UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Defining The Teratoma as a Model for Multi-Lineage Human Development

Permalink

https://escholarship.org/uc/item/0d58w2qt

Author

McDonald, Daniella Nicole

Publication Date

2021

Supplemental Material

https://escholarship.org/uc/item/0d58w2qt#supplemental

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Defining the Teratoma as a Model for Multi-Lineage Human Development

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biomedical Sciences

by

Daniella Nicole McDonald

Committee in Charge:

Professor Prashant Mali, Chair
Professor Karl Willert, Co-Chair
Professor Louise Laurent
Professor Alysson Muotri
Professor Kun Zhang

2021

Copyright

Daniella Nicole McDonald, 2021

All rights reserved.

The dissertation of Daniella Nicole McDonald is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

This dissertation is dedicated to my transgender brothers and sisters. Girls like us don't achieve things like this so I have to pinch myself every day that I have made it this far. I could not have done it without the support of all of you. This one is for you! Cheers!

EPIGRAPH

“She was fierce, she was strong, she wasn’t simple.

She was crazy and sometimes she barely slept.

She always had something to say.

She had flaws and that was ok.

And when she was down, she got right back up.

She was a beast in her own way, but one idea described her best.

She was unstoppable and she took anything she wanted with a smile.”

— R.M. Drake

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
ESC	Embryonic Stem Cell
FISH	Fluorescent In Situ Hybridization
GCV	Ganciclovir
GFP	Green Fluorescent Protein
HSC	Hematopoietic Stem Cell
HSV-tk	Human Simplex Virus Thymidine Kinase
HUVEC	Human Umbilical Vein Endothelial Cell
iPSC	Induced Pluripotent Stem cell
miRNA	micro Ribonucleic Acid
MSC	Mesenchymal Stem Cell
NTC	Non-Targeting Control
RPE	Retinal Pigmented Epithelium
PSC	Pluripotent Stem Cell
scRNA-seq	Single Cell RNA-sequencing
sgRNA	Single Guide RNA
SWNE	Similarly Weighted Nonnegative Embedding
UMAP	Uniform Manifold Approximation and Projection

LIST OF SUPPLEMENTAL TABLES

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Prashant Mali for his support as the chair of my committee and my mentor. He has shown me what hard work and determination can achieve. He let me explore a wild and risky idea. There were some serious highs and big time lows, but I value the growth I have gained under his mentorship. Thank you!

I would like to acknowledge Professor Karl Willert for his support as my co-advisor. He always asked exciting questions about my research and I thoroughly enjoyed his stem cell course where I learned much of my stem cell knowledge.

I would like to acknowledge Professor Kun Zhang for his support on my committee. I would sneak into his lab at odd hours performing late night RNAScope studies and appreciated his constant upbeat enthusiasm along the way.

I would like to acknowledge Professor Alysson Muotri for his support on my committee. He is always full of optimism and enthusiasm and his presentation on brain organoids still blows me away to this day.

I would like to acknowledge Professor Louise Laurent for her support on my committee. She asked the tough questions and educated me on fetal development. She also filled that missing slot for my committee that the department made overtly meticulous for me. My last member "…must be a NON-BMS Faculty who is UCSD Prof or UCSD Assoc PROF and NOT Adjunct and NOT affiliated with Bioengineering and/or CMM." I also value her as the female voice on my committee. Thank you!

I would like to acknowledge Dr. Yan Wu for being a hard-working co-author on our impressive Cell Press publication. This would not have been possible without his expertise in

computational conundrums. I would take months to get data to him which he would analyze in days flat. Your help was tremendous. Thank you!

I would like to acknowledge my mentees Justin Tat and Aravind Anand for helping me out with all my protocols. Thank you for dealing with my crazy schedule and not only learning from me, but also teaching me. I could not have asked for a better undergrad and master's student to work with. Thank you!

I would like to acknowledge Dr. Udit Parekh for being my first mentor in my PhD lab as I followed him around for weeks learning stem cell culture and basic PCRs. Whenever I had a question he was my first go-to person. I trusted his advice over anyone. I was always happy to help him out as well when it came down to animal surgery.

I would like to acknowledge Amir Dailamy for being my first true grad school friend and lab mate. His mind always inspires me and he always knew exactly what to say to move me. I never would have thought I could become best friends with a guy until I met him. Of course our constant collaboration was a bonus as well. I really hope to work with him again in the future when we open up a lab or company together.

I would like to acknowledge Michael Hu for being another grad school friend and helping me out in all my microscopy needs. I'm going to miss "It's mouse time!" as he somehow got tricked into weekly teratoma growth measurements on all my mice. I also value him in our final collaboration that is ongoing.

I would like to acknowledge Dr. Ana Moreno for being another grad school friend. She showed me so much kindness and even taught me how to start a company. I was more than happy to help her on her big final publication with some minor validation experiments.

I would like to acknowledge Kyle Ford for being the first grad student I “collaborated” with as we wrote a full CRISPR review together when we both just joined the lab.

I would like to acknowledge Marianna Yusupova for showing me all she knows on teratomas. To me she is the original teratoma lady, not me. I was more than honored to take on her work when she left. Thank you!

I would like to acknowledge Dr. Ann Tipps for all my histology questions. I would bombard her office constantly to validate cell types and structures in my teratomas and appreciated all her insight as she chatted about her cats and passion for photography.

I would like to acknowledge Professor Paul Insel for showing me what an MD/PhD program was and opening up an entire world of research for me during my master’s. His lab was the first I ever worked in and is what inspired me to continue to pursue science.

I would like to acknowledge Professor Fiona Murray who was my first mentor ever in a research setting. Her kindness really motivated me to constantly try to impress her. I also appreciate her supporting me personally as it was during a time when I was undergoing a lot of self-discovery. Thank you!

I would like to acknowledge Nakon Aroonsakool who is no longer with us. He was the first human to show me what a pipette was. We worked side-by-side for years during my undergraduate and master’s research. He had the “golden hands.” You are deeply missed.

I would like to acknowledge Dr. Marci Bowers for always being an inspiration. She is one of the sole reasons I applied to medical school. She made me realize trans girls can be doctors too. Thank you!

Finally, I would like to acknowledge my family, friends, and partner. I won’t get them all but mom, dad, grandma, Foxie, Avi, Sarah, Christina, Claire, Dione, Cydney, Jakobee, Chay, Josh

B., Ren, Leland, Rebecca, Saeed, Siddiq, Alexis, Jonathan, Bethany, Lola, and my partner Josh thank you! You all supported me throughout this process immensely. All of you were always so stunned by my achievements that it made me realize I really was doing something cool.

Chapters 1-5 are in part reprints of the following materials of which the dissertation author was one of the primary investigators and authors of this paper:

Chapter 1, 3, 4, 5: McDonald D*, Wu Y*, Dailamy A, Tat J, Parekh U, Zhao D, Hu M, Tipps A, Zhang K, Mali P. Defining the Teratoma as a Model for Multi-lineage Human Development. Cell. 2020 Nov 25;183(5):1402-1419.e18. doi: 10.1016/j.cell.2020.10.018.

Chapter 2: Ford K*, McDonald D*, Mali P. Functional Genomics via CRISPR-Cas. J Mol Biol. 2019 Jan 4;431(1):48-65. doi: 10.1016/j.jmb.2018.06.034. Epub 2018 Jun 28. PMID: 29959923; PMCID: PMC6309720

*Both of these authors contributed equally

VITA

2011 Bachelor of Science in Human Biology, University of California San Diego

2011 Bachelor of Arts in Theatre, University of California San Diego

2013 Master of Science in Biology, University of California San Diego

2021 Doctor of Philosophy in Biomedical Sciences, University of California San Diego

PUBLICATIONS

1. Moreno AM, Alemán F, Catroli GF, Hunt M, Hu M, Dailamy A, Pla A, Woller SA, Palmer N, Parekh U, **McDonald D**, Roberts AJ, Goodwill V, Dryden I, Hevner RF, Delay L, Gonçalves Dos Santos G, Yaksh TL, Mali P. (2021) 'Long-lasting analgesia via targeted in situ repression of $Na_V1.7$ in mice', Science Translational Medicine, 13(584), p. eaay9056. doi: 10.1126/scitranslmed.aay9056
2. Parekh U, **McDonald D**, Dailamy A, Wu Y, Cordes T, Zhang K, Tipps A, Metallo C, Mali P (2021) 'Charting oncogenicity of genes and variants across lineages via multiplexed screens in teratomas', bioRxiv, p. 2021.03.09.434648. doi: 10.1101/2021.03.09.434648.
3. Nuyen B, Kandathil C, **McDonald D**, Thomas J, & Most SP (2021) The impact of living with transfeminine vocal gender dysphoria: Health utility outcomes assessment, International Journal of Transgender Health, DOI: 10.1080/26895269.2021.1919277
4. Nuyen B, Kandathil C, **McDonald D**, Chou DW, Shih C, Most SP. The Health Burden of Transfeminine Facial Gender Dysphoria: An Analysis of Public Perception. Facial Plast Surg Aesthet Med. 2020 Oct 15. doi: 10.1089/fpsam.2020.0192. Epub ahead of print. PMID: 33054404.

5. **McDonald D***, Wu Y*, Dailamy A, Tat J, Parekh U, Zhao D, Hu M, Tipps A, Zhang K, Mali P. Defining the Teratoma as a Model for Multi-lineage Human Development. Cell. 2020 Nov 25;183(5):1402-1419.e18. doi: 10.1016/j.cell.2020.10.018.
6. Hu M, Dailamy A, Lei XY, Parekh U, **McDonald D**, Kumar A, Mali P. Facile Engineering of Long-Term Culturable Ex Vivo Vascularized Tissues Using Biologically Derived Matrices. Adv Healthc Mater. 2018 Oct 23
7. Ford K*, **McDonald D***, Mali P. Functional Genomics via CRISPR-Cas. J Mol Biol. 2019 Jan 4;431(1):48-65. doi: 10.1016/j.jmb.2018.06.034. Epub 2018 Jun 28. PMID: 29959923; PMCID: PMC6309720.
8. Insel PA, Wilderman A, Zambon AC, Snead AN, Murray F, Aroonsakool N, **McDonald DS**, Zhou S, McCann T, Zhang L, Sriram K, Chinn AM, Michkov AV, Lynch RM, Overland AC, Corriden R. G Protein-Coupled Receptor (GPCR) Expression in Native Cells: "Novel" endoGPCRs as Physiologic Regulators and Therapeutic Targets. Mol Pharmacol. 2015 Jul;88(1):181-7. doi: 10.1124/mol.115.098129. Epub 2015 Mar 3. PMID: 25737495; PMCID: PMC4468643.
9. **McDonald, D.**, Aroonsakool, N., Kwon, O., Insel, P. and Murray, F. (2014), G protein-coupled receptor expression and function in pulmonary artery smooth muscle cells: new targets in pulmonary arterial hypertension (1090.3). The FASEB Journal, 28: 1090.3. https://doi.org/10.1096/fasebj.28.1_supplement.1090.3
10. **McDonald, D.S.**, Aroonsakool, N., Kwon, O., Insel, P.A. and Murray, F. (2012), G protein-coupled receptor (GPCR) arrays identify physiologically relevant targets in Pulmonary Artery Smooth Muscle Cells (PASMC): mRNA to Function. FASEB J, 26: 669.2-669.2. https://doi.org/10.1096/fasebj.26.1_supplement.669.2

*Both of these authors contributed equally

PATENTS

1. Zhang K, Wu Y, Dailamy A, Mali P, **McDonald D**, Parekh U, Hu M. Novel method to engineer translantable human tissues, *WO 202001024 9A1* (2020).

FIELDS OF STUDY

Major Field: Biomedical Sciences

Studies human development, stem cell engineering, and tissue engineering.
Professor Prashant Mali

ABSTRACT OF THE DISSERTATION

Defining the Teratoma as a Model for Multi-Lineage Human Development

by

Daniella Nicole McDonald

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2021

Professor Prashant Mali, Chair

We propose that the teratoma, a recognized standard for validating pluripotency in stem cells, could be a promising platform for studying human developmental processes. Performing single cell RNA-seq of 179,632 cells across 23 teratomas from 4 cell lines, we found teratomas reproducibly contain approximately 20 cell types across all 3 germ layers, the inter-teratoma cell type heterogeneity was comparable to organoid systems, and that the teratoma gut and brain cell

types correspond well to similar fetal cell types. Cellular barcoding confirmed that injected stem cells robustly engraft and contribute to all lineages. Using pooled CRISPR-Cas9 knockout screens, we showed that teratomas can simultaneously assay the effects of genetic perturbations across all germ layers. Additionally, we demonstrated teratomas can be enriched for specific lineages via a genetic or materials approach. We either molecularly sculpted via miRNA-regulated suicide gene expression or assayed teratomas under multiple matrix conditions. Taken together, the teratoma is a promising platform for modeling multi-lineage development, pan-tissue functional genetic screening, and tissue engineering.

1 Teratoma Characterization

1.1 Abstract

We propose that the teratoma, a recognized standard for validating pluripotency in stem cells, could be a promising platform for studying human developmental processes. Performing single cell RNA-seq of 179,632 cells across 23 teratomas from 4 cell lines, we found teratomas reproducibly contain approximately 20 cell types across all 3 germ layers, the inter-teratoma cell type heterogeneity was comparable to organoid systems, and that the teratoma gut and brain cell types correspond well to similar fetal cell types. Cellular barcoding confirmed that injected stem cells robustly engraft and contribute to all lineages.

1.2 Introduction

Current understanding of early human development heavily relies on inference from animal models. Model systems such as frogs [1], fish [2], and mice [3,4] have demonstrated that many features of early embryogenesis are evolutionarily conserved across species [5–7]. However, several aspects of development are highly species-specific, especially in neural development [8–11] While there have been studies on human embryonic development [12,13], such studies are limited by a scarcity of relevant biological material and key ethical constraints. There has thus been a push to establish models specific to human development.

Human pluripotent stem cells (PSCs), such as embryonic stem cells (ESCs) or induced pluripotent stem cells (iPSCs), have been used as developmental models by directing differentiation of ESCs or iPSCs into various cell types. These studies have shed light on processes such as lineage bifurcation [14] and heterogeneity [15] during human neuronal development, as well as the presence of discrete cell states during early ESC differentiation [16]. Additionally, perturbation screens in these cell culture models have looked at the key regulators of differentiation [17] and

reprogramming [18]. However, true human development takes place in 3-dimensions, which is difficult to capture with a 2-dimensional monolayer [19,20].

Newer methods for modeling human development use organoid systems. Organoids are 3D "mini-organs" derived from PSCs in which the cells spontaneously self-assemble into differentiated, functional cell types which mimic their *in vivo* counterparts structurally and functionally [21–27]. The use of organoids has enabled researchers to model human specific development in a 3D context, which is especially beneficial for modeling rare genetic diseases or cancers [21,22,28–33]. However, tissue types derived from organoids may be immature [34,35] and limited in thickness and scale due to the absence of abundant vasculature. Additionally, most organoid models can only generate a single or few developmental lineages [22,3622,37–39]. In this regard, gastruloids, which model early anteroposterior organization, can recapitulate all germ layers, but they are unable to model later stages of development [40].

In some remarkable newer studies, a group was able to grow a mouse embryo in culture *ex vivo* to stage E11.5 before the embryo suffered from hydrops fatalis [41]. Again this study does not utilize human tissue however and has limitations in maturity.

We propose here the use of teratomas as a model for studying human development [42]. "Teratoma" comes from the Greek root "*teratos*" meaning monster. The earliest depictions of teratomas date back to 600 to 900 BCE on ancient tablets from the Chaldean Royal Library of Nineveh[43]. Initial discoveries of embryonic stem cells (ESCs) and their isolation ought to be credited to the early research in teratomas. Much of this research was pioneered by Dr. Leroy Stevens who realized teratomas derive from pluripotent germ cells which have a resemblance to cells embryonic in origin[44–46]. The initial thorough description of a teratoma was made by Thürlbeck and Scully in 1960[47]. These tumors form in human patients from misguided primordial

germ cell migration during embryogenesis[48,49]. Histologically they can be at varying levels of maturity which can determine prognosis in patients (more immature states equating to higher malignancy). Most teratomas are mature and thus, benign[50]. These tumors are notorious for containing teeth, hair, nails, and bone contributing to their "wow factor" and continuous curiosity by researchers and the public.

The teratoma displays multi-lineage differentiation to all germ layers, has vascularized 3D structure, bears regions of complex tissue-like organization, and is relatively straightforward to implement. PSC-derived teratomas are generated by directly injecting PSCs into immunodeficient mice, where the cells will attach and differentiate in a semi-random fashion into all three germ layers [47,51–53]. In this regard, teratoma formation is the gold standard to validate pluripotency and developmental potential of hPSC lines [54,55].

There has also been some progress in utilizing the inherent differentiation potential of teratomas to derive highly sought-after cell types. For instance, teratomas were recently utilized to derive skeletal myogenic progenitors by injecting PSCs into the *tibialis anterior* muscle of mice to enrich for muscle cell types in the teratomas that formed in those muscles [56]. Additionally, some groups have successfully enriched for hematopoietic stem cells (HSCs) from teratomas utilizing strategies such as human umbilical vein endothelial cell (HUVEC) pooling [57–60]. However, the semi-random nature of teratoma development has previously made characterization of teratomas difficult, as the different lineages can often be found in close spatial proximity.

We hypothesized that the advent of high-throughput single cell gene expression profiling via droplet based methods [61–67] coupled with histology, and RNA *in situ* hybridization, we established a comprehensive experimental and computational framework to systematically analyze human PSC-derived teratomas to evaluate their potential for modeling human development.

1.3 Materials and Methods

1.3.1 Experimental Model and Subject Details

1.3.1.1 Cell Culture

The H1 (P30), H9 (P36), PGP1 (P39), and HUES62 (P20) hESC cell line was maintained under feeder-free conditions in mTeSR medium (Stem Cell Technologies). Prior to passaging, tissue-culture plates were coated with growth factor-reduced Matrigel (Corning) diluted in DMEM/F-12/Glutamax medium (Thermo Fisher Scientific), and incubated for 30 minutes at 37ºC, 5% CO2. Cells were dissociated and passaged using the dissociation reagent Versene (Thermo Fisher Scientific). Cells were passaged a maximum of 4 times for proper expansion prior to injection. HEK 293T and HeLa were maintained in high glucose DMEM supplemented with 10% fetal bovine serum (FBS) and passaged every couple days upon confluency with .05% Trypsin-EDTA (Gibco). HUVECs were maintained in EGM-2 (Lonza).

1.3.1.2 Animals

Animals used in this study were male NOD-scid IL2Rgammanull mice 8-10 weeks of age. Housing, husbandry and all procedures involving animals used in this study were performed in compliance with protocols (#S16003) approved by the University of California San Diego Institutional Animal Care and Use Committee (UCSD IACUC). Mice were group housed (up to 4 animals per cage) on a 12:12 hr light-dark cycle, with free access tso food and water in individually ventilated specific pathogen free (SPF) autoclaved cages. All mice used were healthy and were not involved in any previous procedures nor drug treatment unless indicated otherwise.

1.3.2 Method Details

1.3.2.1 Library Preparation

The lentiviral backbone plasmid for the barcode vector was constructed containing the EF1α promoter, mCherry transgene flanked by BamHI restriction sites, followed by a P2A peptide and hygromycin resistance enzyme gene immediately downstream (ECIH). The backbone was digested with HpaI, and a pool of 20 bp long barcodes with flanking sequences compatible with the HpaI site, was inserted immediately downstream of the hygromycin resistance gene by Gibson assembly. The vector was constructed such that the barcodes were located only 200 bp upstream of the 3'-LTR region. This design enabled the barcodes to be transcribed near the poly-adenylation tail of the transcripts and a high fraction of barcodes to be captured during sample processing for scRNA-seq.

The Gibson assembly reactions were set up as follows: 1:10 molar ratio of digested backbone to sgRNA insert, 2X Gibson assembly master mix (New England Biolabs), H20 up to 20 µl. After incubation at 50°C for 1 h, the product was transformed into One Shot Stbl3 chemically competent Escherichia coli (Invitrogen). A fraction (150 µL) of cultures was spread on carbenicillin (50 µg/ml) LB plates and incubated overnight at 37°C for 15-18hrs (miRNA constructs required longer incubation times). Individual colonies were picked, introduced into 5 ml of carbenicillin (50 µg/ml) LB medium and incubated overnight in a shaker at 37°C. The plasmid DNA was then extracted with a QIAprep Spin Miniprep Kit (Qiagen), and Sanger sequenced to verify correct assembly of the vector and to extract barcode sequences.

1.3.2.2 Viral Production

HEK 293T cells were maintained in high glucose DMEM supplemented with 10% fetal bovine serum (FBS). Cells were seeded in a 15 cm dish 1 day prior to transfection, such that they

were 60-70% confluent at the time of transfection. For each 15 cm dish 36 µl of Lipofectamine 2000 (Life Technologies) was added to 1.5 ml of Opti-MEM (Life Technologies). Separately 3 µg of pMD2.G (Addgene #12259), 12 µg of pCMV delta R8.2 (Addgene #12263) and 9 µg of an individual vector or pooled vector library was added to 1.5 ml of Opti-MEM. After 5 minutes of incubation at room temperature, the Lipofectamine 2000 and DNA solutions were mixed and incubated at room temperature for 30 minutes. Medium in each 15 cm dish was replenished with 25 ml of fresh medium. After the incubation period, the mixture was added dropwise to each dish of HEK 293T cells. Supernatant containing the viral particles was harvested after 48 and 72 hours, filtered with 0.45 µm filters (Steriflip, Millipore), and further concentrated using Amicon Ultra-15 centrifugal ultrafilters with a 100,000 NMWL cutoff (Millipore) to a final volume of 600-800 µl, divided into aliquots and frozen at -80°C.

1.3.2.3 Viral Transduction

For viral transduction, virus was added at a low MOI (ensuring a single barcode/cell or a single sgRNA/cell) to stem cells at 20% confluency alongside polybrene (5 µg/ml, Millipore) in fresh mTeSR medium. The following day, medium was replaced with fresh mTeSR. Appropriate selection reagent was added 48 hrs after transduction (hygromycin [50µg/µL] for barcode) (Thermo Fisher Scientific) and was replaced daily.

1.3.2.4 Teratoma Formation

A subcutaneous injection of 5-10 million PSCs in a slurry of Matrigel® and mTeSR medium (1:1) was made in the right flank of anesthetized Rag2-/-;γc-/- immunodeficient mice. Weekly monitoring of teratoma growth was made by quantifying approximate elliptical area (mm2) with the use of calipers measuring outward width and height.

1.3.2.5 Teratoma Processing

After growth for 70 days on average mice were euthanized by slow release of CO2 followed by secondary means via cervical dislocation. Tumor area was shaved, sprayed with 70% ethanol, and then extracted via surgical excision using scissors and forceps. Tumor was rinsed with PBS, weighed, and photographed. Tumors were inspected for external heterogeneity to ensure proper tumor representation. Representative tumors were cut in a semi-random fashion in ≤ 22 mm diameter pieces and frozen in OCT for sectioning and H&E staining courtesy of the Moore's Cancer Center Histology Core. Remaining tumor was cut into small pieces 1-2mm in diameter and subjected to standard GentlaMACS™ protocols: Human Tumor Dissociation Kit (medium tumor settings), Red Blood Cell Lysis Kit, and Dead Cell Removal Kit. Single cells were then resuspended in .04% BSA for 10X Genomics chromium[63] platform.

1.3.2.6 Histology and RNAScope®

Sectioning and H&E staining was performed by the Moore's Cancer Center Histology Core. In brief, Optimal Cutting Temperature (O.C.T.) blocks were sectioned with a cryostat into 10 micron sections onto a positively charged glass slide. The slide was then stained with Harris hematoxylin and then rinsed in tap water and treated with an alkaline solution. The slide was then de-stained to remove non-specific background staining with a weak acid alcohol. The section was then stained with an aqueous solution of eosin and passed through several changes of alcohol, then rinsed in several baths of xylene. A thin layer of polystyrene mountant was applied, followed by a glass cover slip. Sections from teratomas were confirmed to have the presence of all 3 germ layers: ectoderm, mesoderm, and endoderm via microscopy identification courtesy of pathologist Dr. Ann Tipps. Further detailed identification also performed by Dr. Tipps.

Fresh frozen sections were subjected to standard RNAScope® Fluorescent Multiplex Reagent Kit protocols following fresh frozen tissue requirements. In brief, sections were fixed with chilled 200 mL of 4% PFA in 1X PBS in 4°C for 15 min. The slides were then placed in 50% EtOH for 5 min at RT, then placed in 70% EtOH for 5 min at RT, and then finally placed in 100% EtOH for 5 min at RT twice. After the slides had dried, we drew a hydrophobic barrier around the tissue. We then placed the dried slides on a HybEZ™ Slide Rack, and added Pretreat 4 to entirely cover each section and then incubated for 30 min at RT. Slides were then washed with 1X PBS. We then added the appropriate probe to cover each section. Slides were then placed in the slide rack and then placed in a HybEZ™ Oven for 2 hrs at 40°C. After 2 hrs, slides were taken out and slides were washed with 1X Wash Buffer for 2 min at RT twice. AMP 1-FL was then added to entirely cover each section. The slides were then placed on the slide rack and inserted into the oven for 30 min at 40°C. The slides were then taken out and slides were washed with 1X Wash Buffer for 2 min at RT twice. AMP 2-FL was then added to entirely cover each section. The slides were then placed on the slide rack and inserted into the oven for 15 min at 40°C. The slides were then taken out and slides were washed with 1X Wash Buffer for 2 min at RT twice. AMP 3-FL was then added to entirely cover each section. The slides were then placed on the slide rack and inserted into the oven for 30 min at 40°C. The slides were then taken out and slides were washed with 1X Wash Buffer for 2 min at RT twice. AMP 4-FL (Alt A, B, or C) was then added to entirely cover each section. The slides were then placed on the slide rack and inserted into the oven for 15 min at 40°C. The slides were then taken out and slides were washed with 1X Wash Buffer for 2 min at RT twice. The slides were then counterstained with DAPI (30 sec at RT) and mounted with ProLong™ Gold Antifade Mountant (Cat# P10144). We then placed a 24 mm x 50 mm coverslip over the tissue section and stored them in the dark at 4C.

