nature portfolio

Corresponding author(s): Caroline Uhler and GV Shivashankar

Last updated by author(s): 06/30/2024

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 *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: N/A

Data analysis	The code is available in the Github repository: https://github.com/uhlerlab/DCISprogression. The following Python packages are used in our code: Python 3.10.6 pytorch 1.9.1 scanpy 1.9.1 scipy 1.9.1 numpy 1.23.3 scikit-learn 1.1.2 matplotlib-base 3.6.0 matplotlib 3.6.0 seaborn 0.12.0 pandas 1.5.3 umap-learn 0.5.3 anndata 0.8.0 Cell segmentation was performed using the StarDist model available at https://github.com/stardist/stardist.

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 chromatin imaging data generated in this study have been deposited in the PSI Public Data Repository under accession code: https://doi.org/10.16907/f2030c03-231e-4cb9-9b69-446714b51d26.

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	All tissue samples are from females.
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	All tissue microarrays were purchased from US Biomax, Inc., Derwood, USA, who obtained consent from both the hospitals and the individuals with discrete legal consent forms.
Ethics oversight	All experiments were performed in accordance with relevant guidelines and regulations at PSI/ETH.

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	We imaged 560 tissue microarray samples from 122 patients at 3 different stages of DCIS progression with 11 phenotypic categories. The sample sizes were chosen such that there is a large number of samples at each DCIS stage.
Data exclusions	557 samples out of the total 560 samples were used in our analysis. Since there was only one core labeled as “hyperplasia with saccular dilatation”, this core was removed from further analysis, which explains the removal of 3 samples.

Replication	Some tissue microarray samples were randomly selected to be held out during model training and analysis. All results were reproduced on these held-out samples. Cross validation was used for the DCIS stage prediction from neighborhood and/or cluster composition.
Randomization	We did not have different experimental groups.
Blinding	We did not have different experimental groups.

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	Collagen Type 1: Abcam, ab6308, 1:200. gH2AX: CST, 2577S, 1:50. Ki67: Millipore, AB9260, 1:50. aSMA: Abcam, ab5694, 1:200. Cytokeratin AE1/AE3: Leica Biosystems, PA0094, 1:100. Secondary antibodies: Alexa Fluor 647 (ThermoFisher, A32728 and A32733, 1:1000) and Alexa Fluor 555 (ThermoFisher, A32773 and A32794, 1:1000).
Validation	The relevant validations of the primary antibodies can be found on the manufacturers' websites.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*