mRNAsi were associated with specific genomic alterations (e.g., *TP53, CTNNB1, AXIN1*, among others).

Detailed analyses of ACC samples revealed an association between high mRNAsi and defined molecular subtypes (Zheng et al., 2016), clinical stage, and mutations in *PRKAR1A* and *TP53* genes (Figure 4D). We found a positive correlation between mRNAsi and adrenal differentiation score, that is based on expression of 25 genes that are important for adrenal function (Zheng et al., 2016) (Figure 4D).

Stemness indices are correlated with tumor pathology and predictive of clinical outcome—We observed a positive correlation between tumor histology and pathology grading and both stemness indices for the majority of the TCGA cases (Figures S5A, S5B, and Figures 3A, 3C, 4C, 4D, and S1B). For mRNAsi, the most significant correlations were found for BRCA (IDC and ILC), CESC, LIHC, PAAD, UCEC (Figure S5A). Interestingly, mRNAsi shows low values in GBM and STAD. On the other hand, mDNAsi strongly stratifies glioma by the pathology grade culminating with the highest value for GBM (Figure S5B). The reversed values of mDNAsi and mRNAsi in case of gliomas were also evident in the clinical data analyses. An adverse association between the mRNAsi and survival was detected (Figure 4E), which was significant for OS and PFS after adjusting for clinical factors (Figures S5C). In contrast, the mDNAsi had no significant association with OS and PFS after correcting for clinical factors. We found a positive correlation between previously published glioma subtypes and mDNAsi suggesting that mDNAsi might recapitulate prognostic molecular subtypes (Figure 3C). The discordance between the mRNAsi and the mDNAsi for gliomas may be explained in part by the dominant genomic alteration associated with the LGG tumor type. Roughly 80% of LGG tumors carry an *IDH1/2* mutation and, as demonstrated by our group and others, tconfers a genome-wide hypermethylator phenotype (G-CIMP) (Noushmehr et al., 2010; Turcan et al., 2012). Given that the mDNAsi is driven primarily by low methylation levels associated with the stemness phenotype, the LGG tumors might resemble non-stem like phenotypes, which are predominantly hypermethylated. The subgroup of G-CIMP with the lowest overall DNA methylation levels (G-CIMP-low) is associated with the worst outcomes. Compared to G-CIMP-high tumors, G-CIMP-low tumors are known to be more proliferative, express cell-cycle-related genes, and have various stem cell-like genomic features (Ceccarelli et al., 2016).

Cancer stemness indices are higher in tumor metastases and reveal intratumor heterogeneity

The TCGA samples are derived mostly from primary tumors except for skin melanoma for which tissues are mostly metastatic lymph nodes. We used the mRNAsi to interrogate the MET500 dataset comprising expression profiles from 500 metastatic samples obtained from 22 different organs (Robinson et al., 2017). In most cases, mRNAsi was significantly higher in e metastatic samples compared to primary TCGA tumors (Figure 5A). Prostate and pancreatic adenocarcinomas metastases had the most dedifferentiated phenotypes, and are also more aggressive and resistant to therapies in contrast to primary tumors. Weaker association with the mRNAsi was due to a small number of available samples (n<20). Interestingly, TGCT presents the less differentiated phenotype in primary tumors when compared to distant metastases. Primary TGCT tuor cells have high mRNAsi and may differentiate when metastasizing to distant organs. A similar trend was observed for STAD.

Using another dataset, we found that mDNAsi was significantly higher in glioma samples obtained at first recurrence in contrast to primary gliomas (Figure 5B). Our results reveal significant dedifferentiation of glioma cancer cells that contribute to glioma recurrence which is frequently associated with poor prognosis and resistance to treatment (de Souza et al., 2017).

By taking advantage of single-cell transcriptome datasets, we used mRNAsi to probe tumor heterogeneity for oncogenic dedifferentiation of individual cancer cells (Chung et al., 2017; Tirosh et al., 2016). We revealed high variation of stemness in the glioma and breast primary tumors. Individual glioma cells showed higher variegation of oncogenic dedifferentiation in comparison to breast cancer cells (Figure 5C). Single cells from metastases had higher stemness index in breast cancer (Figure 5D). Interestingly, the negative correlation of EMT signature and stemness that we observed in TCGA primary tumors was also found in metastatic samples (Figure 5E).

Stemness index evaluated in the context of immune response

We found that, for many tumors, higher stemness indices are associated with a reduced leukocyte fraction and lower PD-L1 expression (Figure 6A). For mDNAsi, the most distinctive negative correlations were found in the PanCan-12 squamous cluster (LUSC, HNSC, BLCA) (Hoadley et al., 2014) and in GBM (Figures 6A [left panel] and S6B). For the mRNAsi, the highest negative correlation values were seen in GBM/LGG, prostate adenocarcinoma (PRAD), LICH, and UCS tumors (Figures 6A [right panel] and S6A). We expect that such tumors will be less susceptible to immune checkpoint blockade treatments, due to insufficient immune cell infiltration or e pre-existing downregulation of the PD-L1 pathway, which makes further inhibition ineffective. Our findings are consistent with previous reports showing a strong correlation between PD-L1 protein expression and infiltration of CD8+ cytotoxic lymphocytes (Zaretsky et al., 2016).

We further explored correlations between stemness and immune microenvironment variables in the context of molecular subtypes of tumors. Figure 6B highlights several tumor types with the strongest (positive or negative) correlations. Except for KIRC, the association between stemness and PD-L1 expression and leukocyte fraction is readily apparent from the increasing and decreasing trends of individual variables across the molecular subtypes. For example, we found mesenchymal tumors to have the highest PD-L1 expression levels, the most significant leukocyte fractions, and lowest mDNAsi compared to other HNSC subtypes, suggesting potential susceptibility to checkpoint blockade inhibitors. The use of immunotherapy for HNSC tumors is under active investigation (Economopoulou et al., 2016; Fuereder, 2016), with the recent FDA approval of pembrolizumab; however, whether the effectiveness of therapy is limited to specific HNSC molecular subtypes is not clear from those reports.

To assess other relationships between stemness and tumor microenvironment, we computed correlations between stemness indices and individual types of immune cells. By applying CIBERSORT, we scored 22 immune cell types for their relative abundance in TCGA tumor samples. These cell types included NK cells, monocytes, macrophages, dendritic and mast cells, eosinophils, and neutrophils. We also obtained absolute estimates by scaling their relative abundance by overall leukocyte infiltration in each tumor, as determined by ESTIMATE applied to DNA methylation data. For any given TCGA sample, we calculated the correlation between mDNAsi/mRNAsi and the estimated fraction of individual immune cell types. In addition to individual immune subpopulation fractions, we considered the functional activation of distinct cells by measuring the difference between activated and resting fractions of NK cells, CD4+ T cells, and macrophages. This approach was motivated by recent observations that activation of peripheral CD4+ T cells triggered by immunotherapy is responsible for the specific killing of tumor cells (Spitzer et al., 2017).

Although the squamous cluster tumors had a negative correlation between stemness and the fraction of CD4+ T cell populations, the activation state of the CD4+ T cells was higher in dedifferentiated tumors. This finding is consistent with our observation that PD-L1 protein expression is lower in these tumors, suggesting again that immune checkpoint blockade might be ineffective and an additional mechanism of immune evasion may be operative. The opposite trend is present in thymomas, where PD-L1 protein expression and the fraction of the CD4+ T cell population are positively correlated with tumor dedifferentiation. Likewise, the activation state of CD4+ T cells is lower in dedifferentiated tumors, suggesting that they might be more susceptible to immunotherapy treatments (Figures S6AB).

Connectivity map (CMap) analysis identifies potential compounds/inhibitors capable of targeting the stemness signature

We employed CMap, a data-driven, systematic approach for discovering associations among genes, chemicals, and biological conditions, to search for candidate compounds that might target pathways associated with stemness. We found enrichment for compounds associated with stemness in at least three cancer types Figure 7A. Five compounds are significantly enriched in more than ten cancer types and have been reported to inhibit stemness-related tumorigenicity: the dopamine receptor antagonists thioridazine and prochlorperazine (Cheng et al., 2015; Lu et al., 2015, (Dolma et al., 2016)), the WNT signaling inhibitor pyrvinium (Xu et al., 2016), the HSP90 inhibitor tanespimycin and the protein synthesis inhibitor puromycin. Further, telomerase inhibitor gossypol induced apoptosis and growth inhibition of CSCs (Volate et al., 2010), and histone deacetylase inhibitors such as trichostatin A (SAHA) reduced glioblastoma stem cell growth (Chiao et al., 2013). According to our analysis, pyrvinium and puromycin could inhibit stemness in LUAD. We found several candidates with recognized anti-CSC activity for HNSCC, including the aforementioned compounds. For LIHC, thioridazine, a prospective inhibitor of lung cancer stem cells (Yue et al., 2016), pyrvinium, puromycin, prochlorperazine, and others are potential compounds targeting undifferentiated tumors (Figure 7).

CMap Mode-of-action (MOA) analysis of the 74 compounds revealed 56 mechanisms of action shared by the above compounds (Figure 7B and Table S4B). Five compounds (fluspirilene, pimozide, prochlorperazine, thioridazine, and trifluoperazine) shared the MoA of Dopamine receptor antagonist. We observed that entinostat, trichostatin-a, vorinostat shared MoA as HDAC inhibitors, and LY-294002, zaprinast, zardaverine as Phosphodiesterase inhibitors.

CMap Target analysis revealed 212 distinct drug-target genes shared by the mentioned compounds (Figure S7 and Table S4C). Eight genes are targets of five different compounds, namely DRD2 (8 drugs), HTR2A (7 drugs), HRH1 (6), ADRA1A (5), CALM1 (5), CHRM3 (5), HTR1A (5), HTR2C (5).

Recent polypharmacology studies suggest the need to design compounds that act on multiple genes or molecular pathways. In this study, we observed similar mechanisms of action among different compounds suggesting selective therapies can target the undifferentiated phenotypes for selected cancer types.

DISCUSSION

This study is based on integrated analysis of cancer stemness in almost 12,000 primary human tumors of 33 different cancer types. We interrogated TCGA data for mutations, DNA methylation, expression of mRNA and miRNA, expression and post-translational modification of proteins, histopathological grade, and clinical outcome. Applying CIBERSORT, we gained insight into the tumor microenvironment and composition of immune cell infiltrates. By applying a machine-learning algorithm to molecular datasets from normal stem cells and their progeny, we developed two different molecular metrics of stemness and then used them to assess epigenomic and transcriptomic features of TCGA cancers according to their grade of oncogenic dedifferentiation. Ultimately, the analyses led us to potentially actionable targets (and their modes-of-action), as candidates for possible differentiation therapy of solid tumors and metastases. Our approach could be applied to longitudinal study of samples from primary, recurrent, and metastatic cancers and gene expression signatures identified in the tumor samples can be used to interrogate CMap to suggest actionable targets and inhibitors for further analysis.

To the best of our knowledge, this is the first study in which molecular PCBC datasets comprised of stem cells and defined populations of their differentiated progeny have been leveraged to develop a classification tool and machine-learning algorithm for analysis of a spectrum of human malignancies. A number of cancer stemness scores, based on genes that are differentially expressed between CSCs and non-CSCs, have been published and are relevant to clinical outcomes in AML (Eppert et al., 2011; Gentles et al., 2010; Ng et al., 2016). In those studies, gene sets enriched in ESCs (e.g., targets of NANOG, OCT4, SOX2, and c-MYC) were frequently overexpressed in poorly differentiated tumors compared with well-differentiated ones. In breast cancers, those gene sets were associated with high-grade estrogen receptor-negative, basal-like tumors and poor clinical outcome (Ben-Porath et al., 2008). Another web-based tool, StemChecker, uses a curated set of 49 published *stemness signatures* defined by gene expression, RNAi screens, transcription factor binding sites, text-mining of the literature, and other computational approaches. But it has been tested only for pancreatic ductal adenocarcinoma. In that case, high expression of stemness genes correlated with poor prognosis (Pinto et al., 2015). All previous studies were transcriptome-based and limited to a narrow set of genes and a small number of tumor types.