1.3.2.7 Microscopy

Following 24 hrs of incubation with RNAScope® probes in 4ºC, slides were imaged using Zeiss 880 Airyscan Confocal microscope with special thanks to Michael Hu for image processing utilizing the UC San Diego Microscopy Core. Raw images on the Leica DMi8 were obtained with 16bit bit-depth per color, and highlights and shadows were adjusted in the LASX software. Raw images on the Zeiss 880 were obtained with 16bit bit-depth per color, and highlights and shadows were adjusted in the ZEN software. RNAScope images were dilated using ImageJ's MorphoLib by splitting the image into the composite channels and dilating the dots in the appropriate channel. Dots were dilated to 3 pixels as disks.

1.3.2.8 Cost Analysis

Overall, the cost of profiling a single teratoma with the 10X RNA-seq system runs at about $1,300, including sequencing costs for ~8,000 cells (the output of a single 10X RNA-seq run) at a sequencing depth of 50,000 reads per cell. Mouse husbandry and reagents related to teratoma formation (cells, Matrigel, media) are relatively cheap in comparison. During teratoma growth, the researcher needs to only monitor the mice for health concerns, weights, and tumor measurements if desired. The teratoma can be extracted at any time after 3 weeks of growth. It is also theoretically possible to inject both flanks of the mouse to generate 2 teratomas per animal. With the availability of easy to use analysis tools such as Seurat/PAGODA2, as well as methods for integrating datasets (such as CONOS), running a basic clustering and cell type annotation of scRNA-seq data is fairly straightforward.

1.3.3 Quantification and Statistical Analysis

1.3.3.1 Overview

For all figures, we used the CellRanger pipeline as described in the *Single Cell RNA-Seq Processing* section to generate counts matrices [63]. We also used the Seurat R package for clustering, data integration, and classification for all figures as described in the *Seurat Data Integration* and *H1 Teratoma Clustering and Validation* methods sections [68]. For assigning lentiviral barcodes, we used the genotyping-matrices method as described in the *Lentiviral Barcode and CRISPR Guide Assignment* section [17]. For Figure 1.5/1.6, we used Similarity Weighted Nonnegative Embedding (SWNE) as described in the *Developmental Staging Analysis* section [69]. The remaining analysis was done using custom R scripts.

For the heterogeneity analysis in Figure 1.3/1.4, we treated each teratoma as an individual data replicate. In other analyses each cell was treated as a replicate.

A brief summary of the analysis details for each figure can be found in the results and figure legends. Below we also provide a mapping between each figure and the relevant methods sections:

- Figure 1.1/1.2: *Seurat Data Integration* and *H1 Teratoma Clustering and Validation*
- Figure 1.3/1.4: *Quantitative Assessment of Teratoma Heterogeneity and Cell Type Bias* and *Lentiviral Barcode and CRISPR Guide Assignment*
- Figure 1.5/1.6: *Developmental Staging Analysis*

All analysis code as well as instructions on how to reproduce our analyses can be found at the Github repository: yanwu2014/teratoma-analysis-code.

1.3.3.2 Single Cell RNA-seq Processing

Using the 10X Genomics CellRanger (v2.01) pipeline [63], we aligned Fastq files to a combined hg19 and mm10 reference using STAR aligner [70], counted UMIs to generate human and mouse gene-expression counts matrices, and aggregated samples across 10X runs with the cellranger aggr command. All cellranger commands were run using default settings.

1.3.3.3 Seurat Data Integration

Data integration was performed on the aggregated counts matrices for each of the following datasets: the 7 H1 teratomas and the 3 cell line teratomas. We used the Seurat v3 data integration pipeline [68,71]. Briefly, we first filtered the counts matrix for genes that are expressed in at least 0.1% of cells, and cells that express at least 200 genes. We then normalized the counts matrix using total-counts normalization, and log-transformed the result. Log-transforming RNA-seq counts results in the data following an approximately normal distribution, which is the assumption that Seurat makes for the remainder of the analysis [72]. For each teratoma, we identified highly variable genes, and selected the top 4000 genes that appeared as overdispersed across the most teratomas. We then identified anchor cells, and integrated the teratomas to create a batch-corrected gene expression matrix. After batch correction, we used a linear model to regress away library depth, and mitochondrial gene fraction, and ran Principal Components Analysis (PCA) [73], keeping the first 30 principal components. We then used the PCs to generate a k Nearest Neighbors (kNN) graph, setting k = 10, and then used the kNN graph to calculate a shared nearest neighbors (SNN) graph [74]. We ran modularity optimization algorithm with a resolution of 0.4 on the SNN graph to find clusters [71].

1.3.3.4 H1 Teratoma Clustering and Validation

H1 clusters were assigned to cell types using a two-stage strategy. First, we trained a kNN classifier on the Mouse Cell Atlas dataset using k = 40 [75], mapping mouse genes to their human orthologs. We projected each cell in the teratoma dataset onto the first 40 Principal Components (PCs) of the Mouse Cell Atlas and classified each cell in the H1 teratoma dataset using this kNN classifier to generate a rough set of cell type assignments for each cluster. We then manually inspected the marker genes for each cluster and adjusted the cell type based on the expression of canonical markers (**Table S2A**). We also specifically looked at transcription factor markers using the TRRUST database (**Table S2A**)[76]. We computed differential gene expression in Seurat using the default Wilcoxon rank-sum test, which does not make any assumptions about the distribution of the data being tested, otherwise known as a non-parametric test [77]. Clusters that mapped to the same MCA cell type, and expressed similar marker genes were merged. Finally, we ran UMAP on the first 30 PCs as input in order to visualize the results [78,79]. We validated each annotated cell type by computing the Pearson correlation between the average expression of each cell type and the average expression of each broad cell type in the Mouse Organogenesis Cell Atlas [4]. We used the union of all marker genes for the teratoma cell types and Mouse Organogenesis Cell Atlas cell types to perform the correlation analysis.

In some cases, it was necessary to sub-cluster the cells to achieve greater cell type resolution. Specifically, we noted that the ciliated epithelium cluster had both retinal and airway markers so we sub-clustered the all cells mapping to ciliated epithelium in order to separate retinal epithelium and airway epithelium. Additionally, we sub-clustered the neuro-ectoderm in order to identify interneurons, peripheral neurons, retinal progenitors, and early neuro-ectoderm. In both

cases we simply subsetted the gene expression matrix with the cells of interest and reran the Seurat analysis pipeline, identifying sub-clusters using known marker genes (**Table S3**).

1.3.3.5 Quantitative Assessment of Teratoma Heterogeneity and Cell Type Bias

In order to quantify the level of heterogeneity between teratomas we used the Normalized Relative Entropy metric from CONOS [80].

$$1 - \frac{\sum_{k=1}^{n_{clust}} s_k \times KL(f_k, F)}{log(n_{teratomas}) \sum_{k=1}^{n_{clust}} s_k}$$

Where f_k is a vector with the number of cells in each teratoma from cluster *k,* $KL(f_k, F)$ is the empirical KL divergence between f_k and the total number of cells in each teratoma, F. Higher Normalized Relative Entropy means the cell types are more mixed across the teratomas and thus the teratomas are less heterogeneous.

There was only one replicate per non-H1 cell line teratoma as our main goal was to assess the heterogeneity across cell lines versus the heterogeneity within the H1 cell line, while also demonstrating that we could generate teratomas using multiple cell lines.

To quantify the heterogeneity/bias of individual cell types across teratomas we simply take the KL divergence of the number of cells in each teratoma from that cell type/cluster and the total number of cells in each teratoma and then scale by the number of cells in each cell type. For each cell type *k*:

$$s_k \times KL(f_k, F)$$

1.3.3.6 Lentiviral Barcode Assignment

To assign one or more lentiviral barcode to each cell, we extracted each barcode by identifying its flanking sequences, resulting in reads that contain cell, UMI, and barcode tags. To remove potential chimeric reads, we used a two-step filtering process. First, we only kept barcodes that made up at least 0.5% of the total amount of reads for each cell. We then counted the number

of UMIs and reads for each plasmid barcode within each cell, and only assigned that cell any barcode that contained at least 10% of the cell's read and UMI counts. The code for assigning barcodes to each cell can be found on GitHub at: https://github.com/yanwu2014/genotyping-matrices [17].

1.3.3.7 H1 Cell Barcoding Analysis

We extracted lentiviral barcodes from the genomic DNA fastq files before and after teratoma formation for the 3 barcoded H1 teratomas. We counted the number of unique barcodes that were supported by at least 10 reads (the reads requirement is to mitigate overcounting unique barcodes due to minor sequencing errors) and then computed the fraction of unique barcodes that remain after teratoma formation to assess the approximate number of cells that are involved in the teratoma formation process.

We also identified lentiviral barcodes at the single cell level, using the barcode assignment strategy described in the Lentiviral Barcode and CRISPR Guide Assignment section. For each cell type, we computed its bias for specific barcodes using the same relative entropy metric we used to compute teratoma bias.

$$s_k \times KL(b_k, B)$$

Where b_k is a vector with the number of cells in each barcode from cluster k, $KL(b_k, B)$ is the empirical KL divergence between b_k and the total number of cells in each barcode, B.

1.3.3.8 Developmental Staging Analysis

In order to assess the developmental maturity of the teratoma cell types, we computed the average expression of all cells related to neuro-ectoderm (Radial Glia, Intermediate Neuronal Progenitors, Early Neurons) and gut (Oral/Esophageal, Stomach, Intestine) cell types and calculated the cosine similarity of the teratoma average expression to the average expression of

fetal human cells across different time points. We used all genes that were detected in both the fetal and teratoma data.

For the neuro-ectoderm cells, we then sub-clustered those cells and identified additional cell types using canonical marker genes (**Table S2**). We then matched those neuro-ectoderm sub-clustered cell types to cell types in a larger fetal week 17-18 single cell prefrontal cortex dataset.

We next generated Similarity Weighted Nonnegative Embeddings (SWNE) [81] for the neuronal and gut cell types using the top 3000 overdispersed genes in each tissue type. Briefly, SWNE uses nonnegative matrix factorization (NMF) [82] to decompose a gene expression matrix into component factors, embeds the factors in 2D using sammon mapping [83], and embeds the cells and key genes in the 2D space relative to the factors. The cell positions are smoothed using a shared nearest neighbors (SNN) network. For the neuronal SWNE embedding, we used 30 NMF factors and 20 nearest neighbors when computing the SNN. For the gut SWNE embedding, we used 20 NMF factors and 30 nearest neighbors. We projected teratoma data onto the fetal SWNE, by first projecting the teratoma data onto the fetal NMF factors and generating embedding coordinates. We then smooth the projected coordinates by projecting the teratoma data onto the fetal SNN.

We then compared the expression of key neuronal/gut marker genes in each neuronal and gut cell type by correlating the expression of those markers between the teratoma data and the fetal human data. We used the scaled gene expression for both the teratoma and fetal data, which involves subtracting the average expression and dividing by the standard deviation. We selected the cell type markers for the neuro-ectoderm and gut comparisons using published studies of the developing human cortex and developing gut. Specifically, we selected VIM/SOX2 as markers for Radial Glia, DLX1 as a marker for Interneurons, and HMGB2 as a marker for Cycling Progenitors

using the markers from the single-cell RNA-seq study of week 17 – 18 developing human cortex [84]. HES5 is known to be a key regulator of the neural progenitor state while DCX and NEUROD1 are essential for early neuronal differentiation [85–87]. For the developing gut, we selected CDX1/CDX2 as Mid/Hindgut markers and PAX9 as a foregut marker from the single-cell RNA-seq study of the developing human digestive tract [88]. HHEX regulates midgut development, specifically the formation of the pancreas from the gut tube [89]. SOX2 is a known foregut marker that regulates gut patterning while FOXJ1 marks foregut cells primed for the lung epithelial lineage [90,91].

1.3.3.9 Figure Generation

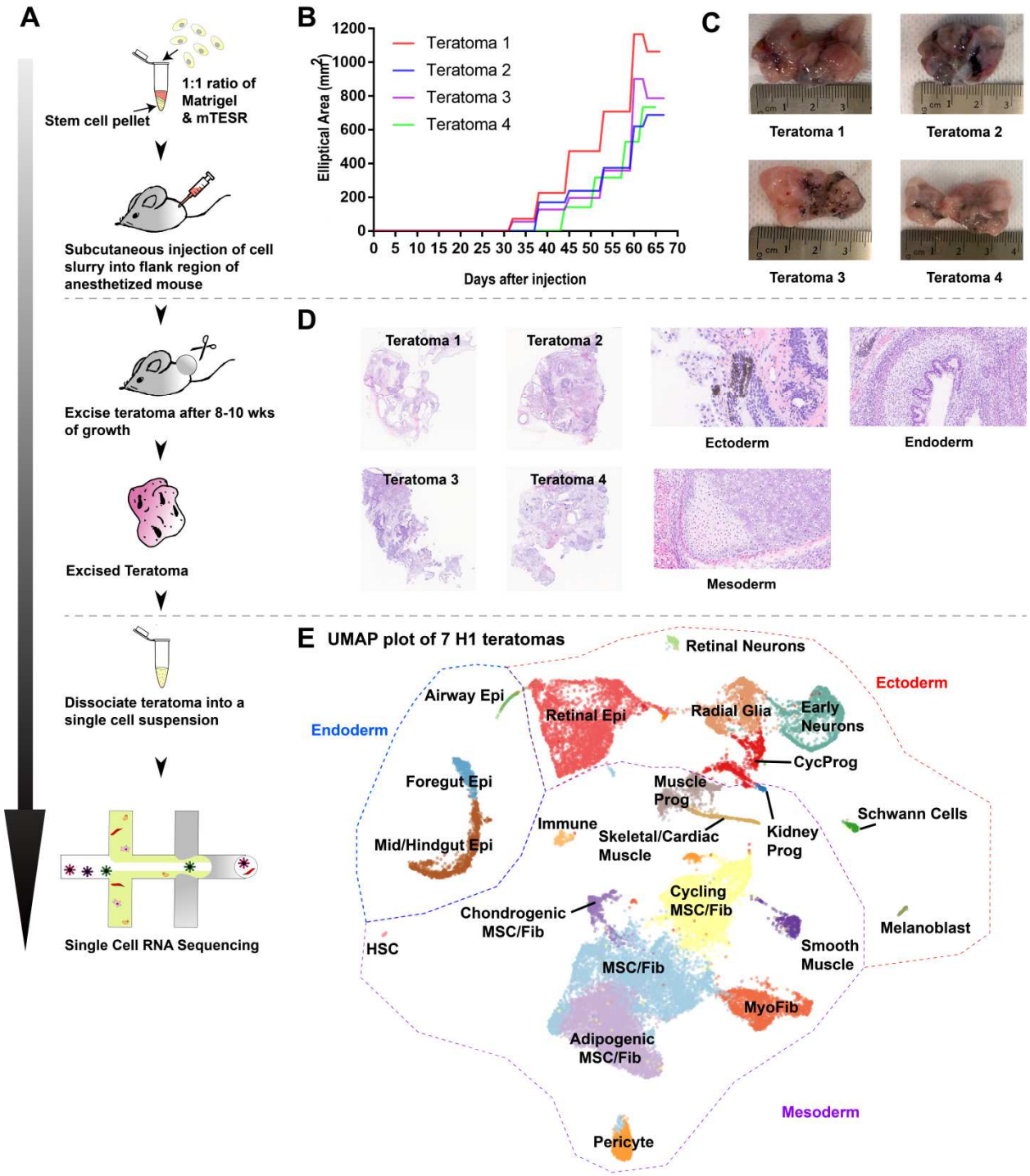
All figures were generated using original artwork or open source with InkScape, Adobe Illustrator®, and ImageJ.

1.4 Results

1.4.1 Teratoma Characterization

We first characterized the teratoma to better understand its growth kinetics, constituent cell types, and spatial organization. Towards this we generated 7 teratomas using H1 ESCs, identified cell types using single cell RNA-seq, and validated these cell types and assessed their spatial organization with histology and RNA FISH. To generate a teratoma, we made a subcutaneous injection of 5-10 million hESCs into $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice (**Figure 1.1A**, **1.3. Materials and Methods**). Kinetic trajectories show that it takes an average of around 37 days until we can begin to outwardly see and measure tumor size. We grew the teratomas for up to 70 days until the tumors were of a sufficient size for extraction and downstream analyses (~820 mm^2, **Figure 1.1B**). Post-extraction, tumors were weighed, inspected, and sectioned **(Figure 1.1C, 1.3. Materials and Methods)**.

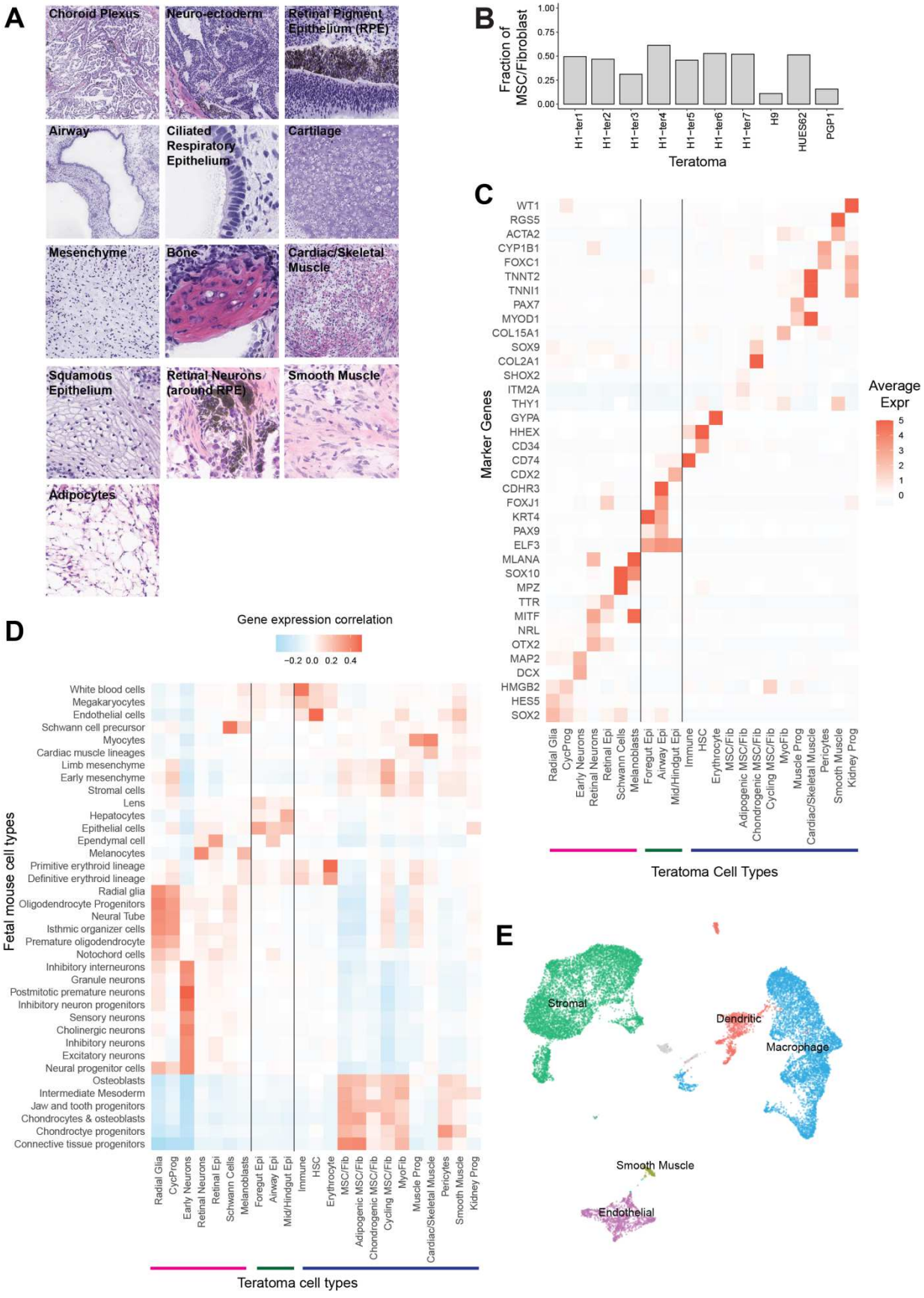
Figure 1.1. Comprehensive teratoma characterization. (A) Schematic of general workflow. Subcutaneous injection of H1 PSCs in a slurry of Matrigel® and embryonic stem cell medium was made in the right flank of $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice. Weekly monitoring of teratoma growth was quantified by approximating elliptical area (mm^2). Tumors were then extracted after 8-10 wks of growth and observed for external heterogeneity before small sections were frozen for H&E staining. The remaining tumor was dissociated into a single cell suspension via standard GentleMACS protocols. Single cell suspension was used for scRNA-seq (10x Genomics). **(B)** Growth kinetics of four H1 teratomas. **(C)** Images of four teratomas generated from H1 cells. **(D)** H&E stains of the four teratoma histology sections. The presence of ectoderm, mesoderm, and endoderm confirmed for pluripotency and developmental potential. **(E)** UMAP visualization of cell types identified from single cell RNA-sequencing of the seven H1 teratomas. Dotted lines separate the cell types originating from each of the 3 germ layers.


A
1:1 ratio of Matrigel & mTESR
Stem cell pellet
Subcutaneous injection of cell slurry into flank region of anesthetized mouse
Excise teratoma after 8-10 wks of growth
Excised Teratoma
Dissociate teratoma into a single cell suspension
Single Cell RNA Sequencing
B
Elliptical Area (mm2)
Teratoma 1
Teratoma 2
Teratoma 3
Teratoma 4
Days after injection
C
Teratoma 1
Teratoma 2
Teratoma 3
Teratoma 4
D
Teratoma 1
Teratoma 2
Ectoderm
Endoderm
Teratoma 3
Teratoma 4
Mesoderm
E
UMAP plot of 7 H1 teratomas
Retinal Neurons
Ectoderm
Airway Epi
Endoderm
Retinal Epi
Radial Glia
Early Neurons
CycProg
Foregut Epi
Muscle Prog
Schwann Cells
Immune
Skeletal/Cardiac Muscle
Kidney Prog
Mid/Hindgut Epi
Cycling MSC/Fib
Chondrogenic MSC/Fib
Melanoblast
HSC
Smooth Muscle
MSC/Fib
MyoFib
Adipogenic MSC/Fib
Mesoderm
Pericyte

We used histology to validate the presence of all 3 germ layers (ectoderm, mesoderm, endoderm) to confirm pluripotency (**Figure 1.1D**, **1.3. Materials and Methods**). An independent histology analysis also revealed structures such as developing airways, retinal pigment epithelium and neurons, fetal cartilage and bone, muscle, vasculature, GI tract, connective tissue, adipocytes and neuroectoderm (**Figure 1.2A**). Remaining tissue was dissociated for single cell RNA sequencing with the droplet-based 10X Genomics Chromium platform [63].

To analyze the resulting sequencing data, we generated single cell gene expression matrices across the 7 teratomas for both human and mouse cells using the CellRanger [63] pipeline from 10X Genomics (**1.3. Materials and Methods**, **Figure 1.1A, Table S1**). We removed any teratoma specific batch effects by using the Seurat data integration pipeline [92], and then clustered the cells using Louvain clustering [74]. We generated a rough biological annotation of the clusters using a k-nearest neighbors classifier trained on the Mouse Cell Atlas, and refined the cluster annotations manually using canonical cell type markers [92,93] (**Table S2**). We sub-clustered a cell type expressing ciliated epithelial markers with divergent expression of Airway and Retinal markers and identified Airway Epithelium, Retinal Epithelium, and erythrocytes (**Table S3**). We then visualized both the human and mouse cells with a Uniform Manifold Approximation and Projection (UMAP) [79] scatterplot (**Figure 1.1E**).

Figure 1.2. Comprehensive teratoma characterization. Related to Figure 1.1 and Table 1.1. (A) H&E stains (left to right, top to bottom): Choroid Plexus, Fetal Neuro-ectoderm, Retinal Pigment Epithelium (RPE), Developing Airway, Ciliated Respiratory Epithelium, Fetal Cartilage, Mesenchyme, Bone, Developing Cardiac/Skeletal Muscle, Squamous epithelium, Retinal Neurons (around RPE), Smooth Muscle, Adipocytes. **(B)** The fraction of cells that are classified as MSC/Fibroblast across each teratoma. **(C)** Heatmap of key marker genes for each cell type (guidelines separate cell types from different germ layers) **(Table S3C)**. **(D)** Correlation of the average expression of each human teratoma cell type with the average expression of each fetal mouse cell type. **(E)** UMAP plot of mouse cell types in the H1 teratomas.


