nature portfolio

Corresponding author(s): Qin Ma

Last updated by author(s): 07172026

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 *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 *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	No software was used for data collection
Data analysis	https://github.com/mtduan/emitgcl Other tools and packages used in this paper include: Python: 3.8.0, PyTorch: 1.12.0, torch_cluster: 1.6.0, torch_scatter: 2.1.0, torch_sparse: 0.6.15, Seurat: 4.3.0, CellSIUS: 1.0.0, GiniClust: 3.0, MarsGT: 0.2.1, CellOracle: 0.15.0.

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

Previously published scRNA-seq datasets were used for benchmarking and case studies in this study. Breast cancer scRNA-seq datasets were obtained from GEO

under accession numbers [GSE167036] and [GSE180286]. Lung cancer datasets were obtained from GEO under accession number [GSE198099]. Head and neck cancer datasets were obtained from GEO under accession number [GSE188737]. Pancreatic cancer datasets were obtained from GEO under accession number [GSE197177]. Papillary thyroid cancer datasets were obtained from GEO under accession number [GSE214148]. Nasopharyngeal carcinoma scRNA-seq datasets were obtained from the GSA for Human database under accession number [HRA000036]. Additional previously published RNA-seq datasets from UCSC were used for breast cancer studies, including datasets from Chin (2006), Miller (2005), Desmedt (2007), Wang (2005), and Caldas (2007) [https://xenabrowser.net/datapages/]. All datasets were reprocessed and analyzed as described in the Methods section. Details of data information can be found in Supplementary Data 1. Source data are provided with this paper.

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	This study does not involve primary data collection on human participants.
Reporting on race, ethnicity, or other socially relevant groupings	This study relies on publicly available datasets, which do not contain individual-level sex, gender, race, or ethnicity information.
Population characteristics	As this study is based on secondary analysis of existing data, we do not have access to disaggregated demographic details.
Recruitment	Demographic information is not applicable to this study, as no human participant data was directly collected or analyzed.
Ethics oversight	This study does not involve human participants, their data, or biological materials requiring ethical approval.

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	For in vitro migration studies, at least three independent replicates were used. For the animal study in Fig 5J, we did a preliminary power analysis based on the differences observed in the Yy1 KO migration experiments that suggested a sample size of 4-6 mice would be sufficient to detect statistically meaningful differences. We chose on the higher end of this range considering we were moving to an in vivo model and we were overexpressing Yy1 rather than performing knock out.
Data exclusions	No data was excluded from in vitro or in vivo experiments.
Replication	In vitro experiments were completed with at least three independent replicate experiments. The animal study utilized 6 independent mice per group.
Randomization	For the animal study, 12 mice were injected with either control (n=6) or Yy1-expressing cells (n=6). The mice that were injected with control vs. Yy1-expressing cells was randomized. No further intervention was applied in this study that required any further randomization.
Blinding	Personnel were not blinded to the groups for in vitro or animal studies.

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	Yy1 antibody (Cell Signaling Technology #46395) was used to validate Yy1 expression in EO771 cells.
Validation	This antibody has been published with (63 citations) and validated by the manufacturer, Cell Signaling Technology.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	EO771 cells were purchased from American Tissue Cell Collection (Cat #CRL-3461)
Authentication	Authentication is performed on EO771 cells annually by STR profiling.
Mycoplasma contamination	EO771 cells are routinely checked for mycoplasma contamination before being used for experiments.
Commonly misidentified lines (See ICLAC register)	The EO771 cell line shares characteristics with the human-derived MCF-7 cell line, as both serve as models for luminal B subtype breast cancer. They exhibit similar hormone receptor expression profiles and sensitivity to anti-estrogen therapies. However, molecular differences between EO771 and MCF-7 have been reported in specific research contexts.

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	C57BL/6NHsd (source: Inotiv) were used for the animal study.
Wild animals	This study did not involve wild animals.
Reporting on sex	Only female mice were included as the study focused on female breast cancer.
Field-collected samples	This study did not involve field-collected samples.
Ethics oversight	The animal study was approved by the Indiana University Bloomington IACUC committee (protocol #23-025)

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

Plants

Seed stocks	This study did not involve plants
Novel plant genotypes	This study did not involve plants
Authentication	This study did not involve plants