nature research

Corresponding author(s): Nicola Aceto

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	R (a list of individual packages is provided in the methods section)
Data analysis	R (a list of individual packages is provided in the methods section), FastQC, FastQ Screen, MultiQC, Trim Galore, STAR, RSeQC, BWA, SAMtools, Picard, GATK, QualiMap, Monovar, SnpEff, SnpSift. All the software used for the analysis is free and publicly available.

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Data availability. Data analysis, statistical testing and visualization were conducted in R (version 3.4.0; R Foundation for Statistical Computing, Vienna, Austria). RNA and exome sequencing data have been deposited to Gene Expression Omnibus (GEO, NCBI; accession number GSE109761) and to the European Nucleotide Archive

(ENA, EMBL-EBI; accession number PRJEB24623), respectively. Original R scripts to reproduce data analysis have been deposited to GitHub (CA, USA; accession URL https://github.com/CMETlab/CTC-WBC). Source data for all mouse experiments are provided. All data are available from the corresponding author upon reasonable request.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	Sample size for each experiment was determined based on the level of expected heterogeneity of the samples, the significance threshold (chosen at 0.05) and the expected or observed difference.
Data exclusions	No data were excluded from the manuscript. Regarding samples, single cells that did not pass the QC step for RNA and exome sequencing were removed prior to the analysis.
Replication	All attempts at replication were successful. Two independent observers have collected the data (e.g. CTC counts, neutrophil infiltration counts).
Randomization	All mice were randomized before tumor inoculation.
Blinding	Investigators were blinded during data collection (e.g. during quantification of CTC numbers).

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants

Methods

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

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	All unique materials used are readily available from the authors or from standard commercial sources (for a list of all commercial sources, refer to the methods section).

Antibodies

Antibodies used	All antibodies used in the manuscript were purchased from commercial sources (details for each antibody are provided in the methods section). Antibodies were all validated with positive and negative controls.
Validation	Antibodies were validated individually using positive and negative controls.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	BR16 CTC-derived cell line was developed in our laboratory upon CTC isolation from a patient with metastatic breast cancer. Brx50 CTC-derived cells were obtained from Dr. Daniel Haber and Dr. Shyamala Maheswaran (Massachusetts General

	Hospital, HMS, Boston, MA) MDA-MB-231 LM2 human breast cancer cells were obtained from Dr. Joan Massague (MSKCC, NY). 4T1 murine breast cancer cells were purchased from ATCC.
Authentication	None of the cell lines have been authenticated.
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	NOD.Cg-PrkdcSCID Il2rgTM1WJL / SzJ (commonly referred to as "NSG") BALB/cJRj (commonly referred to as "BALB/c") Friend Virus B NIH Jackson (commonly referred to as "FVB") All mice used in the manuscript were females between 8-10 weeks old at the moment of tumor inoculation.
Wild animals	Our study did not involve wild animals.
Field-collected samples	Our study did not involve samples collected from the field.

Human research participants

Policy information about studies involving human research participants

Population characteristics	We obtained blood samples from 70 female patients with invasive breast cancer that discontinued their treatment due to progressive disease. Clinical features of these patients are provided in Supplementary Table 1. Patients donated blood only upon signing the written informed consent.
Recruitment	Involved patients were characterized by invasive breast cancer (all subtypes), high tumor load and progressive disease at the time of blood sampling. No biases occurred during patient selection.