A
Choroid Plexus
Neuro-ectoderm
Retinal Pigment Epithelium (RPE)
Airway
Ciliated Respiratory Epithelium
Cartilage
Mesenchyme
Bone
Cardiac/Skeletal Muscle
Squamous Epithelium
Retinal Neurons (around RPE)
Smooth Muscle
Adipocytes
B
Fraction of MSC/Fibroblast
1.00
0.75
0.50
0.25
0.00
H1-ter1
H1-ter2
H1-ter3
H1-ter4
H1-ter5
H1-ter6
H1-ter7
H9
HUES62
PGP1
Teratoma
C
Marker Genes
WT1
RGS5
ACTA2
CYP1B1
FOXC1
TNNT2
TNNI1
PAX7
MYOD1
COL15A1
SOX9
COL2A1
SHOX2
ITM2A
THY1
GYPA
HHEX
CD34
CD74
CDX2
CDHR3
FOXJ1
KRT4
PAX9
ELF3
MLANA
SOX10
MPZ
TTR
MITF
NRL
OTX2
MAP2
DCX
HMGB2
HES5
SOX2
Average Expr
5
4
3
2
1
0
Radial Glia
CycProg
Early Neurons
Retinal Neurons
Retinal Epi
Schwann Cells
Melanoblasts
Foregut Epi
Airway Epi
Mid/Hindgut Epi
Immune
HSC
Erythrocyte
MSC/Fib
Adipogenic MSC/Fib
Chondrogenic MSC/Fib
Cycling MSC/Fib
MyoFib
Muscle Prog
Cardiac/Skeletal Muscle
Pericytes
Smooth Muscle
Kidney Prog
Teratoma Cell Types
D
Gene expression correlation
−0.2 0.0 0.2 0.4
Fetal mouse cell types
White blood cells
Megakaryocytes
Endothelial cells
Schwann cell precursor
Myocytes
Cardiac muscle lineages
Limb mesenchyme
Early mesenchyme
Stromal cells
Lens
Hepatocytes
Epithelial cells
Ependymal cell
Melanocytes
Primitive erythroid lineage
Definitive erythroid lineage
Radial glia
Oligodendrocyte Progenitors
Neural Tube
Isthmic organizer cells
Premature oligodendrocyte
Notochord cells
Inhibitory interneurons
Granule neurons
Postmitotic premature neurons
Inhibitory neuron progenitors
Sensory neurons
Cholinergic neurons
Inhibitory neurons
Excitatory neurons
Neural progenitor cells
Osteoblasts
Intermediate Mesoderm
Jaw and tooth progenitors
Chondrocytes & osteoblasts
Chondroctye progenitors
Connective tissue progenitors
Radial Glia
CycProg
Early Neurons
Retinal Neurons
Retinal Epi
Schwann Cells
Melanoblasts
Foregut Epi
Airway Epi
Mid/Hindgut Epi
Immune
HSC
Erythrocyte
MSC/Fib
Adipogenic MSC/Fib
Chondrogenic MSC/Fib
Cycling MSC/Fib
MyoFib
Muscle Prog
Cardiac/Skeletal Muscle
Pericytes
Smooth Muscle
Kidney Prog
Teratoma cell types
E
Stromal
Dendritic
Macrophage
Smooth Muscle
Endothelial

In the human cells, we identified 23 putative cell types across all three germ layers, including endodermal cell types (gut epithelium), ectodermal cell types (early neurons), and an abundance of mesoderm-like cell types that expressed Mesenchymal Stem Cell (MSC)/Fibroblast markers, most notably the canonical MSC marker *THY1* [94] (**Figure 1.1E**, **Figure 1.2B, Table S2**). We annotated these putative MSC/Fib cell types as Adipogenic (*ITM2A, SHOX2*), Chondrogenic (*COL2A1, SOX9*), MyoFibroblasts (*COL15A1*), or Cycling (*HMGB2*) **(Table 1.1, Table S2)**. We visualized the expression of canonical marker genes for each cell type to assess the robustness of our preliminary cell type annotations (**Table 1.1, Figure 1.2C, Table S2, 1.3. Materials and Methods**).

We further validated the cell type annotations by correlating the expression of each teratoma cell type with the expression of cell types from the Mouse Organogenesis Cell Atlas [4], demonstrating that each teratoma cell type generally correlates with at least one fetal mouse cell type (**Figure 1.2D**). While most of the teratoma cell types correlate to the expected mouse cell type, there are some discrepancies that may be due to differences in developmental stage, mouse/human specific expression, as well as the fact that a broad correlation analysis may not be able to distinguish closely related cell types **(Figure 1.2D)**. For example, Hematopoietic Stem Cells (HSCs) from the teratoma correlate with fetal mouse endothelial cells, reflecting the endothelial origin of HSCs [95]. The MSC/Fib subtypes, as well as Pericytes, all broadly correlate to the same block of mesenchymal fetal mouse cell types which reflects their similar developmental origins [96]. Retinal Pigment Epithelia are a type of Ependymal Cell, and thus correlate accordingly [97]. Melanoblasts and Retinal Neurons are also both derived from the neural crest and may share some marker genes such as *MITF*, although they are not as closely related as the other cell type correlations discussed previously [98,99]. And finally, Kidney Progenitors do not correlate well with

any fetal mouse cell type, although there were no Kidney cell types in the fetal mouse data at the level of annotation we used (**Figure 1.2D**).

Overall, we used canonical marker genes and mouse cell atlases to generate a preliminary annotation of the cell types found in the teratoma scRNA-seq datasets. We provide a summary table of the key marker genes, and the experimental and computational validations performed on each cell type **in Table 1.1**. In the mouse cells, we primarily observed invading immune cells, endothelial cells, and stromal cells (**Figure 1.2E**).

Table 1.1. Summary of Cell Type Validations

Germ Layer	Broad Cell Type (used for CRISPR screen & miRNA analysis)	Cell Type	Cells (H1 & cell line teratomas)	Minimal Marker Set	RNA FISH marker validation	Identified in histology analysis	Mapped to fetal human data
Ecto	Neural Prog	Radial Glia	2579	SOX2, HES5	HES5		Yes
		CycProg (Cycling Neural Prog)	1619	SOX2, HMGB2			Yes
	Neurons	Early Neurons	6010	DCX, MAP2	DCX		Yes
		Retinal Neurons	493	OTX2, NRL		Yes	
	Retinal Epi	Retinal Epi	7238	OTX2, MITF, FOXJ1		Yes	
	Schwann Cell Prog (SCP)	Schwann Cells	174	MPZ			
		Melanoblasts	200	MITF, SOX10, MLANA			
Endo	Foregut Epi	Foregut Epi	584	ELF3, PAX9, KRT4			Yes
		Airway Epi	76	FOXJ1, CDHR3	FOXJ1	Yes	
	Mid/Hindgut Epi	Mid/Hindgut Epi	1742	ELF3, CDX2	CDX2		Yes

Table 1.1. Summary of Cell Type Validations (continued)

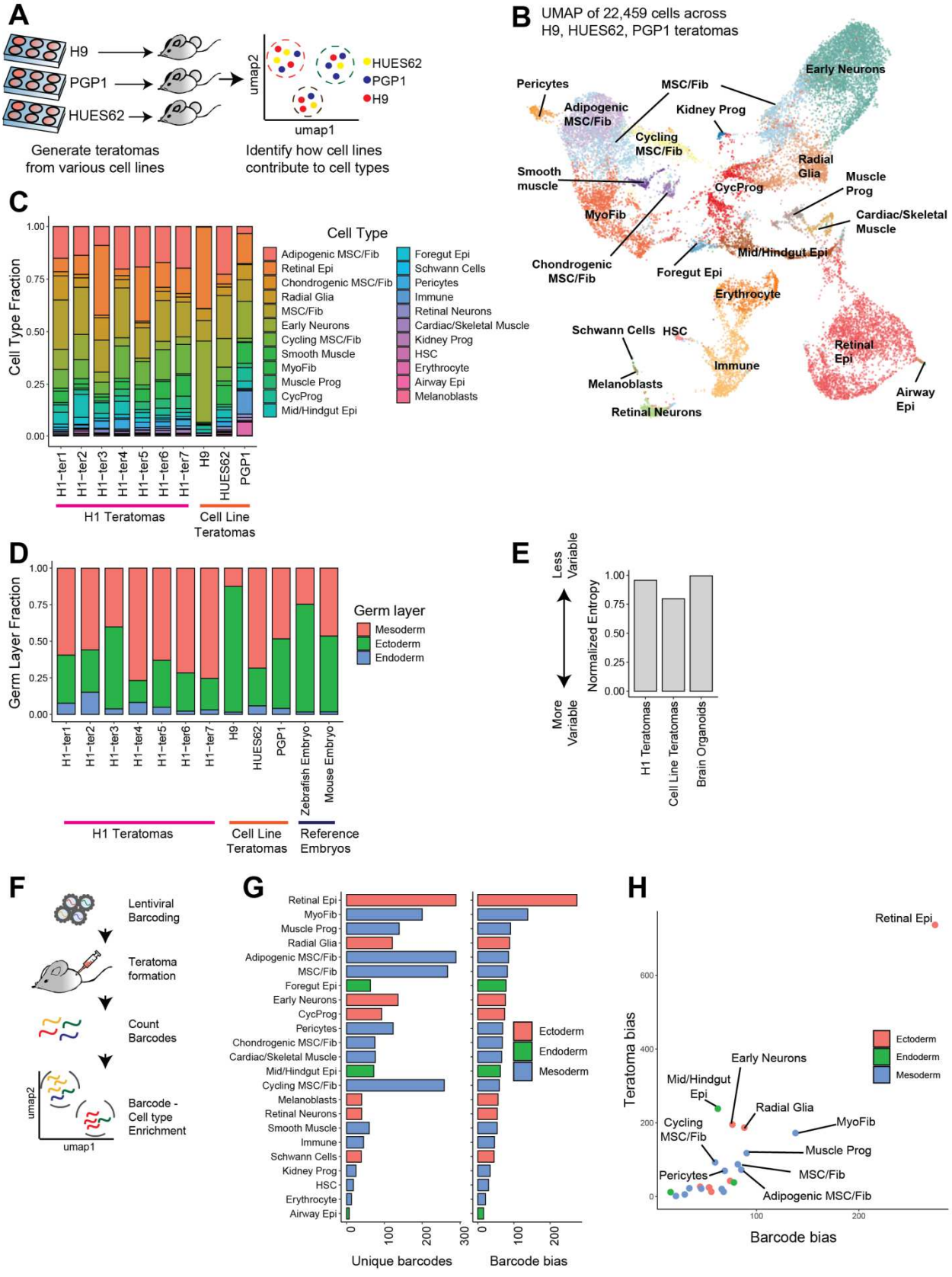
Germ Layer	Broad Cell Type (used for CRISPR screen & miRNA analysis)	Cell Type	Cells (H1 & cell line teratomas)	Minimal Marker Set	RNA FISH marker validation	Identified in histology analysis	Mapped to fetal human data
Meso	Hematopoietic	Immune	1490	CD74			
		HSC	140	CD34, HHEX			
		Erythrocyte	834	GYPA			
	MSC/Fib	Adipogenic MSC/Fib	6487	THY1, ITM2A, SHOX2		Yes	
		Chondrogenic MSC/Fib	587	THY1, COL2A1, SOX9			
		MSC/Fib	8046	THY1, COL14A1	THY1		
		Cycling MSC/Fib	4010	THY1, HMGB2		Yes	
		MyoFib	3329	THY1, COL15A1			
	Muscle	Muscle Prog	1276	MYOD1, PAX7		Yes	
		Cardiac/ Skeletal Muscle	528	MYOD1, TNNI1, TNNT2	TNNT2	Yes	
	Pericytes	Pericytes	1053	FOXC1, CYP1B1			
	Smooth Muscle	Smooth Muscle	550	ACTA2, RGS5		Yes	
		Kidney Prog	153	WT1			

1.4.2 Assaying Teratoma Heterogeneity

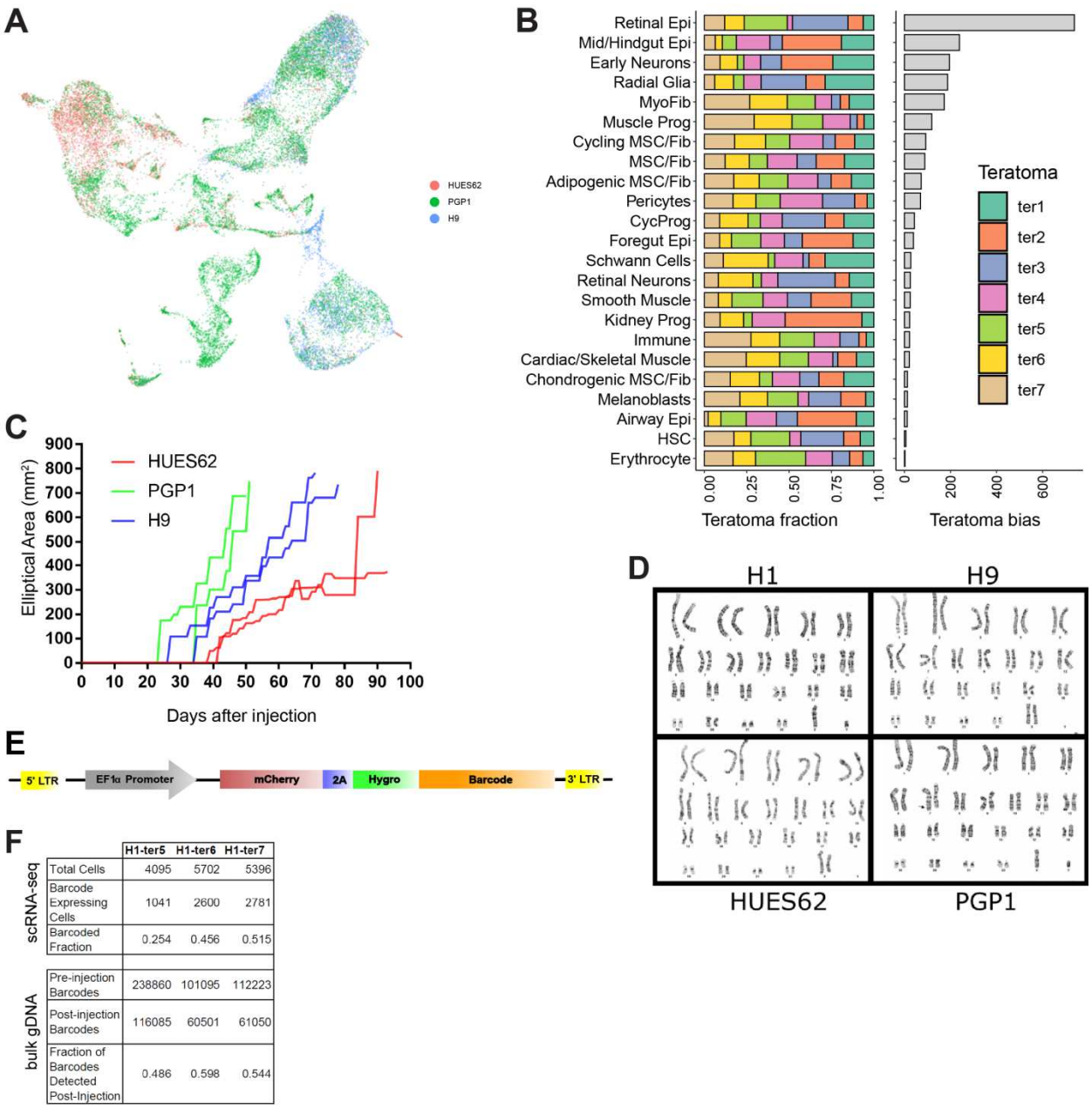
Assessing heterogeneity between teratomas (especially between teratomas generated from different stem cell lines) is critical for assessing the reproducibility and utility of this model. Towards this, we generated additional teratomas (per **Figure 1.1A**) with H9 ESCs, HUES62 ESCs, and PGP1 iPSCs, and assessed the cell type composition of the teratomas (**Figure 1.3A, Table S4**). We ran 10X sequencing on each teratoma, integrated the expression profiles, classified cell types using the H1 teratomas as reference, and visualized the cell types with aUMAP scatterplot (**Figure 1.3B**) while also showing the relative contribution of each cell line teratoma to the UMAP embedding (**Figure 1.4A**). We also assessed the distribution of cell types represented in each individual H1 teratoma alongside the H9, HUES62, and PGP1 teratomas (**Figure 1.3C**, **Figure 1.4B**). We then compared the germ layer representation between all teratomas using zebrafish and Mouse Organogenesis Cell Atlas single-cell datasets for reference [3,100] (**Figure 1.3D**). Teratomas are comprised mostly of mesoderm and neuroectoderm, with less endoderm (**Figure 1.3D**). The mesoderm is primarily from MSC/Fibroblasts in H1 teratomas, while teratomas from different cell lines show more variability in terms of the MSC/Fibroblast fraction (**Figure 1.3D**, **Figure 1.2B**). The relatively low fraction of endoderm in both the teratomas as well as the zebrafish and mouse embryo models indicate that endoderm is prevalent during development (**Figure 1.3D**). Qualitatively, while there is variability in cell type representation among the different teratomas, every teratoma contains most of the major cell types (**Figure 1.3C**). By computing the scaled mutual information between cell type assignments and teratoma assignments, we can compute a quantitative metric of this heterogeneity across teratomas (**Figure 1.3E**) [101]. We find that the cell type heterogeneity across the H1 teratomas is similar to that of patterned brain organoids [102], while the teratomas generated from different cell lines have a much higher level of heterogeneity (**Figure**

1.3E). Interestingly, line-specific kinetics were present in regard to teratoma growth with PGP1 teratomas growing the fastest and HUES62 the slowest (**Figure 1.4C**). Some of this accelerated growth may be due to chromosomal abnormalities as karyotyping has shown the PGP1 line has material translocated to 7q34 (BRAF) (**Figure 1.4D**).

Figure 1.3. Assaying teratoma heterogeneity. (A) Schematic portraying generation of teratomas from multiple cell lines and process for identifying how lines contribute to cell types. **(B)** UMAP scatterplot of all cell types present across 3 PSC lines (H9, HUES62, and PGP1) **(C)** Distribution of cell types represented in each individual teratoma **(D)** Distribution of germ layer representation in each individual teratoma (along with zebrafish and mouse comparison). **(E)** The Normalized Entropy represents how well cell type assignments are mixed with teratoma/organoid/cell line identities. A higher Normalized Entropy implies less cell type variation between teratomas/organoids/cell lines. The Cell Line Teratomas include one teratoma generated from each of HUES62, H9, and PGP1 lines. **(F)** H1 cells were uniquely barcoded at low MOI with lentiviral vectors before teratoma formation. The barcodes were counted and assessed for lineage/cell type priming of cells. **(G)** Number of unique barcodes detected in each cell type plotted alongside the cell type bias for specific barcodes (computed using the KL divergence of cell type identities with barcode identities scaled by the number of cells in each cell type). **(H)** Teratoma bias for each cell type plotted against barcode bias.


A
H9
PGP1
HUES62
Generate teratomas from various cell lines
umap2
umap1
HUES62
PGP1
H9
Identify how cell lines contribute to cell types
B
UMAP of 22,459 cells across H9, HUES62, PGP1 teratomas
Pericytes
Adipogenic MSC/Fib
MSC/Fib
Kidney Prog
Early Neurons
Cycling MSC/Fib
Smooth muscle
Radial Glia
CycProg
Muscle Prog
MyoFib
Cardiac/Skeletal Muscle
Chondrogenic MSC/Fib
Foregut Epi
Mid/Hindgut Epi
Erythrocyte
Schwann Cells
HSC
Immune
Retinal Epi
Melanoblasts
Airway Epi
Retinal Neurons
C
Cell Type Fraction
1.00
0.75
0.50
0.25
0.00
Cell Type
Adipogenic MSC/Fib
Retinal Epi
Chondrogenic MSC/Fib
Radial Glia
MSC/Fib
Early Neurons
Cycling MSC/Fib
Smooth Muscle
MyoFib
Muscle Prog
CycProg
Mid/Hindgut Epi
Foregut Epi
Schwann Cells
Pericytes
Immune
Retinal Neurons
Cardiac/Skeletal Muscle
Kidney Prog
HSC
Erythrocyte
Airway Epi
Melanoblasts
H1−ter1
H1−ter2
H1−ter3
H1−ter4
H1−ter5
H1−ter6
H1−ter7
H9
HUES62
PGP1
H1 Teratomas
Cell Line Teratomas
D
Germ Layer Fraction
1.00
0.75
0.50
0.25
0.00
Germ layer
Mesoderm
Ectoderm
Endoderm
H1−ter1
H1−ter2
H1−ter3
H1−ter4
H1−ter5
H1−ter6
H1−ter7
H9
HUES62
PGP1
Zebrafish Embryo
Mouse Embryo
H1 Teratomas
Cell Line Teratomas
Reference Embryos
E
Less Variable
More Variable
Normalized Entropy
1.00
0.75
0.50
0.25
0.00
H1 Teratomas
Cell Line Teratomas
Brain Organoids
F
Lentiviral Barcoding
Teratoma formation
Count Barcodes
Barcode - Cell type Enrichment
umap2
umap1
G
Retinal Epi
MyoFib
Muscle Prog
Radial Glia
Adipogenic MSC/Fib
MSC/Fib
Foregut Epi
Early Neurons
CycProg
Pericytes
Chondrogenic MSC/Fib
Cardiac/Skeletal Muscle
Mid/Hindgut Epi
Cycling MSC/Fib
Melanoblasts
Retinal Neurons
Smooth Muscle
Immune
Schwann Cells
Kidney Prog
HSC
Erythrocyte
Airway Epi
0
100
200
300
Unique barcodes
Ectoderm
Endoderm
Mesoderm
0
100
200
Barcode bias
H
Teratoma bias
600
400
200
0
Retinal Epi
Early Neurons
Mid/Hindgut Epi
Radial Glia
Cycling MSC/Fib
MyoFib
Muscle Prog
Pericytes
MSC/Fib
Adipogenic MSC/Fib
100
200
Barcode bias
Ectoderm
Endoderm
Mesoderm

Another key question in teratoma formation is how many cells engraft after stem cell injection. To determine this, for 3 out of the 7 H1 ESC teratomas, prior to PSC injection, cells were transduced with an integrating lentiviral ORF barcode that can be detected by scRNA-seq [103] (**Figure 1.3F, Figure 1.4E**). With this barcoding scheme, cells can be individually labeled prior to teratoma formation and their descendants can be captured after formation via scRNA-seq. Transduced PSCs were evenly split: half for teratoma formation and half were frozen down for DNA sequencing. By comparing unique barcodes extracted from genomic DNA in these two cell populations we can calculate the proportion of cells that engraft. Results showed that across the three teratomas, over 25% of cells engraft, out of a total of 10 million injected cells, which suggests that no major bottlenecking occurs during teratoma formation (**Figure 1.4F**). This is especially important in the context of using teratomas in high-throughput genetic screens, as one must ensure that there are enough cells contributing to the final tumor so that none of the elements of the genetic screen are lost.



		H1-ter5	H1-ter6	H1-ter7
scRNA-seq	Total Cells	4095	5702	5396
	Barcode Expressing Cells	1041	2600	2781
	Barcoded Fraction	0.254	0.456	0.515
bulk gDNA	Pre-injection Barcodes	238860	101095	112223
	Post-injection Barcodes	116085	60501	61050
	Fraction of Barcodes Detected Post-Injection	0.486	0.598	0.544

Figure 1.4. Assaying teratoma heterogeneity. Related to Figure 1.3. (A) UMAP scatterplot showing how each line (HUES62, PGP1, and H9) contributes to the various cell type clusters. **(B)** Left: the normalized proportion of each teratoma in every cell type. Right: the bias each cell type shows towards specific teratomas. A low bias score means the cell type is well mixed across all 7 teratomas. **(C)** Growth kinetics of 6 teratomas based on cell line (HUES62, PGP1, and H9). **(D)** Karyotyping of all 4 PSC lines. **(E)** Lentiviral barcode construct map. **(F)** Barcoding summary statistics for both bulk and single cell assays across the three barcoded teratomas.

We next tracked barcodes in individual cells by amplifying the expressed barcode from the scRNA-seq library. Since cells from the teratoma with the same barcode originated from the same PSC, we were able to track whether certain PSCs were primed to develop into certain lineages. For each cell type, we computed a barcode bias score, which reflects the level to which barcodes tend to be enriched or depleted in that cell type and plotted this barcode bias, alongside the total number of barcodes detected in each cell type (**Figure 1.3G**, **1.3. Materials and Methods**). We also computed a teratoma bias score for each cell type, which reflects how much the proportion of that cell type varies across teratomas and plotted the correlation of the teratoma bias score with the barcode bias score (**Figure 1.3H, 1.3. Materials and Methods**). We found that retinal epithelium is an outlier with both a high teratoma bias, and a high barcode bias (**Figure 1.3H**). Myofibroblast cells also have a relatively high barcode and teratoma bias score while Early Neurons, Radial Glia, Mid/Hindgut have high teratoma bias score (**Figure 1.3H**). Both the barcode bias and teratoma bias scores are scaled by the number of cells in each cell type (**1.3 Materials and Methods**).

