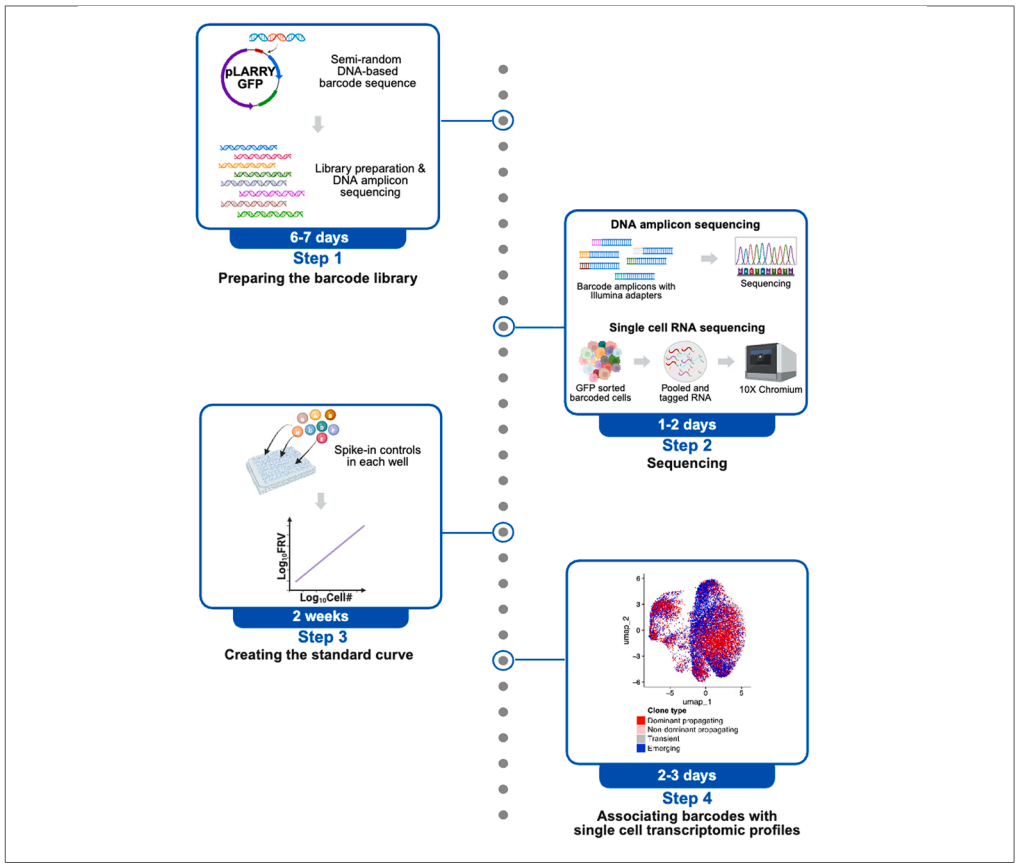
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol to track single-cell-derived clones using DNA barcoding combined with single-cell RNA sequencing



Clonal fitness and plasticity drive tumor heterogeneity contributing to disease progression and treatment resistance. Here, we provide a protocol to track the clonal outputs from uniquely marked cells. We describe steps for constructing and validating lentiviral DNA barcode libraries. We then detail procedures for single-cell RNA sequencing. This approach links clonal growth activity with transcriptomic profiles and enables permanent cellular labeling to track clonal dynamics in various model systems, which can provide insight into molecular processes underlying clone function.

Publisher's note: Undertaking any experimental protocol requires adherence to local institutional guidelines for laboratory safety and ethics.

Hee Jin Shin, Sahil Sharma, Sarah Kronheim, May Cheung, Long V. Nguyen

long.nguyen@uhn.ca

Highlights

Protocol to construct, validate, and utilize DNA barcoding for clonal tracking studies

Sample preparation for barcode DNA amplicon and single-cell RNA sequencing

Analysis pipeline to associate clone function with transcriptomic profiles

Shin et al., STAR Protocols 6, 104229
December 19, 2025
https://doi.org/10.1016/j.xpro.2025.104229

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol to track single-cell-derived clones using DNA barcoding combined with single-cell RNA sequencing

Hee Jin Shin,[1,4] Sahil Sharma,[2,4] Sarah Kronheim,[2] May Cheung,[2] and Long V. Nguyen[1,2,3,5,6,*]

[1]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada

[2]Division of Medical Oncology and Hematology, Princess Margaret Cancer Centre, University Health Network, Toronto, ON M5G 2M9, Canada

[3]Department of Medicine, University of Toronto, Toronto, ON M5S 3H2, Canada

[4]These authors contributed equally

[5]Technical contact

[6]Lead contact

*Correspondence: long.nguyen@uhn.ca
https://doi.org/10.1016/j.xpro.2025.104229

SUMMARY

Clonal fitness and plasticity drive tumor heterogeneity contributing to disease progression and treatment resistance. Here, we provide a protocol to track the clonal outputs from uniquely marked cells. We describe steps for constructing and validating lentiviral DNA barcode libraries. We then detail procedures for single-cell RNA sequencing. This approach links clonal growth activity with transcriptomic profiles and enables permanent cellular labeling to track clonal dynamics in various model systems, which can provide insight into molecular processes underlying clone function.

For complete details on the use and execution of this protocol, please refer to Nguyen et al.[1]

BEFORE YOU BEGIN

Tumors consist of heterogenous cell populations where certain cellular clones, due to their high proliferative activity, contribute disproportionately to tumor growth.[2–5] The identification and molecular characterization of these clones is crucial to devise new approaches to target these populations specifically. Single-cell RNA sequencing (scRNAseq) has enabled the study of tumor transcriptional landscapes at single cell resolution; however, it does not inherently allow tracking of clonal dynamics over time. Combining DNA barcoding with scRNAseq simultaneously enables the tracking of clonal growth within tumors and the association of these clones with their single cell transcriptomic profiles. The protocol below describes how to construct, validate and apply a lentiviral-based DNA barcode library to track the clonal outputs from single and uniquely marked cells and to associate these with their single cell transcriptomic profiles obtained from scRNAseq.

