Title

A dataset of chromosomal instability gene signature scores in normal and cancer cells from the human breast.

Authors

Baba, Shahnawaz

Labhsetwar, Shreyas

Klemke, Richard

et al.

Publication Date

2023-12-01

DOI

10.1016/j.dib.2023.109647

Data in Brief 51 (2023) 109647

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A dataset of chromosomal instability gene signature scores in normal and cancer cells from the human breast

Shahnawaz A. Baba, Shreyas Labhsetwar, Richard Klemke, Jay S. Desgrosellier*

Department of Pathology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

ARTICLE INFO

Article history:
Received 18 September 2023
Accepted 29 September 2023
Available online 8 October 2023

Dataset link: A dataset of chromosomal instability gene signature scores in normal and cancer cells from the human breast (Original data)

Keywords:
Cancer stem cell
Chromosomal missegregation
Signaling state
Stress tolerance
Breast cancer subtype
Tumor initiation

ABSTRACT

These data show the relative amount of chromosomal instability (CIN) in a diverse array of human breast cell types, including non-transformed mammary epithelial cells as well as cancer cell lines. Additional data is also provided from human embryonic and mesenchymal stem cells. To produce this dataset, we compared a published chromosomal instability gene signature against publicly available datasets containing gene expression information for each cell. We then analyzed these data with the Python GSEAPY software package to provide a CIN enrichment score. These data are useful for comparing the relative amounts of CIN in different breast cell types. This includes cells representing the major clinical (ER/PR$^+$, HER2$^+$ & Triple-negative) as well as intrinsic breast cancer subtypes (Luminal B, HER2+, Basal-like and Claudin-low). Our dataset has a great potential for re-use given the recent surge in interest surrounding the role of CIN in breast cancer. The large size of the dataset, coupled with the diversity of the cell types represented, provides numerous possibilities for future comparisons.

DOI of original article: 10.1016/j.heliyon.2023.e20182
* Corresponding author.
E-mail address: jdesgrosellier@ucsd.edu (J.S. Desgrosellier).

https://doi.org/10.1016/j.dib.2023.109647

Specifications Table

Subject	Cancer Research
Specific subject area	Our work focuses on breast cancer stem cells and their role in tumor progression and metastasis.
Data format	Raw
Type of data	Excel spreadsheets
Data collection	Data were collected after comparing each cell type present in the gene expression data files downloaded from NCBI with the published CIN gene signature. The provided CIN enrichment score for each cell type is listed in the Excel spreadsheets deposited. Data was normalized using the trimmed mean of M-values (TMM).
Data source location	• *Institution:* University of California, San Diego • *City/Town/Region:* La Jolla, CA • *Country:* USA • *Latitude and longitude:* 32.876328, −117.236067
Data accessibility	Repository name: UC San Diego Library Digital Collections Data identification number: doi:10.6075/J0R78FDG Direct URL to data: https://library.ucsd.edu/dc/object/bb12054874
Related research article	S.A. Baba, Q. Sun, S. Mugisha, S. Labhsetwar, R. Klemke, J.S. Desgrosellier, Breast cancer stem cells tolerate chromosomal instability during tumor progression via c-Jun/AXL stress signaling, *Heliyon*. In Press.

1. Value of the Data

- While CIN is a defining hallmark of cancer, little is known about its relationship with stemness. High levels of CIN are associated with breast cancer evolution and metastasis [1] suggesting that it may be a unique feature of aggressive cells. Tumor-initiating cancer stem cells (CSCs) bearing similarities to adult mammary stem cells are a highly aggressive tumor cell subset that contribute to metastasis and disease progression [2–7]. This suggests that CSCs may contain high levels of CIN, enhancing their aggressive properties. Thus, we generated this dataset to better understand how CIN levels compare in stem and non-stem cell types.
- These data add value to our original research article by allowing us to assess potential associations between CIN levels and stem-like breast cancer cells and draw important conclusions regarding the impact of CIN on tumor initiation.
- The CIN enrichment scores provided in this dataset are useful for comparing the relative amounts of CIN present in different non-transformed and cancer cell lines from the breast.
- This dataset may be useful to anyone associated with breast cancer research. CIN is a hallmark of cancer and is well described to play a role in tumor evolution, especially in regards to therapeutic resistance [8]. Recent high-profile work has further characterized CIN as a driver of antiviral innate immune signaling in breast cancer cells [9], resulting in tumor progression and metastasis [1]. Thus, CIN is an important area of research in the breast cancer field and this dataset may aid future studies of this topic.
- These data can be used/reused for further insights into associations between CIN levels and numerous additional variables, including cell-of-origin, mutational status, subtype, gene expression, etc.