Taken together, we found teratomas to generally contain the same major cell types at 10 weeks of growth: a large fraction of MSC/Fibroblast and neuronal cell types, and a small fraction of endoderm. RPE shows both a high degree of variability across teratomas and a high level of lineage priming. Notably, the level of heterogeneity between teratomas generated from H1 stem cells is comparable to that observed in organoids [102,104,105], but there is a much higher level of heterogeneity among teratomas derived from different PSC lines. This reflects known epigenetic variability across those lines [106].

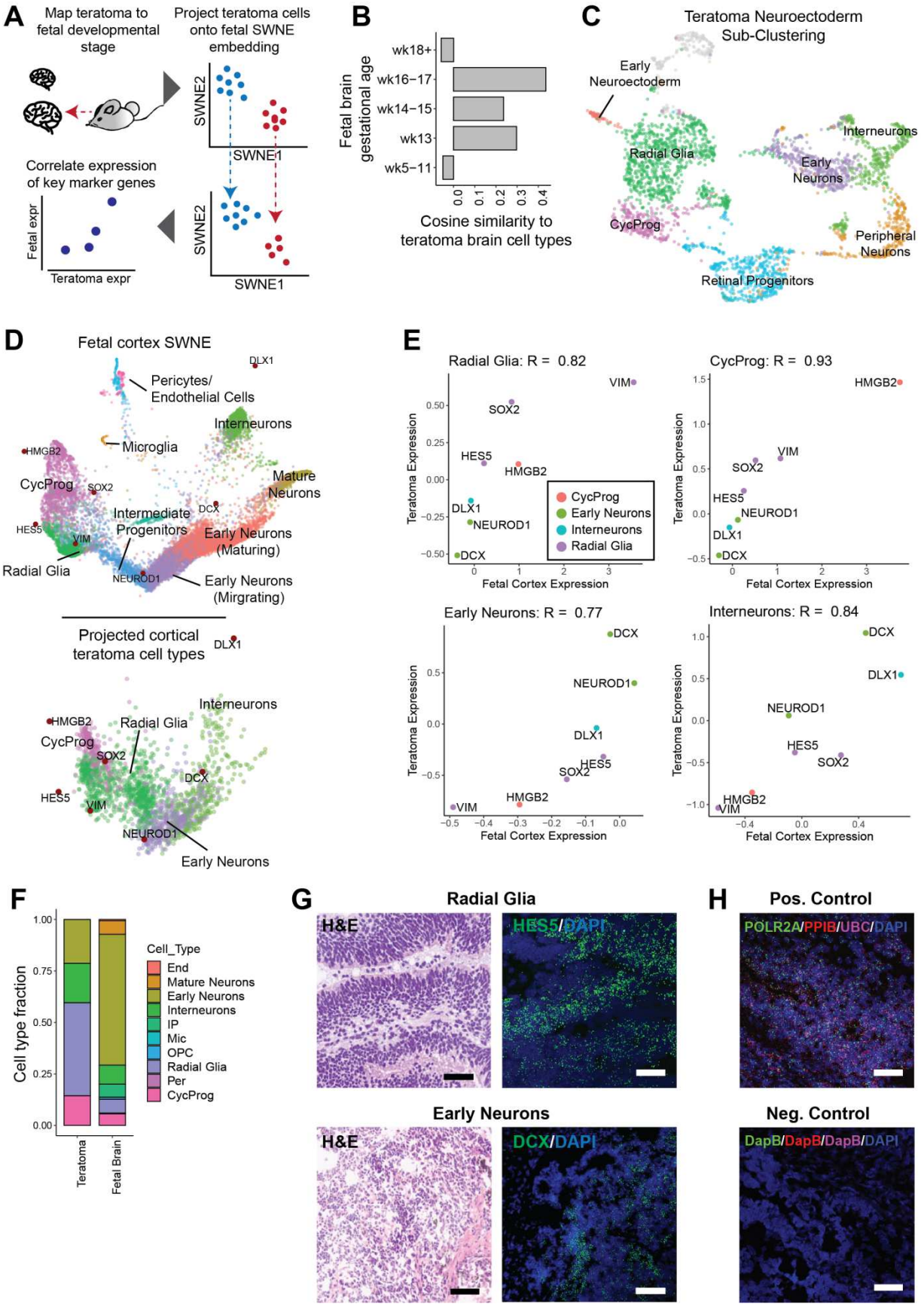
1.4.3 Assaying Teratoma Maturity

We next assessed the transcriptional similarity of the teratoma cell types to human fetal cell types, using published single-cell RNA-seq datasets from the human neuroectoderm and gut,

to determine their utility as a tool for modeling human development. We looked at which human embryonic stage the 10-week teratoma cell types most resemble, projected the teratoma data onto the fetal data to assess global transcriptional similarity, and compared the expression of key cell type marker genes (**Figure 1.5A**).

Due to the semi-random nature of teratoma differentiation, it is possible that different cell types will resemble different stages of embryonic development. Thus, we analyzed individual tissue types separately, looking specifically at the teratoma neuro-ectoderm and gut cell types in-depth. We first sub-clustered the neuro-ectoderm cells and identified additional subtypes, including a cluster of early interneurons (**Figure 1.5C**, **Table S3**). We then compared the average expression of all cells belonging to neural subtypes with the average expression of the same subtypes in a (2,300 cell) fetal brain dataset at different stages of development [107] (**Figure 1.5A**, **Figure 1.5B**). We found that the teratoma neuronal cells had high similarity scores to the human prefrontal cortex at gestational week 13 – 17 with the highest score for weeks 16 – 17 (**Figure 1.5B**). Due to the high similarity with week 16 – 17 human data, we identified the teratoma subtypes (Radial Glia, Cycling Progenitors, Early Neurons, Early Interneurons) that matched with the cell types seen in a larger 40,000+ cell week 17 – 18 dataset also from the human prefrontal cortex for further analysis [84] (**Figure 1.5A**, **Figure 1.5C**).

Figure 1.5. Assaying teratoma maturity. (A) Teratoma neuro-ectoderm cell types were mapped to fetal cortical cell types and the corresponding teratoma cell types were projected onto SWNE embeddings of fetal cells. Key marker genes were correlated across matching teratoma/fetal cell types, and average expression of teratoma cell types was correlated with fetal cell types from different stages of development. **(B)** Cosine similarity of teratoma brain cells with fetal brain cells of different ages. **(C)** UMAP embedding of teratoma neuro-ectoderm sub-clusters **(Table S2G)**. **(D)** Projection of teratoma neuro-ectoderm cell types onto the SWNE embedding of fetal cortical cells. **(E)** Correlation of the scaled expression of key marker genes across Radial Glia, Cycling Progenitors, Early Neurons, and Interneurons. **(F)** Fraction of brain related cell types in the teratoma and fetal cortex. **(G)** H&E stain (left) and RNAScope image (right) of HES5 (radial glia marker, top) and DCX (early neuron, bottom) expression. DAPI is a nuclear stain. 4-10 punctate dots/cell is a positive result. Dots were dilated using ImageJ. Scalebar = 50μM. **(H)** Positive (top) and negative (bottom) RNAScope® control staining. DAPI is a nuclear stain. 4-10 punctate dots/cell is a positive result. Scalebar = 50μM.

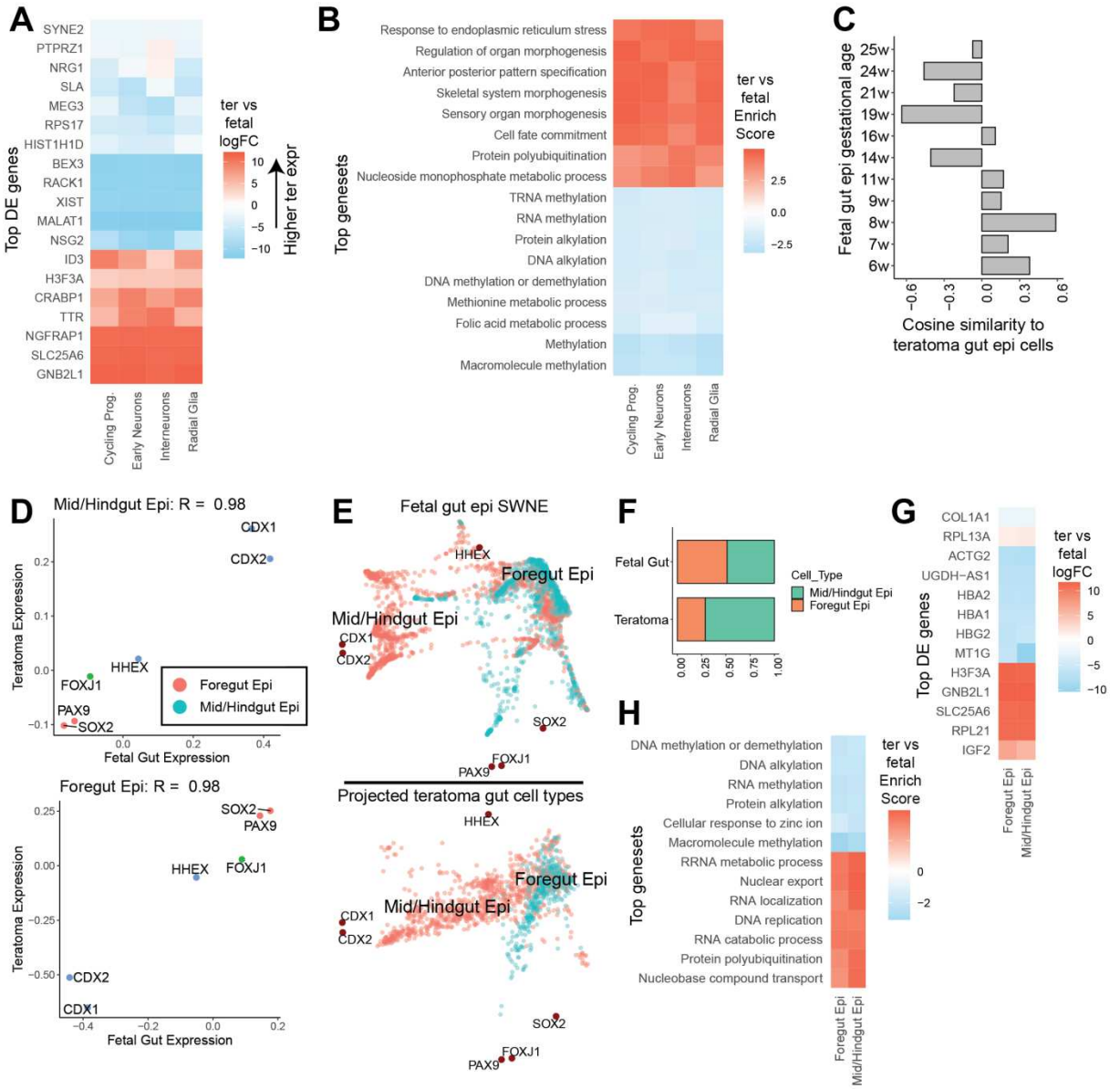

A
Map teratoma to fetal developmental stage
Project teratoma cells onto fetal SWNE embedding
SWNE2
SWNE1
Correlate expression of key marker genes
Fetal expr
Teratoma expr
B
Fetal brain gestational age
wk18+
wk16–17
wk14–15
wk13
wk5–11
Cosine similarity to teratoma brain cell types
C
Teratoma Neuroectoderm Sub-Clustering
Early Neuroectoderm
Radial Glia
Interneurons
Early Neurons
CycProg
Peripheral Neurons
Retinal Progenitors
D
Fetal cortex SWNE
Pericytes/ Endothelial Cells
Interneurons
Microglia
Mature Neurons
Intermediate Progenitors
Early Neurons (Maturing)
Early Neurons (Mirgrating)
Radial Glia
Projected cortical teratoma cell types
E
Radial Glia: R = 0.82
CycProg: R = 0.93
Early Neurons: R = 0.77
Interneurons: R = 0.84
Teratoma Expression
Fetal Cortex Expression
CycProg
Early Neurons
Interneurons
Radial Glia
F
Cell type fraction
Cell_Type
End
Mature Neurons
Early Neurons
Interneurons
IP
Mic
OPC
Radial Glia
Per
CycProg
Teratoma
Fetal Brain
G
Radial Glia
H&E
HES5/DAPI
Early Neurons
DCX/DAPI
H
Pos. Control
POLR2A/PPIB/UBC/DAPI
Neg. Control
DapB/DapB/DapB/DAPI

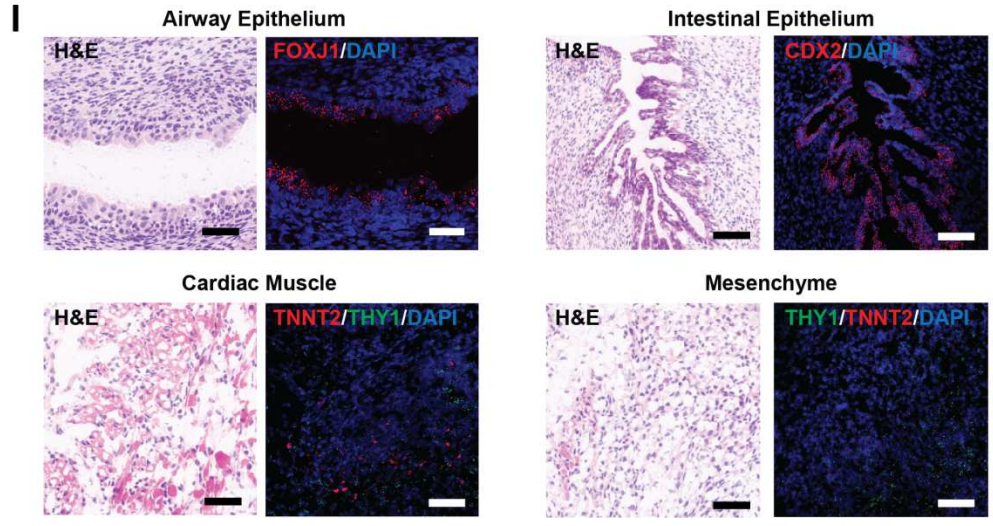
We then generated a Similarity Weighted Nonnegative Embedding (SWNE) of the week 17 – 18 human prefrontal cortex cells and projected the teratoma cells from the matching subtypes onto the fetal human SWNE (**Figure 1.5A**, **Figure 1.5D**) [69]. We found similar cell types map to similar spatial positions in the SWNE embedding, suggesting overall similar expression patterns, although the teratoma SWNE embedding shows some overlap between cycling progenitors and radial glia as well as early interneurons and excitatory neurons (**Figure 1.5D**). Additionally, the teratoma radial glia cells project onto the fetal intermediate progenitors (**Figure 1.5D**).

To further assess the similarity of the teratoma neuro-ectoderm cell types to the fetal prefrontal cortex cell types, we defined a panel of neuronal cell type marker genes: DCX, NEUROD1, HES5, SOX2, HMGB2, VIM, DLX1 and then correlated the expression of these marker genes between the teratoma cells and fetal brain cells for every matched cell type (**Figure 1.5A**, **Figure 1.5E**). We found a fairly high correlation overall, with R = 0.82 for Radial Glia, R = 0.93 for Cycling Progenitors, R = 0.84 for Interneurons, and R = 0.77 for Early Neurons (**Figure 1.5E**). We also looked at the cell type proportions in the fetal prefrontal cortex versus the teratoma, showing that the teratoma has far more progenitor cells such as Radial Glia, and fewer early neurons with no detectable mature neurons (**Figure 1.5F**). We also ran a differential expression as well as a geneset enrichment analysis between the matched teratoma and fetal prefrontal cortex cell types to assess the differences between the teratoma and fetal cells (**Figure 1.6A, 1.6B**). All four cell types showed similar top differentially expressed genes as well as genesets, suggesting that the main differences between the teratoma and fetal cells are global and not cell type specific (**Figure 1.6A, 1.5B**). The teratoma cells have a higher expression of genes related to organ morphogenesis while the fetal cells express genes related to methylation, suggesting the teratoma cells may not have the same epigenetic signatures as fetal cells (**Figure 1.6A, 1.6B**).

This analysis was repeated with teratoma gut subtypes using a published fetal gut dataset as reference [88]. The teratoma gut cells were most similar to gestational week 8-11 gut age (**Figure 1.6C**). We compared marker genes for gut cell types (CDX1, CDX2, HHEX, FOXJ1, PAX9, SOX2) between teratoma and fetal cells and found a high overall correlation, with an R = 0.98 for foregut and R = 0.98 for mid/hindgut (**Figure 1.6D**). Projecting fetal gut data onto the teratoma SWNE again resulted in relatively similar spatial positioning (**Figure 1.6E**). We see that the teratoma produces less foregut and more mid/hindgut than the fetal gut (**Figure 1.6F**). When looking at the differences between the teratoma and fetal gut cells, we again see that the fetal cells express more methylation related genes (**Figure 1.6G, 1.6H**). In this case, the teratoma cells express more genes related to RNA/DNA metabolism (**Figure 1.6G, 1.6H**).

Figure 1.6. Assaying teratoma maturity. Related to Figure 3 and Table 1.1. (A) A heatmap of log fold-changes for the top differentially expressed genes between matched teratoma neuro-ectoderm and fetal cortical cell types. **(B)** A heatmap of the enrichment scores for top differential genesets (via Geneset Enrichment Analysis) between matched teratoma neuro-ectoderm and fetal cortical cell types. **(C)** Cosine similarity of teratoma gut cells with fetal gut cells of different ages. **(D)** Projection of fetal gut epithelium cell types onto a teratoma gut epithelium SWNE embedding. **(E)** Correlation of the scaled expression of key marker genes across mid/hindgut epithelium and foregut epithelium between teratoma and fetal cell types. **(F)** Proportion of foregut and mid/hindgut cells in the teratoma and fetal gut. **(G)** A heatmap of log fold-changes for the top differentially expressed genes between matched teratoma gut epithelium and fetal gut epithelium cell types. **(H)** A heatmap of the enrichment scores for top differential genesets (via Geneset Enrichment Analysis) between matched teratoma gut epithelium and fetal gut epithelium cell types. **(I)** H&E stains (left) as well as RNA FISH staining (right) of FOXJ1 (Airway epithelium), CDX2 (Intestinal epithelium), TNNT2 (Cardiac muscle), and THY1 (mesenchymal stem cell/fibroblast). Scalebar = 50μM (20x). Dots were dilated using ImageJ.


A
Top DE genes
SYNE2
PTPRZ1
NRG1
SLA
MEG3
RPS17
HIST1H1D
BEX3
RACK1
XIST
MALAT1
NSG2
ID3
H3F3A
CRABP1
TTR
NGFRAP1
SLC25A6
GNB2L1
Cycling Prog.
Early Neurons
Interneurons
Radial Glia
ter vs fetal logFC
Higher ter expr
B
Top genesets
Response to endoplasmic reticulum stress
Regulation of organ morphogenesis
Anterior posterior pattern specification
Skeletal system morphogenesis
Sensory organ morphogenesis
Cell fate commitment
Protein polyubiquitination
Nucleoside monophosphate metabolic process
TRNA methylation
RNA methylation
Protein alkylation
DNA alkylation
DNA methylation or demethylation
Methionine metabolic process
Folic acid metabolic process
Methylation
Macromolecule methylation
ter vs fetal Enrich Score
C
Fetal gut epi gestational age
25w
24w
21w
19w
16w
14w
11w
9w
8w
7w
6w
Cosine similarity to teratoma gut epi cells
D
Mid/Hindgut Epi: R = 0.98
Teratoma Expression
Fetal Gut Expression
CDX1
CDX2
HHEX
FOXJ1
PAX9
SOX2
Foregut Epi
Mid/Hindgut Epi
Foregut Epi: R = 0.98
E
Fetal gut epi SWNE
Foregut Epi
Mid/Hindgut Epi
Projected teratoma gut cell types
F
Fetal Gut
Teratoma
Cell_Type
G
Top DE genes
COL1A1
RPL13A
ACTG2
UGDH-AS1
HBA2
HBA1
HBG2
MT1G
H3F3A
GNB2L1
SLC25A6
RPL21
IGF2
H
Top genesets
DNA methylation or demethylation
DNA alkylation
RNA methylation
Protein alkylation
Cellular response to zinc ion
Macromolecule methylation
RRNA metabolic process
Nuclear export
RNA localization
DNA replication
RNA catabolic process
Protein polyubiquitination
Nucleobase compound transport


I
Airway Epithelium
H&E
FOXJ1/DAPI
Intestinal Epithelium
CDX2/DAPI
Cardiac Muscle
TNNT2/THY1/DAPI
Mesenchyme
THY1/TNNT2/DAPI

To further validate these results, we used RNAScope In-Situ Hybridization (ISH) to probe for the radial glia marker HES5 and the early excitatory neuron marker DCX, which both showed high abundance in regions of neuro-ectoderm in fixed teratoma tissue sections (**Figure 1.5G**). POLR2A, PPIB, and UBC were used as positive controls and bacterial marker DapB as a negative control (**Figure 1.5H**). Additionally, we probed for FOXJ1 (cilia), CDX2 (intestine epithelium), TNNT2 (cardiac), and THY1 (mesenchyme/fibroblast) in ciliated airway epithelium, intestinal villi, developing cardiac muscle, and mesenchyme, respectively (**Figure 1.6I**). We were able to visualize a high abundance of the respective RNA transcripts, as well as confirm the identity of the respective tissue using H&E staining and histology (**Figure 1.6I**). Overall, we were able to show that the teratoma neuro-ectoderm and gut cell types are transcriptionally similar to their fetal counterparts, while also identifying the developmental stage of the teratoma cells. We validated the presence of six cell types (2 per germ layer) using RNAScope ISH and histology, which also showed that these cell types contain some degree of spatial organization (**Figure 1.5G**, **Figure 1.6I, Table 1.1**). Thus, we were able to further validate the teratoma neuro-ectoderm and gut cell types by mapping them onto reference fetal human scRNA-seq datasets and probing the spatial expression of canonical marker genes DCX, HES5, and CDX2 **(Table 1.1, Table S2)**. We also probed the spatial expression of FOXJ1, TNNT2, and THY1, adding more evidence to the Ciliated Epithelium, Cardiac Muscle, and MSC/Fibroblast cell type annotations **(Table 1.1, Table S2)**.

1.5 Discussion

The teratoma has the potential to be a fully vascularized, multi-lineage model for human development. Its major advantages are that it can grow to a large size due to its vascularization, and it can produce a wide array of relatively mature cell types from all major developmental lineages.

Future studies with this model could explore increasing tissue maturity with extended growth/larger animal hosts. Benchmarking with human patient-derived teratomas would also be valuable, especially as many of these often can become quite mature. Another critical future study is assessing the impact of different dissociation methods on teratoma cell type proportion. The ability to achieve greater cell numbers with the most current single cell RNA sequencing protocols, such as SPLiT-seq [61] and sci-RNA-seq [62], will be vital for identifying additional cell types. A time series analysis of teratomas at multiple stages of maturity could help uncover developmental pathways that the cell types follow. Additionally, pooling different cell types together with PSCs prior to injection may help aid in cellular enrichment/maturity in the teratoma (i.e. HUVECs to enrich for HSC populations) [59] or enriching for desired cell types based on injection site [56]. Growing patient-specific teratomas could benefit disease research through isogenic iPSC lines aiding in understanding the disease state in various tissues that otherwise may be inaccessible with current technologies. Taken together, we believe the teratoma is a promising platform for modeling multi-lineage human development.

Any model system has its intrinsic strengths and weaknesses, and below we discuss some of the limitations of the teratoma system and also considerations towards improving it for enabling basic science and engineering studies. One issue with the teratoma system (and organoids) is the intrinsic degree of heterogeneity [26,104,105,108].

While the teratoma has regions of organization and maturity, these may develop in an asynchronous manner. This lack of synchronization may prove to be a barrier in accessing certain mature cell types that need a highly ordered cellular context to develop.

Also, since the teratoma contains cell types from all lineages, finding a single dissociation protocol that captures as many cell types as possible is a challenge. The choice of dissociation

method can drastically change the cell types profiled in single cell RNA-seq, and it is likely that the set of cell types we see in our data is biased by our dissociation protocol [109]. It may be the case that no single dissociation method can capture all cell types, and it will be necessary to design specific dissociation protocols to capture specific tissues.

Additionally, our cell type annotations are still preliminary. While we validated key cell types by comparison to fetal human/mouse reference datasets and RNA FISH, we were not able to validate all cell types due to limited developmental human reference scRNA-seq datasets, as well as cost constraints. Thus, some cell types, such as the neuro-ectoderm cell types, have more validation than others, giving us greater confidence in their identity **(Table 1.1)**. We may also still be underpowered in detecting less abundant cell types and additional single cell RNA-seq could enable us to resolve some missing cell types, as under sampling could result in smaller cell types being collapsed into a larger cell type during analysis.

1.6 Acknowledgements

We thank members of the Mali lab for advice and help with experiments, Marianna Yusupova for help with initial studies, Alexander Militar for assistance in schematic generation, in loving memory of Nakon Aroonsakool, and to the Moore's Cancer Center Histology Core, UC San Diego Microscopy Core, Sanford Consortium Flow Cytometry Core, and IGM Genomics Center for help with sample processing. This work was generously supported by UCSD Institutional Funds and NIH grants (R01HG009285, RO1CA222826, RO1GM123313).