In the present study, we found oncogenic dedifferentiation to be associated with several characteristics: mutations in genes that encode oncogenes and epigenetic modifiers, perturbations in specific mRNA/miRNA transcriptional networks, and deregulation of signaling pathways. Cancer stemness also appeared to involve core expression of Myc, Oct4, Sox2, and other genes involved in the regulatory circuitry that underlies normal and malignant self-renewal potential. Our indices derived from mRNA expression and DNA methylation signatures reliably stratified tumors of known stemness phenotype. High mRNAsi was associated with basal breast carcinomas but also Her2 and lumB subtypes that are more aggressive than the hormone-dependent lumA group. In contrast, high mDNAsi was strongly associated with high-grade glioblastomas, poor overall and progression-free survival. The association between stemness signatures and adverse outcome for some tumor types, including gliomas, may reflect malignant cell origins or the impact of their microenvironment.

Dedifferentiated cells can arise from different sources: from long-lived stem or progenitor cells that accumulate mutations in oncogenic pathways, or via dedifferentiation from non-stem cancer cells that convert to CSCs through deregulation of developmental and/or non-developmental pathways. It is important to distinguish between the inherent stemness of CSCs and dedifferentiation induced by the tumor microenvironment. However, addressing that issue would require further validation beyond the scope of this study using other genomic datasets and/or laboratory experiments.

Both stemness indices were lowest in normal cells, increased in primary tumors, and highest in metastases, consistent with the idea that tumor progression generally involves oncogenic dedifferentiation. Interestingly, we observed negative associations between stemness and EMT gene signatures. The relationship between EMT and stemness remains a hotly debated topic, with several studies showing that EMT is necessarily associated with stemness (Fabregat et al., 2016). However, most TCGA data are obtained from primary tumors, which exist in a pre-EMT state, since EMT is strongly associated with tumor progression and with metastasis for many tumor types. Cancer cells in many primary solid tumors are basically epithelial regardless of their degrees of dedifferentiation, but some cells in such contexts could acquire mesenchymal characteristics, either by accumulating additional mutations or by undergoing epigenetic changes shaped by the tumor microenvironment. Those mesenchymal cells can traverse the underlying tissue, enter the bloodstream and seed distant organs where they reacquire an epithelial phenotype to form metastatic tumors.

We observed epithelial phenotype and increased stemness index in molecular profiles of tumor type-matched metastatic samples in the MET500 cohort. This portends an association between dedifferentiation and spread of tumor cells to distant organs. The observation is further supported by high mDNAsi in samples from recurrent gliomas. It appears that tumor growth *de novo*, or at recurrence/metastasis, is associated with an increased stemness phenotype. Decreased mRNAsi levels seen in TGCT suggest its possible differentiation as a germ cell tumor type induced by the microenvironment of liver or lung parenchyma, the organs it most often colonises. Clinically, in general, tumor progression is associated with greater aggressiveness and resistance to therapy of almost all types.

The mRNAsi was high for individual primary glioma and breast cancer cells. Interestingly, when applied to transcriptomic profiles obtained from analysis of single cancer cells in bulk tumors, stemness indices revealed a high degree of intratumor heterogeneity with respect to dedifferentiation phenotype. The heterogeneity was greater in gliomas than in breast cancer, suggesting that intratumor environment, including stromal cells, hypoxia, and infiltration of immune cells, may play a role in shaping CSC niches, and affect cancer cell developmental plasticity. Further molecular analyses of cancer cells stratified by the stemness phenotype would provide novel insights into the biology of primary tumors.

We found that, for a number of tumor types (GBM, LUSC, HNSC, and BLCA), higher mDNAsi was associated with reduced leukocyte fraction and/or lower PD-L1 expression.

Such tumors are expected to be less susceptible to immune checkpoint blockade, due either to insufficient immune cell infiltration of tumors or to inherent downregulation of the PD-L1 pathway. Both factors can render immune checkpoint immunotherapy ineffective. The interaction between PD-L1 on cancer cells and PD1 receptor on T-cells helps cancer cells elude the immune system by preventing activation of cytotoxic T cells in lymph nodes and subsequent recruitment of other immune cell types to the tumor site (Chen and Mellman, 2013). The presence of tumor-infiltrating lymphocytes and/or PD-L1 expression correlates with aggressiveness in gastrointestinal stromal tumors (Bertucci et al., 2015) and breast carcinomas (Polónia et al., 2017).

Common features shared between cancer cells and stem cells in the context of the immune response are being highlighted by a growing number of studies showing that vaccination with ESC or iPSC can raise specific immune response against cancer cells (Kooreman et al., 2018). That finding may indicate that both cell populations use protein networks that, in tumors, result in uncontrolled self-renewal and de-differentiated phenotypes histopathologically defined by loss of architecture specific to the tissue of origin. We speculate that the indices described here may help predict the efficacy of stem-cell based immunotherapies and contribute to the identification of patients who will respond to such therapies.

We interrogated CMap using the gene expression signatures from tumor samples with the highest and lowest mRNAsi levels. Surprisingly perhaps, the Cmap analysis, which is based on only a limited number of treated cell lines, very precisely selected drugs that have been shown to affect cancer stem cells with specificity. These translational analyses may ultimately pave the way for implementation of differentiation therapies for solid tumors.

Here, we have also shown that cancer hallmarks can be extracted from datasets on cells with defined phenotypes and used to train machine-learning methods applicable to index molecular profiles of cancer. Our mRNAsi and mDNAsi can be translated into stemness scores (e.g., STEM50) that stratify tumors based on their dedifferentiation features, thus providing biomarkers for prediction of patient outcomes and response to to differentiation therapies.

By defining new metrics of cancer stemness and using them to interrogate TCGA datasets, our results provide a comprehensive characterization of dedifferentiation as new and significant hallmarks of cancer. The strengths of the approach are that it leverages features of dedifferentiated cells across a spectrum of tumor types that reflect tumor pathology and, in some cases, clinical outcome. This study also provides strategies for integrated analysis of cancer genomics based on machine-learning methods trained on molecular profiles obtained from cells with defined phenotypes. The findings based on those methods may advance the development of objective diagnostics tools for quantitating cancer stemness in clinical tumors, perhaps leading eventually to new biomarkers that predict tumor recurrence, guide treatment selection, or improve responses to therapy.

STAR★Methods

Detailed methods are provided in the online version of this paper and include the following:

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Workflow to reproduce the stemness index	This paper	https://bioinformaticsfmrp.github.io/PanCanStem_Web/
CIBERSORT	(Gentles et al., 2015)	https://precog.stanford.edu/
R 3.3.1	(R Core Team, 2017)	https://www.R-project.org
ggplot2 (v2.2.1)	(Wickham, 2009)	https://CRAN.R-project.org/package=ggplot2
gelnet (v1.2.1)	(Sokolov et al., 2016)	https://CRAN.R-project.org/package=gelnet
GSEA	(Subramanian et al., 2005)	https://software.broadinstitute.org/gsea/index.jsp
TCGAbiolinks (v2.4.3)	(Colaprico et al., 2016)	http://bioconductor.org/packages/TCGAbiolinks/
ELMER (v1.4.1)	(Yao et al., 2015)	http://bioconductor.org/packages/ELMER/
fgsea (v1.2.1)	(Sergushichev, 2016)	http://bioconductor.org/packages/fgsea/
Methylumi (v2.20.0)	(Davis et al., 2015)	http://bioconductor.org/packages/methylumi/
MoonlightR (v1.2.0)	(Colaprico et al., 2018)	http://bioconductor.org/packages/MoonlightR/
Amaretto	(Gevaert et al., 2013)	http://med.stanford.edu/gevaertlab/software.html
STATA (v13)	(StataCorp, 2013)	http://www.stata.com/
Deposited Data		
TCGA data	NIH Genomic Data Commons (GDC)	https://gdc.cancer.gov/about-data/publications/PanCanStemness-2018
Progenitor Cell Biology Consortium (PCBC)	(Daily et al., 2017; Salomonis et al., 2016)	https://www.synapse.org/pcbc
Chromatin State (ChromHMM)	(Roadmap Epigenomics Consortium et al., 2015)	http://www.roadmapepigenomics.org
Stem Cells Validation set	(Nazor et al., 2012)	GEO: mRNA expression (GSE30652) and DNA methylation (GSE30654)
Glioma validation set	(Sturm et al., 2012)	GEO: mRNA expression (GSE36245) and DNA methylation (GSE36278)
Glioma validation set	(Turcan et al., 2012)	GEO: GSE30339
BRCA validation set	(Reyngold et al., 2014)	GEO: GSE59000
Deconvolution of breast cancer (BRCA)	(Onuchic et al., 2016)	http://genboree.org/theCommons/projects/edec
MET500 - Metastatic solid tumors	(Robinson et al., 2017)	Database of Genotypes and Phenotypes (dbGaP) accession number phs000673.v2.p1
Gliomas Single Cell RNA expression	(Tirosh et al., 2016)	GEO: GSE70630

REAGENT or RESOURCE	SOURCE	IDENTIFIER
BRCA Single Cell RNA expression	(Chung et al., 2017)	GEO: GSE75688

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by Maciej Wiznerowicz (maciej.wiznerowicz@iimo.pl).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Clinical and molecular data were collected from the NIH Genomic Data Commons (GDC) of 11,392 participants from The Cancer Genome Atlas PanCancer Atlas cohort (https://gdc.cancer.gov/about-data/publications/PanCanStemness-2018).

METHODS DETAILS

DNA Methylation Data—A total of 9,627 PanCancer TCGA samples across 33 different tumor types were available for DNA methylation using the robust Illumina HumanMethylation 450 (HM450) platform. TCGA samples included primary (8,471), recurrent (41), and metastatic tumor (394) tissues and a set of 721 non-tumor tissues.

Level 3 data were downloaded from TCGA Data Portal using TCGAbiolinks functions GDCquery, GDCdownload and GDCprepare importing into R (http://www.r-project.org) for further analysis (Colaprico et al., 2016).

DNA methylation level 3 data are β-values that were calculated from pre-processed raw data using the methylumi Bioconductor package (Davis et al., 2015). Pre-processing steps included background correction, dye-bias normalization, and calculation of β-values and detection p-values. β-values range from zero to one, with zero indicating no DNA methylation and one indicating complete DNA methylation. A detection p-value compares the signal intensity difference between the analytical probes and a set of negative control probes on the array. Any data point with a corresponding p-value greater than 0.01 is deemed not statistically significantly different from background and is thus masked as "NA" in TCGA level 3 data. The data levels and the files contained in each data level package are on the NIH Genomic Data Commons (GDC).