Innovation

Clone size quantification and comparison between samples and sequencing batches can be an inherent challenge of using sequencing-based readout for clonal analysis. Our approach allows for the quantification of clone size in absolute cell number. Use of spike-in controls allows for normalization of clone size based on sequence read count between multiplexed samples and even between different sequencing runs. Furthermore, these spike-in controls allow for the sensitivity and specificity of clone detection to be calculated to enable more accurate interpretation of clonal tracking analyses. While preserving the quantitative aspects of this approach, clonal growth activity can be associated with single cell transcriptomic profiles to provide molecular insight into clone function. This protocol can be applied to a wide range of studies from clonal dynamics in tumors to lineage

tracking of stem and progenitor cells in normal tissues. We have recently published a study that used this protocol to characterize the heterogeneity of clonal fitness and transcriptional plasticity of human breast cancer single-cell-derived clones in patient-derived tumor xenograft models.[1]

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
pLARRY-EGFP plasmid	Rodriguez-Fraticelli et al.[6] and Weinreb et al.[7]	RRID: Addgene_140025
psPAX2 plasmid	Didier Trono lab	RRID:Addgene_12260
pMD2.G plasmid	Didier Trono lab	RRID:Addgene_12259
ElectroMAX DH10B cells	Invitrogen Life Technologies	Cat # 18290-015
Chemicals, peptides, and recombinant proteins		
Lipofectamine 3000 reagent	Invitrogen	Cat # L3000001
Q5 high-fidelity DNA polymerase	New England Biolabs	Cat # M0491S
Lenti-X concentrator	Takara Bio	Cat # 631232
Glycogen, RNA grade	Thermo Scientific	Cat # R0551
3 M sodium acetate solution, pH 5.2	Thermo Scientific	Cat # R1181
Ethanol	Commercial Alcohols	Cat # P016EAAN
Critical commercial assays		
QIAEX II gel extraction kit	QIAGEN	Cat # 20021
PicoPure DNA extraction kit	Applied Biosystems	Cat # KIT0103
prepGEM Universal DNA extraction Kit	ForenteQ Limited	Cat # PUN0500
IDT for Illumina UD indexes	Illumina	Cat # 20091654
KAPA library quantification kit	Roche	Cat # 07960140001/KK4824
QIAGEN plasmid mini kit	QIAGEN	Cat # 12123
Qubit 1X dsDNA high sensitivity (HS) assay kits	Invitrogen	Cat # Q33230
Deposited data		
Bulk RNA sequencing raw count matrices	https://doi.org/10.5281/zenodo.10978989	RawCounts.csv
Bulk RNA sequencing normalized count matrices	https://doi.org/10.5281/zenodo.10978989	LogCPMNormCounts.csv
scRNAseq MetaCell processed count matrix	https://doi.org/10.5281/zenodo.10978989	mat.pdx_LN_v2_filt.Rda
scRNAseq MetaCell partitions	https://doi.org/10.5281/zenodo.10978989	mc.pdx_LN_v2_filt.Rda mc2d.pdx_LN_v2_filt.Rda
scRNAseq Seurat processed count matrices for 4 patient-derived tumor xenograft models	https://doi.org/10.5281/zenodo.10978989	STG139.rds STG201.rds AB040.rds IC07.rds
Experimental models: cell lines		
HEK293T cells	ATCC	CRL-11268
MDA-MB-231 cells	ATCC	HTB-26
Oligonucleotides		
Barcode sequence FWD: 5′phos – TCGACAGACTGNNATCNNGAT SSAAANNGGTNNAACNNTGTAAAACGA CGGCCAGTGAGT – 3′	Integrated DNA Technologies	N/A
Barcode sequence REV: 5′phos-CTAGACTCACTGGCCGTCGTTTT ACANNGTTNNACCNNTTTSSATCNNGA TNNCAGTCTG-3′	Integrated DNA Technologies	N/A
Staggered Primers with vector specific sequences in bold		
Pair 1 FWD primer: 5′-TCGTCGGCAGCG TCAGATGTGTATAAGAGACAG **TAGAAGGCACAGGTCGACAG**-3′	Integrated DNA Technologies	N/A
Pair 1 REV primer: 5′-GTCTCGTGGGCTC GGAGATGTGTATAAGAGACAG **GTCTAGACTCACTGGCCGTC**-3′	Integrated DNA Technologies	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Pair 2 FWD primer: 5′-TCGTCGGCAGCG TCAGATGTGTATAAGAGACAG **GCAACTAGAAGGCACAGGTC**-3′	Integrated DNA Technologies	N/A
Pair 2 REV primer: 5′-GTCTCGTGGGCT CGGAGATGTGTATAAGAGACAG **GACTCACTGGCCGTCGTTTT**-3′	Integrated DNA Technologies	N/A
Pair 3 FWD primer: 5′-TCGTCGGCAGC GTCAGATGTGTATAAGAGACAG **CAACTAGAAGGCACAGGTCG**-3′	Integrated DNA Technologies	N/A
Pair 3 REV primer: 5′-GTCTCGTGGGCT CGGAGATGTGTATAAGAGACAG **AGACTCACTGGCCGTCGTTT**-3′	Integrated DNA Technologies	N/A
Forward primer for Sanger sequencing: 5′-**ATGACACCTACTCAGACAATGC**-3′	Integrated DNA Technologies	N/A
Software and algorithms		
MetaCell version 0.3.7	Baran et al.[8]	https://tanaylab.github.io/metacell/
Seurat version 5.1.0	Satija et al.[9]	https://satijalab.org/seurat/
Custom code	This protocol	https://github.com/lvnguyen-lab/Barcode-protocol
RStudio version 2023.12.0+369	Posit Software, PBC	https://posit.co/download/rstudio-desktop/
FastQC version 0.11.9	Andrews[10]	N/A
Cutadapt version 1.10	Martin[11]	N/A
10x Genomics Cell Ranger version 7.0.1	Zheng et al.[12]	https://www.10xgenomics.com/
Clustree version 0.5.1	Zappia and Oshlack[13]	https://github.com/lazappi/clustree
scDblFinder version 1.18.0	Germain et al.[14]	https://github.com/plger/scDblFinder
dittoSeq version 1.16.0	Bunis et al.[15]	https://github.com/dtm2451/dittoSeq
Other		
Nunc square bioassay dishes	Thermo Fisher Scientific	Cat # 240835
LB agar	WISENT INC	Cat # 800-011-LG
10X TBE buffer	WISENT INC	Cat # 880-545-CL
Agarose D1-LE	WISENT INC	Cat # 800-015-EG
XbaI digestion enzyme	New England Biolabs	Cat # R0145T
SalI-HF digestion enzyme	New England Biolabs	Cat # R3138S
10X rCutSmart digestion buffer	New England Biolabs	Cat # B6004S
T4 DNA ligase with reaction buffer	New England Biolabs	Cat # M0202S
ElectroLigase	New England Biolabs	Cat # M0369
Glycerol	BioShop Canada	CAS # 56-81-5
Dulbecco's modified Eagle's medium 1X (DMEM)	WISENT INC	Cat # 319-005-CL
Fetal bovine serum (FBS)	Gibco	Cat # 12483020
10X TBE buffer, liquid	WISENT INC	Cat # 880-545 CL
Nuclease free water	VWR	Cat # SH30538.02
Ampicillin solution 0.1 g/mL	WISENT INC	Cat # 450-110 XL
Glycerol for bacterial stock	MilliporeSigma	Cat # G2025

MATERIALS AND EQUIPMENT

Cell culture medium: Supplemented DMEM		
Reagent	Final concentration	Amount
Dulbecco's Modified Eagle Medium (DMEM)	N/A	450 mL
Fetal bovine serum (FBS)	10%	50 mL
Total	**N/A**	**500 mL**

Note: Storage conditions: 4°C, up to 4 weeks.