Chapters 1 is in part reprints of the following materials of which the dissertation author was one of the primary investigators and authors of this paper:

Chapter 1, in part, is a reprint of the material as it appears in McDonald D*, Wu Y*, Dailamy A, Tat J, Parekh U, Zhao D, Hu M, Tipps A, Zhang K, Mali P. Defining the Teratoma as a Model for Multi-lineage Human Development. Cell. 2020 Nov 25;183(5):1402-1419.e18. doi: 10.1016/j.cell.2020.10.018.

*Both of these authors contributed equally

2 Functional Genomics via CRISPR-Cas

2.1 Abstract

RNA-guided CRISPR (clustered regularly interspaced short palindromic repeat)-associated Cas proteins have recently emerged as versatile tools to investigate and engineer the genome. The programmability of CRISPR-Cas has proven especially useful for probing genomic function in high-throughput. Facile single guide RNA (sgRNA) library synthesis allows CRISPR-Cas screening to rapidly investigate the functional consequences of genomic, transcriptomic, and epigenomic perturbations. Furthermore, by combining CRISPR-Cas perturbations with downstream single cell analyses (flow cytometry, expression profiling, etc.), forward screens can generate robust data sets linking genotypes to complex cellular phenotypes. In the following review, we highlight recent advances in CRISPR-Cas genomic screening while outlining protocols and pitfalls associated with screen implementation. Finally, we describe current challenges limiting the utility of CRISPR-Cas screening as well as future research needed to resolve these impediments. As CRISPR-Cas technologies develop, so too will their clinical applications. Looking ahead, patient centric functional screening in primary cells will likely play a greater role in disease management as well as therapeutic development.

2.2 Introduction

Prior to further discussion on the teratoma and engineering teratomas via genetic perturbations, it is key to first delve into a deeper discussion about functional genomics utilizing the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas9 system. An ongoing challenge in biology is comprehensively mapping genotype-phenotype relationships. With this objective in mind, functional genomics makes use of data from all levels of biology (genome, transcriptome, epigenome, proteome, metabolome, etc.) to better define genetic and

protein functions and interactions. In this way, researching functional genomics is essential for better understanding the human genome and its intricate interactions in healthy, as well as pathophysiologic states. Characterizing the functional consequences of genomic variation is crucial for many aspects of biomedical research including cancer screening methodologies, drug-drug interactions, drug sensitivity and resistance, gene therapy, regenerative medicine applications, infectious disease, and general understanding of human physiology.

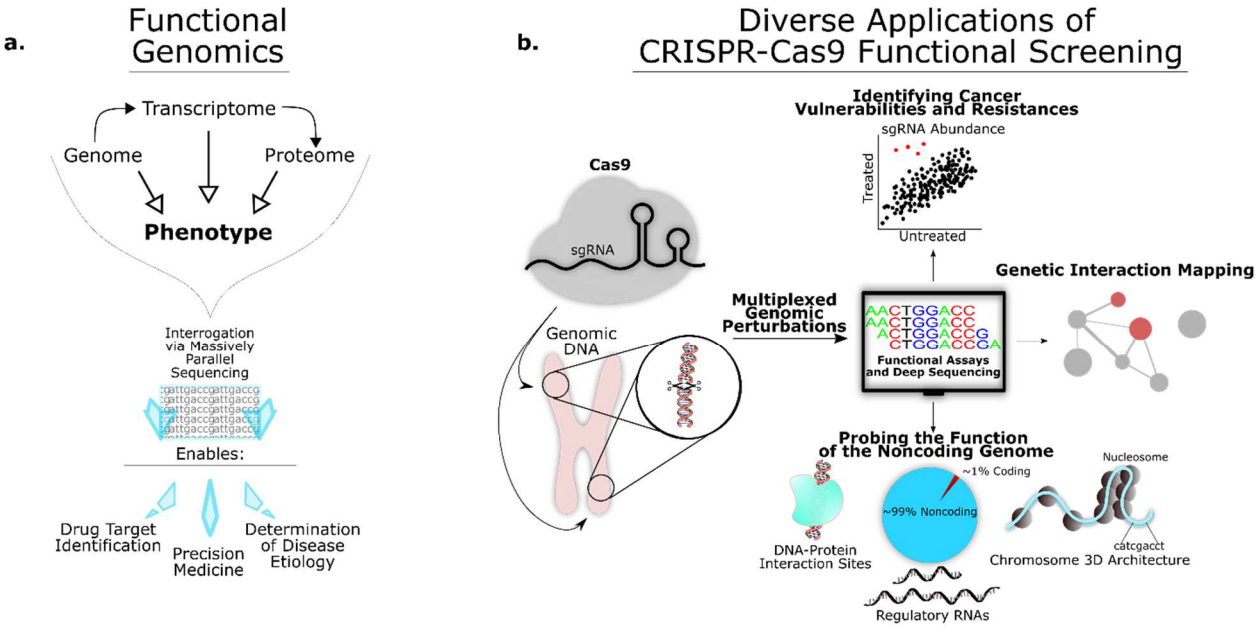


Figure 2.1. Functional Genomics and CRISPR-Cas: (a) The goal of functional genomics is to better understand how the genome informs diverse biological phenotypes. To this end, functional genomics makes use of mass data sets spanning the genome, the transcriptome, and the proteome. The declining cost of massively parallel sequencing platforms has made genome wide functional screens broadly achievable and economically viable for academic labs of all sizes. **(b)** CRISPR-Cas9 has made multiplexed functional screening with single cell resolution more robust than ever more. The ease of sgRNA design has led to accelerated functional mapping of the genome with extensive consequences for medicine and biotechnology. Because sgRNA targeting almost any region of the genome can be designed *in silico*, CRISPR-Cas screens can be rapidly designed and executed. Functional screens using Cas9 have been used for a wide variety of applications, such as identifying novel cancer therapeutics and vulnerabilities, quantifying genetic interactions, and exploring the function of the non-coding genome.

It has become increasingly clear that the volume and complexity of genomic information necessitates rapid screening methodologies. Utilizing large scale and high-throughput assays, researchers can more quickly map the function of a multitude of genes and/or proteins in parallel. To this end, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated (Cas) proteins have been utilized to help interrogate and realize functional outputs based on targeted editing strategies. CRISPR-Cas systems are powerful tools for targeted genome editing, that have dramatically impacted genomic research and screens since their first mammalian applications in 2013[110,111]. This technology has revolutionized the field with its ease, speed, and targeting versatility, allowing for facile genetic perturbations and resulting functional output analysis in a multiplexed fashion. It has allowed for a large number of high-throughput functional genomic screens to be performed which have, in turn, identified key genes involved in a broad range of human health and disease including cancers, infections, immune regulators and responses, and metabolic diseases[112].

2.3 CRISPR-Cas Toolsets

CRISPR-Cas systems are divided into different classes, types, and subtypes. Class 1 utilizes multi-protein effector complexes and class 2 utilizes single protein effectors. Class 1 includes types I, III, and IV. Class 2 includes types II, V, and VI. There are a further 19 subtypes and this will likely continue to expand as new CRISPR-Cas systems are identified[113,114].

The most common Cas protein used in functional screening is a type II single protein effector derived from *Streptococcus pyogenes* (SpCas9). The SpCas9 uses a guide RNA to assist in effectively cleaving the target gene. Once Cas9 successfully finds a target sequence with proper pairing of the complement guide RNA and an appropriate protospacer adjacent motif (PAM), the endonuclease will cleave the phosphodiester bonds upstream of the PAM forming a double-strand

break[115]. When the double strand break occurs, Non-Homologous End Joining (NHEJ) or Homology-Directed Repair (HDR) will attempt to repair the damage. NHEJ often results in a small insertion-deletion mutation (indel). If targeted to a gene, this may result in a knockout due to generation of a frameshift resulting in a premature stop codon and nonsense-mediated decay of the transcript. NHEJ is often the repair process of choice for mutagenesis. HDR, however, is a templated repair process most commonly recognized for its natural use in the body during gamete formation allowing for genetic recombination. Its use in the cell is restricted to the S and G2 phase[116]. Due to its high fidelity, HDR can be utilized to insert a new custom region into the genome creating knock-ins or specific gene mutations (or corrections) if desired[117]. Increasing the efficiency and utility of HDR is still necessary to fully apply its uses for CRISPR-Cas systems.

Over the last several years the versatility of CRISPR-Cas systems has increased dramatically. There currently exist Cas9 effector fusions with the ability to modify specific histones, edit particular DNA base pairs, activate or inhibit the transcription of certain genes (CRISPRa/CRISPRi), or effect DNA methylation/demethylation at user determined loci[118]. This wide array of effector functions enables a variety of genomic elements to be probed systematically in a high-throughput fashion.

The dominant method of generating Cas9 variants with novel functions consists of fusing a catalytically inactive Cas9 (dCas9) protein to an effector moiety[119]. In this way, the dCas9 serves only as a DNA targeting platform, which guides the effector moiety to the location of interest in the human genome. The benefit of this design strategy is that it enables rapid development of new dCas9 functionalities due to its modularity. However, optimizing the efficacy and off-target effects of novel Cas9 fusions is a laborious undertaking which increases rapidly as the protein engineering search space is expanded. Furthermore, because the effector moiety is fused permanently to dCas9,

orthogonal parallel perturbations require the co-delivery of multiple fusion constructs to the cells of interest. This, coupled with the large size of dCas9 fusions, imposes significant delivery challenges limiting their use in functional screens. Nevertheless, dCas9 fusions represent a robust set of tools with which to probe genome function.

The choice of appropriate Cas9 variant will depend heavily on what functionality is being investigated. The broad array of available Cas9 based perturbation systems are summarized in **Table 2.1**. While **Table 2.1** includes the most common Cas9 based perturbation choices, it is far from exhaustive.

Table 2.1. Cas9 Perturbation Options for Functional Screens

Perturbation Choice	Effect on the Genome	Mechanism	References
wtCas9	Loss of function and deletions	Double stranded DNA cleavage at the target locus	[120,121]
CRISPRa	Transcriptional activation	Fusion of dCas9 to various activating domains (ex. VP64 or the p65 subunit of nuclear factor kappa B (NF-κB))	[119,122–124]
CRISPRi	Transcriptional repression	Fusion of dCas9 to domains which inhibit transcription (ex. Krüppel-associated box (KRAB))	[119,124,125]
Base editors	Catalyze a nucleotide base pair substitution without DNA cleavage	Fusion of dCas9 to enzymes which catalyze nucleobase conversion (ex. activation-induced cytidine deaminase (AID) for C->T edits)	[126–131]
DNA methylation and demethylation	Cas9 guided DNA methylation and demethylation modifies chromosome structure and subsequent gene transcription	Fusion of dCas9 to DNA (cytosine-5)-methyltransferase 3A (DNMT3A) and ten-eleven translocation (TET) proteins respectively	[132,133]
Histone modification	Cas9 guided control of histone acetylation and methylation	Fusion of dCas9 to histone modifying enzymes (Ex. Histone deacetylase 3 (HDAC3), p300 acetyltransferase, or lysine-specific histone demethylase 1A (KDM1A/LSD1)	[134–137]

The wild type Cas9 protein functions as a targeted endonuclease, catalyzing DNA double stranded breaks[115]. These double stranded breaks often lead to indels via the error prone NHEJ. Frameshifts resulting from these mutations can knockout the function of protein coding genes,

making wtCas9 ideal for loss of function studies[120,121]. Knockout studies are often used to determine the *essentiality* of genes in high-throughput, and simplifies downstream validation and data analysis due to the binary nature of the perturbation. However, this simplification in some ways limits the translational relevance of knockout screening. Although knockouts can inform our understanding of what genes are essential for specific biological processes *in vitro,* there is no guarantee that small molecule or protein-mediated inhibition *in vivo* will have the same effect. Furthermore, knockout studies fail to recapitulate gain-of-function mutations and transcriptional dysregulation which play a key role in many pathologies[138–140]. In this way, Cas9 knockout experiments should not be considered a surrogate for drug studies, but rather a parallel set of tools with which interrogate the user's model. For these reasons, knockout screening requires extensive downstream target validation before any significant conclusions can be drawn.

As an alternative to knockout experiments, CRISPRa/i systems use enhancer/repressor proteins fused to dCas9 as a way of modulating gene transcription at particular loci[119,141,142]. Because CRISPRa/i functions at the transcriptional level, it enables investigation of genome function without permanently modifying genomic structure. Unlike wtCas9, activation of target genes by CRISPRa can facilitate complex gain-of-function screening from endogenous genomic loci. In addition, CRISPRi can perform loss of function screening without the confounding effects of off target nuclease activity[141]. For even more robust genetic studies, the combination of CRISPR effector functions can generate complementary data sets with which researchers can generate conclusions with greater confidence[124]. As a recent example, by co-delivering both CRISPRa and wtCas9, researchers were able to interrogate the directionality of genetic interactions in high-throughput[143]. However, CRISPRa/i experiments suffer from their own set of limitations. First and foremost is the limited correlation between mRNA levels and protein expression[144]. While

CRISPRa/i can reduce or increase the levels of a particular mRNA transcript, protein expression is subject to post-transcriptional regulation which has the potential to obfuscate the perturbations' actual effect[144]. As well, the CRISPRa/i systems require sgRNAs targeting the promoter region or transcriptional start site of the gene of interest[123,125]. Promoter regions and transcriptional start sites can be rendered inaccessible to sgRNA due to chromatin structure or may not have an appropriate PAM sequence nearby, limiting the pool of genes for which CRISPRa/i is effective. In addition, some genes are controlled by multiple functional promoters, further confounding screens using CRISPRa/i. Ideally, these limitations ought to inform the experimental design of CRISPRa/i genomic screens to ensure output data is reproducible and conclusions justifiable.

Functional studies using DNA and Histone modifying Cas9 fusion constructs operate in a similar fashion to CRISPRa/i[134]. By modifying the structure of DNA/Histones (via acetylation or methylation), these Cas9 fusions vary gene accessibility to transcriptional machinery and consequently gene expression[132,133,135]. A key difference is the mechanism underlying these structural perturbations. Whereas CRISPRa/i can modulate gene expression without leaving a scar on the target site, DNA/Histone modifications affect gene expression via lasting structural changes. The choice of perturbation is largely dependent on the nature of the biological question being asked. For probing the function of protein coding genes, CRISPRa/i and CRISPR knockout are well validated systems with a spectrum of reagents available commercially, enabling a powerful toolset for genome wide screening. However, if the goal of the experiment is mapping chromosomal structure-function relationships the DNA/Histone epigenetic modifiers may be a more fitting choice. Several groups have used these DNA/Histone modifying Cas9 variants to probe how chromosomal chemical structure and 3D architecture controls gene regulation through diverse mechanisms of action[136,137]. Nevertheless, DNA/Histone modifying Cas9 variants are not

the only way to perturb chromosomal structure. Deletions and chromosomal rearrangements induced by wtCas9 have also been used to explore how structural variation in the human genome impacts nearby gene function[145].

In contrast with wtCas9, CRISPRa/i, and Cas9 based structural modifiers, CRISPR base editing constructs have recently been developed as novel tools for functional genomic screens. CRISPR base editors work by modifying individual nucleic acid base pairs within the target genes in a precise, or pseudo random manner[126]. These systems function by fusing a cytidine deaminase or an adenosine deaminase to dCas9 to effect C→T mutations or A→G mutations respectively.[127,128] These novel systems represent a versatile avenue with which to model gain or loss-of-function mutations in an endogenous context[129–131].

Engineered sgRNAs have also been explored as an alternative way to impart novel function to the Cas9 system[146]. By incorporating protein binding RNA aptamers (PP7, MS2, etc.) into the sgRNA structure, Cas9 can recruit orthogonal proteins with a variety of functionalities. Because the perturbation choice is encoded in the sgRNA itself, multiple perturbation types can be explored in the same pooled screen using unmodified dCas9. This system has been used to effect multiplexed gene activation and interference in parallel (via sgRNA modified to recruit vp64 and KRAB respectively) as well as perform multiplexed fluorescent labelling of specific genomic loci[147,148].

2.4 Genomics Screens

The use of the CRISPR-Cas systems has many implications for functional genomics and has been the topic of much excitement. Functional screens, in turn, are typically performed in an arrayed or pooled format, and rely equally on three integral ingredients: a perturbation, a model, and an assay. In an arrayed screen, the reagents are added into a multi-well plate so that one reagent

or a small pool is added to each well allowing for a single perturbation per well. Because each well will contain a population of cells with identical genomic perturbations, a wider array of phenotypic data can be assayed simultaneously (proteomics data, functional assays, tissue level phenotypes, etc.) without limitation to growth phenotypes. Furthermore, arrayed screening precludes any paracrine mediated cell-cell interactions which may obscure the effects of individual perturbations. Unfortunately, this arrayed format is significantly more expensive to perform and lower throughput[120]. Arrayed library screening often requires specialized automation for cell culture due to the need to culture large quantities of cells in isolation from one another[149]. These challenges have typically limited the widespread adoption of high-throughput arrayed screening to the biopharmaceutical industry. Because of this, pooled screening has rapidly become a key method of probing genome elements using Cas9. Pooled screens involve testing thousands of genetic perturbations in a single assay and have become increasingly popular over the past decade. Pooled screens allow for massive libraries of gene targets to be investigated in a single cell culture dish, accelerating the process of functional screening. However, pooled screens are somewhat limited in the output data they can reliably produce. Because each cell in the dish will have a unique sgRNA delivered to it, only measurements with single cell resolution (Next Generation Sequencing [NGS], fluorescence-activated cell sorting [FACS], etc.) can be used to quantitate the effect of the perturbations. Harnessing CRISPR-Cas systems effectively allows for a library of perturbations (sgRNA targeting a particular locus) to be performed in a cell population either in the arrayed or pooled format via typically lentiviral transduction. Cells successfully transduced with the perturbation must then be selected for by some means (e.g. drug resistance, FACS). Follow-up assays are then performed to help delineate which perturbations caused which functional phenotypic changes. This can be done through multiple means either by high-content

imaging (HCI) or through NGS[150–153]. HCI is beneficial for arrayed screens, allowing for quantification of spatially or temporally resolved images. This allows for a large output of phenotypic measurements while visualizing the biology. NGS is the high-throughput sequencing of DNA and RNA that performs quicker and cheaper than Sanger sequencing with the ability to quantitate reads. Massively parallel sequencing has helped revolutionize the study of functional genomics and molecular biology. In earlier years, identifying the causal mutations that led to functional changes would have been costly and labor intensive. With the advent of NGS platforms, mapping such mutations can be achieved quickly and with less costly streamlined protocols. Because of this, NGS has helped fuel pooled screens at a rapid pace. NGS enables single molecule DNA quantitation and readout of library population dynamics. Thus, a quantification can be made on the proportion of uniquely integrated library constructs in the population of cells while assessing cell viability to determine which genes after being perturbed are enriched and/or depleted. To ensure the screen results are reproducible, it is critical to validate the top hits identified from the pooled screen using an arrayed screen, preferably selecting additional sgRNAs targeting similar genes. Further biological assays should also be performed to confirm top candidates[154,155]

Although there are many diverse CRISPR tools, their use in genome scale functional screening is relatively conserved. Rather than isolating a trait and investigating what in the genome causes that phenotype, Cas9 screens function by perturbing the genome and measuring the subsequent change in a phenotype of interest. A common example of the former would be The Cancer Genome Atlas (https://cancergenome.nih.gov/). This massive research effort attempts to determine the genomic etiology of cancer through mass sequencing of patient cancer samples (phenotype→genotype). Cas9 genetic screening inverts this protocol. By purposefully introducing

a genomic perturbation with Cas9, the resulting trait can be recorded and genotype-phenotype relationships mapped.

The primary benefit of screening with Cas9 (or other CRISPR-Cas effectors) is the throughput. Rapid screening with Cas9 is made possible by the ability to perturb multiple parallel targets in the genome via a library of sgRNA. The declining cost of DNA synthesis (<1 cent/nucleotide) has enabled academic labs to construct these genome scale sgRNA libraries at low costs and with relatively low error rates, spurring Cas9's widespread adoption[156–158].

Cas9 genetic screening has most frequently been applied to screening various cancer cell lines (https://portals.broadinstitute.org/achilles)[121,159]. Cancer cell lines have several features which make them ideal for Cas9 screening. Unlike many primary cells, cancer cell lines grow well *in vitro* and can be expanded to large numbers. This is necessary to effectively screen large genome scale libraries with proper coverage[118]. Furthermore, immortalized cancer cell lines can be genetically modified to constitutively express Cas9 from a stable location in their genome, obviating the challenge of delivering the Cas9 protein in the screen. Because Cas9 is expressed in every cell being screened, only the much smaller sgRNA constructs need to be delivered. Consequently, constitutive Cas9 expression enables simplified delivery of the sgRNA library resulting in typically higher perturbation efficiencies (albeit with greater off-target rates)[160]. However, this workaround is not feasible when studying primary cells, which require the co-delivery of Cas9 and sgRNA. In addition to providing many procedural benefits, screening in cancer cell lines is often performed to identify cancer specific genetic vulnerabilities. Mapping how genomic perturbations affect cell fitness can be used to circumvent drug resistances, as well as understand underlying genetic polymorphisms driving cancer growth[121,159,161].

However, CRISPR-Cas screening is not limited to just cancer research. Screening with Cas9 has shown great utility in the study of infectious diseases[162,163]. By perturbing the target cells with libraries of sgRNA before infection with the pathogen of interest, researchers can identify genes regulating susceptibility and resistance to an infectious disease. Alternatively, the genome of the pathogen itself can be the target of CRISPR-Cas perturbations to identify essential genes controlling pathogenesis. In this way, functional screening with CRISPR-Cas can provide key information regarding the critical role host and pathogen genetics play in disease progression. This data can then be used to help determine new molecular targets for drug development, and better understand the genetic basis of divergent responses to existing therapeutics[162]. For example, several groups have recently applied Cas9 functional screening to the study of HIV, Malaria, and Tuberculosis, identifying critical genetic host factors as well as essential genes regulating infection within the genomes of pathogenic viruses and bacteria[164–166].

2.4.1 Library Design and Synthesis

The first step in developing a genomic screen using Cas9, is identifying what genomic loci to perturb. Genome wide Cas9 screens are increasingly popular due to their relatively unbiased interrogation of genome function. That being said, the choice of which genomic targets to perturb is primarily determined by the researcher's own personal interest. Regardless of what genes are perturbed there are several key library design considerations that are universally relevant.

Nearly every gene (and non-coding region) can be considered a potential target, although the endonuclease activity of Cas9 is limited to sequences with an adjacent PAM motif (NGG for SpCas9). However, recent efforts to engineer Cas9 variants which tolerate expanded PAM sequences indicate this barrier will not be a long term impediment[167]. Many *in silico* tools are

available to facilitate rapid guide RNA design, enabling large libraries of guide RNA to be designed efficiently[168].

Targeting a large library of sequences enables higher throughput interrogation of genomic elements, while a small library of genomic perturbations will lend results greater accuracy due to better library coverage[118]. The theoretical max library size is limited by several factors. DNA synthesis is an inherently error prone process itself, increasing the likelihood of inaccurate synthesis at high library size[156]. Furthermore, researchers are limited by the amount of DNA they can effectively introduce to both bacterial and mammalian cells. While libraries of greater than 10^7 sgRNAs can be easily transformed and maintained in bacteria for DNA production, the sheer number of mammalian of cells required to screen such a large library serves as a practical limit to the library search space[118,169]. Because of this, libraries greater than ~100,000 sgRNAs often require cells to be grown in large-scale cell culture setups or bio-reactors.

After choosing what genomic elements to study and how to perturb them, the library of sgRNA needs to be synthesized. There currently are a wide variety of premade sgRNA libraries available for purchase, ranging from genome wide libraries with ~10^5 sgRNAs, to more targeted libraries focused on single pathways on gene families[121,159,170]. This is often the simplest option for many labs, but limits researchers to preselected gene targets which may be irrelevant to their study. Alternatively, custom sgRNA libraries can also be generated via commercial chip based DNA synthesis[158]. This allows researchers to preselect a curated library of genomic elements for perturbation, facilitating the development of more precise experiments.

2.4.2 Delivery Systems

Choice of delivery of the CRISPR-Cas reagents is key for high editing efficiencies, proper cell uptake, reduced off-target effects, and large cargo capacities. The advantages and challenges of these different methods are outlined in **Table 2.2**.

Table 2.2. Advantages and disadvantages of different CRISPR-Cas delivery systems

Delivery Method	Advantages	Disadvantages	References
Lentivirus	-Stable gene expression -High transfection efficiency -Good for difficult-to-transfect cells (primary cells) -Large cargo capacity	-Not ideal for *in vivo* delivery	[171–176]
AAV	-High transduction efficiency -Low cytotoxicity -Relevant for *in vivo* screens	-Limited cargo capacity (4.7 kb) -Expensive	[177,178]
Electroporation	-High transfection efficiency -Good for difficult-to-transfect cells (primary cells)	-High cytotoxicity -Limited to arrayed screens	[175,179]
Lipid nanoparticles	-Low cost -Easy handling	-Low transfection efficiency -Highly dependent on cell type -Limited to arrayed screens	[175,179]
***piggyBac* transposon**	-Stable gene expression	-Potential for off-target effects -Limited scalability in pooled formats	[180,181]
Gold nanoparticles	-High transfection efficiency -Large cargo capacity -Less off-target effects	-Limited to arrayed screens	[179,182,183]

The choice of delivery method is important and should be catered to the unique needs of the experimental screen being run dependent on if it is an arrayed or pooled screen, cells being used, and cargo size. Standard delivery for most screening applications is viral, specifically lentivirus[171–174]. There are many advantages to utilizing lentivirus. It is a retrovirus with the ability

to integrate into dividing and non-dividing cells thus, creating stable transductions that can later be read via NGS. This ability also makes lentiviral transduction ideal for delivery to primary cells that are notorious for being difficult to transfect. Lentivirus is also beneficial for large gene or multiple gene cassette deliveries with its large cargo capacity[175]. One study utilized a lentiviral vector library in human cells to identify the key genes that contribute to the intoxication of cells by anthrax and diphtheria toxins[174]. Some drawbacks to this system is the random integration which poses a risk for insertional mutagenesis and unexpected off-target effects making it not ideal for *in vivo* delivery[176]. *In vitro* screens are typically fine by having 500–1000x coverage and using multiple sgRNAs per gene, reducing any risk of off-target effects. The benefits of being able to stably transduce a variety of cell types easily and quickly have ensured the continued use of lentivirus in screens.