In addition to TCGA data, we used a dataset of 99 human stem/progenitor cells from the Progenitor Cell Biology Consortium (PCBC) (https://www.synapse.org/pcbc) to define stem cell signatures (Daily et al., 2017; Salomonis et al., 2016). PCBC samples were profiled using the Illumina HumanMethylation 450 (HM450) platform and consisted of 4 embryonic stem cells (ESC), 40 induced pluripotent stem cells (iPSC), 22 stem cell (SC)-derived embryoid bodies (EB), 11 SC-derived mesoderm day 5 (MESO), 11 SC-derived ectoderm (ECTO), and 11 SC-derived definitive endoderm (DE). We downloaded raw IDAT files from PCBC Genomic Data Commons and processed the data according to the TCGA standard level 3 protocol described above.

RNA Expression Data—PanCancer TCGA RNA sequence level 3 normalized data were downloaded from the GDC Data Portal using TCGAbiolinks functions GDCquery, GDCdownload and GDCprepare importing into R (http://www.r-project.org) for further analysis (Colaprico et al., 2016). A total of 10,852 samples across 33 tumor types were available, including primary (9,702), recurrent (45) and metastatic tumor (395) tissues and a set of 710 non-tumor tissues.

We also downloaded PCBC RNA sequence data from the PCBC Synapse Portal (https://www.synapse.org/pcbc), consisting 16 ESC, 77 iPSC, 66 SC-derived EB, 29 SC-derived MESO, 29 SC-derived ECTO, and 36 SC-derived DE (Daily et al., 2017; Salomonis et al., 2016).

Stemness Index Derived Using OCLR—To calculate a stemness index (si) based on mRNA expression or DNA methylation, we built a predictive model using one-class logistic regression (OCLR) (Sokolov et al., 2016) on the pluripotent stem cell samples (ESC and iPSC) from the PCBC dataset (Daily et al., 2017; Salomonis et al., 2016).

For mRNA expression-based signatures, to ensure compatibility with the TCGA PanCancer cohort, we first mapped the gene names from Ensembl IDs to Human Genome Organisation (HUGO), dropping any genes that had no such mapping. The resulting training matrix contained 12,945 mRNA expression values measured across all available PCBC samples. For DNA methylation-based signatures, we used each of the signatures (probe set) described below.

We mean-centered the data, then applied OCLR to just the samples labeled SC (which included both ESC and iPSC). We chose to use the one-class framework because of its robustness in the absence of the a “negative” class. The PCBC data does not have data for fully differentiated cells, and progenitor cell types might exhibit some of the stemness signals.

Once the signature is obtained, it can be applied to score new samples. For RNA expression data, we computed Spearman correlations between the model’s weight vector and the new sample’s expression profile. We advocate for the use of Spearman correlation over the more traditional dot product operation because it is more robust with respect to potential cross-dataset batch effects that may arise. For DNA methylation data, which follow the beta distribution, the samples were scored using the standard application of a linear model: f(x) = w^T x + b.

We validated our approach using leave-one-out cross-validation by withholding each SC sample in turn. A separate signature was then trained on all other SC samples and used to score the withheld sample as well as all the non-SC samples. The performance was measured using the area under the curve (AUC) metric, which can be interpreted as the probability that the model correctly ranks a positive sample above a negative (Agarwal et al., 2005). In our cross-validation experiment, every withheld SC sample was scored higher than all the non-SC samples, yielding an overall AUC of 1.0.

We performed additional validation of the stemness signature by applying it to an external dataset composed of pluripotent stem cells (ESC and iPSC), somatic cells (17 distinct tissue types and several primary cell lines of diverse origin), and hydatidiform mole samples (Nazor et al., 2012). The mRNA expression data for the study were downloaded from GEO (GSE30652) as were DNA methylation data (GSE30654). We observed that all of the SC samples were correctly scored above all of the somatic samples by both platforms (Figure 1B). This is particularly striking for mRNA expression, because mRNA expression in study by the Nazor et al. was measured using microarrays, whereas the signature was trained using RNA-seq data.

Having validated the signature by using cross-validation and external SC data, we then applied it to score the TCGA PanCancer cohort using the same Spearman correlation (RNA expression) or linear model (DNA methylation) operators. The indices were subsequently mapped to the [0,1] range by using a linear transformation that subtracted the minimum and divided by the maximum. The mapping was done to assist with interpretation as well as integration with the stemness indices derived from other data platforms (*i.e.,* DNA methylation and mRNA expression).

Additionally, we downloaded independent, non-TCGA datasets of gliomas [(Sturm et al., 2012) (GSE36245, GSE36278) and (Turcan et al., 2012) (GSE30339)] and BRCA samples (Reyngold et al., 2014) (GSE59000) and applied our metrics to measure the stemness in the validation data. For mRNA expression, the preprocessing consisted of mapping the Illumina probe IDs (Illumina HumanHT-12 V3.0 platform) to HUGO symbols, and then reducing the signature and the external dataset to a common set of genes. We then computed the Spearman correlation between the signature and the external samples. For DNA methylation, we applied the linear model.

DNA Methylation Stemness Signatures—Due to the magnitude of the available DNA methylation platform Infinium HumanMethylation450 (HM450), we defined DNA methylation-based stemness signatures as a reduced input to the OCLR machine learning algorithm. For the DNA methylation-based stemness indices, three signatures were utilized, each defining a distinct, biologically relevant, molecular phenotype of stemness. First, we performed a supervised analysis between human pluripotent stem cells (ESC and iPSC) and stem cell-derived progenitors (embryoid bodies [EB], mesoderm [MESO], ectoderm [ECTO], and definitive endoderm [DE]) (β value mean difference < −0.4 and false discovery rate [FDR] < 10e-22; β value mean difference > 0.3 and false discovery rate [FDR] < 10e-17).All ‘rs’ and ‘ch’ probes were removed prior to analyses. To eliminate somatic tissue-specific probes, we removed probes that were consistently methylated (standard deviation β value > 0.05) in non-tumor adult tissues available through TCGA. This resulted in a set of 62 pluripotent cell-specific and differentially methylated regions, which was then used as input for the OCLR to determine the stemness index for each TCGA tumor sample, named “differentially methylated probes-based stemness index” (DMPsi). Interestingly, most of these probes (85%) were positioned within intergenic regions known as open seas (Figure S2A).

Second, we defined a stem cell signature associated with genomic enhancer elements. Enhancers have been shown to be a critically relevant functional element for defining gene target expression and chromatin organization. For this, we downloaded Chromatin State data (ChromHMM) from the NIH Roadmap Epigenomics Consortium (http://www.roadmapepigenomics.org), which defined 18 chromatin states (based on 6 different histone marks: H3K4me3, H3K27ac, H3K4me1, H3K36me3, H3K9me3, and H3K27me3) across 98 different cell types (Roadmap Epigenomics Consortium et al., 2015). Briefly, by using ChromHMM data we mapped the HM450 probes to the chromatin states in each individual cell type; then we identified genomic regions corresponding to active enhancers that are specific to pluripotent stem cell states (ESC and iPSC), meaning that each region was defined as active enhancers (according to their states: 9-EnhA1 and 10-EnhA2 (Roadmap Epigenomics Consortium et al., 2015)) in all pluripotent stem cells (n=9) whereas not enhancer (enhancer in less than 25% of non-pluripotent stem cells (n= 89)) in non-pluripotent stem cells. We identified 82 DNA methylation probes of the HM450 platform that mapped to enhancer elements and considered them to be a DNA methylation-based pluripotent stem cell enhancer signature, which was then used as input for the OCLR to evaluate stemness signatures for TCGA samples, named “enhancer-based stemness index” (ENHsi) (Figure S2A).

Third, we applied ELMER (Enhancer Linking by Methylation/Expression Relationships), an R/Bioconductor package (Yao et al., 2015) that uses DNA methylation to identify enhancer elements and correlates enhancer state with expression of nearby genes to identify putative transcriptional targets. Using ELMER, we compared pluripotent stem cells (ESC and iPSC) to stem cell-derived progenitors (EB, MESO, ECTO, DE) from PCBC and identified 87 CpGs that were hypomethylated in the pluripotent state (ESC and iPSC) compared to stem cell-derived progenitors and that potentially regulate 103 genes. We confirmed the importance of these probe-gene pair targets by identifying that the SOX2-OCT4 transcription factor binding motif was among the most highly enriched signatures within these elements (+/−250 bp from the center). The SOX2-OCT4 complex is an important master regulator of pluripotency and stemness. We then derived a new set of signatures using the OCLR and defined TCGA samples’ stemness as “epigenetically regulated stemness indices” for each molecular feature (RNA expression-based Epigenetically regulated-mRNAsi [EREG-mRNAsi] and DNA methylation-based [EREG-mDNAsi]).

Because there was high concordance among the three DNA methylation-based indices (DMPsi, ENHsi, and EREG-mDNAsi) (not shown) and each contributes important and complementary biological relevance to stemness, we combined the three stemness signatures (total of 219 probes) and derived a comprehensive DNA methylation index, named **mDNAsi** (Figure S2A). The lists of probes and genes used to derive the stemenss indices are provided on the publication portal accompanying this publication (https://gdc.cancer.gov/about-data/publications/PanCanStemness-2018).

Stemness vs Molecular and Clinical Features—To evaluate the performance of our stemness indices across the entire TCGA cohort, we performed an enrichment analysis by sorting TCGA samples by stemness index for each tumor type and looked for associations with all available genomic features (by using comprehensive mutation data [MC3]), molecular features (previously published TCGA molecular subtypes available at TCGAbiolinks package (http://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html) (Colaprico et al., 2016; Silva et al., 2017), through the function “PanCancerAtlas_subtypes()”, which provides full access to the curated matrix used for this study), and clinical features (more than 10,000 features). We used the fgsea R/Bioconductor package to compute the enrichment scores (Sergushichev, 2016). Briefly, for each tumor type we ranked the TCGA samples according to their stemness index (from -low to -high stemness index) and tested if any particular genomic/molecular/clinical feature was associated with either -low or -high stemness index in a non-random behavior. We performed 10,000 permutations for each parameter analyzed to calculated our enrichment score. We then normalized the enrichment scores to mean enrichment of random samples of the same size (NES - normalized enrichment score). Tables containing all the results can be accessed at https://gdc.cancer.gov/about-data/publications/PanCanStemness-2018. In addition, an interactive portal with the results across all tumor samples/types vs. mDNAsi and mRNAsi can be accessed at https://bioinformaticsfmrp.github.io/PanCanStem_Web/where the user can search for any gene or molecular/clinical feature of interest.

Stemness versus Clinical Predictors—The associations between the three stemness indices and overall survival (OS) and progression free survival (PFS) in different tumors were evaluated in two stages. First, the proportional hazard (PH) model with the index as a single continuous covariate was used to test whether there was a statistically significant effect on OS or PFS. Given that, for each outcome, the effects of the three indices were tested for 33 cancer types. The significance level of the tests was adjusted for multiple testing to control the overall type I error probability at 5%. In the next stage, the cancer types for which at least one index showed a statistically significant association with either OS or PFS were analyzed in more detail by using a multivariable PH model that included relevant clinical factors. Moreover, the model included a functional form of the index obtained by using degree-2 fractional polynomials (Royston and Altman, 1994). The plausibility of the PH assumption was checked by using the test based on the scaled Schoenfeld residuals (Therneau and Grambsch, 2000). The analyses were conducted using STATA v13 software.

To select the clinical factors for inclusion in the PH model used in the second stage of the OS/PFS analysis for selected cancer types, a detailed analysis of the association between the stemness indices and demographic and clinical features (such as sex, age, race, stage, grade, etc.) was carried out by using linear models. mRNAsi and EREG-mRNAsi were analyzed on the original scale, while mDNAsi was transformed logarithmically to make its distribution more symmetric. The fit of the constructed models was assessed by using residual plots. The analyses were conducted using STATA v13 software.