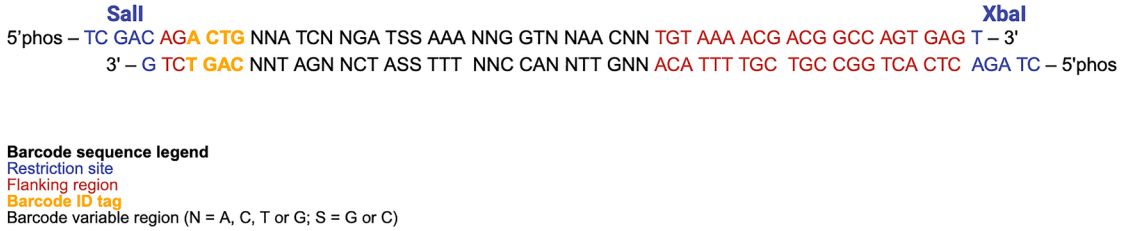


Figure 1. Semi-random barcode design

Schematic of semi-random barcode oligonucleotide design with restriction sites used for insertion and a barcode ID tag.

STEP-BY-STEP METHOD DETAILS

Constructing the DNA barcode library and validating barcode diversity

Timing: 6–7 days

This major step describes a step-by-step protocol for preparing and ligating DNA barcode oligonucleotides into a lentiviral vector pLARRY-EGFP to generate a plasmid DNA barcode library.

1. Prepare DNA oligonucleotides (Integrated DNA Technologies) for ligation into the lentiviral plasmid.

Note: This 27 bp semi-random DNA barcode was adapted from a previous design[16] to which a 4 bp unique library ID sequence was added (Figure 1). Additional libraries can be created by modifying the 4 bp library ID. This unique library ID allows for multiple libraries to be generated, and when used together, can be sequenced simultaneously and then computationally resolved.

⚠ CRITICAL: The forward and reverse oligonucleotides should be dual PAGE and HPLC purified with a 5′ phosphorylation modification.

a. Anneal the forward and reverse oligonucleotides by first dissolving the oligonucleotides in nuclease-free water to a concentration of 200 μM. Mix 25 μL of each oligonucleotide separately with 25 μL of 1X T4 DNA ligase buffer (1:1 mix; final concentration of T4 DNA ligase buffer should be 0.5X).
b. Combine 50 μL of forward and reverse oligonucleotide mix (100 μL in total).
c. In a PCR thermocycler, heat the oligonucleotide mixture to 95°C for 5 min, and cool gradually at a rate of 0.5°C per min to reach 10°C.
d. Run the annealed oligonucleotides on a 4% agarose-TBE gel (120 V constant).

Note: Unannealed oligonucleotides should be run as control to accurately predict the size of annealed oligonucleotides (60 bp). Leave a gap between samples if more than one library is being prepared to avoid cross-contamination.

e. Perform gel extraction as per manufacturer's instructions using the QIAEX II Gel extraction kit (Qiagen) and elute the annealed oligonucleotides into 100 μL nuclease-free water.
f. Perform ethanol precipitation of the DNA by adding the following reagents to the 100 μL eluate in the order as indicated in Table 1, then:
   i. Vortex 3 sec to mix and freeze at −80°C for 16–24 h to precipitate the DNA.

Pause point: The reaction can be stored at −80°C for longer, but it is recommended to proceed with the next steps after incubation for 16–24 hrs.

Table 1. DNA ethanol precipitation

Reagent	Volume	Remarks
Glycogen (5 mg/mL)	1 uL	50 μg/mL (final)
3M Sodium acetate (pH 5.2)	10 uL	1/10 volume
Ethanol (100%)	300 uL	3 volumes

ii. The next day, centrifuge at maximum speed (~18,000 x *g*) at 4°C for 30 min to obtain a white pellet of the precipitated DNA.
iii. Carefully discard the supernatant without disturbing the DNA pellet.
iv. Add 1 mL of 70% cold ethanol to wash the pellet.
v. Centrifuge again at maximum speed (~18,000 x *g*) at 4°C for 10 min.
vi. Carefully remove the supernatant completely and let the pellet dry to room air for 5–15 min.

Note: Circle around the pellet with a marker pen as the pellet may become invisible when dry.

vii. Resuspend the pellet in 20 μL nuclease-free water and store at −20°C.
viii. Quantify the DNA using a Qubit fluorometer (Thermo Fisher Scientific) and the Qubit 1X dsDNA high sensitivity assay kit.

Note: This step yields approximately 10–15 μg of total DNA.

Pause point: The annealed oligonucleotides can be stored at −20°C until ready to proceed.

2. Prepare the pLARRY-EGFP vector (Addgene) for ligation with annealed barcode oligonucleotides as follows:
   a. Double-digest the pLARRY-EGFP vector using SalI and XbaI restriction endonuclease sites in the 3′ UTR of EGFP as indicated in Table 2.
   b. Incubate the reaction at 37°C for 20 min followed by heat inactivation of enzymes at 65°C for 20 min.

   Note: Only use 1 μg of plasmid DNA per reaction to achieve complete digestion. Additional reactions can be included and subsequently pooled together to increase the final yield of the digested vector as needed.

   c. Run the double-digested vector and the following controls on a 0.7% agarose-TBE gel (120 V constant):
      i. Undigested pLARRY-EGFP plasmid.
      ii. pLARRY-EGFP plasmid digested with either SalI or XbaI (single cut linearized).
   d. Perform gel extraction of the double-digested vector using QIAEX II Gel Extraction Kit and elute the DNA into 100 μL nuclease-free water.
   e. Perform ethanol precipitation of the DNA as described in step 1f in constructing the DNA barcode library and validating barcode diversity.

Table 2. Restriction digestion reaction

Reagent	Amount
Vector (pLARRY-EGFP)	1 ug
10X rCutSmart buffer	5 uL
SalI	1 uL
XbaI	1 uL
Water (nuclease-free)	Up to 50 uL

Table 3. Ligation reaction

Reagent	Amount
Annealed oligonucleotides: double-digested vector (3:1)	10 uL
10X T4 DNA ligase buffer	2 uL
T4 DNA ligase	1 uL
Water	Up to 20 uL

f. Resuspend the double-digested vector in 20 μL of nuclease-free water and quantify the concentration of DNA using a Qubit fluorometer (Thermo Fisher Scientific) and the Qubit 1X dsDNA high sensitivity assay kit.