A few studies have more recently looked at utilizing viruses for screens that do not integrate into the host genome such as the Adeno-associated virus (AAV). The idea to use AAVs for functional screens is novel and somewhat limited, but could allow functional screening of tissue level phenotypes *in vivo*. This is of great value because much of the data sets obtained from *in vitro* screens need to be taken with some amount of skepticism. There is not true physiologic representation in a dish, meaning the results of *in vitro* screens require rigorous validation. *In vivo* screening could help circumvent some of these issues, obtaining phenotypic outputs from a screen that was performed in live animals. One such study utilized the AAV to develop a unique *in vivo* CRISPR screen in conditional-Cas9 mice[177]. This study screened 49 genes known to be tumor suppressing with 5 sgRNAs for each gene. These guides were engineered into AAVs to allow for direct *in vivo* delivery into the lateral ventricle of immunocompetent living mice. Mice grew glioblastomas over time and whole-brains were then homogenized to perform downstream

analyses at the DNA, RNA, and protein level. The largest obstacle to overcome with this study was sequencing which tumors received which gene knockouts as the AAVs do not integrate into the host genome. This study designed probes to target-capture the predicted sequences of interest where expected gene knockouts would occur. This complex capture sequencing technique successfully could determine which tumors received which gene knockouts and follow up with multiple phenotypic metrics. More studies like this need to be emphasized in future research to truly recapitulate physiologic conditions during a screen. AAVs however cannot be utilized in *in vitro* screens because as cells divide the AAV will be diluted out and NGS studies that rely on genome integration could not be performed. Using clever tactics like targeted-capture sequencing as mentioned prior or reading the viral episome are possible strategies to help circumvent some of these issues for *in vivo* screening methodologies specifically. Another barrier with AAV usage is their limited cargo capacity. The cargo must be less than 4.7 kb and SpCas9 alone is encoded by a 4.2 kb sequence[178]. Utilizing conditional-Cas9 animals would be key for *in vivo* screening applications with AAVs. Other studies have performed *in vivo* screens utilizing lentiviral transduction of cancer cells *in vitro*, followed by transplantation into a mouse[184]. This simplifies downstream NGS analysis due to the integrated guides in the genomes of cell transplants.

There are also many non-viral delivery methods in place that are not frequently used, but could be useful for arrayed screens performed in multi-well plates. For non-viral delivery, because the sgRNA is not stably integrated into the target cells, an arrayed format is necessary to track which cells received which sgRNA. These methods often deliver the reagents either as mRNA or as ribonucleoprotein (RNP) complexes via electroporation or lipid nanoparticles and further summarized in Table 2[175,179]. Another effective way to introduce Cas9 and/or sgRNA into cells, and of particular benefit to functional pooled screens, is utilizing a piggyBac transposon system[180].

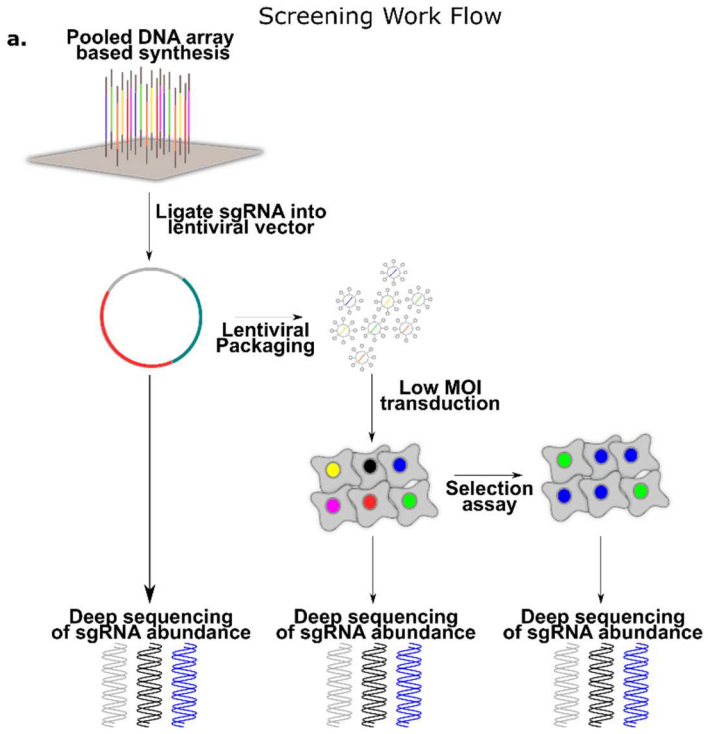
The piggyBac transposon system is a "cut and paste" mechanism and during transposition, the PB transposase will recognize inverted terminal repeat sequences (ITRs) flanking the end of a transposon vector and then move those contents and integrate them into TTAA sites on the host's DNA. This allows for creating stable cell lines. One study effectively used the piggyBac system to perform an *in vivo* CRISPR library screen utilizing PB sgRNAs in mice looking at tumorigenesis[181]. Creating an inducible Cas9 cell line with this system would be beneficial for screens and then subsequently add the pooled sgRNA library of choice. Cas9 can then be selectively turned on via doxycycline to limit off-target effects. There have also been further developments in novel ways to introduce CRISPR-Cas reagents into cell types to improve efficiency, reduce off-target effects, and increase cargo capacities such as the use of gold nanoparticles[179,182,183]. However, additional benchmarking of these non-viral delivery methods is needed to determine what screening application they are most suited for.

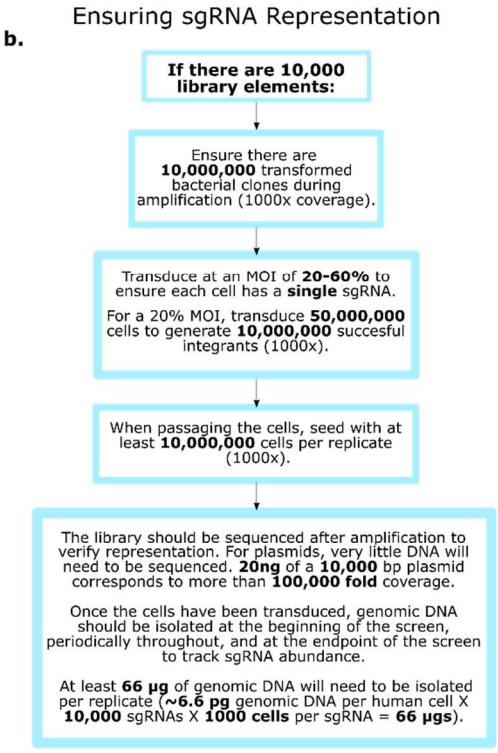
2.4.3 Library Transduction and Maintenance

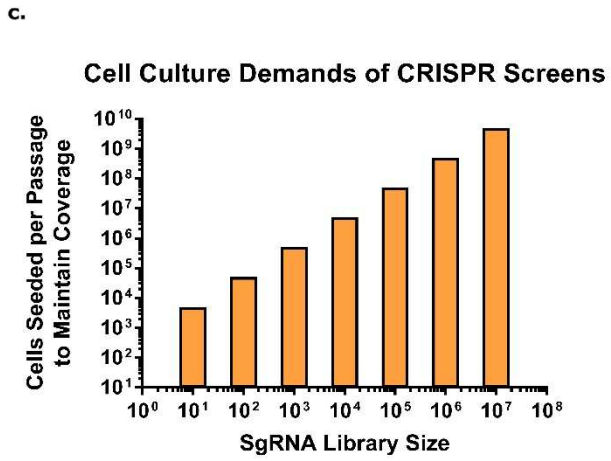
Due to the size of the Cas9 protein as well as the need to co-deliver sgRNAs, a large amount of payload must be delivered to cells to effectively perturb them. In response to these delivery challenges, lentiviral gene delivery has emerged as the primary method for delivering the sgRNA library to cells, facilitated by the virus's high genetic capacity and broad tropism[118,120,185].

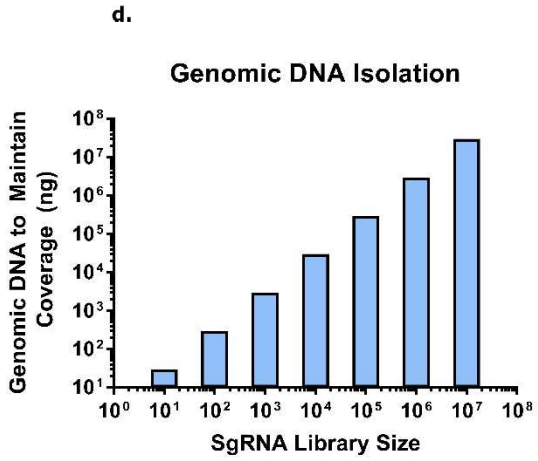
After identifying target genes and synthesizing the library of sgRNAs, the next step is ligating them into an appropriate lentiviral vector.

Figure 2.2. Mechanics of CRISPR-Cas screens: (**a-b**) shows the key steps in performing a CRISPR screen in mammalian cells. Initially the sgRNA library is ordered as a pooled tube of DNA oligonucleotides, typically synthesized commercially via chip based DNA synthesis. The library is then amplified via PCR and cloned into an appropriate lentiviral vector, insuring library coverage is maintained throughout. If the library is obtained in plasmid form (ex. pooled sgRNA libraries available from Addgene), the library simply needs to be transformed into bacteria, expanded, and sequenced to confirm sgRNA representation. Once the library is in a suitable lentiviral vector, the next step is packaging the DNA into lentivirus. Standard lentiviral packaging protocols will suffice, so long as coverage is maintained throughout the packaging. After packaging the lentivirus, a test transduction should be performed to quantify the functional titer (i.e. the actual number of cells transduced per lentiviral particle delivered). This can then be used to determine the amount of lentivirus needed to achieve an MOI of 20-60%. The transduced cells are then passaged with at least 500-1000 fold coverage of the library at each step to ensure accurate sgRNA quantitation. As the cells are passaged, it also is beneficial to store freeze and store aliquots of the library for subsequent massively parallel sequencing. At the end of the functional assay, the library is sequenced a final time to determine the relative enrichment and depletion of specific sgRNA, corresponding to target gene fitness. **(c-d)** Maintaining library coverage throughout the protocol is essential for insuring statistical confidence and preventing arbitrary library skewing. However, maintaining high coverage of the library imposes significant practical challenges for researchers attempting to implement a CRISPR/Cas9 screen. The figures above highlight the technical challenges of large library screening, and can serve as a reference for future screen design (bar plots calculated assuming 500 fold coverage of the library). As the number of sgRNA in the library increases, the scale of the experiment may outpace available resources and become untenable. Correspondingly, when planning a CRISPR/Cas9 genetic screen it is important to determine if the screen is executable in terms of lab equipment, reagents, and manpower. Once the screen has been started, the same mindfulness needs to be directed at insuring there are no library bottlenecking points which could artificially influence the results of the assay.


a.
Screening Work Flow
Pooled DNA array based synthesis
Ligate sgRNA into lentiviral vector
Lentiviral Packaging
Low MOI transduction
Selection assay
Deep sequencing of sgRNA abundance
Deep sequencing of sgRNA abundance
Deep sequencing of sgRNA abundance


b.
Ensuring sgRNA Representation
If there are 10,000 library elements:
Ensure there are 10,000,000 transformed bacterial clones during amplification (1000x coverage).
Transduce at an MOI of 20-60% to ensure each cell has a single sgRNA.
For a 20% MOI, transduce 50,000,000 cells to generate 10,000,000 succesful integrants (1000x).
When passaging the cells, seed with at least 10,000,000 cells per replicate (1000x).
The library should be sequenced after amplification to verify representation. For plasmids, very little DNA will need to be sequenced. 20ng of a 10,000 bp plasmid corresponds to more than 100,000 fold coverage.
Once the cells have been transduced, genomic DNA should be isolated at the beginning of the screen, periodically throughout, and at the endpoint of the screen to track sgRNA abundance.
At least 66 µg of genomic DNA will need to be isolated per replicate (~6.6 pg genomic DNA per human cell X 10,000 sgRNAs X 1000 cells per sgRNA = 66 µgs).


c.
Cell Culture Demands of CRISPR Screens
Cells Seeded per Passage to Maintain Coverage
SgRNA Library Size


d.
Genomic DNA Isolation
Genomic DNA to Maintain Coverage (ng)
SgRNA Library Size

This ligation is the first of many potential bottlenecks where it is important to maintain coverage of the library (typically 500-1000x or more)[118]. To effectively screen a large library of gene targets with confidence, adequate representation of the library elements is key. After packaging the library of sgRNAs into lentivirus, the target cells are then transduced at a low multiplicity of infection (MOI), typically 20-60%[118,120]. The transduction is carried out at a low MOI to ensure each cell in the screen receives a single sgRNA. The cells are then routinely passaged, ensuring at least 500-1000x library representation each passage. This high coverage is used to limit false positives and negatives due to erroneous library skewing[118]. As they grow, the cells are then assayed to physically isolate cells displaying the phenotype of interest.

2.4.4 Data Outputs

The simplest form of output data obtainable from a CRISPR screen comes from cell growth and viability assays. Because the sgRNA is genetically encoded into the cell via lentiviral transduction, NGS enables analysis of the library population dynamics. In this way, the sgRNA a cell receives both causes the genetic perturbation and functions as a unique barcode to determine through sequencing how the population is evolving in response to the screen conditions. This method of determining perturbation effects vis-à-vis sgRNA abundance is especially suited for investigating cancer cell fitness and gene essentiality. For example, in a CRISPR knockout fitness screen enriched sgRNAs indicate their target genes are nonessential or antithetical to growth. In the same way, sgRNAs that are depleted at the end of the screen indicate their target genes are essential for cell growth under the assay conditions. Using this protocol, groups have mapped novel synthetically lethal genetic interactions, investigated how particular genes affect cancer cell drug resistance, and explored how key genes impact the efficacy of immune checkpoint blockers[123,184,186].

While fitness based screening assays (to probe drug resistance or otherwise) are the simplest Cas9 screens to perform, there exist creative workarounds to probe diverse cell phenotypes independent of growth rate in a pooled format. Using an engineered fluorescent reporter system, one group utilized CRISPR screening to investigate the unfolded protein response. This pooled screen used an mCherry transcriptional reporter of IRE1α activation to facilitate cytometric isolation of cells with an activated unfolded protein response, thus enabling the enrichment of a unique phenotype separate from growth rate[187]. Utilizing similar methods researchers have been able to quantitate how genomic perturbations affect diverse cellular processes such as protein stability and the innate immune response[188,189]. However, FACS analysis is limited to predetermined targets that have fluorescently labeled antibodies commercially available, or to genetically encoded fluorescent reporter systems.

After isolating cells with the phenotype of interest in a pooled screen, the data output from CRISPR screens is not limited to simply measuring sgRNA abundance. Advancements in single cell RNA sequencing have made it possible to analyze the transcriptome of thousands of single cells utilizing a unique barcoding strategy[65]. By associating a unique barcode with each cell's transcriptome, CRISPR perturbations can be tracked and associated with transcriptomic signatures[190–192]. This enables researchers to identify (on a cell-by-cell basis) the effect of unique perturbations on the gene expression profile of a cell, and determine clusters of perturbations that may function through similar mechanisms. Unfortunately, the throughput of single cell RNA sequencing is currently not amenable for large genome scale libraries. As the cost per cell of single cell RNA sequencing decreases, this method will likely become more ubiquitous.

In contrast, when performing an arrayed screen the user is not limited to data outputs with single cell resolution. Since each unique sgRNA is physically separated from the onset of the

screen, traditional RNA sequencing (using cDNA isolated from many cells) can be performed to analyze the effect of a given perturbation on the gene expression of a cell. As well, in an arrayed format HCI can be used to examine the impact of a perturbation on cell morphology, cellular processes, as well as tissue level phenotypes. This gives arrayed screening a much wider set of cellular phenotypes which can be examined, albeit at much lower throughputs.

2.4.5 Bioinformatic Analysis of Screening Results

At the conclusion of a standard pooled CRISPR screen, the user will have a set of sequencing data representing sgRNA abundances. This raw sequencing data corresponds to which genetic perturbations are enriched or depleted for the phenotype of interest. Fortunately, there are many well validated bioinformatics tools with which to analyze this sequencing data and generate relevant conclusions. Before getting involved in design packages and computational pipelines, it is wise to perform some manual examination to identify possible outliers or mislabeled samples. This vital information could be lost if a cut and paste data dump into a statistical tool is performed too quickly. Additionally, the user should manually average the effect of multiple sgRNAs targeting one gene to compile a preliminary list of top hits. If multiple sgRNAs targeting the same gene rank highly, that gene can be listed as a hit.

After these initial steps have been taken, the user can perform a more complete in-depth analysis using a wide array of design packages. Picking the proper statistical package for the user's needs is key. Many factors must be accounted for in addition to identifying sgRNAs that are significant. Most screens typically have little to no replicates which can be a potential setback when trying to estimate the variance of reads in addition to statistical significance between treatments and controls. Additionally, researchers must utilize a computational tool that takes sgRNA variability into account in terms of specificities and efficiencies. Finally, knockout screens

often result in only a few sgRNAs that tend to dominate the reads in positive selection. A successful algorithm will require robust read normalization. Some older algorithms such as baySeq, DESeq, edgeR, and NBPSeq have been used with some success[193–196]. They are commonly used algorithms for RNA-seq analysis, but limited to the sgRNA level in terms of statistical significance of hits.

Some of the more common tools for pooled screens that show robust results are MAGeCK, caRpools, and CRISPRcloud[197–199]. In brief, MAGeCK robustly identifies positively and negatively selected sgRNAs and genes simultaneously in genome-scale CRISPR-Cas9 knockout screens. Its four steps include read count normalization, mean-variance modeling, sgRNA ranking, and finally gene ranking. Interestingly, MAGeCK can assess relevant biological pathways by reporting positively and negatively selected pathways based on gene rankings in the pathway. This algorithm has been shown to outperform existing methods with its high sensitivity and low false discovery rate[197]. In addition there is now MAGeCK-VISPR which was developed for quality control and visualization of CRISPR screens[200]. CaRpools is a user-friendly R package that does not require prior programming knowledge. CaRpools provides the user with biological information for every hit with external links to databases. This package incorporates screening documentation into the analysis process to generate a comprehensive report. CRISPRcloud uniquely allows the user to deposit sequencing files confidentially and analyze them in a cloud-based online system.

Arrayed screens analyze more advanced phenotypes than simply growth and thus, often utilize HCI. The vendors for many of these HCI platforms provide their own statistical packages for analysis. The largest challenge with these packages is they require extensive user interaction and can often lack statistical power as the data return from HCI is rich. Many packages are available and have been reviewed[201]. A few common open-source ones are CellProfiler and EBImage[202,203]. Commercial software is available as well such as Columbus or MetaXpress. After

features have been measured and collected with imaging software, this data must be analyzed for statistical significance. Statistical packages for R are commonly used such as *cytominer* (https://github.com/ CellProfiler/cytominer/) to assess morphological cell features.

When looking at combinatorial screens, the user must assess the phenotypic effect when a combination of sgRNAs target the same cell. The initial combinatorial studies were performed in yeast in mass arrays known as synthetic genetic arrays (SGA) where a gene deletion could be crossed systematically with a deletion mutant array that contains all possible knockout ORFs in the genome[204]. More recently groups have scaled up this technology utilizing CRISPR-Cas for *de novo* mapping of genetic interactions in mammalian cells[186,205]. This requires additional statistical packages such as the dual CRISPR software pipeline constructed from Python, R, and Jupyter Notebooks (http://ideker.ucsd.edu/papers/rsasik2017/)[206]. Other tools are also available such as TOPS which is another open-source package to analyze and visualize data from functional genomic gene-gene and gene-drug interaction screens[207].

Single-cell screens have benefited greatly from the Seurat pipeline (http://satijalab. org/seurat/)[208]. Seurat is an R package designed to analyze single cell RNA-seq data. This package uses canonical correlation analysis to determine shared correlation structures across data sets. After alignment, cells are transposed on a 2D plot (i.e. t-SNE) into clusters with shared transcriptomic reads. Clustering can identify cell types across conditions looking at shifts and cell-specific transcriptomic responses. Seurat allows users to identify and interpret sources of heterogeneity at the single cell transcriptomic level.

2.4.6 Validating Results

CRISPR-Cas genome wide screening is valuable because it provides an unbiased way to probe genome function, but the screen is only the first step in identifying functional genomic

elements. After identifying potential genes of interest via a perturbation screen and subsequent bioinformatics analysis, significant work must be done to validate these targets. In this way, CRISPR-Cas genome wide screening can be thought of as hypothesis generating experiments, which guide future genomic characterization efforts.

Initial validation is focused on ensuring the effects of the perturbations are consistent and reproducible. To this end, CRISPR screens often utilize multiple sgRNAs targeting each genomic element[170,209]. Ideally, one would expect all sgRNAs targeting the same gene to have similar phenotypic effects. This redundancy provides researchers with a way to ensure that the hits identified from the screen are due to the intended sgRNA mediated genetic perturbation, rather than off-target effects or random noise Beyond that, potential hits can be sub-screened in a smaller more focused library[161]. This step provides researchers with greater confidence in their results, and helps narrow down target genes for further biological analysis. New sgRNAs targeting potential genes of interest can also be designed and used to verify reproducibility[210]. Furthermore, it can be informative to analyze data sets with different perturbational technologies (CRISPR, CRISPRi, RNAi) to ensure the data is reproducible across multiple systems[210]. However, each of these perturbations will have their own unique biases and limitations which may affect the reproducibility of data across different systems[118].

After several top hits have been established, a key validation step is checking the effects of the sgRNA of interest individually, outside of the context of the pooled screen, to remove any confounding paracrine effects. At the same time, if the gene of interest is protein coding, a western blot can be used to ensure the gene is completely knocked out by its cognate sgRNA[210]. To generate further confidence in top hits, Cas9 can also be used to generate a clonal population of cells with identical genetic perturbations. Genotyping of this clonal population should then be performed to

ensure the gene of interest is effectively knocked out via frame shifts or the introduction of stop codons. After establishing the clonal cell line, robust phenotypic data can be collected to fully interrogate the functional role of the gene of interest. The ultimate step in verifying the effect a gene has on cell phenotype is to restore gene function in the knockout cell line via delivery of cDNA encoding the gene of interest[211]. If the gene of interest is truly the cause of the phenotypic change, cDNA delivery should restore the wild type phenotype to the knockout cell line. If necessary, researchers can also begin testing the perturbation in multiple cell types. While genotype-phenotype relationships may not be consistent across multiple cell types, this step can provide a way to better understand the biology underlying the phenotypic effect of the genetic perturbation[118]. As well, small molecules or monoclonal antibodies targeting the gene(s) of interest can serve to verify the biological mechanism underlying the effect of the perturbation.

2.5 Challenges and Limitations

Although Cas9 based genetic screening is a rapidly maturing technology, there are still many technical challenges that have yet to be resolved. One large obstacle when it comes to performing pooled library screens in a dish are the potential effects of paracrine signaling. In a pooled format it is difficult to assess and eliminate cross-talk between neighboring cells in a dish that may all have unique genomic knockouts. Because of this, the importance of certain genes can be easily missed if the gene function can be rescued by nearby cells. For example, if a growth factor is knocked out in a specific cell its neighbor may continue to release the growth factor, preventing a true knockout phenotype from appearing. In this way, a pooled genome wide screen may still not identify all genes that are vital for a given phenotype.

Another issue with pooled approaches is the limit to phenotypic outputs that can be read. The researcher is typically restricted to measuring cell proliferation or survival. Additionally, there

can be efforts to look at phenotypes that FACS can select and sort through such as fluorescence or cell surface markers. More complex phenotypes will be difficult to measure in a pooled screen with reliability. In the future, cheaper robotics that can perform arrayed screens with unique perturbations in each well of multi-well plates will likely allow for more complex tissue level phenotypes to be assayed. In addition, this sort of high-throughput arrayed screening would remove many of the paracrine effects that may confound results as mentioned previously. If a gene that is being studied is known to be essential for cell viability, it cannot be studied in a complete CRISPR knockout screen when assessing for additional phenotypes. Performing a knockdown study utilizing dCas9 would be more appropriate. Additionally, genes that retain their function at low expression levels may easily be missed in knockdown studies and be better performed with a complete knockout screen.