The screening of the association between the stemness indices and OS (Figure 4E) by using univariable proportional hazard (PH) models indicated a statistically significant (using p values adjusted for multiple testing) effect of mRNAsi on OS for LGG ($p < 0.0001$) and STAD ($p = 0.005$) and on PFS for GBM ($p = 0.04$), LGG ($p < 0.0001$), LIHC ($p = 0.05$), STAD ($p = 0.04$), and UCEC ($p = 0.03$). For mDNAsi, an effect on OS was found for LGG ($p < 0.0001$) and on PFS for KIRP ($p = 0.04$) and LGG ($p < 0.0001$). Finally, for EREG-mRNAsi, a statistically significant effect on OS was found for ACC ($p = 0.005$), KIRC (p =

0.008), and LGG (p = 0.03), and on PFS for ACC (p = 0.03), LGG (p = 0.03), and UCEC (p = 0.04). In these selected cases, multivariable analyses were conducted (using STATA v13 software), which took into account the effect of clinical factors. The analyses confirmed (by using unadjusted p values) the effect of mRNAsi on OS for STAD (p = 0.0001) and for GBM/LGG (p = 0.002) and the effect on PFS for GBM/LGG (p = 0.008) and LIHC (p = 0.002). For mDNAsi, the effect on PFS in KIRP was confirmed (p = 0.0001), while for EREG-mRNAsi, the effect on PFS in UCEC was confirmed (p = 0.05). These confirmed results indicate that the indices have a potential role as novel, independent prognostic factors for the indicated tumor types.

Compounds Targeting with Cancer Stemness—To determine which target drugs might be useful against cancer stem cells, we used the Broad Institute's Connectivity Map build 02 (CM) (Lamb et al., 2006), a public online tool (https://portals.broadinstitute.org/cmap/) (with registration) that allows users to predict compounds that can activate or inhibit based on a gene expression signature.

To further investigate about mechanism of actions (MoA) and drug-target we performed specific analysis within Connectivity Map tools (https://clue.io/) (Subramanian et al., 2017).

Using Connectivity Map (Query) in May 2017 having data available from a collection of cell lines (MCF, PC3, HL60 and SKMEL5) and 164 compounds as small molecules perturbagens. We obtained 33 mRNA expression signatures (one for each cancer type) by applying a differential expression analysis to samples with high mRNAsi and low mRNAsi, using the function TCGAanalyze_DEA from the the R/Bioconductor package TCGAbiolinks version 2.5.9 (Colaprico et al., 2016), carrying edgeR pipeline. The table with differentially expressed genes is reported as Table S3. Due to a limitation of the Connectivity Map tool that matches gene symbol and HG-U133A probe set (eg 200800_s_at) GPL96 platform ID, we had to remove duplicate IDs after sorting by decreasing |logFC|. We selected the top 1000 genes (500 up regulated and 500 downregulated) where the number of differentially expressed genes was enough or considering the aggregation of up-regulated or down-regulated genes.

Connectivity MAP is a method similar to GSEA analysis and follows a 4 step approach: (i) looking for similarity between a query signature (diff.expr. genes) and expression profiles present in the dataset using pattern-matching strategy based on Kolmogorov-Smirnov test (ii) rank-ordering the list of genes according their diff.expr. relative to the control from the above expression profiles with significantly similarity (iii) comparison of each rank-ordered list with a query signature to specify when up-regulated query genes appear in the proximity of the top of the list or near the bottom ("positive connectivity") or vice versa ("negative connectivity") producing an Enrichment Score (ES) from −1 to 1. (iv) All instances in the database are then ranked according to their connectivity scores; those at the top are most strongly correlated to the query signature, and those at the bottom are most strongly anticorrelated.

For each cancer type we obtained two tables that applied the Connectivity Map's findings to stemness mRNA expression signatures, namely, "detailed results" and "permuted results".

We used the permuted results and filter (with $p < 0.05$), to identify an average of 74 compounds per tumor type that are predicted to repress or activate the stemness signature (Table S4A).

Connectivity Map (CMap) was recently updated (September 2017) (Subramanian et al., 2017), providing the end-users new functionalities and new graphical interface as web-server, previous registration (https://clue.io/) allowing easily the extraction of drug-interaction knowledge using as input a signature of genes or compounds.

The new interface (https://clue.io/), provided 7 different analysis (query, touchstone, proteomics query, command, data library, repurposing, morpheus).

In particular CMap Query it is a tool for perturbagens that give rise to similar (or opposing) expression signatures, for a technical limit, the CMap Query 2017 allows only to upload 150 genes max for up-regulated genes and 150 genes for down-regulated genes. For this reason we considered the results analysed in May 2017 using 500 genes for up-down regulated genes.

STATISTICAL ANALYSIS

R version 3.3.1 was used for all statistical analyses, unless specified otherwise. The statistical details of all experiments are reported in the figure legends and figures, including statistical analysis performed, statistical significance and exact n values.

To identify differentially methylated DNA methylation probes, we used the Wilcoxon test followed by multiple testing using the Benjamini-Hochberg (BH) method to estimate false discovery rate (Benjamini and Hochberg, 1995).

To identify proteins and microRNAs differentially expressed between tumors with low vs. high stemness index, we used a t-test followed by multiple testing using BH.

P values for the association between stemness index and continuous clinical data were also computed using a t-test followed by multiple testing using BH.

DATA AND SOFTWARE AVAILABILITY

All data are available on the NIH Genomic Data Commons (GDC), https://gdc.cancer.gov/about-data/publications/PanCanStemness-2018.

The workflow to reproduce the stemness index, including downloading PCBC and TCGA PanCan33 datasets, training a stemness signature, and applying it to score TCGA samples can be accessed at https://bioinformaticsfmrp.github.io/PanCanStem_Web/.

An interactive portal with the results for enrichment of molecular and clinical features and Stemness Indices across all tumor samples/types can be accessed at https://bioinformaticsfmrp.github.io/PanCanStem_Web/ where the user can search for any gene or molecular/clinical feature of interest.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Marcin Cieślik from Michigan Center for Translational Pathology at University of Michigan for providing the MTE500 dataset. This work was supported by the following grants: NIH grants: U54 HG003273, U54 HG003067, U54 HG003079, U24 CA143799, U24 CA143835, U24 CA143840, U24 CA143843, U24 CA143845, U24 CA143848, U24 CA143858, U24 CA143866, U24 CA143867, U24 CA143882, U24 CA143883, U24 CA144025, P30 CA016672; NCI grants: 5R01CA180778, 3U24CA143858, 1U24CA210990, 5U54HG006097; NIGMS grant 5R01GM109031; Henry Ford Cancer Institute's Early Career Investigator Award; Sao Paulo Research Foundation (FAPESP) grants: 2014/02245-3, 2014/08321-3, 2015/07925-5, 2016/01389-7, 2016/10436-9, 2016/06488-3, 2016/12329-5, 2016/01975-3, 2016/15485-8; institutional grants from Henry Ford Hospital; Spanish Institute of Health Carlos III: CP14/00229; Mary K. Chapman Foundation gift to JNW; CA143883; CPRIT# RP13039; the Michael & Susan Dell Foundation; CA016672; Polish Science Foundation Welcome grant 2010/3-3 to MW.

References

Agarwal, Graepel S., Herbrich, T., Har-Peled, R., Roth, S., Dan. Generalization Bounds for the Area Under the ROC Curve. The Journal of Machine Learning Research. 2005

Bai X-F, Ni X-G, Zhao P, Liu S-M, Wang H-X, Guo B, Zhou L-P, Liu F, Zhang J-S, Wang K, et al. Overexpression of annexin 1 in pancreatic cancer and its clinical significance. World J Gastroenterol. 2004; 10:1466–1470. [PubMed: 15133855]

Bao B, Wang Z, Ali S, Kong D, Banerjee S, Ahmad A, Li Y, Azmi AS, Miele L, Sarkar FH. Over-expression of FoxM1 leads to epithelial-mesenchymal transition and cancer stem cell phenotype in pancreatic cancer cells. J Cell Biochem. 2011; 112:2296–2306. [PubMed: 21503965]

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; 57:289–300.

Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, Weinberg RA. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet. 2008; 40:499–507. [PubMed: 18443585]

Bertucci F, Finetti P, Mamessier E, Pantaleo MA, Astolfi A, Ostrowski J, Birnbaum D. PDL1 expression is an independent prognostic factor in localized GIST. Oncoimmunology. 2015; 4:e1002729. [PubMed: 26155391]

Bradner JE, Hnisz D, Young RA. Transcriptional addiction in cancer. Cell. 2017; 168:629–643. [PubMed: 28187285]

Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, Morozova O, Newton Y, Radenbaugh A, Pagnotta SM, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell. 2016; 164:550–563. [PubMed: 26824661]

Chen DS, Mellman I. Oncology meets immunology: the cancer-immunity cycle. Immunity. 2013; 39:1–10. [PubMed: 23890059]

Chung W, Eum HH, Lee H-O, Lee K-M, Lee H-B, Kim K-T, Ryu HS, Kim S, Lee JE, Park YH, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun. 2017; 8:15081. [PubMed: 28474673]

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016; 44:e71. [PubMed: 26704973]

Colaprico A, Olsen C, Cava C, Terkelsen T, Silva TC, Olsen A, Cantini L, Bertoli G, Zinovyev A, Barillot E, et al. Moonlight: a tool for biological interpretation and driver genes discovery. BioRxiv. 2018:265322.

Daily K, Ho Sui SJ, Schriml LM, Dexheimer PJ, Salomonis N, Schroll R, Bush S, Keddache M, Mayhew C, Lotia S, et al. Molecular, phenotypic, and sample-associated data to describe pluripotent stem cell lines and derivatives. Sci Data. 2017; 4:170030. [PubMed: 28350385]

Davis, S., Bilke, S., Triche, T., Jr, Bootwalla, M. R Package Version 2.18.0. 2015. methylumi: Handle Illumina methylation data.

Dolma S, Selvadurai HJ, Lan X, Lee L, Kushida M, Voisin V, Whetstone H, So M, Aviv T, Park N, et al. Inhibition of dopamine receptor D4 impedes autophagic flux, proliferation, and survival of glioblastoma stem cells. Cancer Cell. 2016; 29:859–873. [PubMed: 27300435]

Economopoulou P, Perisanidis C, Giotakis EI, Psyrri A. The emerging role of immunotherapy in head and neck squamous cell carcinoma (HNSCC): anti-tumor immunity and clinical applications. Ann Transl Med. 2016; 4:173. [PubMed: 27275486]

Eppert K, Takenaka K, Lechman ER, Waldron L, Nilsson B, van Galen P, Metzeler KH, Poeppl A, Ling V, Beyene J, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. Nat Med. 2011; 17:1086–1093. [PubMed: 21873988]

Fabregat I, Malfettone A, Soukupova J. New Insights into the Crossroads between EMT and Stemness in the Context of Cancer. J Clin Med. 2016:5.