Note: This step should yield approximately 0.2–0.5 μg of total DNA per reaction.

Pause point: The double digested vector can be stored at −20°C until ready to proceed.

3. Ligate annealed barcode oligonucleotides into 3′ UTR of EGFP in the pLARRY-EGFP vector.
   a. Prior to ligation, check the purity of the annealed oligonucleotides and the double-digested vector by running the Agilent Bioanalyzer.
   b. Mix the annealed barcode oligonucleotides with double-digested pLARRY-EGFP vector in a 3:1 molar ratio (10 ng of double-digested vector with 3X molar ratio of annealed oligonucleotides) to a final reaction volume of 10 μL.
   c. Perform ligation with T4 DNA ligase (New England Biolabs) as indicated in Table 3, then:
      i. Gently mix by pipetting, spin down and incubate at 16°C for 16–24 hrs.

⚠ CRITICAL: Do not pause here. After incubation, proceed to the next step.

      ii. Heat inactivate at 65°C for 10 min and chill on ice.
      iii. For the ligation step, include a negative control containing the double digested vector without annealed oligonucleotides.

Note: The negative control tests the frequency of single digested vector present that can re-circularize even without annealed oligonucleotides.

4. Transform the ligated product and negative control into ElectroMAX DH10B electrocompetent cells (Invitrogen) as per manufacturer's instructions.
   a. For each transformation reaction, add 1 μl of ligated product per 20 μL of electrocompetent cells and perform electroporation. Prepare control conditions as below:
      i. **Positive control:** undigested pLARRY-EGFP plasmid.
      ii. **Negative control:** double digested pLARRY-EGFP plasmid with no insert.

⚠ CRITICAL: Do not discard remaining ligated product as it will be used for additional transformations in following steps.

   b. Add 1 mL of SOC media to each transformation reaction and incubate at 37°C for 1 h.
   c. Immediately following incubation for 1 h, plate different volumes of transformed cells (i.e., 50 μL, 100 μL, 150 μL and 200 μL) onto 100 mm LB-agar plates containing 100 μM ampicillin and incubate at 37°C for 20 h.

⚠ CRITICAL: Do not discard remaining transformed cells (store at 4°C) as they will be pooled with additional transformations for plating on 245 mm × 245 mm square LB-agar plates.

d. 20 h after plating, examine plate density and proportionally calculate volume for plating on 245 mm × 245 mm square LB-agar plates containing 100 μM ampicillin using the volume that resulted in dense, single colonies on the 100 mm plates.
e. Perform 19 additional transformations with the remaining 19 μL of the ligated product to maximize the barcode library diversity.
f. Pool all transformation reactions together and plate on an appropriate number of 245 mm × 245 mm square LB-agar plates containing 100 μM ampicillin based on the plating volume determined in steps 4c and 4d in constructing the DNA barcode library and validating barcode diversity. Incubate the plates at 37°C for 20 h.

Note: Increasing the number of plates used, and thus the number of bacterial colonies formed will increase the diversity of the barcode library.

g. After 20 h incubation, count colonies on a representative section of the square plates to approximate the potential diversity of the barcode library. To prepare a homogenous bacterial mixture and to prepare glycerol stocks:
   i. Use 100–800 μL of LB broth containing 100 μM ampicillin for each 245 mm × 245 mm square LB-agar plate to facilitate scraping and combining of the colonies.
   ii. Measure the final volume of the resulting bacterial mixture. Add 1:1 volume of 50% glycerol to the bacterial mixture to prepare glycerol stocks. Gently mix thoroughly, aliquot the glycerol stock into 1 mL cryovials, and store at −80°C.
   iii. Extract plasmid DNA from each glycerol stock using the QIAGEN Plasmid Mini Kit (Qiagen) according to the manufacturer's instructions.

5. Prepare the plasmid DNA library for DNA amplicon sequencing as described in steps 8–15 in multiplexed barcode DNA amplicon sequencing. Sequence plasmid libraries to a depth of at least 10X the anticipated diversity (e.g., 10 million reads per library for an expected diversity of 1 million unique barcodes).
6. Package lentiviruses with HEK293T cells and the barcode library plasmid DNA using the Lipofectamine 3000 kit (Invitrogen) as per manufacturer's instructions.
   a. Concentrate lentiviral supernatant with Lenti-X concentrator (Takara Bio).
   b. Test and calculate lentiviral titer using an appropriate cell line.

⚠ CRITICAL: It is important to optimize lentiviral transduction conditions to achieve a MOI of less than 0.3 to minimize the likelihood for multiple barcode integrations.

7. Validate the diversity of the barcode library by transducing an appropriate cell line (we used MDA-MB-231 cells, a human breast cancer cell line).
   a. In triplicate, transduce 10^5 to 10^6 MDA-MB-231 cells with the lentivirus to target a MOI of less than 0.3.
   b. Culture the cells for 24 h after transduction and harvest for DNA amplicon sequencing as described in steps 8–15 in multiplexed barcode DNA amplicon sequencing.

Multiplexed barcode DNA amplicon sequencing

Timing: 1–2 days

This major step describes the preparation of plasmid DNA or barcoded cells for DNA amplicon sequencing.

8. If starting from cells, extract and purify genomic DNA (gDNA) from the cells (maximum of 2×10^5 cells per sample using the prepGEM Universal DNA extraction kit (ForenteQ Limited) or PicoPure DNA extraction kit (Thermo Fisher). Ethanol precipitate the purified genomic DNA as described

Table 4. PCR #1 reaction

Reagent	Amount	Final concentration
DNA template	50 ng	2 ng/uL
Q5 DNA polymerase	0.25 uL	N/A
10 μM FWD primer	1.25 uL	0.5 uM
10 μM REV primer	1.25 uL	0.5 uM
5X Q5 buffer	5 uL	1X
10 mM dNTPs	0.5 uL	0.2 mM
Nuclease-free water	Up to 25 uL	N/A

in step 1f in constructing the DNA barcode library and validating barcode diversity. If starting from plasmid DNA, use 50 ng DNA per sample, which can be used directly without ethanol precipitation.

9. Perform PCR to amplify the barcode sequence of interest with primers that contain the i7 and i5 linker sequence compatible with Illumina sequencing (see key resources table for information on primers).
   a. Carry out PCR #1 (Tables 4 and 5) to amplify each sample using one of the three staggered primer sets to increase the diversity of nucleotides represented in a given sequencing cycle to improve cluster recognition and reduce the amount of PhiX spike-in DNA required.