Other issues may arise with false positives and false negatives. In particular, although uncommon, an in-frame repair could occur during a standard positive selection knockout screen resulting in a gain-of-function mutation[212,213]. This issue is rare enough to not cause vast concern, but something to still be mindful of. More commonly false positives can occur with genes that have a high copy number such as oncogenes. When performing a standard Cas9 knockout screen, these genes will consistently be cleaved leading to multiple double strand breaks and eventually too many will cause cells to apoptose thus, mistakenly assuming that gene was essential for cell fitness. A gene that may not truly have much of an effect on fitness can falsely appear to if the target site is in one of these amplified regions with a high gene copy number thus, inducing many more double strand breaks by Cas9 than is typical[214–216]. This can be problematic when performing cancer screens. Many groups have looked at this in detail looking at several cancer cell lines, genes, and sgRNAs for analysis of this amplification effect[214–216]. Aneuploid cell lines produced false

positives that mapped to amplified regions of the genome. CRISPR-mediated lethality of cells was independent of transcriptional halting, thus showing this is due to double strand breaks and not gene knockout. Previous studies have shown similar discoveries such as targeting the oncogenic BCR-ABL gene fusion that is present in high copy number in K562 cells and notorious for making up the Philadelphia chromosome in chronic myelogenous leukemia. Cas9 targeting resulted in decreased cell viability independent of the target genes function themselves[217]. Ways to prevent these false positives would be to use CRISPRi which do not cut the genome and only offer transcriptional repression. However even with CRISPRi, other errors can occur especially when dealing with bidirectional promoters causing silencing of multiple genes instead of just the gene of interest. Attempts can be made to remove sgRNAs with massive off-target effects or exclude them from analysis[218]. Utilizing an inducible Cas9 can also be an effective solution to select specifically when to turn on Cas9 with the use of doxycycline.

False negatives come with their own share of complications. If a sgRNA has relatively low activity it can inadvertently be read as a negative result in a screen. Machine learning approaches can help circumvent some of these issues to design and include only sgRNAs with high activity which has been actively utilized by groups[170,219,220]. However, *in silico* sgRNA design has its own share of challenges. When utilizing available online tools, the researcher needs to be aware of the underlying rules to limit off-target effects and increase effectiveness applied by the tool developers. There are also constant updates to gene annotations that need to be ensured for their accuracy and quality. In addition to using computational tools to predict guide efficacy, efforts can also be made to modify the sgRNA scaffold itself to improve activity[221].

One of the large concerns with the use of CRISPR-Cas systems for screens is the possibility of off-target effects. Because sgRNA libraries can contain more than 10^5 different guides,

comprehensive individual sgRNA validation and testing is not possible. Multiple studies have shown that Cas9 can tolerate some mismatches between the sgRNA and target sequence allowing for targeting of the wrong gene[110,222–224]. The farther these mismatches are from the PAM sequence the more likely these mismatches will be tolerated[225]. It has also been shown that small insertions and deletions are somewhat tolerated as well leading to bulging of the sgRNA or target sequence[224]. Predictive scores have been developed to help the researcher in picking appropriate sgRNAs[226]. Additional Cas9 options are the high fidelity Cas9 (SpCas9-HF1) or the enhanced specificity Cas9 (eSpCas9)[227,228]. Many benefits have been shown by delivering Cas9 as a protein instead of a gene in a plasmid as the protein will act immediately and then be quickly degraded which eliminates the constant peaks in expression from a promoter[229]. One strategy to ensure a positive is true and not from an off-target effect is through validation and ensuring that other reagents targeting that same gene have that same phenotype. However, when performing large pooled screens there will be multiple sgRNAs targeting the same gene or noncoding region. Effects of a single sgRNA will be less problematic when multiple sgRNAs are targeting that region allowing for some consistency and realization of an off-target effect.

Another challenge is working with PAM sequence restrictions. SpCas9 has a PAM sequence that is more abundant in exons and thus coding regions of the genome which tend to be more GC rich. Other nucleases such as Cpf1 has a PAM sequence that is more abundant in introns which are more AT rich[230]. This is an important factor to keep in mind when selecting a nuclease for screening applications. Performing noncoding functional screens utilizing CRISPR-Cas systems to tile sgRNAs may benefit more from a nuclease such as Cpf1 than SpCas9. One group effectively engineered SpCas9 to recognize different PAM sequences[167]. This can increase

specificity and reduce off-target effects while selecting a PAM that is appropriate and unique for the researcher's screening needs.

One often untapped tool for CRISPR-Cas screening is harnessing HDR to insert exogenous genes of interest into the host genome. With HDR's relatively low efficiency compared to NHEJ, it has proven to be difficult to benefit from this technology and perform large knock-in screens at endogenous loci. Knock-in screens can provide valuable information when assessing the roles of knocked-in promoters or repressors on gene function or knocking in mutated genes to mimic disease states. As well, knock-in screens using HDR would preclude the possibility of random lentiviral integration causing confounding effects on cell phenotype. Because of this, more research should be done on pushing the cell to favor HDR over NHEJ. One such study used blocking mutations to increase HDR efficiency[231]. They introduced silent mutations in either the PAM or sgRNA target sequence of the donor strand. These mutations prevented Cas9 from re-cutting the target sequence once the desired donor was introduced. Greatest efficiency of this is achieved when the mutation is closest to the cut site. This distance can also be optimized to focus on either a homozygous edit or heterozygous edit in the cell depending on the researcher's specific needs (homozygous edits are more likely when the mutation is closest to the cut site and heterozygous edits are more likely when further). Utilizing this blocking method, another study successfully performed a large screen utilizing HDR and saturation mutagenesis to determine function of regulatory elements[232]. They utilized a library of all possible 6-bp combinations to insert into exon 18 of the breast cancer susceptibility gene BRCA1 to measure transcript abundance. They had a similar approach for the lariat debranching enzyme gene DBR1 to measure the relative effects on growth and function. Interestingly, HDR could also be harnessed to create a knock-in pooled library of sgRNAs in place of typical lentiviral delivery creating cells with stably

integrated guides[233]. This could circumvent issues with off-target effects from lentivirus and avoid gene shuffling. Highlighting the potential of HDR based screening approaches, one group recently performed a large-scale multiplexed HDR CRISPR screen in yeast, utilizing a fusion protein to enhance HDR efficiency[234]. They increased editing efficiency more than 5-fold with use of the fork head protein homolog 1 transcription factor (Fkh1p) fused with the DNA binding protein LexA creating a LexA-Fkh1p fusion protein. This fusion protein recruits donor DNA to the double-strand break site. Utilizing HDR, they incorporated unique barcodes into cells. In addition, they performed saturation editing of a gene encoding for the phospholipid transfer protein SEC14. They incorporated all possible amino acid combinations to identify amino acids critical for chemical inhibition of lipid signaling. Ideally, combining multiple strategies will improve HDR at the greatest efficiency when performing knock-in functional screens. Additionally, a researcher could use base-editing techniques to perform a targeted knock-in screen instead of HDR. CRISPR base-editing techniques can modify individual nucleic acid base pairs within the target genes. This is especially beneficial to edit single nucleotide polymorphisms (SNPs). Groups have used this technique to identify novel mutations in drug resistance[129,130]. Overall, screening from endogenous loci using HDR or base editors, although limited to unique screening needs, has significant unexplored potential for investigating genomic function.

Another challenge lies in the large reliability researchers place on cell lines to perform many of these pooled functional screens. Many of these cell lines may not adequately model human disease and functional genomics. Additionally, unless kept at a low passage number, cells can begin to change over time with varying mutations, epigenetic changes, and chromosomal changes. Ideally primary cells, human tissues, or *in vivo* screens should be the gold standard. In the next chapter, we will perform an *in vivo* screen in the teratoma. Validating findings in multiple

model systems with different techniques is critical. However, with this is the caveat that obtaining different results in different cell lines is permissible if it further explains a critical phenotype unique to the biology of these different systems. Additionally, plating cells with the correct growth medium and environmental parameters can be a challenge or whether they even properly plate in 2D. Studies have shown that many human cell types change their physiology in 2D or cannot be cultured at all. For instance, pancreatic cells are notorious for being difficult to culture in 2D and have lasted at most a mere week before huge losses in cell viability[235]. More efforts need to be placed in 3D culture systems and biomimetic environments to ideally model true physiology—this will be more realized in the next chapter as we perform a three-dimensional screen in the teratoma.

2.6 Future Directions

As technical challenges limiting Cas9 based genomic screens are resolved, their ability to inform our understanding of disease progression and treatment will rapidly evolve. By utilizing the expanding toolbox of genetic perturbations and better integrating multiomics data for downstream validation, screens will be able to identify functional elements in the genome more rapidly and accurately. At the same time, expanding screens to patient derived cell types (iPSCs, tumor biopsies, etc.) will better model human pathologies while providing a potential way to identify patient specific disease vulnerabilities.

Because the majority of human diseases are polygenic (rather than mendelian) there is a clear need for screens which investigate multigene interactions[236,237]. Towards this end, investigators have recently developed dual knockout Cas9 vectors which deliver two unique sgRNA to identify synthetically lethal genetic interactions in cancer cell lines[186,238]. In parallel, other researchers have developed alternative dual knockout systems, using a combination of orthogonal Cas9 variants from different bacteria. By utilizing both SpCas9 and SaCas9 (each with

their own cognate sgRNAs) they effectively reduce interference between delivered sgRNAs in a dual knockout screen[239]. Moving forward, characterizing a greater number of gene combinations will generate an improved understanding of the genetic basis of non-mendelian diseases. In addition, expanding combination gene perturbations beyond knockouts will provide scientists with a better understanding of directional genetic interactions. In order to characterize these directional interactions, researchers have recently implemented a dual knockout and activation screen in cancer cells to better understand therapeutically relevant genetic interactions networks[143]. Looking forward, integrating multiple different perturbation types in combination has the potential to generate unique datasets with which to probe genomic interactions. For example, integrating inducible Cas9/sgRNA constructs with pooled screening could elucidate temporal dependencies underlying dynamic genetic interactions[240].

Beyond probing exon function, there is an increasing understanding that the noncoding region of the human genome plays a significant role in disease progression across a wide variety of pathologies[241]. In order to better understand this relationship, there have recently been several parallel efforts to map the function of the noncoding portion of the genome using Cas9[117]. While wtCas9 is ideal for inducing frameshift mutations in the coding regions of exons, probing the noncoding portion of the genome is more challenging because insertions and deletions are less likely to impact structure and function. To overcome this challenge, CRISPR pooled screening of noncoding loci has primarily focused on using multiple tiled sgRNA to create indels across entire noncoding regulatory sections of the genome to determine functional hotspots. These strategies have identified critical components of endogenous enhancers, as well as novel regulatory elements in unannotated regions of the genome[242–244]. Combining this approach with novel downstream single cell assays (single cell RNA seq, etc.) should further aid in rapidly characterizing the

structure-function relationship of the noncoding genome. Furthermore, screens utilizing the full CRISPR perturbation tool box will provide researchers with even more novel data sets with which to assay the noncoding genome.

While Cas9 genetic screening has enabled systematic characterization of a broad range of cancer cell lines (via the Broad Institute's Project Achilles among other work), screening primary cells is still in its infancy. Although there is a wealth of information to be gained from screening cancer cell lines, as discussed above they are not ideal models for healthy cells or diseases other than cancer. Screening in primary cells would better model the *in vivo* genetic and epigenetic profile of the cells of interest, while simultaneously allowing for patient-specific screening strategies to be developed. Because primary cells can be obtained from individuals (or mice) afflicted with nearly any disease, a broader range of disease-specific screening strategies can be developed. As well, screening in primary cells would allow scientists to unravel the genomic mechanisms underlying the function of various healthy cell types. Primary cell screening has so far been limited to immune cell types which grow sufficiently *in vitro*. As a proof of principal, two groups have recently described a protocol for lentiviral knockout CRISPR screens in mouse primary immune cells, identifying key regulators of the innate immune response and plasma cell differentiation[188,245]. To push this technology forward, the editing efficiency of Cas9 in primary cells needs to be further optimized to allow for large library screening in many primary cell types. In parallel, improving *in vitro* primary cell culture techniques will drastically improve the ease of primary cell screening protocols. Looking ahead, transitioning this technology toward screening iPSCs could provide a novel method to understand biological development and patient-specific pathological phenotypes. Although iPSC CRISPR screens are still in their infancy, one group recently published a method using Cas9-mediated homologous recombination to fluorescently tag

endogenous proteins in developing iPSCs[246]. This method would allow researchers to track the temporal expression and localization of diverse cellular proteins over the course of iPSC differentiation.

As an alternative way to more accurately model cell phenotypes, several groups have independently developed *in vivo* CRISPR screening protocols. *In vivo* CRISPR screening typically involves delivering a library of sgRNAs to a tumor cell line *ex vivo*, implanting the cells into a mouse model, and then tracking which sgRNAs are enriched or depleted as the tumor grows. This method has been used to effectively identify genetic vulnerabilities to immune checkpoint blockers, as well as track genetic drivers of metastasis[184,247]. These *in vivo* screening methods represent a more robust contextual model with which to analyze cell function, and warrant additional investigation. Other efforts to screen cells in a context that better matches their native environment have utilized 3D culture systems and organoid models. While 3D and organoid models necessitate arrayed screening due to their multicellular architecture, the ability to investigate tissue level phenotypes has immense implications for functional screens. In 2015, one study described a small scale CRISPR knockout screen in an organoid model, investigating genetic elements controlling the differentiation of unpolarized basal progenitors into airway epithelium[248]. Although screens involving 3D culture models will certainly be restricted to small libraries of perturbations, their ability to dissect tissue level phenotypes guarantees their utility to the biomedical community.

As CRISPR screens become more commonplace, it is necessary to stress the importance of using diverse output data to validate results. While sgRNA abundance provides valuable information regarding which genes are essential for a cellular phenotype, it provides little to no mechanistic data with which to understand gene function. To better understand the biology

underlying CRISPR screen results, future research needs to be done on how to best integrate multiomics data with pooled CRISPR screens. Utilizing advances in proteomic and metabolomic measurements has great potential to complement next generation DNA and RNA sequencing technologies already common place in CRISPR screens. As mass spectrometry pushes closer toward single cell resolutions, this data will only become more robust, opening up new avenues for understanding the results of pooled screens[249,250].

Although CRISPR knockout screening via the NHEJ repair pathway has seen widespread adoption, knock-in screening via the HDR templated repair mechanism has been less utilized due to its relatively low efficiency. Many parallel efforts are currently underway to improve the efficacy of HDR mediated gene editing, paving the way for library scale knock-in screening[234,251,252]. Knock-in screening using HDR to scarlessly insert a mutagenized DNA sequence at its endogenous locus has many unexplored applications. In the future, researchers could use HDR to perform site directed mutagenesis of complex mammalian proteins in their endogenous loci, enabling the engineering of post-translationally modified proteins which may not be amenable to production in yeast or bacteria. This same method could also be used to engineer mammalian cell lines with novel metabolic pathways for use in biopharmaceutical production.

The past half-decade has seen rapid development of novel CRISPR-Cas based tools with which to investigate genomic function. At the same time, *de novo* DNA synthesis and *in silico* sgRNA design tools have quickly become mature technologies, resolving many of the technical challenges preventing the widespread adoption of CRISPR-Cas genetic screens. Consequently, CRISPR-Cas genetic screening has transitioned from exciting new academic research, to a ubiquitous technology with few barriers to use. Looking forward, it now seems plausible that the many functional screens ongoing in immortalized cancer cell lines will lead to a complete mapping

of cancer specific gene function and genetic interactions. While this research has great potential to inform our understanding of cancer etiology and drug candidate efficacy, the immense genetic variation in patient cancer samples limits the translational relevance of cell line based genetic screening. In addition, conclusions drawn from screens performed in cancer cell lines may have limited relevance to other disease phenotypes. This genetic variation between patients and cancer cell lines necessitates the development of patient-specific CRISPR-Cas screening protocols. Building off existing cancer mapping initiatives, CRISPR-Cas functional screening efforts in patient-derived cells should one day help oncologists predict treatment efficacy and inform drug choice. In parallel, future screens in patient derived iPSCs will allow researchers to expand the range of disease phenotypes CRISPR-Cas functional screening can investigate. In this way, CRISPR-Cas screening can contribute to a growing body of research underlying precision medicine and personalized therapeutics. In the next chapter we will utilize a CRISPR-Cas9 genomic screen in the context of the teratoma.

2.7 Acknowledgements

Chapters 2 is in part reprints of the following materials of which the dissertation author was one of the primary investigators and authors of this paper:

Chapter 2, in part, is a reprint of the material as it appears in Ford K*, McDonald D*, Mali P. Functional Genomics via CRISPR-Cas. J Mol Biol. 2019 Jan 4;431(1):48-65. doi: 10.1016/j.jmb.2018.06.034. Epub 2018 Jun 28. PMID: 29959923; PMCID: PMC6309720

*Both of these authors contributed equally

3 Engineering Teratomas via Genetic Perturbations

3.1 Abstract

We propose that the teratoma, a recognized standard for validating pluripotency in stem cells, could be a promising platform for studying human developmental processes. Using pooled CRISPR-Cas9 knockout screens, we showed that teratomas can simultaneously assay the effects of genetic perturbations across all germ layers. We found that TWIST1, RUNX1, ASCL1, CDX2, and KLF6 knockouts resulted in reproducible shifts in lineage abundance consistent with known biology. All of these knockouts had effects on cell types from multiple germ layers. Additionally, we used the teratoma to model human neural disorders, specifically Pitt-Hopkins, Rett, and L1 syndromes. We generated a CRISPR knockout library targeting the genes responsible for these syndromes in teratomas: TCF4, MECP2, and L1CAM respectively and assessed their downstream differential gene expression which were consistent with known pathways related to these genes. Taken together, the teratoma is a promising platform for modeling multi-lineage development and pan-tissue functional genetic screening.

3.2 Introduction

We propose here the use of teratomas as a model for studying human development [42]. The teratoma displays multi-lineage differentiation to all germ layers, has vascularized 3D structure, bears regions of complex tissue-like organization, and is relatively straightforward to implement. Teratoma formation is the gold standard to validate pluripotency and developmental potential of hPSC lines [54,55].

There has also been some progress in utilizing the inherent differentiation potential of teratomas to derive highly sought-after cell types. However, the semi-random nature of teratoma

development has previously made characterization of teratomas difficult, as the different lineages can often be found in close spatial proximity.

After rigorous characterization of the teratoma, we deem it essential to validate the use of the teratoma as a model for studying developmental biology. Do the cells within the teratoma develop as would be expected in normal development and follow standard differentiation trajectories or are they developing in non-canonical ways? Performing a developmental knockout screen within the teratoma will help elucidate some of these questions as well as potentially find novel biology previously difficult to study in standard 2D stem cell culture systems or organoid systems without the use of human embryos.

We hypothesized that the advent of high-throughput single cell gene expression profiling via droplet based methods [61–67], and simple genetic perturbation toolsets such as CRISPR-Cas9 could enable us to address this challenge by enabling systematic analysis and perturbation of teratomas at the single cell level [187,192,253–258]. We established a comprehensive experimental and computational framework to systematically analyze, perturb and modulate human PSC-derived teratomas to evaluate their potential for modeling human development.

3.3 Materials and Methods

3.3.1 Method Details

3.3.1.1 PGP1-Cas9 Clone Generation

The PGP1 human induced pluripotent stem cell line was a kind gift of Dr. George Church at Harvard Medical School. The sgRNA targeting AAVS1 locus of the human genome (spacer sequence GGGCCACTAGGGACAGGAT) was cloned into the Lenti-guide-puro plasmid (Addgene #52963). To generate the knockin donor plasmid, we cloned the CAG promoter followed by a cassette of co-expression of spCas9 and EGFP splitting via the P2A sequence into

the pCR4-Blunt-TOPO vector (Thermo Fisher Scientific). Two homology arms were amplified from upstream (804 bp) and downstream (837 bp) of the sgRNA targeting site in AAVS1 genomic locus and constructed into the donor plasmid flanking the CAG-spCas9-P2A-EGFP cassette. Between the upstream homology arm and the CAG promoter, we inserted a splice acceptor sequence following by a T2A linked blasticidin resistance gene.

Human iPSC PGP1 cells were electroporated using 4D-Nucleofector system and P3 Primary Cell X kit (Lonza) according to the manufacturer's instruction. Briefly, the PGP1 cells were dissociated into single cells. $1x10^6$ cells were mixed with 100 µl nucleofection reagents and 10 µg DNA (5 µg Cas9 donor + 5 µg sgRNA) and electroporated. The cells were recovered with pre-warmed medium and then cultured on inactivated MEF feeders in 10 cm dishes with mTeSR medium supplemented with 0.5 µM ROCK-inhibitor. Afterward, the mTeSR medium without ROCK-inhibitor was refreshed daily. 2 µg/ml blasticidin were added into the culture medium 7 days after electroporation. The cells were cultured without passage until clones emerged on the plate. The clones were checked under the microscope and those with EGFP expression were picked up and expanded individually.

To detect genomic integration, the genomic DNA from cultured cells was extracted using DNeasy Blood & Tissue Kits (Qiagen). Approximately 500 ng of genomic DNA was used for each PCR reaction using KAPA HiFi HotStart Ready Mix (Kapa Biosystems). The PCR amplification of the left and right arm utilized primers that amplified regions spanning both the PGP1 AAVS1 endogenous locus and the engineered cassette (**Figure 3.2B**).

The primer sequences are listed below.

Left_arm_forward	ACTTCCCCTCTTCCGATGTTG
Left_arm_reverse	ATTGTAGCCGTTGCTCTTTCA

Right_arm_forward	GAGCAAAGACCCCAACGAGAAGC
Right_arm_reverse	CTGCCTGGAGAAGGATGCAGGA

This was further validated by direct Sanger sequencing of the arms (**Figure 3.2A),** The activity of Cas9 in the PGP1-Cas9 cells was validated by the generation of indels at the expected position when guide RNAs were introduced.

3.3.1.2 sgRNA Design

The CRISPR-KO sgRNA sequences targeting transcription factor genes were obtained from the GPP sgRNA Designer web tool (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design, accessed February 2018) as follows. The 24 gene symbols in the table below were converted to Entrez gene IDs using Bioconductor package org.Hs.eg.db_3.5.0, and the resulting IDs were submitted together with the following parameters: enzyme Sp, taxon human, quota 50, include unpicked. From the resulting output, the two guide sequences with the highest "pick order" were selected for each target gene. To check the validity of each guide sequence, the corresponding context sequence was compared to the human reference genome at the predicted cut location using Bioconductor package BSgenome.Hsapiens.UCSC.hg38_1.4.1, and the cut location was confirmed to be fully within the target gene coding sequence determined using Bioconductor package TxDb.Hsapiens.UCSC.hg38.knownGene_3.4.0.

Gene symbol	Entrez ID	sgRNA-1	sgRNA-2
SOX17	64321	GGCAACGGGTAGCCG TCGAG	AGGGCGAGTCCCGT ATCCGG
CDX2	1045	CCGCAGTACCCGGAC TACGG	CAAATATCGAGTGG TGTACA
HNF4A	3172	GGGACCGGATCAGCA CTCGA	GCAATGACTACATTG TCCCT

GATA4	2626	TGTGGGCACGTAGACTGGCG	CCGGCTTACATGGCCGACGT
GATA6	2627	CGGGACGCCTCAGCTCGACA	GCCGACAGCGAGCTGTACTG
RUNX1	861	CTGATCGTAGGACCACGGTG	TGCTCCCCACAATAGGACAT
FOXA2	3170	ATGAACATGTCGTCGTACGT	TCCGTGAGCAACATGAACGC
PDX1	3651	GGAGAACAAGCGGACGCGCA	TATTCAACAAGTACATCTCA
NKX2-1	7080	GCGAGCGGCATGAACATGAG	GGTTGGCGCCGTACCATCCG
NKX2-5	1482	GTAGGCACGTGGATAGAAGG	GAAGACAGAGGCGGACAACG
SOX9	6662	ACGTCGCGGAAGTCGATAGG	TTCACCGACTTCCTCCGCCG
PROX1	5629	AGTGTCCACAACTTGCGACA	CGGGTTGAGAATATAATTCG
SNAI1	6615	GGGACTCTCCTGGAGCCGAA	TGTAGTTAGGCTTCCGATTG
TWIST1	7291	CGGGAGTCCGCAGTCTTACG	AGCGGGTCATGGCCAACGTG
ASCL1	429	CCAGGTTGACCAACTTGACG	AAACGCCGGCTCAACTTCAG
NEUROG1	4762	CCGCATGCACAACTTGAACG	TTGGTGTCGTCGGGGAACGA
KLF6	1316	TCTGAGGCTGAAACATAGCA	GCTGACCAAAACTTCGCCAA
KLF2	10365	GGTTCGGGGTAATAGAACGC	CTTCGGTCTCTTCGACGACG

HES1	3280	GTGCGAGGGCGTTAATACCG	AGCCAGTGTCAACACGACAC
FOXG1	2290	AGCGCGTTGTAGCTGAACGG	CCGCGCCACTACGACGACCC
TULP3	7289	GGAGTATGACAGTTCACCAA	TGAAAGTGTGAACTTCGATG
MYOG	4656	TTACACACCTTACACGCCCA	TCGAACCACCAGGCTACGAG
GATA3	2625	TCCAAGACGTCCATCCACCA	CAGGGAGTGTGTGAACTGTG
FGFR2	2263	CTTAGTCCAACTGATCACGG	TGACCAAACGTATCCCCCTG

Gene symbol	Entrez ID	sgRNA-1	sgRNA-2	sgRNA-3
TCF4	925	GTGGACATCGGAGGAAGAC	TGTCCACTTTCCATCGTAG	CAAACGTTCATGTGGATGC
MECP2	204	GCTCCATCATCCGTGACCG	AAAGCCTTTCGCTCTAAAG	TTGCGTACTTCGAAAAGGT
L1CAM	897	GCGTCCGGTGTCATTGGCC	GCGTACTATGTCACCGTGG	GCCAGTACCGAACTGGATG

3.3.1.3 Library Preparation

The lentiviral backbone plasmid for the sgRNAs was the CROPseq-Guide-Puro vector (Addgene #86708). To create the sgRNA library, individual sgRNAs were PCR amplified utilizing overlapping forward and reverse primers custom designed with flanking sequences compatible with the BSMBI restriction sites **(Table S7)**. The lentiviral backbone was digested with BSMBI (New England Biolabs) at 55°C for 3 hours in a reaction consisting of: CROPseq-Guide-Puro backbone, 5 µg, Buffer NEB 3.1, 5 µl, BSMBI, 5 µl, H20 up to 50 µl. After digestion, the vector

was purified using a QIAquick PCR Purification Kit (Qiagen). Each sgRNA was then individually assembled via Gibson assembly.