Friedmann-Morvinski D, Verma IM. Dedifferentiation and reprogramming: origins of cancer stem cells. EMBO Rep. 2014; 15:244–253. [PubMed: 24531722]

Fuereder T. Immunotherapy for head and neck squamous cell carcinoma. Memo. 2016; 9:66–69. [PubMed: 27429658]

Ge Y, Gomez NC, Adam RC, Nikolova M, Yang H, Verma A, Lu CPJ, Polak L, Yuan S, Elemento O, et al. Stem cell lineage infidelity drives wound repair and cancer. Cell. 2017; 169:636–650.e14. [PubMed: 28434617]

Gentles AJ, Plevritis SK, Majeti R, Alizadeh AA. Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. JAMA. 2010; 304:2706–2715. [PubMed: 21177505]

Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, Nair VS, Xu Y, Khuong A, Hoang CD, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. 2015; 21:938–945. [PubMed: 26193342]

Gevaert O, Villalobos V, Sikic BI, Plevritis SK. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. Interface Focus. 2013; 3:20130013. [PubMed: 24511378]

Gingold J, Zhou R, Lemischka IR, Lee DF. Modeling Cancer with Pluripotent Stem Cells. Trends Cancer. 2016; 2:485–494. [PubMed: 27722205]

Gonzalez DM, Medici D. Signaling mechanisms of the epithelial-mesenchymal transition. Sci Signal. 2014; 7:re8. [PubMed: 25249658]

Gregory PA, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G, Vadas MA, Khew-Goodall Y, Goodall GJ. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nat Cell Biol. 2008; 10:593–601. [PubMed: 18376396]

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–674. [PubMed: 21376230]

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MDM, Niu B, McLellan MD, Uzunangelov V, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014; 158:929–944. [PubMed: 25109877]

Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka IR. Dissecting self-renewal in stem cells with RNA interference. Nature. 2006; 442:533–538. [PubMed: 16767105]

Kang W-Y, Chen W-T, Huang Y-C, Su Y-C, Chai C-Y. Overexpression of annexin 1 in the development and differentiation of urothelial carcinoma. Kaohsiung J Med Sci. 2012; 28:145–150. [PubMed: 22385607]

Kim J, Orkin SH. Embryonic stem cell-specific signatures in cancer: insights into genomic regulatory networks and implications for medicine. Genome Med. 2011; 3:75. [PubMed: 22126538]

Kim J, Woo AJ, Chu J, Snow JW, Fujiwara Y, Kim CG, Cantor AB, Orkin SH. A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs. Cell. 2010; 143:313–324. [PubMed: 20946988]

Kooreman NG, Kim Y, de Almeida PE, Termglinchan V, Diecke S, Shao N-Y, Wei T-T, Yi H, Dey D, Nelakanti R, et al. Autologous iPSC-Based Vaccines Elicit Anti-tumor Responses In Vivo. Cell Stem Cell. 2018

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, et al. The Connectivity Map: using gene-expression signatures to connect small molecules genes, and disease. Science. 2006; 313:1929–1935. [PubMed: 17008526]

Lu C, Ward PS, Kapoor GS, Rohle D, Turcan S, Abdel-Wahab O, Edwards CR, Khanin R, Figueroa ME, Melnick A, et al. IDH mutation impairs histone demethylation and results in a block to cell differentiation. Nature. 2012; 483:474–478. [PubMed: 22343901]

Lyssiotis CA, Kimmelman AC. Metabolic interactions in the tumor microenvironment. Trends Cell Biol. 2017; 27:863–875. [PubMed: 28734735]

Mak MP, Tong P, Diao L, Cardnell RJ, Gibbons DL, William WN, Skoulidis F, Parra ER, Rodriguez-Canales J, Wistuba II, et al. A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. Clin Cancer Res. 2016; 22:609–620. [PubMed: 26420858]

Mathur D, Danford TW, Boyer LA, Young RA, Gifford DK, Jaenisch R. Analysis of the mouse embryonic stem cell regulatory networks obtained by ChIP-chip and ChIP-PET. Genome Biol. 2008; 9:R126. [PubMed: 18700969]

Nazor KL, Altun G, Lynch C, Tran H, Harness JV, Slavin I, Garitaonandia I, Müller FJ, Wang YC, Boscolo FS, et al. Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. Cell Stem Cell. 2012; 10:620–634. [PubMed: 22560082]

Ng SWK, Mitchell A, Kennedy JA, Chen WC, McLeod J, Ibrahimova N, Arruda A, Popescu A, Gupta V, Schimmer AD, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. Nature. 2016; 540:433–437. [PubMed: 27926740]

Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP, Bhat KP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010; 17:510–522. [PubMed: 20399149]

Onuchic V, Hartmaier RJ, Boone DN, Samuels ML, Patel RY, White WM, Garovic VD, Oesterreich S, Roth ME, Lee AV, et al. Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. Cell Rep. 2016; 17:2075–2086. [PubMed: 27851969]

Palmer NP, Schmid PR, Berger B, Kohane IS. A gene expression profile of stem cell pluripotentiality and differentiation is conserved across diverse solid and hematopoietic cancers. Genome Biol. 2012; 13:R71. [PubMed: 22909066]

Papillon-Cavanagh S, Lu C, Gayden T, Mikael LG, Bechet D, Karamboulas C, Ailles L, Karamchandani J, Marchione DM, Garcia BA, et al. Impaired H3K36 methylation defines a subset of head and neck squamous cell carcinomas. Nat Genet. 2017; 49:180–185. [PubMed: 28067913]

Pinto JP, Kalathur RK, Oliveira DV, Barata T, Machado RSR, Machado S, Pacheco-Leyva I, Duarte I, Futschik ME. StemChecker: a web-based tool to discover and explore stemness signatures in gene sets. Nucleic Acids Res. 2015; 43:W72–7. [PubMed: 26007653]

Polónia A, Pinto R, Cameselle-Teijeiro JF, Schmitt FC, Paredes J. Prognostic value of stromal tumour infiltrating lymphocytes and programmed cell death-ligand 1 expression in breast cancer. J Clin Pathol. 2017

R Core Team, R.C.T. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.

Reyngold M, Turcan S, Giri D, Kannan K, Walsh LA, Viale A, Drobnjak M, Vahdat LT, Lee W, Chan TA. Remodeling of the methylation landscape in breast cancer metastasis. PLoS ONE. 2014; 9:e103896. [PubMed: 25083786]

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]

Robinson DR, Wu YM, Lonigro RJ, Vats P, Cobain E, Everett J, Cao X, Rabban E, Kumar-Sinha C, Raymond V, et al. Integrative clinical genomics of metastatic cancer. Nature. 2017; 548:297–303. [PubMed: 28783718]

Royston P, Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. Appl Stat. 1994; 43:429.

Salomonis N, Dexheimer PJ, Omberg L, Schroll R, Bush S, Huo J, Schriml L, Ho Sui S, Keddache M, Mayhew C, et al. Integrated Genomic Analysis of Diverse Induced Pluripotent Stem Cells from the Progenitor Cell Biology Consortium. Stem Cell Reports. 2016; 7:110–125. [PubMed: 27293150]

Sato N, Sanjuan IM, Heke M, Uchida M, Naef F, Brivanlou AH. Molecular signature of human embryonic stem cells and its comparison with the mouse. Dev Biol. 2003; 260:404–413. [PubMed: 12921741]

Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. BioRxiv. 2016:060012.

Shibue T, Weinberg RA. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. Nat Rev Clin Oncol. 2017; 14:611–629. [PubMed: 28397828]

Silva TC, Colaprico A, Olsen C, Bontempi G, Ceccarelli M, Berman BP, Noushmehr H. TCGAbiolinksGUI: A graphical user interface to analyze cancer molecular and clinical data. BioRxiv. 2017:147496.

Sokolov A, Paull EO, Stuart JM. One-class detection of cell states in tumor subtypes. Pac Symp Biocomput. 2016; 21:405–416. [PubMed: 26776204]

de Souza CF, Sabedot TS, Malta TM, Stetson L, Morozova O, Sokolov A, Laird PW, Wiznerowicz M, Iavarone A, Snyder J, et al. Distinct epigenetic shift in a subset of Glioma CpG island methylator phenotype (G-CIMP) during tumor recurrence. Cell Reports. 2017 in press.

Spitzer MH, Carmi Y, Reticker-Flynn NE, Kwek SS, Madhireddy D, Martins MM, Gherardini PF, Prestwood TR, Chabon J, Bendall SC, et al. Systemic immunity is required for effective cancer immunotherapy. Cell. 2017; 168:487–502.e15. [PubMed: 28111070]

StataCorp. Stata Statistical Software: Release 13. College Station TX: StataCorp LP; 2013.

Sturm D, Witt H, Hovestadt V, Khuong-Quang DA, Jones DTW, Konermann C, Pfaff E, Tönjes M, Sill M, Bender S, et al. Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. Cancer Cell. 2012; 22:425–437. [PubMed: 23079654]

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102:15545–15550. [PubMed: 16199517]

Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017; 171:1437–1452.e17. [PubMed: 29195078]

Tellez CS, Juri DE, Do K, Bernauer AM, Thomas CL, Damiani LA, Tessema M, Leng S, Belinsky SA. EMT and stem cell-like properties associated with miR-205 and miR-200 epigenetic silencing are early manifestations during carcinogen-induced transformation of human lung epithelial cells. Cancer Res. 2011; 71:3087–3097. [PubMed: 21363915]

Therneau, TM., Grambsch, PM. Modeling Survival Data: Extending the Cox Model. New York, NY: Springer New York; 2000. Estimating the survival and hazard functions; p. 7-37.

Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, Fisher JM, Rodman C, Mount C, Filbin MG, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016; 539:309–313. [PubMed: 27806376]

Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015; 19:A68–77. [PubMed: 25691825]

Tsai S-C, Lin C-C, Shih T-C, Tseng R-J, Yu M-C, Lin Y-J, Hsieh S-Y. The miR-200b-ZEB1 circuit regulates diverse stemness of human hepatocellular carcinoma. Mol Carcinog. 2017

Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, Campos C, Fabius AWM, Lu C, Ward PS, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature. 2012; 483:479–483. [PubMed: 22343889]

Venezia TA, Merchant AA, Ramos CA, Whitehouse NL, Young AS, Shaw CA, Goodell MA. Molecular signatures of proliferation and quiescence in hematopoietic stem cells. PLoS Biol. 2004; 2:e301. [PubMed: 15459755]

Visvader JE, Lindeman GJ. Cancer stem cells: current status and evolving complexities. Cell Stem Cell. 2012; 10:717–728. [PubMed: 22704512]

Wang MD, Wu H, Fu GB, Zhang HL, Zhou X, Tang L, Dong LW, Qin CJ, Huang S, Zhao LH, et al. Acetyl-coenzyme A carboxylase alpha promotion of glucose-mediated fatty acid synthesis enhances survival of hepatocellular carcinoma in mice and patients. Hepatology. 2016; 63:1272–1286. [PubMed: 26698170]

Wen N, Wang Y, Wen L, Zhao S-H, Ai Z-H, Wang Y, Wu B, Lu H-X, Yang H, Liu W-C, et al. Overexpression of FOXM1 predicts poor prognosis and promotes cancer cell proliferation, migration and invasion in epithelial ovarian cancer. J Transl Med. 2014; 12:134. [PubMed: 24885308]

Wickham, H. ggplot2 - Elegant Graphics for Data Analysis. New York, NY: Springer New York; 2009.