   Note: Use one primer set per sample. Evenly distribute the 3 primer sets across the samples per sequencing batch.

10. Top up the 25 μL product from PCR #1 up to 100 μL with nuclease-free water and ethanol precipitate as described in step 1f in constructing the DNA barcode library and validating barcode diversity. Resuspend precipitated DNA in 15 μL nuclease-free water.
11. Perform PCR #2 (Tables 6 and 7) to multiplex the samples with Illumina unique dual indexes.

    Note: Each well or sample should include a unique pair of indexes for multiplexing; choose indexes for each based on the Illumina pooling guide.

12. Run PCR #2 product on a 1.5% agarose gel (120 V constant).
13. Gel extract and ethanol precipitate each PCR product as described in step 1f in constructing the DNA barcode library and validating barcode diversity and resuspend in 15 μL nuclease-free water.
14. Quantify the PCR product by qPCR using the KAPA library quantification kit (Roche) as per manufacturer's instructions. Perform qPCR (in triplicate) using 3 dilutions ($1/10^4$, $1/10^5$ and $1/10^6$).
15. Pool together each multiplexed library in equimolar ratio and sequence on the Illumina MiSeq or other sequencing platform using 2×75 bp paired-end sequencing with at least 10% PhiX spike-in.

    Note: We recommend a sequencing depth of 1 million reads per experimental sample expected to contain hundreds to thousands of barcodes. For validation of the barcode library diversity, we recommend at least 10X read depth based on number of unique barcodes expected (i.e., at least 10 million reads for an expected library diversity of 1 million barcodes).

Table 5. PCR #1 cycling conditions

Steps	Temperature	Time	Cycles
Initial denaturation	98 °C	30 sec	1
Denaturation	98 °C	10 sec	25
Annealing	67 °C[a]	30 sec	
Extension	72 °C	30 min	
Final extension	72 °C	5 min	1
Hold	10 °C	Infinite	

[a]Annealing temperature can vary based on the primer design.

Table 6. PCR #2 reaction

Reagent	Amount	Final concentration
DNA template	1.5 µL of PCR #1 product	N/A
Q5 high-fidelity DNA polymerase	0.25 uL	N/A
IDT for Illumina UD Indexes	2.5 uL	1 uM
5X Q5 buffer	5 uL	1x
10 mM dNTPs	0.5 uL	0.2 mM
Nuclease-free water	Up to 25 uL	N/A

⚠ CRITICAL: 10% PhiX spike-in was used on the Illumina MiSeq platform. Other sequencing platforms may require a larger amount up to 25%. It is recommended to review this requirement with your sequencing center.

Preparing spike-in controls for clone size normalization

Timing: 2 weeks

This major step describes the preparation and validation of spike-in controls for multiplexed DNA amplicon sequencing.

Note: Spike-in controls should be generated using a separate barcode library with a unique 4 bp library ID.

16. Clone the spike-in control barcode oligonucleotides into pLARRY-EGFP as described in steps 1–3 in constructing the DNA barcode library and validating barcode diversity.
17. Perform electroporation for transformation of plasmid DNA into ElectroMAX DH10B electrocompetent cells as described in steps 4a-4b in constructing the DNA barcode library and validating barcode diversity.
18. Pick single colonies using a sterile pipette tip, inoculate in 5 mL of LB broth with 100 µM ampicillin and grow at 37°C for 16–20 h for expansion and plasmid DNA extraction.
    a. Extract plasmid DNA using the QIAGEN Plasmid Mini Kit (Qiagen) according to the manufacturer's instructions.
    b. Quantify DNA concentration using the Qubit fluorometer and the Qubit 1X dsDNA high sensitivity assay kit.

⚠ CRITICAL: Pick at least twice the number of colonies as the number of desired spike-in controls (i.e., if 6 controls desired, pick 12 colonies) to be able to choose the barcode clones with the greatest edit distance from each other.

Pause point: Freeze DNA at −20°C until use.

19. Submit an aliquot of each mini-prep product for Sanger sequencing; follow guidelines of the appropriate sequencing center for submission. Forward primer 5′-ATGACACCTACTCAGA CAATGC-3′ should be used to sequence the region containing the barcode.

Table 7. PCR #2 cycling conditions

Steps	Temperature	Time	Cycles
Initial denaturation	98 °C	30 sec	1
Denaturation	98 °C	10 sec	5
Annealing	62 °C	30 sec	
Extension	72 °C	30 min	
Final extension	72 °C	5 min	1
Hold	10 °C	Infinite	

20. From the Sanger sequencing results, identify and record the barcode sequence for each control plasmid. Looking at the chromatograph, make a note of and discard any sequences that either do not contain a barcode or contain multiple barcodes.
21. Compare all retained barcode sequences and determine their average edit distance from each other. Rank the sequences by decreasing similarity and use the most diverse sequences for the spike-in controls.
22. Prepare lentivirus and transduce MDA-MB-231 or an appropriate cell line of choice as described in steps 6–7 in constructing the DNA barcode library and validating barcode diversity.

Note: Ensure a MOI of less than 0.1 for this step to further minimize the possibility of multiple integrations which can affect downstream data normalization and quantification steps.

23. FACS sort for GFP+ transduced cells, expand in tissue culture and freeze aliquots for future use.
24. FACS sort spike-in controls into a 96-well conical bottom PCR plate at different cell numbers corresponding to each control.
    a. To each well, add known cell numbers of each spike-in control.

Note: For example, each well should have: 10 cells with control barcode 1; 20 cells with control barcode 2; 50 cells with control barcode 3; 100 cells with control barcode 4; 500 cells with control barcode 5; and 2,500 cells with control barcode 6.

Pause point: Freeze plate at −20°C until ready to proceed with next steps.

25. Include 3–6 spike-in control-only samples for each round of multiplexed DNA amplicon sequencing, to establish a standard curve to associate read counts from sequencing with clone size in absolute cell number.

Note: These spike-in control-only samples can also be used to estimate the sensitivity and specificity of clone detection and provide a representation of the amount of noise (i.e., unexpected sequence reads) from library preparation and sequencing.

26. For spike-in control-only samples, extract gDNA from the cells in each well directly and proceed with multiplexed barcode DNA amplicon sequencing as described in steps 8–15 in multiplexed barcode DNA amplicon sequencing.
27. For experimental samples, use the buffer from the gDNA extraction kit to resuspend the cells.