2. Data Description

The deposited dataset consists of two distinct Excel spreadsheets [10]. Each spreadsheet lists the cell names or identifiers across the top of each column and the corresponding enrichment scores (ES) and normalized enrichment scores (NES) underneath. To obtain the scores, we compared a published CIN gene signature [1] to the following gene expression datasets downloaded from NCBI:

Breast cancer cell lines [11]:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50470
HCC38 sorted cells [12]:
https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA750073

We compared these datasets to the CIN signature by first converting FASTQ files to gene-expression matrices and processing so that only the data were retained. Scores were obtained for each cell type by performing single sample gene set enrichment analysis (GSEA) with Python GSEAPY Library software, and the resulting data were processed into Excel format.

The resulting spreadsheets are labelled and described as follows:

Breast cell lines – Excel spreadsheet listing the CIN enrichment scores for an assortment of breast cancer cell lines as well as non-transformed human cells. Cell names are listed across the top of each column with the corresponding ES and NES scores underneath.

HCC38 sorted breast cancer cells – Excel spreadsheet containing the CIN enrichment scores for HCC38 breast cancer cells sorted according to their cell surface EpCAM (Ep) and integrin $\alpha v\beta 3$ (b3) status from three independent experiments. Sorted cells from each experiment were divided into four categories: EpCAM low (lo) versus high (hi) and $\alpha v\beta 3$ positive (pos) versus negative (neg). Each cell type belonging to a particular experiment was given a unique Run number (SRR) listed across the top of each column. For quick reference, the sorted cell type is included underneath, followed by the ES and NES scores.

3. Experimental Design, Materials and Methods

Publicly available RNA sequencing data from breast cancer and normal mammary cell lines was obtained from NCBI GEO (GSE50470), while data from sorted HCC38 cell populations was downloaded from the NCBI Sequence Read Archive (PRJNA750073). FASTQs were converted to gene-expression matrices and the files were processed to remove all the header information and only retain the data. The CIN gene signature was acquired from Bakhoum et al. [1] and CIN scores were obtained by examining enrichment for the CIN associated gene signature in each cell type represented in the sequencing datasets according to Barbie et al. [13].

To generate the CIN scores for each cell type, we analyzed data with the Python GSEAPY Library (https://gseapy.readthedocs.io/en/latest/). First, input files were read using Python's Pandas library and joined with each other using the ID & amp columns before deleting any unnecessary columns. Ensemble Gene IDs were mapped to their HGNC Symbols using Python's BioMart API (https://pypi.org/project/biomart). Any Ensemble ID which did not have a corresponding HGNC Symbol was dropped. Once we obtained the data frame having HGNC Symbols as rows, samples as columns, and their feature counts as values in all rows, this data frame, along with the CIN gene set was passed to the Single Sample GSEA Python library. The final data comprised 36,866 rows and 106 columns before feeding it into GSEAPY. To determine the enrichment scores (ES), we applied Single Sample GSEA to the final data frame. The experiment was repeated with a normalized version of the data frame, but the normalized enrichment scores (NES) were identical to the ES. GSEAPY output was then processed into Excel format and saved as final results files.

Limitations

None.

Ethics statement

We confirm that the authors have read and follow the ethical requirements for publication in Data in Brief and that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

A dataset of chromosomal instability gene signature scores in normal and cancer cells from the human breast (Original data) (UC San Diego Library Digital Collections)

CRediT Author Statement

Shahnawaz A. Baba: Conceptualization, Methodology, Writing – original draft; **Shreyas Labhsetwar:** Methodology, Formal analysis, Software; **Richard Klemke:** Methodology, Formal analysis, Software; **Jay S. Desgrosellier:** Conceptualization, Methodology, Formal analysis, Visualization, Investigation, Supervision, Funding acquisition.