Wong CM, Wei L, Au SLK, Fan DNY, Zhou Y, Tsang FHC, Law CT, Lee JMF, He X, Shi J, et al. MiR-200b/200c/429 subfamily negatively regulates Rho/ROCK signaling pathway to suppress hepatocellular carcinoma metastasis. Oncotarget. 2015; 6:13658–13670. [PubMed: 25909223]

Xu L, Zhang L, Hu C, Liang S, Fei X, Yan N, Zhang Y, Zhang F. WNT pathway inhibitor pyrvinium pamoate inhibits the self-renewal and metastasis of breast cancer stem cells. Int J Oncol. 2016; 48:1175–1186. [PubMed: 26781188]

Yan X, Ma L, Yi D, Yoon J, Diercks A, Foltz G, Price ND, Hood LE, Tian Q. A CD133-related gene expression signature identifies an aggressive glioblastoma subtype with excessive mutations. Proc Natl Acad Sci USA. 2011; 108:1591–1596. [PubMed: 21220328]

Yao L, Shen H, Laird PW, Farnham PJ, Berman BP. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. Genome Biol. 2015; 16:105. [PubMed: 25994056]

Young RA. Control of the embryonic stem cell state. Cell. 2011; 144:940–954. [PubMed: 21414485]

Yue H, Huang D, Qin L, Zheng Z, Hua L, Wang G, Huang J, Huang H. Targeting Lung Cancer Stem Cells with Antipsychological Drug Thioridazine. Biomed Res Int. 2016; 2016:6709828. [PubMed: 27556038]

Zaretsky JM, Garcia-Diaz A, Shin DS, Escuin-Ordinas H, Hugo W, Hu-Lieskovan S, Torrejon DY, Abril-Rodriguez G, Sandoval S, Barthly L, et al. Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. N Engl J Med. 2016; 375:819–829. [PubMed: 27433843]

Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, Lerario AM, Else T, Knijnenburg TA, Ciriello G, et al. Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. Cancer Cell. 2016; 29:723–736. [PubMed: 27165744]

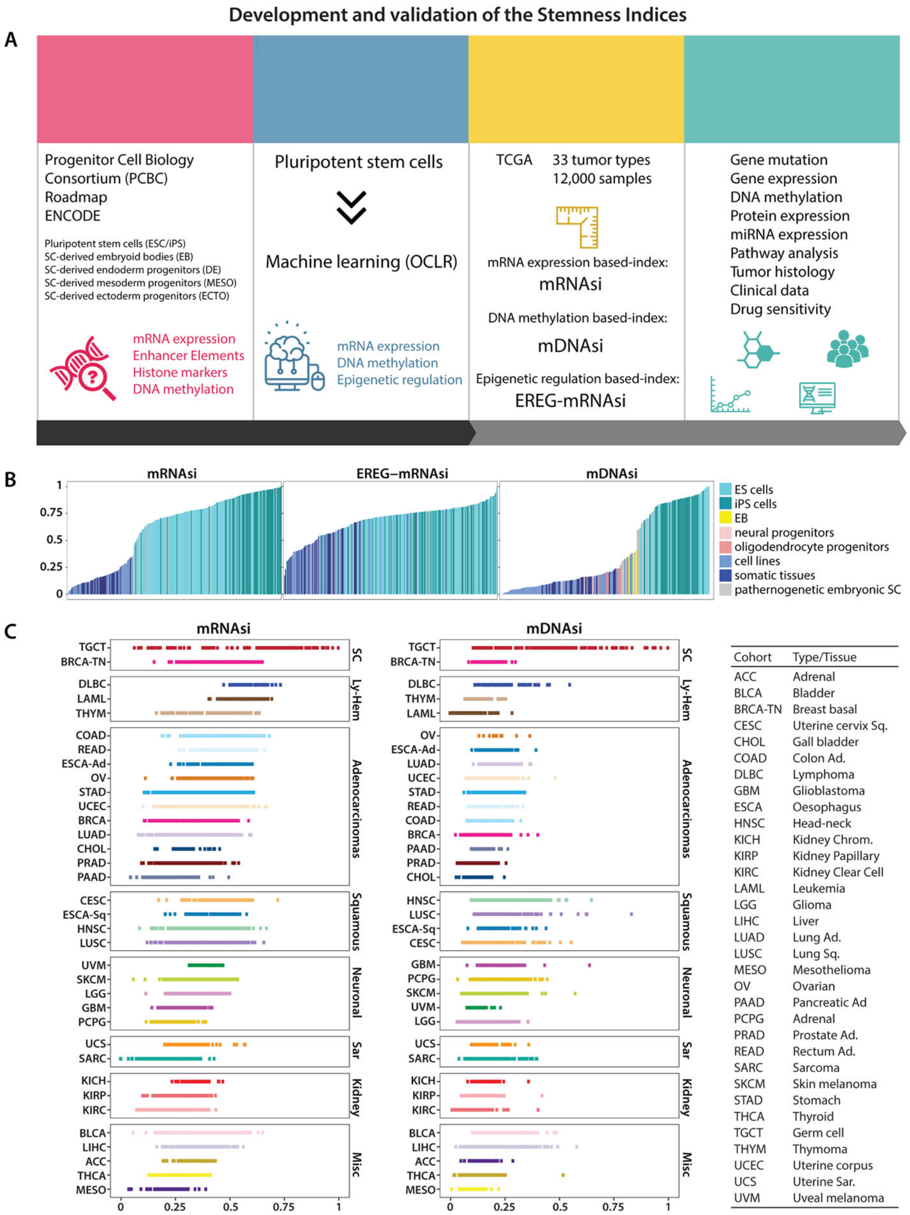
ADDITIONAL RESOURCES - Abbreviations of the TCGA Tumor Types

ACC	Adrenocortical carcinoma
AML	Acute myeloid leukemia
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma

ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and neck squamous cell carcinoma
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LGG	Brain lower grade glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin cutaneous melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular germ cell tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine corpus endometrial carcinoma
UCS	Uterine carcinosarcoma
UVM	Uveal melanoma

HIGHLIGHTS

- Epigenetic and expression-based stemness indices measure oncogenic dedifferentiation
- Immune microenvironment content and PD-L1 levels associate with stemness indices
- Stemness index is increased in metastatic tumors and reveals intratumor heterogeneity
- Applying stemness indices reveals potential drug targets for anti-cancer therapies



Cohort	Type/Tissue
ACC	Adrenal
BLCA	Bladder
BRCA-TN	Breast basal
CESC	Uterine cervix Sq.
CHOL	Gall bladder
COAD	Colon Ad.
DLBC	Lymphoma
GBM	Glioblastoma
ESCA	Oesophagus
HNSC	Head-neck
KICH	Kidney Chrom.
KIRP	Kidney Papillary
KIRC	Kidney Clear Cell
LAML	Leukemia
LGG	Glioma
LIHC	Liver
LUAD	Lung Ad.
LUSC	Lung Sq.
MESO	Mesothelioma
OV	Ovarian
PAAD	Pancreatic Ad
PCPG	Adrenal
PRAD	Prostate Ad.
READ	Rectum Ad.
SARC	Sarcoma
SKCM	Skin melanoma
STAD	Stomach
THCA	Thyroid
TGCT	Germ cell
THYM	Thymoma
UCEC	Uterine corpus
UCS	Uterine Sar.
UVM	Uveal melanoma

Figure 1. Development and validation of the Stemness Indices

(A) Overall methodology. Highlighted are data sources Progenitor Cell Biology Consortium (PCBC), Roadmap and ENCODE databases, OCLR machine learning algorithm, and the resulting stemness indices mRNAsi, mDNAsi and EREG-mRNAsi. The indices for each TCGA tumor sample were correlated with known cancer biology, tumor pathology, clinical information, and drug sensitivity.

(B) Stemness indices of the validation set derived using our stemness signature.

(C) TCGA tumor types sorted by the stemness indices obtained from transcriptomic (mRNAsi) and epigenetic features (mDNAsi); indices were scaled from 0 (low) to 1 (high). The TCGA tumor types were grouped based on their histology and cell-of-origin into stem cell-like (SC), lympho-hematopoietic (Ly-Hem), Adenocarcinomas, Squamous Cell

Carcinomas (Squamous), Neuronal lineage (Neuronal), Sarcomas (Sar), Kidney tumors (Kidney), and not belonging to any of the above (Misc) (Table S2).

See also Figures S1 and S2; and Tables S1 and S2.

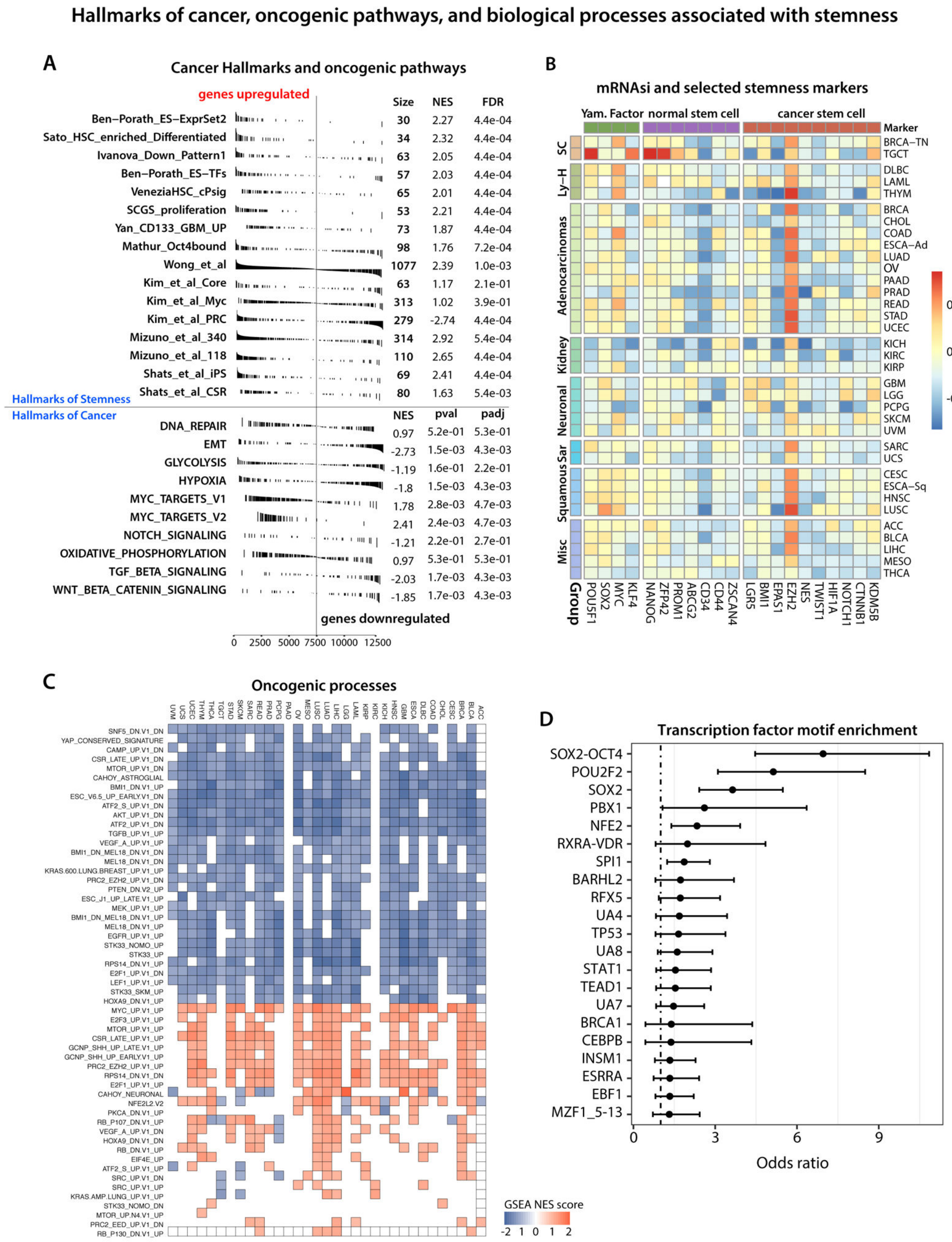


Figure 2. Biological processes associated with cancer stemness

(A) Gene Set Enrichment Analysis showing RNAseq-based stemness signature evaluated in the context of gene sets representative for Hallmarks of Stemness and Cancer.