Note: Use the 1X BLUE buffer if using the prepGEM Universal DNA extraction kit (ForenteQ), or the Extraction Solution if using the PicoPure DNA extraction kit (Thermo Fisher).

    a. Resuspend and transfer the contents to a well containing spike-in controls to combine the experimental sample with spike-in controls.
    b. Proceed with gDNA extraction and multiplexed barcode DNA amplicon sequencing.

Computational analysis for barcode DNA amplicon sequencing

Timing: 2 days

This major step describes the computational analysis steps using the code available on GitHub (https://github.com/lvnguyen-lab/Barcode-protocol/) for assigning sequence reads to each unique barcode sequence and for calculating the clone size of experimental barcode clones in absolute cell number. It is recommended to use a high-performance computing cluster environment for this analysis. Adjust the custom shell scripts to your computational environment, sequencing file names, and

requisite script location. Ensure that the requisite scripts are in the correct location within your environment.

28. After sequencing, computationally demultiplex the multiplexed samples into separate FASTQ files based on Illumina Unique Dual (UD) indexes as per Illumina guidelines.
29. In the working directory, create a folder named "`runs`" and a folder named "`scripts`" as follows:

```
>mkdir runs

>mkdir scripts
```

30. In the "`scripts`" folder, save the following files:
    a. `Pair_amplicon_fastqs.py`
    b. `Levenshtein.sh`
    c. `Levenshtein_txt.R`
31. Navigate into the "`runs`" folder and create a folder for the current analysis, for example:

```
>cd runs

>mkdir Analysis1
```

32. Navigate into the analysis folder and create a folder named "`fastq`" with a subfolder named "`amplicon`" as follows:

```
>cd Analysis1

>mkdir -p fastq/amplicon
```

33. Save the demultiplexed Illumina fastq files in the new "`amplicon`" folder.
34. Edit and save the file "`amplicon_analysis.sh`" in the analysis folder "`Analysis1`" with the following information:
    a. The base directory for where this analysis will be run, for example:

```
BASE_DIR=/cluster/projects/nguyengroup/scRNAseq
```

    b. The run name for this analysis (this will be a separate folder within the "`runs`" folder created in step 29 above), for example:

```
RUN_NAME=Analysis1
```

    c. Indicate whether the sequencing data is paired or unpaired, for example:

```
TYPE=paired
```

    d. Indicate the Phred base quality score with which to filter the sequencing data. We recommend using a score of 30, as follows:

```
QUALITY=30
```

35. Run the script "`amplicon_analysis.sh`" as follows:

```
>sbatch amplicon_analysis.sh
```

36. This custom code is used for quality control, to trim, filter and count reads to determine the number of reads for each unique barcode sequence. Barcodes within 1 edit distance are also grouped together.
    a. This script will output two csv files with sequence and read count for each barcode; one with barcodes within 1 edit distance grouped together and one without grouping.
    b. The output path will be in the "`RUN_NAME`" subfolder called "`Final_counts`" and the naming convention of the output files will have the suffix "`_post-Levenshtein.csv`" for the grouped output and "`pre-Levenshtein.csv`" for the ungrouped output.

Note: We recommend using the grouped output file to reduce noise from errors of 1 edit distance that can be generated during PCR amplification and sequencing.

37. To calculate clone size in absolute cell number for each sequencing run, generate a log-log plot to associate the log (Fractional read value, FRV) with log (cell dose) using all the spike-in controls for all multiplexed samples (Figure 2).

⚠ CRITICAL: There should be one plot generated per sequencing run.

$$\text{Note : FRV} = \frac{\text{read count corresponding to a barcode sequence of interest in a single multiplexed sample}}{\text{sum of read counts for all spike} - \text{in controls in a single multiplexed sample}}$$

    a. Using this association plot, calculate clone size for experimentally detected barcode clones (i.e., not spike-in controls) for which the read count from sequencing is known but clone size is not.

⚠ CRITICAL: This step analyzes the barcode data using absolute clone size. It is not an output of the custom code and will have to be performed manually using the barcode data in the output csv files from step 36 in computational analysis for barcode DNA amplicon sequencing.

Note: In the FRV formula the numerator will differ for each unique barcode sequence in the sample. The denominator will be the same for all FRV calculations within a multiplexed sample and therefore allow for normalization between multiplexed samples based on the number of reads represented by the spike-in controls added similarly to each multiplexed sample.

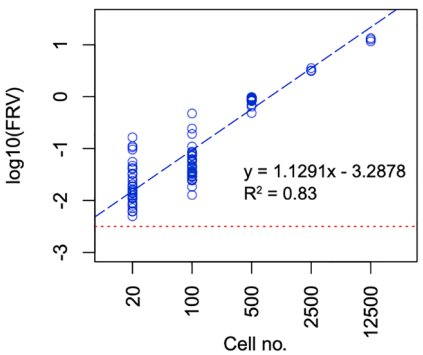


Figure 2. Clone size quantification
A plot of log10(FRV) vs. cell number to calculate absolute clone size for experimentally detected clones. Each data point represents a single spike-in control (for both spike-in control-only samples, and for experimental samples containing spike-in controls). The plot includes spike-in controls across all samples included in a single multiplexed sequencing run. The correlation equation is used to calculate the size of experimentally detected clones from the FRV. The dashed blue line indicates the correlation between log10(FRV) and cell number. The dotted red line indicates the threshold of clone detection, above which only expected spike-in controls are detected (i.e., no noise).

Single-cell RNA sequencing and barcode association

Timing: 2–3 days

This major step associates the single-cell transcriptomic data with the corresponding cellular DNA barcode using the code available on GitHub (https://github.com/lvnguyen-lab/Barcode-protocol/). It is recommended to use a high-performance computing cluster environment for this analysis. Adjust the custom script to your computational environment, sequencing file names, and requisite script location. Ensure that the requisite scripts are in the correct location within your environment.