Acknowledgements

This work was supported by funding from the Tobacco-Related Disease Research Program [Grant #T32IR4741 (to J.S.D.)]; and the California Breast Cancer Research Program [Grant #B28IB5479 (to J.S.D.)].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S.F. Bakhoum, B. Ngo, A.M. Laughney, J.-A. Cavallo, C.J. Murphy, P. Ly, P. Shah, R.K. Sriram, T.B. Watkins, N.K. Taunk, Chromosomal instability drives metastasis through a cytosolic DNA response, Nature 553 (2018) 467–472, doi:10.1038/nature25432.
[2] M. Al-Hajj, M.S. Wicha, A. Benito-Hernandez, S.J. Morrison, M.F. Clarke, Prospective identification of tumorigenic breast cancer cells, Proc. Natl Acad. Sci. 100 (2003) 3983–3988, doi:10.1073/pnas.0530291100.
[3] E. Lim, F. Vaillant, D. Wu, N.C. Forrest, B. Pal, A.H. Hart, M.-L. Asselin-Labat, D.E. Gyorki, T. Ward, A. Partanen, Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers, Nat. Med. 15 (2009) 907–913, doi:10.1038/nm.2000.
[4] A. Prat, C.M. Perou, Deconstructing the molecular portraits of breast cancer, Mol. Oncol. 5 (2011) 5–23, doi:10.1016/j.molonc.2010.11.003.
[5] I. Malanchi, A. Santamaria-Martínez, E. Susanto, H. Peng, H.-A. Lehr, J.-F. Delaloye, J. Huelsken, Interactions between cancer stem cells and their niche govern metastatic colonization, Nature 481 (2012) 85–89, doi:10.1038/nature10694.
[6] A.E. Tijhuis, S.C. Johnson, S.E. McClelland, The emerging links between chromosomal instability (CIN), metastasis, inflammation and tumour immunity, Mol. Cytogenet. 12 (2019) 1–21, doi:10.1186/s13039-019-0429-1.
[7] A.-P. Morel, C. Ginestier, R.M. Pommier, O. Cabaud, E. Ruiz, J. Wicinski, M. Devouassoux-Shisheboran, V. Combaret, P. Finetti, C. Chassot, A stemness-related ZEB1–MSRB3 axis governs cellular pliancy and breast cancer genome stability, Nat. Med. 23 (2017) 568–578, doi:10.1038/nm.4323.
[8] S.F. Bakhoum, D.A. Compton, Chromosomal instability and cancer: a complex relationship with therapeutic potential, J. Clin. Invest. 122 (2012) 1138–1143, doi:10.1172/JCI59954.
[9] C. Hong, M. Schubert, A.E. Tijhuis, M. Requesens, M. Roorda, A. van den Brink, L.A. Ruiz, P.L. Bakker, T. van der Sluis, W. Pieters, cGAS–STING drives the IL-6-dependent survival of chromosomally instable cancers, Nature 607 (2022) 366–373, doi:10.1038/s41586-022-04847-2.
[10] S.A. Baba, S. Labhsetwar, R. Klemke, J.S. Desgrosellier, A Dataset of Chromosomal Instability Gene Signature Scores in Normal and Cancer Cells from the Human Breast, UC San Diego Library Digital Collections, 2023, doi:10.6075/J0R78FDG.
[11] A. Prat, O. Karginova, J.S. Parker, C. Fan, X. He, L. Bixby, J.C. Harrell, E. Roman, B. Adamo, M. Troester, C.M. Perou, Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes, Breast Cancer Res. Treat. 142 (2013) 237–255, doi:10.1007/s10549-013-2743-3.
[12] Q. Sun, Y. Wang, A. Officer, B. Pecknold, G. Lee, O. Harismendy, J.S. Desgrosellier, Stem-like breast cancer cells in the activated state resist genetic stress via TGFBI-ZEB1, NPJ Breast Cancer 8 (2022) 5, doi:10.1038/s41523-021-00375-w.

[13] D.A. Barbie, P. Tamayo, J.S. Boehm, S.Y. Kim, S.E. Moody, I.F. Dunn, A.C. Schinzel, P. Sandy, E. Meylan, C. Scholl, S. Frohling, E.M. Chan, M.L. Sos, K. Michel, C. Mermel, S.J. Silver, B.A. Weir, J.H. Reiling, Q. Sheng, P.B. Gupta, R.C. Wadlow, H. Le, S. Hoersch, B.S. Wittner, S. Ramaswamy, D.M. Livingston, D.M. Sabatini, M. Meyerson, R.K. Thomas, E.S. Lander, J.P. Mesirov, D.E. Root, D.G. Gilliland, T. Jacks, W.C. Hahn, Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1, Nature 462 (2009) 108–112, doi:10.1038/nature08460.