(B) Correlation between mRNAsi and mRNA expression for published hallmarks of stemness.

(C) Correlation between mRNAsi and selected oncogenic processes.

(D) Association between the epigenomic-based stemness signature (EREG-mDNAsi and EREG-mRNAsi) and enrichment in the transcription factor binding sites.

See also Figure S2 and Table S2.

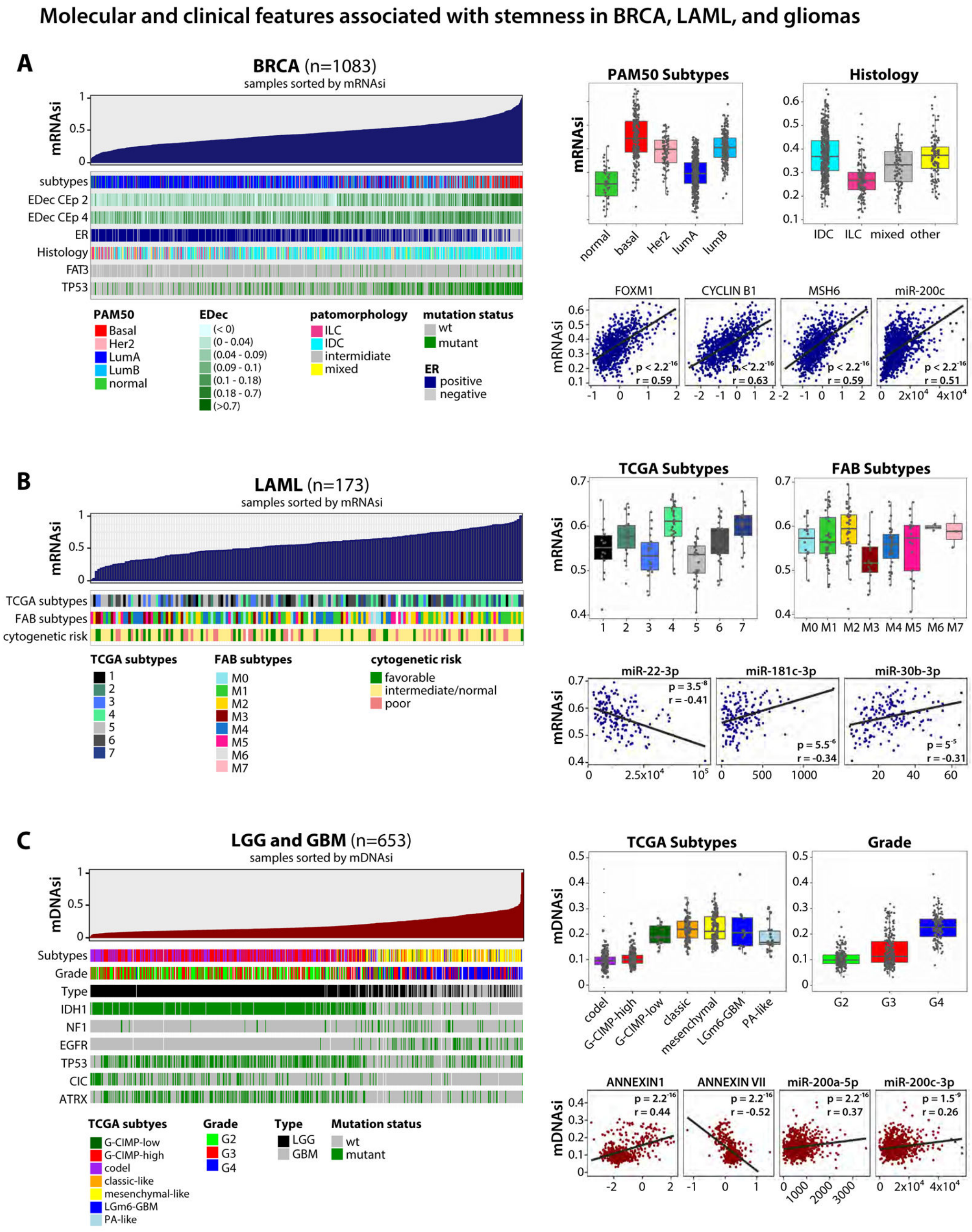


Figure 3. Molecular and clinical features associated with stemness in breast cancer, acute myeloid leukemia, and gliomas

(A) An overview of the association between known molecular and biological processes and stemness in BRCA (Left). Columns represent samples sorted by mRNAsi from low to high (top row). Rows represent molecular and biological processes associated with mRNAsi. Rows named "EDec CEp 2 and 4" represent estimated cell type proportions. Top right, boxplots of mRNAsi in individual samples, stratified by molecular subtype and histology. Bottom right, correlation of mRNAsi and representative protein expression and microRNA.

(B) Similar to A, association of mRNAsi in AML. Top right, mRNAsi by mRNA-based molecular subtype and by FAB classification. Bottom right, correlation scores of mRNAsi and representative microRNA.

(C) As in A and B, GBM and LGG sorted by mDNAsi. Top right, mDNAsi by molecular subtype and grade. Bottom right, correlation scores of mDNAsi and representative protein expression and microRNA. All molecular and clinical features shown are statistically significant.

See also Figures S1, S3, S4, and S5.

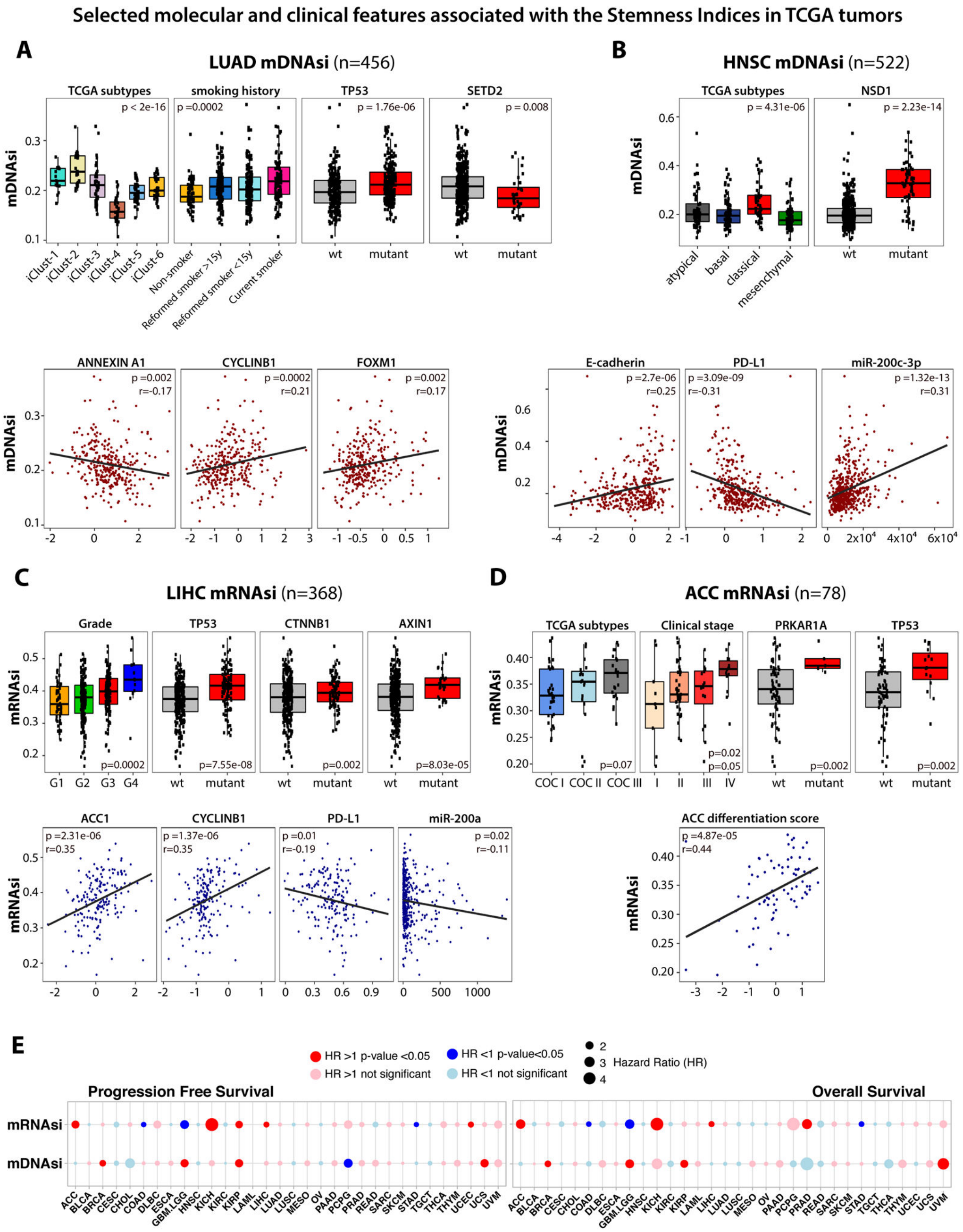


Figure 4. Selected molecular and clinical features associated with the Stemness Indices in TCGA tumors

(A) Association of molecular and clinical features with stemness in LUAD. Top, mDNAsi by integrative molecular subtypes, smoking history, and mutations of TP53 and SETD2. Bottom, correlation scores of mDNAsi and representative protein expression.

(B) Stemness in HNSC. Top, mDNAsi stratified by molecular subtypes and mutation of NSD1. Bottom, correlation scores of mDNAsi and representative protein and microRNA expression.

(C) Stemness in LIHC. Top, mRNAsi stratified by grade and mutations of TP53, CTNNB1, and AXIN1. Bottom, correlation scores of mRNAsi and representative protein and microRNA expression.

(D) Stemness in ACC. Top, mRNAsi stratified by mRNA molecular subtypes, clinical stage, and mutations of PRKAR1A and TP53. Bottom, correlation scores of mRNAsi and adrenal differentiation score.

(E) Cox proportional hazards model analysis. Left, progression-free survival; right, overall survival. Hazard ratio greater than one denotes a trend toward higher stemness index with worse outcome.

See also Figures S3, S4, and S5.