38. FACS sort a single cell suspension of human tumor cells to exclude DAPI+ cells and to enrich for GFP+ barcoded cells.
39. Prepare a library of sorted cells on the 10X Chromium using standard 3′ v3.1 chemistry or 3′ v3.1 HT chemistry to target recovery of 1×0^4 and 2×10^4 cells, respectively, prior to sequencing on the Illumina NovaSeq S4 flowcell targeting 2×10^4 reads per cell.
40. Analyze scRNAseq data as per institutional protocol. Use either MetaCell[8] (https://tanaylab.github.io/metacell/) or Seurat[9,17] (https://satijalab.org/seurat/) for analysis to generate a single cell count matrix.
41. In the "`scripts`" folder created in step 29 in computational analysis for barcode DNA amplicon sequencing, save the following files:
    a. `associate_barcodes_with_10x_cells.R`
    b. `Levenshtein.R`
    c. `fa_csv.R`
42. In the analysis folder "`Analysis1`" created in step 31 in computational analysis for barcode DNA amplicon sequencing, edit and save the "`assoc_bc_steps.sh`" file with the following information:
    a. The base directory for where this analysis will be run (this should be the same as in step 34 in computational analysis for barcode DNA amplicon sequencing), for example:

```
BASE_DIR=/cluster/projects/nguyengroup/scRNAseq
```

    b. The run name for this analysis (this should be the same as the "`RUN_NAME`" in step 34 and the analysis folder "`Analysis1`" created in step 31 in computational analysis for barcode DNA amplicon sequencing), for example:

```
RUN_NAME=Analysis1
```

43. Create subfolders in the analysis folder "`Analysis1`" created in step 31 in computational analysis for barcode DNA amplicon sequencing as follows:
    a. In the "`fastq`" folder created in step 32 above, create two subfolders named "`raw`" and "`combined`" as follows:

```
>cd fastq

>mkdir raw

>mkdir combined
```

    b. Navigate back to the analysis folder "`Analysis1`" and create a "`metadata`" subfolder:

```
>cd ../

>mkdir metadata
```

44. Save all scRNAseq fastq files in the newly created "`raw`" folder.
45. Save full cellranger output folders for each sample in the analysis folder "`Analysis1`" created in step 31 in computational analysis for barcode DNA amplicon sequencing.
46. Create a metadata csv file with information on all the samples you wish to analyze together. It should be saved in "`$BASE_DIR/runs/$RUN_NAME/metadata`" with the following structure:

Pool	Barcode	Sequence	Sample name
SLX-22131	SITTD10	ATGCGAATGG-ACAAGTGTCG	Xenograft1
SLX-22131	SITTF10	CCGGCAACTG-CGGTTTAACA	Xenograft2

⚠ CRITICAL: Ensure there is only one metadata csv file in the metadata folder. All samples for the run can be included in this one file. Also ensure that the sample names in the metadata csv file match the names of the Cell Ranger output folders.

Note: "Pool" indicates the name of the sequencing batch. "Barcode" indicates the UD index used for sample multiplexing. "Sequence" are the UD index sequences. "Sample name" is the name of the sample submitted for sequencing.

47. Run the custom analysis pipeline "`assoc_bcs_steps.sh`". This pipeline associates barcodes with the cell ID in the single cell count matrix and includes the following steps:
    a. Pooling reads for each individual sample from multiple fastq files from sequencing.
    b. Cutadapt to trim read2 sequences based on expected constant sequences flanking the barcode and also filter for size and quality of the reads.
    c. Convert fastq files into csv files then tally unique read counts from csv files.
    d. Create fasta files from the csv files. The fasta files are used for BWA alignment of the original fastq files to the fasta reference, which provides information about cell barcode index. The output for this is a sam file which is used for associating barcode sequences with the cell ID in the single cell count matrix.

```
>sbatch -a 1-4 assoc_bcs_steps.sh
```

Note: The "-a 1-4" component of the above command reflects a batch run with the slurm cluster system. The numbers should match the number of samples to be run, with "1-4" corresponding to samples 1 to 4 inclusive in the order of the metadata csv. To only run sample 1 the command would be "-a 1", to run samples 2 and 4 the command would be "-a 2,4", and to run all 4 but limit the number of concurrent jobs on the cluster to 2 the command would be "-a 1-4%2". This script will output 2 csv files per sample. The file name with suffix "`_cell_bc_umis_count.csv`" contains the number of barcode reads detected in each cell ID in the single cell count matrix. The file name with suffix "`_cell_GFP_umis_count.csv`" contains the number of GFP reads detected in each cell ID in the single cell count matrix. These output files will be saved in the "`report`" folder in the working directory.

EXPECTED OUTCOMES

The barcode design (Figure 1) has the potential to generate up to 4,194,304 unique barcodes (4^N x 2^S). Using a lentiviral plasmid, the barcode sequences will be permanently integrated into the genome of the cell and passed down to all subsequent progenies. From a single vial of 1 mL glycerol stock of the barcode library, approximately 30–40 μg of plasmid DNA can be extracted, but this may vary depending on the plate density as well as DNA extraction efficiency. The spike-in controls will be used as a reference point to quantify clone sizes, however, it is critical that each sample analyzed includes the spike-in controls. These controls provide a measure of sensitivity

and specificity of barcode clone detection and provide a way to normalize sequence read count to calculate absolute clone sizes between different samples and sequencing runs. Computational analyses will ultimately provide transcriptomic profiles of unique clones by pairing scRNAseq data with barcode clones.

LIMITATIONS

To reduce the incidence of multiple barcode integrations per cell, we recommend a transduction efficiency below 30% (or MOI less than 0.3) for clonal tracking experiments. Without a period of *in vitro* culture to express GFP to select for transduced cells, this limits the number of clones tracked in subsequent experiments. However, by culturing first, the uniquely barcoded cells divide, and clonal outputs assayed are no longer representative of the activity from a single cell, but rather the sum of multiple cells. This can lead to different biological findings depending on the experimental context. Another limitation is the low capture efficiency inherent in scRNAseq. Barcode clones can only be associated with single cell RNA profiles where a barcode transcript is detected by sequencing. Some barcoded cells may be misclassified as non-barcoded and the clonal tracking data for those cells lost. Lastly, scRNAseq is generally limited to analysis of a few hundred to a few thousand cells per sample (after filtering low quality cells) making small clones with few barcoded cells less likely detected by scRNAseq.

TROUBLESHOOTING

Problem 1

Low viral titers with the pLARRY-EGFP plasmid.

Potential solution

Ensure that HEK293T cells are healthy and used at a low passage number. Consider obtaining a new vial of HEK293T cells from a reliable source. At the time of transfection, ensure that the cells are in log growth phase. Collect and pool the supernatant at additional time points (e.g., 24, 48 and 72 h after transfection) and concentrate the pooled viral supernatant. Concentration with ultracentrifugation may yield higher titers.

Problem 2

Observing a difference in barcode diversity between aliquots of the barcode library.

Potential solution

Perform plasmid purification of all glycerol stocks for the DNA barcode library together and pool the plasmid to generate a single large barcode plasmid library. Perform DNA amplicon sequencing on this library that will be representative of the single pooled library rather than individual aliquots that may show variably distributed barcode representation.

Problem 3

Overgrowth of colonies on large square 245 mm × 245 mm plates.

Potential solution

Prior to plating, test colony forming efficiency (number of colonies per volume plated) on 100 mm plates. Test different volumes such as 50 μL, 100 μL, and 200 μL. Leave plates at 37°C for 16–24 h and count the number of colonies. Calculate the surface area for each plate to determine the ideal volume for plating onto large square plates.

