nature portfolio

Corresponding author(s): Francesca Nadalin, Francesco Nicassio

Last updated by author(s): 6/06/2024

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Sequencing data were collected on a Novaseq 6000 sequencer, using Novaseq control software v1.8<br>Flow cytometry experiments were performed with a BD FACSAria Sorter, using the FACSDiva software.<br>Data collection from Biorad CFX96 or CFX384 was performed using CFX Manager 3.0. |
| Data analysis | JMP PRO (v16); Microsoft Excel (v16.72); R (v4.2.1); Cell Ranger (v6); Cell Ranger ARC (v2.0.2); seqkit (v2.1.0); deMULTIplex (v1.0.2); Seurat (v4.0.5); limma (v3.49.5); clusterProfiler (v4.1.4); Signac (v1.7.0); GenomicRanges (v1.46.1); cisTopic (v0.3.0); idr (v2.0.3); pROC (v1.18.0); ROGUE (v1.0); GREAT (4.0.4); ChEA (v3); BWA (v0.7.17); bedtools (v2.30.0); CNVkit (v0.9.8); STAR (v2.7.3); Galaxy (v3.36); Deseq2 (v1.34); Scissor (v2.0.0); IPA (v1.22.01); Cytoscape (v3.9.1); cBio Cancer Genomics Portal.<br>The code used to reproduce the analysis reported in this study is available on github at https://github.com/nicassiolab/GBC_SUM159PT_paper and https://github.com/nicassiolab/GBC_SUM159PT_paper_figures.<br>All code has been deposited in Zenodo<br>https://doi.org/10.5281/zenodo.10979191<br>https://doi.org/10.5281/zenodo.10979121 |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The raw data generated in this study have been deposited in the ArrayExpress database under accession codes
E-MTAB-13064 [https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-13064],
E-MTAB-13066 [https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-13066] and
E-MTAB-13069 [https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-13896],
in the GEO database under accession code GSE222596 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE222596] and
in the SRA database under accession code PRJNA922938 [https://www.ncbi.nlm.nih.gov/bioproject/ PRJNA922938].

The processed data generated in this study have been deposited in Zenodo and are provided in the Supplementary Information/Source Data file.
https://doi.org/10.5281/zenodo.10912157

Metanalysis from previous studies was performed using :
raw data (Pal et al 2021, PMID: 33950524) available in the GEO database under accession GSE161529 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161529].
processed data available at (Simeonov et al. 2021, PMID: 34115987) https://doi.org/10.1016/j.ccell.2021.05.005,
(Gavish et al. 2023, PMID: 37258682) https://doi.org/10.1038/s41586-023-06130-4
and (LaFave et al. 2020, PMID: 32707078) https://doi.org/10.1016/j.ccell.2020.06.006

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | *Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.* |
| Reporting on race, ethnicity, or other socially relevant groupings | *Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status). Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.) Please provide details about how you controlled for confounding variables in your analyses.* |
| Population characteristics | *Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."* |
| Recruitment | *Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.* |
| Ethics oversight | *Identify the organization(s) that approved the study protocol.* |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | No statistical method was used to predetermine the sample size, which was comparable to that reported in previous studies.<br>Regarding single-cell experiments, the required number of individual single cell profiles was determined to capture a sizable portion of the total number of GBCs (genetic barcodes), which was first assessed by bulk DNA sequencing. |
| Data exclusions | In vivo: no data were excluded from the analysis.<br>Single cell experiments: one single-cell time-course batch (MULTI-Seq) was removed due to poor library quality. |
| Replication | All experiments were replicated at least twice.<br>Time-points for single-cell experiments where not exactly matched across replicates, but this is vial-dependent. GBCs are comparable across replicates. |
| Randomization | Randomization was not relevant, as the study was performed on uniform biological material (i.e. a commercial cell line).<br>For comparative in vivo experiments (e.g. TM4SF1 high vs bulk), animals were allocated randomly into the experimental groups. |
| Blinding | Blinding was not applicable in our study. We performed characterizations on single cells obtained from a cancer population using barcodes therefore there was no case/test vs control group. For the in vivo experiments (TM4SF1+ vs BULK) blinding was not possible due to technical reasons. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☐ | ☒ Antibodies |
| ☐ | ☒ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☐ | ☒ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☐ | ☒ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Antibodies

| | |
|---|---|
| Antibodies used | TM4SF1-APC: clone: REA851, Miltenyi Biotec. Used at 1:50 dilution as reported in the datasheet. |
| Validation | Validation of the antibody was performed by FACS, analyzing the fraction of TM4SF1 positive cells in untreated cells and upon TM4SF1 perturbation by CRISPR interference, as shown in Supplementary Figure 5. |

## Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

| | |
|---|---|
| Cell line source(s) | SUM159PT were purchased from Asterand (now Bio-IVT) and are available at https://bioivt.com/sum-breast-cancer-cell-lines.<br>MDA-MB-231 TGL were purchased from ATCC ( ATCC® HTB-26TM) |
| Authentication | none of the cell lines were authenticated. |
| Mycoplasma contamination | All lines used were negative for mycoplasma. |
| Commonly misidentified lines<br>(See ICLAC register) | No commonly misidentified lines were used. |

## Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

| | |
|---|---|
| Laboratory animals | Immunodeficient NOD.Cg-PrkdcscidIL2rgtm1Wjl/SzJ (Also Known As NOD/SCID/IL2Rγc-/-). Age: 10 week old. |
| Wild animals | The study did not involve wild animals |
| Reporting on sex | Results were originated only in female (cells, mice), given the distribution of breast cancer between the two sexes. |
| Field-collected samples | The study did not involve field-collected samples. |
| Ethics oversight | Experiments were performed with the approval of Italian Ministry of Health and in compliance with the Italian law (D.lgs. 26/2014), which enforces Dir. 2010/63/EU (Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes) and EU 86/609 directive. Proper permit and consent were granted (Protocol No. 779/2020) by the institutional organism for ethics and animal welfare on experimental procedures (OPBA, Cogentech). |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Flow Cytometry

### Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

| | |
|---|---|
| Sample preparation | SUM159PT_KRAB cells were stained with anti-TM4SF1-APC (clone: REA851 Miltenyi Biotec) antibody for 10' at 4°C, in the dark. |
| Instrument | BD Fusion Aria Sorter. |
| Software | FACSDiva and FlowJo softwares were used for acquisition and analysis, respectively. |
| Cell population abundance | TM4SF1 high purity was estimated by RT-qPCR, using as marker genes TM4SF1 itself or a subset of genes identified as strongly associated with TM4SF1 by single-cell RNA sequencing. |
| Gating strategy | The bulk population was FAC-sorted using FSC/SSC gate, while the TM4SF1high subpopulation was sorted gating the top 5% APC fluorescence intensity (as shown in Figure S4C) |

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.