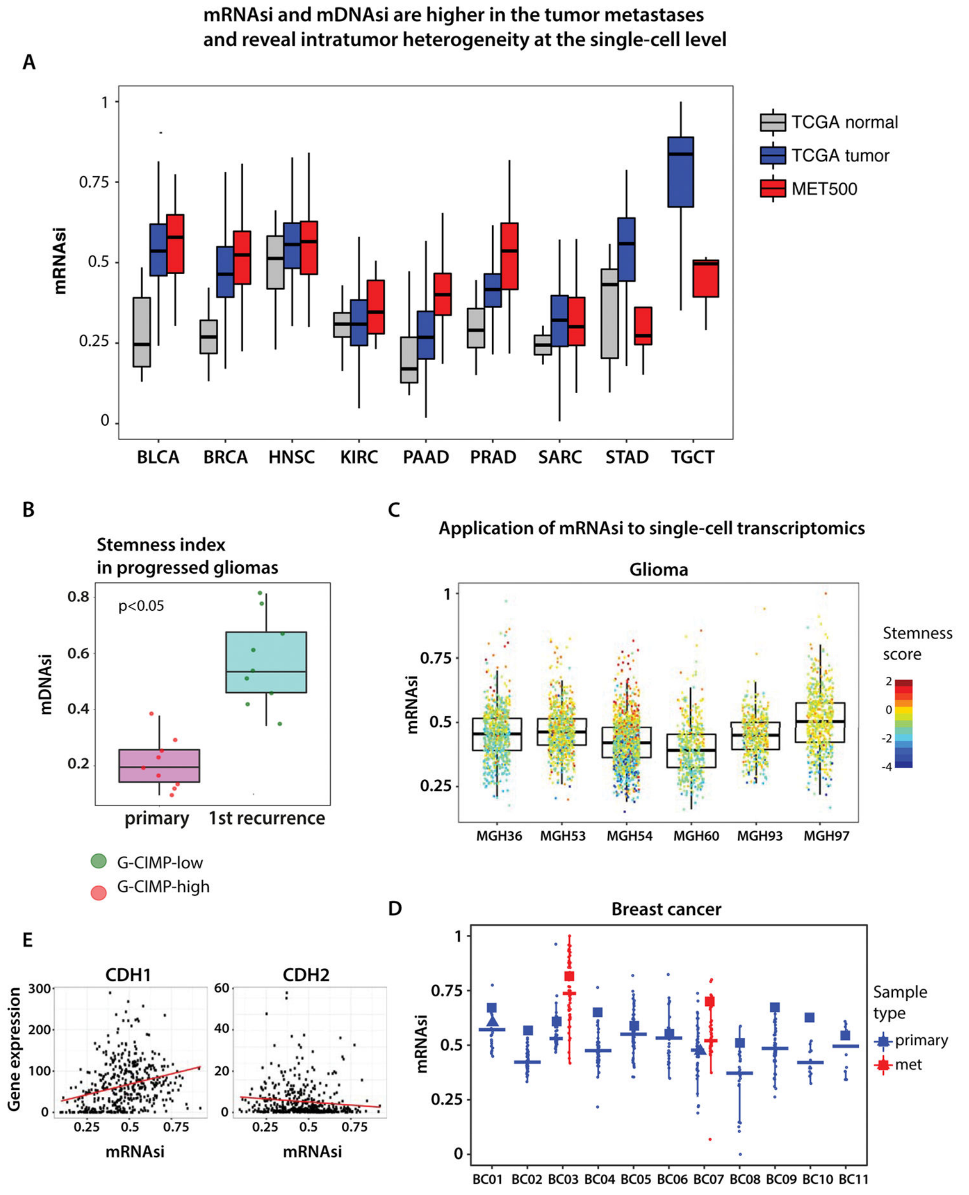


Figure 5. Analysis of cancer stemness in the context of metastatic state and intratumor heterogeneity

(A) mRNAsi is higher in cancer metastases in comparison to the TCGA primary tumors. **(B)** mDNAsi is higher in recurrent glioma samples compared to the primary glioma occurrence from the same patient. G-CIMP - glioma CpG methylator phenotype. **(C)** and **(D)** Application of mRNAsi to single-cell transcriptome of gliomas and breast cancer reveal intratumor heterogeneity and various degrees of the oncogenic dedifferentiation. **(E)** Correlation of mRNAsi and mRNA expression of CDH1 (epithelial marker) and CDH2 (mesenchymal marker) in the cancer metastases.

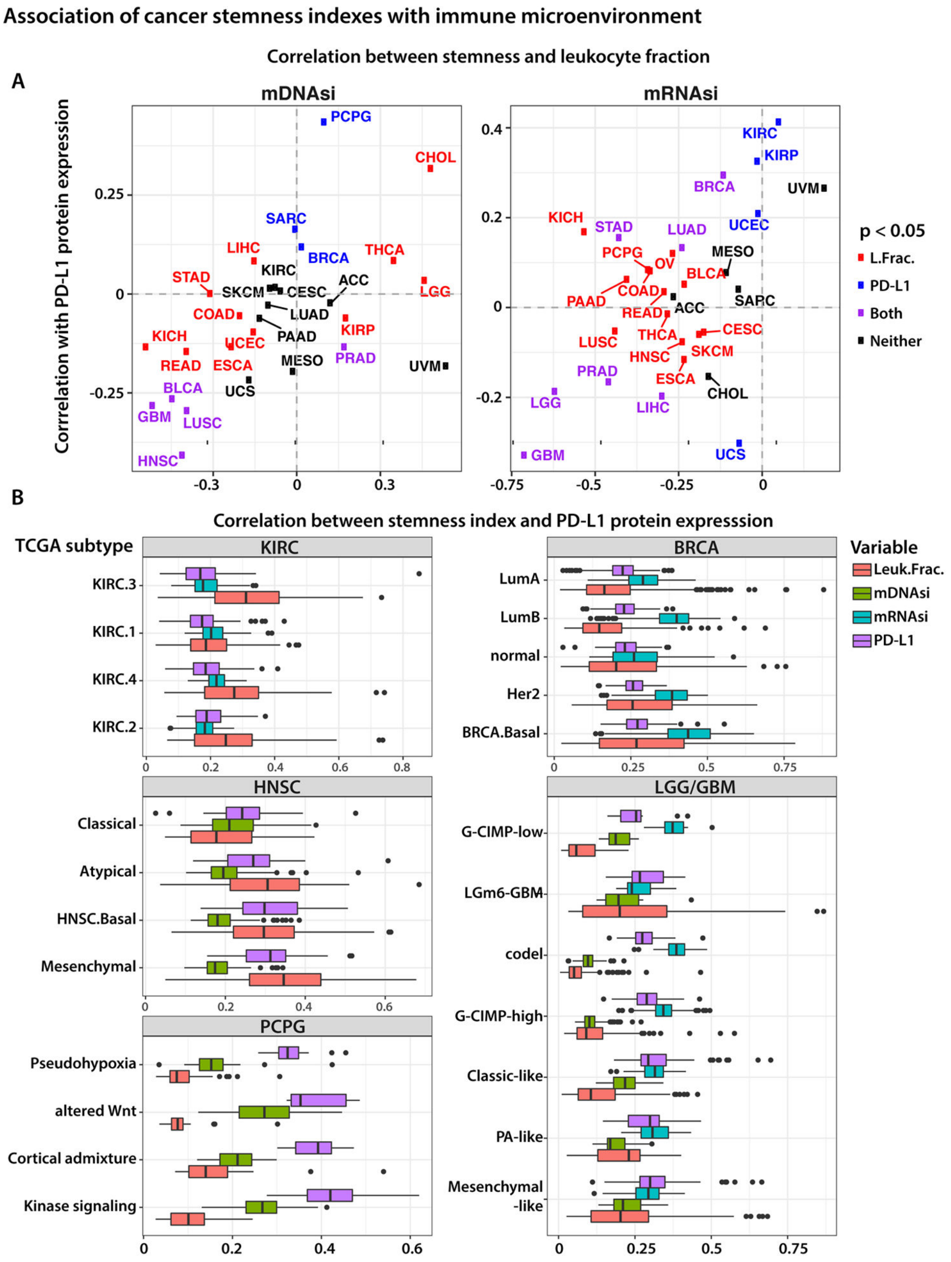


Figure 6. Association of stemness index with immune microenvironment

(A) mDNAsi and mRNAsi in the context of immune microenvironment. Each panel shows the Spearman correlation between the stemness index and PD-L1 protein expression plotted against Spearman correlation between the same stemness index and total leukocyte fraction, as estimated from DNA methylation data.

(B) Highlight of tumor types that exhibit strong correlation between stemness and PD-L1 expression or total leukocyte fraction. See also Figure S6.

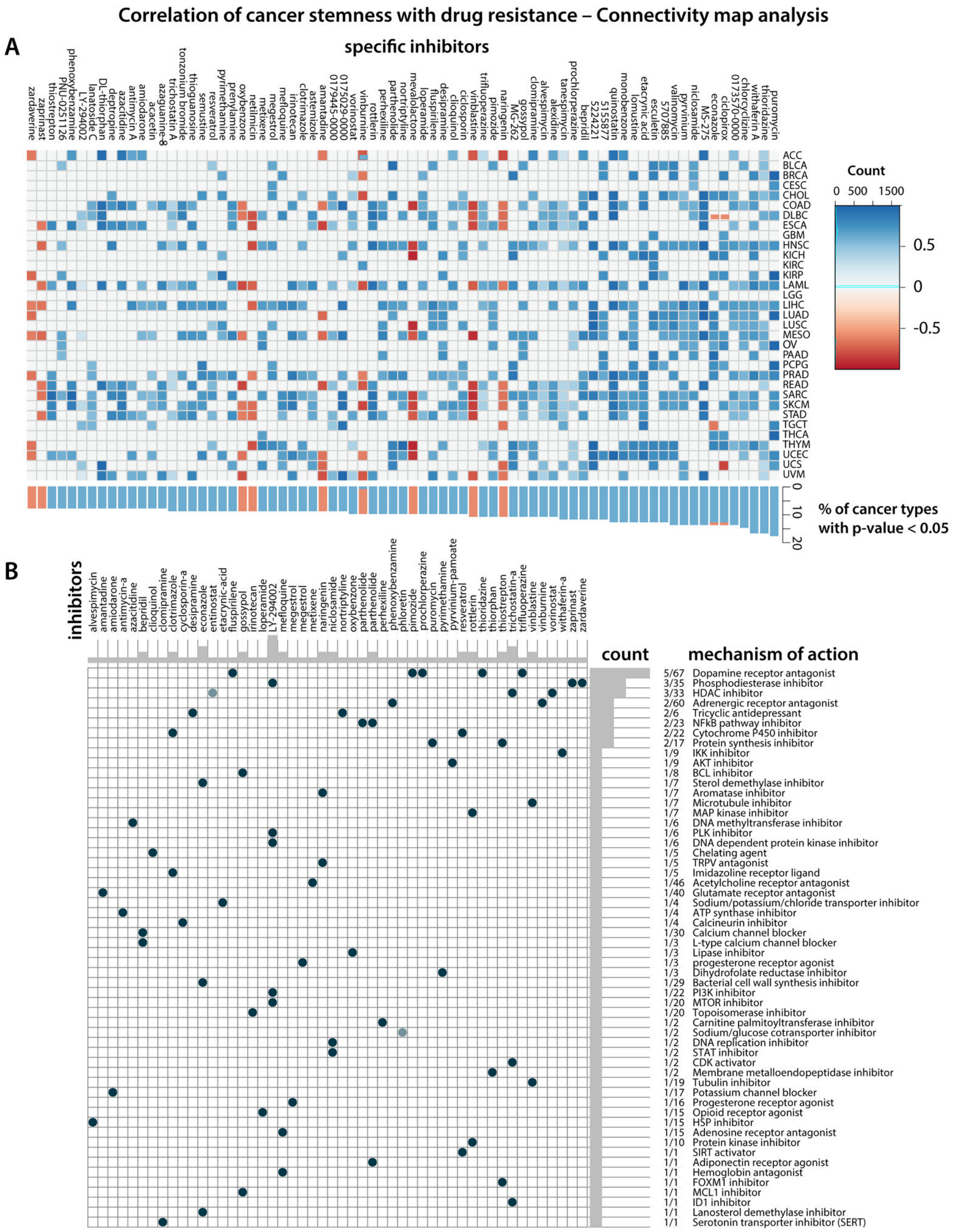


Figure 7. Correlation of cancer stemness with drug resistance – Connectivity map analysis

(A) Heatmap showing enrichment score (positive in blue, negative in red) of each compound from the CMap for each cancer type. Compounds sorted from right to left by descending number of cancer type significantly enriched.

(B) Heatmap showing each compound (perturbagen) from the CMap that share Mechanism of actions (rows). Sorted by descending number of compound with shared mechanism of actions.

See also Figure S7 and Tables S3 and S4.