Problem 4

HEK293T cells detach during lentivirus production.

Potential solution

Use healthy HEK293T cells and consider obtaining a new vial of HEK293T cells from a reliable source. Ensure cells are well-distributed at time of seeding. Consider optimizing the protocol to use solutions such as Poly-L-lysine to coat the plate prior to seeding.

Problem 5

Insufficient production of ligated product after insertion of annealed barcode sequence into the double digested pLARRY-EGFP plasmid.

Potential solution

Replace T4 ligase with ElectroLigase (NEB) and use their recommended protocol for ligation and transformation. The same DH10B ElectroMAX electrocompetent cells can be used for transformation.

Problem 6

Difficulty running the computational analysis pipeline.

Potential solution

We recommend running the computational analysis pipeline on a high-performance computing cluster environment with at least 180 GB memory, and 10 CPU cores, depending on the script being run. This information is included at the top of the custom scripts "`amplicon_analysis.sh`", "`Levenshtein.sh`", and "`assoc_bcs_steps.sh`". Ensure to adjust the custom script based on rules from your local high performance computing cluster environment. Ensure that the requisite scripts are in the correct location within your environment. Ensure that the directory structure is exactly as noted and the fastq and cellranger files are where they should be. If the "`Levenshtein_txt.R`" script does not finish running in the time allotted by "`Levenshtein.sh`", try increasing this time allotment. If the cluster environment has time limits that do not allow the script to finish, downstream analysis can be run on all barcode sequences without grouping sequences within 1 edit distance; these can be found in the "`_Pre-Levenshtein.csv`" amplicon output.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Long V. Nguyen (long.nguyen@uhn.ca).

Technical contact

Technical questions on executing this protocol should be directed to and will be answered by the technical contact, Long V. Nguyen (long.nguyen@uhn.ca).

Materials availability

All unique reagents generated in this study are available from the lead contact with a completed materials transfer agreement.

Data and code availability

- Custom code used for DNA amplicon sequencing and scRNAseq analysis as reported in this protocol are available on Zenodo: https://doi.org/10.5281/zenodo.17372044 and GitHub: https://github.com/lvnguyen-lab/Barcode-protocol.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

L.V.N. is a Hold'em for Life Early Career Professor in Cancer Research in the Temerty Faculty of Medicine, University of Toronto, and was supported by an ESMO Translational Research Fellowship, an ASCO Conquer Cancer Young Investigator Award, a J.P. Bickell Foundation Medical Research Grant, a MOHCCN Clinician-Scientist Award, and the Allan Slaight Breakthrough Fund at The Princess Margaret Cancer Foundation. H.J.S. is supported by an MBP Excellence UTF scholarship from the Department of Medical Biophysics, University of Toronto. S.S. is supported by a Princess Margaret Postdoctoral Fellowship from The Princess Margaret Cancer Foundation and George Knudson Memorial

Fellowship Fund. S.K. is supported by a Research Training Award from the Canadian Cancer Society. The graphical abstract and figures were created using Biorender.com.

AUTHOR CONTRIBUTIONS

L.V.N. conceived the protocol with input from S.K. and M.C. H.J.S. prepared the figures. H.J.S., S.S., S.K., and L.V.N. wrote the manuscript. S.K. and L.V.N. prepared the custom code for analysis. All authors read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Nguyen, L.V., Eyal-Lubling, Y., Guerrero-Romero, D., Kronheim, S., Chin, S.-F., Manzano Garcia, R., Sammut, S.-J., Lerda, G., Lui, A.J.W., Bardwell, H.A., et al. (2025). Fitness and transcriptional plasticity of human breast cancer single-cell-derived clones. Cell Rep. *44*, 115699. https://doi.org/10.1016/j.celrep.2025.115699.

2. Al-Hajj, M., Wicha, M.S., Benito-Hernandez, A., Morrison, S.J., and Clarke, M.F. (2003). Prospective identification of tumorigenic breast cancer cells. Proc. Natl. Acad. Sci. USA *100*, 3983–3988. https://doi.org/10.1073/pnas.0530291100.

3. Nguyen, L.V., Vanner, R., Dirks, P., and Eaves, C.J. (2012). Cancer stem cells: an evolving concept. Nat. Rev. Cancer *12*, 133–143. https://doi.org/10.1038/nrc3184.

4. Bhang, H.e.C., Ruddy, D.A., Krishnamurthy Radhakrishna, V., Caushi, J.X., Zhao, R., Hims, M.M., Singh, A.P., Kao, I., Rakiec, D., Shaw, P., et al. (2015). Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. Nat. Med. *21*, 440–448. https://doi.org/10.1038/nm.3841.

5. Sun, X.x., and Yu, Q. (2015). Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. Acta Pharmacol. Sin. *36*, 1219–1227. https://doi.org/10.1038/aps.2015.92.

6. Rodriguez-Fraticelli, A.E., Weinreb, C., Wang, S.-W., Migueles, R.P., Jankovic, M., Usart, M., Klein, A.M., Lowell, S., and Camargo, F.D. (2020). Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. Nature *583*, 585–589. https://doi.org/10.1038/s41586-020-2503-6.

7. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*, eaaw3381. https://doi.org/10.1126/science.aaw3381.

8. Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., and Tanay, A. (2019). MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. *20*, 206. https://doi.org/10.1186/s13059-019-1812-2.

9. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502. https://doi.org/10.1038/nbt.3192.

10. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

11. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. j. *17*, 10. https://doi.org/10.14806/ej.17.1.200.

12. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.

13. Zappia, L., and Oshlack, A. (2018). Clustering trees: a visualization for evaluating clusterings at multiple resolutions. GigaScience *7*, giy083. https://doi.org/10.1093/gigascience/giy083.

14. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W., and Robinson, M.D. (2021). Doublet identification in single-cell sequencing data using scDblFinder. F1000Res. *10*, 979. https://doi.org/10.12688/f1000research.73600.2.

15. Bunis, D.G., Andrews, J., Fragiadakis, G.K., Burt, T.D., and Sirota, M. (2021). dittoSeq: universal user-friendly single-cell and bulk RNA sequencing visualization toolkit. Bioinformatics *36*, 5535–5536. https://doi.org/10.1093/bioinformatics/btaa1011.

16. Gerrits, A., Dykstra, B., Kalmykowa, O.J., Klauke, K., Verovskaya, E., Broekhuis, M.J.C., de Haan, G., and Bystrykh, L.V. (2010). Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood *115*, 2610–2618. https://doi.org/10.1182/blood-2009-06-229757.

17. Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., and Satija, R. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat. Biotechnol. *42*, 293–304. https://doi.org/10.1038/s41587-023-01767-y.