The Gibson assembly reactions were set up as follows: 1:10 molar ratio of digested backbone to sgRNA insert, 2X Gibson assembly master mix (New England Biolabs), H20 up to 20 µl. After incubation at 50°C for 1 h, the product was transformed into One Shot Stbl3 chemically competent Escherichia coli (Invitrogen). A fraction (150 µL) of cultures was spread on carbenicillin (50 µg/ml) LB plates and incubated overnight at 37°C for 15-18hrs (miRNA constructs required longer incubation times). Individual colonies were picked, introduced into 5 ml of carbenicillin (50 µg/ml) LB medium and incubated overnight in a shaker at 37°C. The plasmid DNA was then extracted with a QIAprep Spin Miniprep Kit (Qiagen), and Sanger sequenced to verify correct assembly of the vector and to extract barcode sequences.

To assemble the library, individual sgRNA vectors were pooled together in an equal mass ratio along with 5 non-targeting control (NTC) sgRNAs which constituted 50% of the final pool.

3.3.1.4 Viral Transduction

For viral transduction, virus was added at a low MOI (ensuring a single barcode/cell or a single sgRNA/cell) to stem cells at 20% confluency alongside polybrene (5 µg/ml, Millipore) in fresh mTeSR medium. The following day, medium was replaced with fresh mTeSR. Appropriate selection reagent was added 48 hrs after transduction (puromycin [0.75µg/µL] for CRISPR KO screen) (Thermo Fisher Scientific) and was replaced daily. For editing in CRISPR KO screen, selection was continued for 5 days prior to use for teratoma formation in mice.

3.3.1.5 sgRNA Editing Rate Validation

We individually transduced each sgRNA into our PGP-Cas9 cell line in an arrayed format and selected with puromycin after 48 hrs and allowed editing to occur for an additional 5 days (7

days total). From there we retrieved the cell pellets from each individual sgRNA and extracted gDNA. We then designed primers (**Table S7**) upstream and downstream of the expected cut site for each individual sgRNA and amplified that region utilizing standard PCR on the gDNA extracted from each cell pellet transduced with each individual sgRNA. Each amplicon for each sgRNA was then sent out for deep sequencing. We used CRISPResso with default parameters to compute the fraction of reads containing mutations, which we split out into an indel rate and an overall mutation rate.

3.3.2 Quantification and Statistical Analysis

3.3.2.1 Overview

For all figures, we used the CellRanger pipeline as described in the *Single Cell RNA-Seq Processing* section to generate counts matrices [63]. We also used the Seurat R package for clustering, data integration, and classification for all figures as described in the *Seurat Data Integration* and *H1 Teratoma Clustering and Validation* methods sections [68]. For assigning lentiviral CRISPR guide RNAs to cells (relevant to Figure 3.1/3.2), we used the genotyping-matrices method as described in the *Lentiviral Barcode and CRISPR Guide Assignment* section [17]. For Figure 3.1, we quantified guide RNA editing using CRISPResso [259]. And for Figure 3.2, we used DESeq2 as described in the *PGP1 Neural Disorder Screen Analysis* section [260]. The remaining analysis was done using custom R scripts.

For Figure 3.2, we collapsed the expression all cells with the same cluster and guide RNA identity into a single replicate in order to run pseudobulk differential expression analysis.

A brief summary of the analysis details for each figure can be found in the results and figure legends. Below we also provide a mapping between each figure and the relevant methods sections:

- Figure 3.1/3.2: *PGP1 Embryonic Lethal Screen Analysis*, *PGP1 Neural Disorder Screen Analysis*

All analysis code as well as instructions on how to reproduce our analyses can be found at the Github repository: yanwu2014/teratoma-analysis-code.

3.3.2.2 Seurat Data Integration

Data integration was performed on the aggregated counts matrices for the 6 PGP1 CRISPR-KO screen teratomas. We used the Seurat v3 data integration pipeline [68,71]. Briefly, we first filtered the counts matrix for genes that are expressed in at least 0.1% of cells, and cells that express at least 200 genes. We then normalized the counts matrix using total-counts normalization, and log-transformed the result. Log-transforming RNA-seq counts results in the data following an approximately normal distribution, which is the assumption that Seurat makes for the remainder of the analysis [72]. For each teratoma, we identified highly variable genes, and selected the top 4000 genes that appeared as overdispersed across the most teratomas. We then identified anchor cells, and integrated the teratomas to create a batch-corrected gene expression matrix. After batch correction, we used a linear model to regress away library depth, and mitochondrial gene fraction, and ran Principal Components Analysis (PCA) [73], keeping the first 30 principal components. We then used the PCs to generate a k Nearest Neighbors (kNN) graph, setting k = 10, and then used the kNN graph to calculate a shared nearest neighbors (SNN) graph [74]. We ran modularity optimization algorithm with a resolution of 0.4 on the SNN graph to find clusters [71].

3.3.2.3 CRISPR Guide Assignment

To assign one or more gRNA barcode to each cell, we extracted each barcode by identifying its flanking sequences, resulting in reads that contain cell, UMI, and barcode tags. To remove potential chimeric reads, we used a two-step filtering process. First, we only kept barcodes

that made up at least 0.5% of the total amount of reads for each cell. We then counted the number of UMIs and reads for each plasmid barcode within each cell, and only assigned that cell any barcode that contained at least 10% of the cell's read and UMI counts. The code for assigning barcodes to each cell can be found on GitHub at: https://github.com/yanwu2014/genotyping-matrices [17].

3.3.2.4 PGP1 Embryonic Lethal Screen Analysis

For each of the six teratomas across the original and replicate screens, we used two technical replicate 10X runs. In order to ensure consistent cell types across teratomas, we merged the 10X runs corresponding to the same teratoma, and then integrated all six teratomas across both the original and replicate screen using Seurat v3 data integration. We used 3000 anchor features and 20 CCA dimensions for the integration. Using the annotated H1 teratoma dataset as the reference, we used Seurat label transfer to identify the cell type for all cells in the screen datasets. Due to the relatively low number of cells per guide RNA in the original screen, we collapsed closely related cell types into broader cell groupings in order to boost the power of our analysis. Specifically, Airway Epithelium was merged into Foregut (Airway epithelium is derived from the foregut epithelium during development), Schwann Cells and Melanoblasts were grouped as Schwann Cell Progenitors (SCP), Immune Cells, Erythrocytes, and Hematopoietic Stem Cells (HSCs) were grouped as Hematopoietic cells, Muscle Progenitors and Cardiac/Skeletal Muscle were grouped as Muscle, all MSC/Fibroblast populations were merged, Intermediate Neuronal Progenitors (INP) and Radial Glia were grouped as Neuronal Progenitors, and Retinal Neurons and Early Neurons were simply grouped as Neurons. In order to visualize the PGP1 data, we projected the integrated screen dataset onto the first 20 PCs from the H1 dataset and ran UMAP on the projected PCs.

We validated the editing efficiencies of all our guide RNAs using PCR amplification of the expected cut site and looking for mutations and indels with CRISPResso. We then selected the top guide targeting each gene with at least a 60% overall editing efficiency and a 40% indel efficiency which resulted in a total of 16 out of 48 guides selected. We then only used these 16 validated guides for further computational analysis. Unfortunately, the TULP3-2 guide was not detected in the replicate screen so we ended up using 15 guides (plus 5 NTC guides) for analysis.

We assigned CRISPR-KO gene perturbations using the barcode assignment strategy described in the Lentiviral Barcode and CRISPR Guide Assignment section. To determine the total effect of each knockout, we computed a normalized Earth Mover's Distance (EMD) between all cells in each gene knockout with all cells belonging to the NTC separately for each screen [261]. EMD computes the difference in cell type composition between two groups of cells, weighted by how transcriptionally distinct the cell types are [261]. Thus, differences in cell type composition between cells belonging to the gene knockouts and NTC that arise from the fact that the label transfer has a hard time distinguishing similar cell types will not be as highly weighted as differences between distinct cell types. We ran the EMD analysis separately for the original and replicate screens, and normalized the EMD metric so that the average EMD for all NTC guides would equal 1.

To assess the effect of gene knockouts on individual cell types, we used a ridge regression model with the R glmnet package as initially described in the PerturbSeq method [190,262]. Briefly, for each CRISPR gRNA, this resulted in regression coefficients for each cell type describing the enrichment or depletion of that gRNA in that cell type. This method assumes that the data is normally distributed, which is approximately true for RNA-seq and scRNA-seq data when log-transformed (insert ref). We permuted the gRNA assignments to assign p-values to each coefficient

representing the probability that coefficient is non-zero by chance. Because we used a non-parametric permutation test, we did not make any assumptions about the distribution of regression coefficients. We then used the Benjamini-Hochberg multiple testing correction [263] to generate False Discovery Rates and visualized coefficients with an FDR < 0.05. For each gRNA, we computed the cell type shift effect size as the average EMD effect across the screens. The reproducibility of the gRNA knockout was assessed by correlating the gRNA knockout effects (regression coefficients) across the original and replicate screen.

3.3.2.5 PGP1 Neural Disorder Screen Analysis

For each of the 2 teratomas across the original and replicate screens, we used two technical replicate 10X runs. In order to ensure consistent cell types across teratomas, we merged the 10X runs corresponding to the same teratoma, and then integrated the teratomas using Seurat v3 data integration. We used the same data integration and label transfer parameters as the embryonic lethal screen. We again collapsed closely related cell types into the broader cell groupings described in the **PGP1 Embryonic Lethal Screen** section, and additionally filtered out any remaining cell types with fewer than 200 cells.

We assigned CRISPR-KO gene perturbations using the barcode assignment strategy described in the Lentiviral Barcode and CRISPR Guide Assignment section. To determine the total effect of each knockout, we again computed the normalized Earth Mover's Distance (EMD) between all cells in each gene knockout with all cells belonging to the NTC separately for each screen [261].

We analyzed differential expression for each broad cell type separately so that cell type specific effects would be captured. For each cell type, we summed the counts for all cells assigned to a specific guide RNA and a specific teratoma to create a pseudobulk expression matrix. This

essentially treats each guide in each teratoma as a biological replicate for a given gene knockout, and enables us to use DESeq2, a well-validated differential expression method [260]. For each gene knockout, we ended up with 6 pseudobulk replicates (3 guides x 2 teratomas). We ran DESeq2 with default parameters, comparing the pseudobulk replicates for each gene with the NTC replicates, and used apeglm to shrink effect sizes. We set a False Discovery Rate cutoff of 0.1 to call a gene differentially expressed. We also ran DESeq2 on each teratoma separately to compute log fold-changes and assess reproducibility.

3.3.2.6 Figure Generation

All figures were generated using original artwork or open source with InkScape, Adobe Illustrator®, and ImageJ.

3.4 Results

3.4.1 Engineering Teratomas via Genetic Perturbations

To establish the utility of the teratoma system as a model for human development, we performed a single-cell genetic knockout screen using CRISPR-Cas9. To identify key developmental genes to include in our screen, we compiled a list of 24 major organ/lineage specification genes that are embryonic lethal upon knockout in mice (**Table S6**). Studying the effects of these genes using cell lines or organoid models would typically require different experiments and different models for each cell lineage, as even a single gene can have functions across cell types, and even different germ layers. With the teratoma model, we can screen the effects of these genetic perturbations in all major cell lineages and germ layers in the same experiment. Using the CROPseq-Guide-Puro vector backbone, we cloned in 48 individual single guide RNAs (sgRNAs) directed at each developmental gene (2 sgRNAs per gene) [192] (**Figure 3.1A, Table S6 and S7**). We also designed a stable Cas9-expressing iPSC line (PGP1) in order to prevent

Cas9 silencing (**Figure 3.2A**, **3.2B**, **3.3 Materials and Methods**). After creating a pooled lentiviral library with our sgRNAs, we transduced our engineered PGP1-Cas9 line at a MOI of 0.1 so that each cell received approximately one perturbation (**Figure 3.1A**). After selection, these cells were injected subcutaneously into 3 $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice for teratoma formation, extraction, and downstream scRNA-seq processing with 10X Genomics (**Figure 3.1A**).

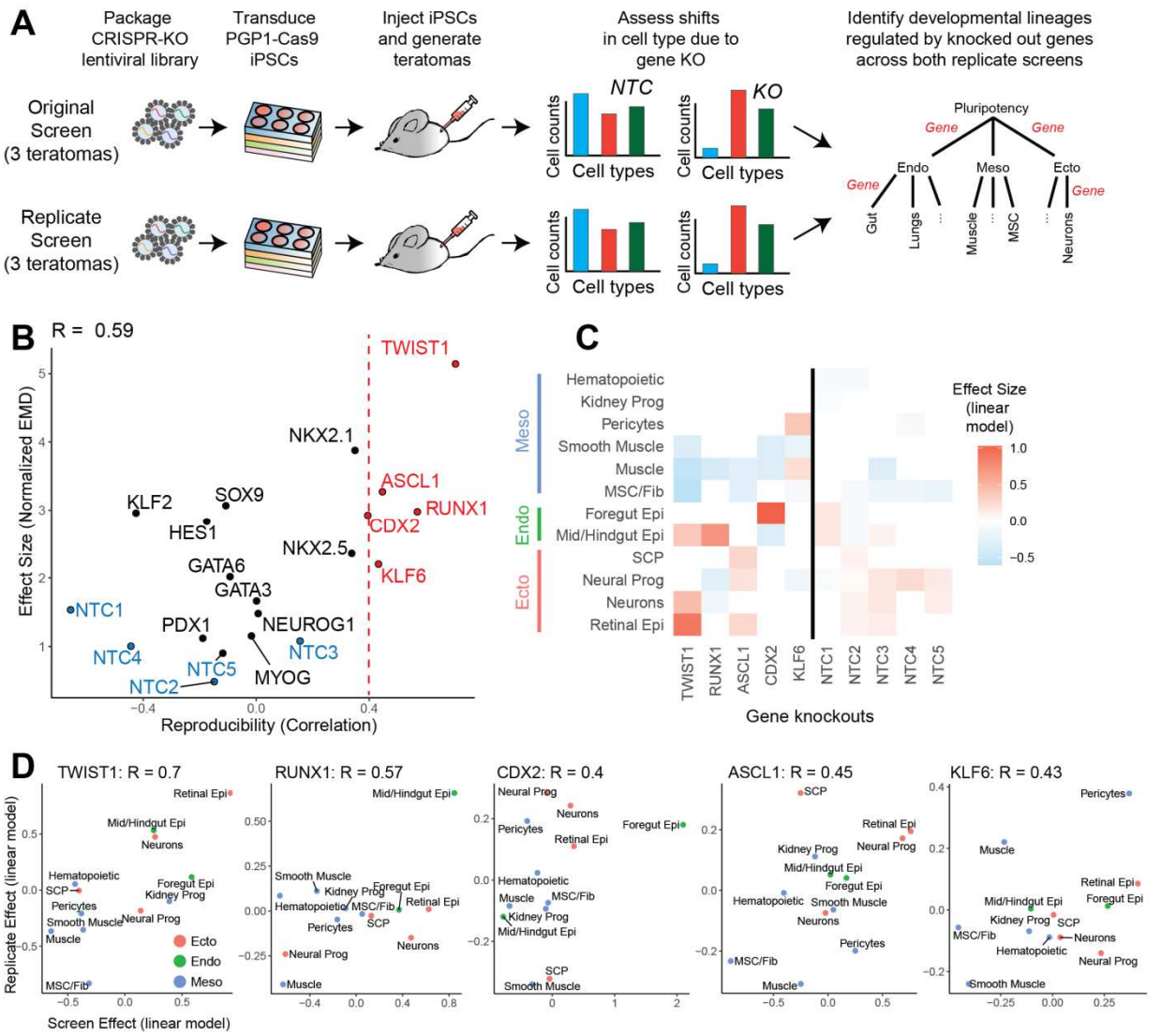
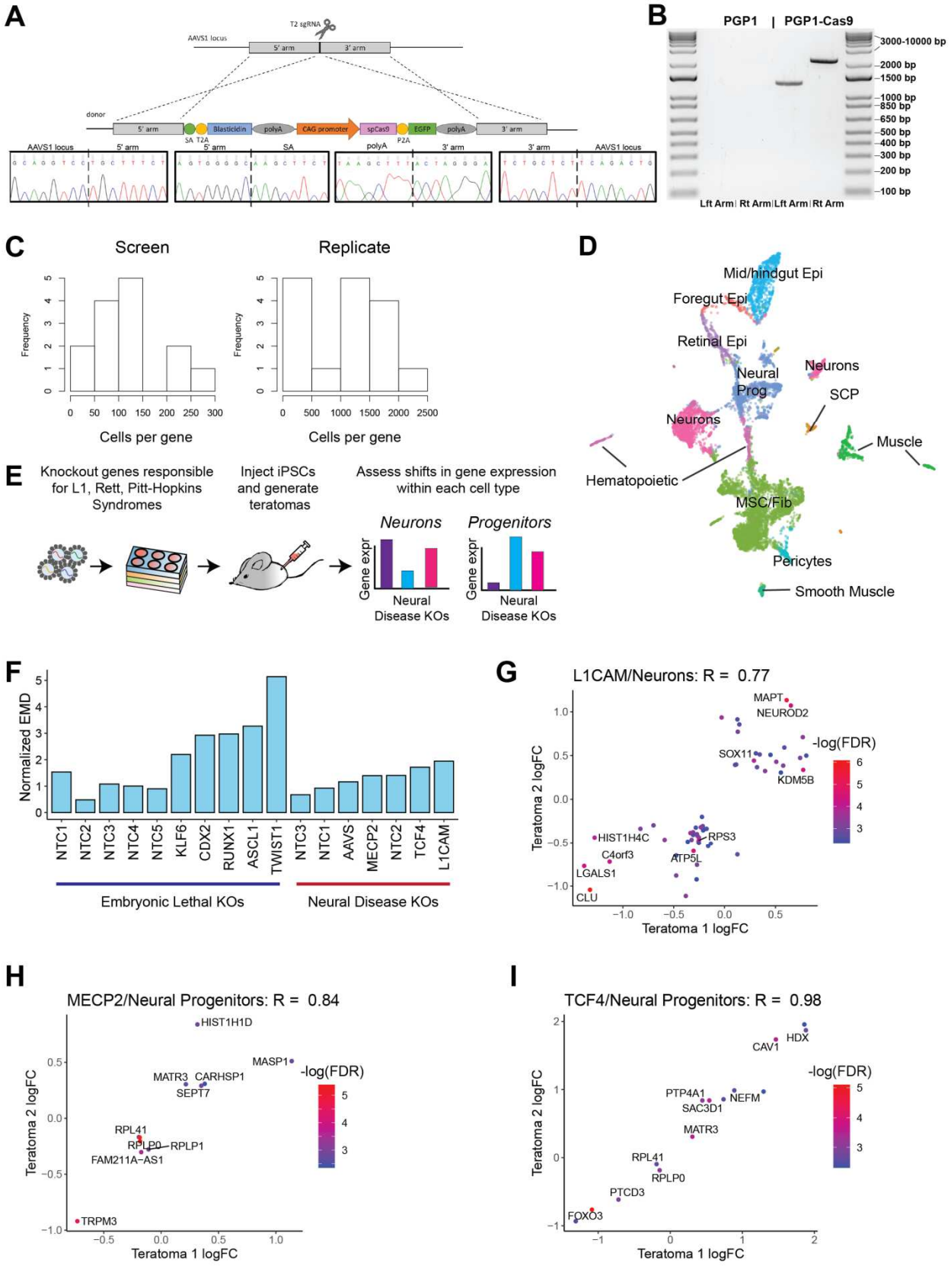


Figure 3.1. Engineering teratomas via genetic perturbations. **(A)** PGP1-Cas9 iPSCs were transduced with a CRISPR library targeting a panel of 16 key developmental genes with 1 gRNA per gene. After generating 3 teratomas with the PGP1-iPSCs, scRNA-seq was used to identify shifts in cell type formation as a result of gene knockouts. We repeated this process with 3 additional teratomas to serve as a replicate screen. (**B)** Average effect of gene knockout on cell type enrichment/depletion versus the correlation of cell type enrichment between the original screen and replicate screen. Genes with a reproducibility greater than 0.4 (**3.3 Materials and Methods**) were selected for further analysis. (**C)** A heatmap of the effect size (regression coefficient) of gene knockout enrichment for cell types and germ layers. **(D)** Scatterplot of individual guide RNA effects on cell type abundance for selected genes TWIST1, RUNX1, CDX2, KLF6, ASCL1.

Figure 3.2. Engineering teratomas via genetic perturbations. Related to Figure 3.1. (A) Schematic showing knock-in of the CAG-spCas9-P2A-EGFP cassette with an upstream T2A linked blasticidin resistance gene into the AAVS1 locus thus, creating the Cas9-expressing PGP1 line (above). Accompanying validated trace sequences of the left and right arms (below). **(B)** 2% agarose gel confirming integration of the CAG-spCas9-P2A-EGFP cassette into the AAVS1 locus of the PGP1 line via PCR amplification of the left and right arm spanning the endogenous locus and the engineered cassette compared to a PGP1 negative control. **(C)** Observed cells per gRNA and cells per gene for the screen. **(D)** UMAP projection of PGP1 cell types classified using the H1 cell types as a reference.


A
T2 sgRNA
AAVS1 locus
5' arm
3' arm
donor
Blasticidin
polyA
CAG promoter
spCas9
EGFP
SA
T2A
P2A
B
PGP1 | PGP1-Cas9
3000-10000 bp
2000 bp
1500 bp
1000 bp
850 bp
650 bp
500 bp
400 bp
300 bp
200 bp
100 bp
Lft Arm | Rt Arm | Lft Arm | Rt Arm
C
Screen
Replicate
Frequency
Cells per gene
D
Mid/hindgut Epi
Foregut Epi
Retinal Epi
Neural Prog
Neurons
SCP
Muscle
Hematopoietic
MSC/Fib
Pericytes
Smooth Muscle
E
Knockout genes responsible for L1, Rett, Pitt-Hopkins Syndromes
Inject iPSCs and generate teratomas
Assess shifts in gene expression within each cell type
Neurons
Progenitors
Gene expr
Neural Disease KOs
F
Normalized EMD
NTC1
NTC2
NTC3
NTC4
NTC5
KLF6
CDX2
RUNX1
ASCL1
TWIST1
NTC3
NTC1
AAVS
MECP2
NTC2
TCF4
L1CAM
Embryonic Lethal KOs
Neural Disease KOs
G
L1CAM/Neurons: R = 0.77
MAPT
NEUROD2
SOX11
KDM5B
HIST1H4C
RPS3
C4orf3
ATP5L
LGALS1
CLU
-log(FDR)
Teratoma 2 logFC
Teratoma 1 logFC
H
MECP2/Neural Progenitors: R = 0.84
HIST1H1D
MASP1
MATR3
CARHSP1
SEPT7
RPL41
RPLP0
RPLP1
FAM211A-AS1
TRPM3
I
TCF4/Neural Progenitors: R = 0.98
CAV1
HDX
PTP4A1
NEFM
SAC3D1
MATR3
RPL41
RPLP0
PTCD3
FOXO3

We validated the editing efficiencies of all our guide RNAs using PCR amplification of the expected cut site and looking for mutations and indels with CRISPResso (**Table S7, S8, 3.3 Materials and Methods).** We then selected the top guide targeting each gene with at least a 60% overall editing efficiency and a 40% indel efficiency which resulted in a total of 16 guides (**Table S7, S8, 3.3 Materials and Methods**). We then only used these validated guides for further computational analysis. To assess the reproducibility of our results, we also reran the CRISPR-KO screen by repooling these validated guides and generated 3 additional teratomas (**Figure 3.1A**, **Table S1, 3.3 Materials and Methods**). We successfully captured a median of 118 cells per gene/guide in the original screen and 1,280 cells per gene/guide in the replicate screen (**Figure 3.2C**). We were able to capture more cells per guide in the replicate screen since we only pooled the top 16 guides, while the original screen had a total of 48 guides (**3.3 Materials and Methods**).

In order to ensure consistent cell types across teratomas, we integrated all six teratomas across both the original and replicate screen using Seruat v3 [68]. We then called cell types in the PGP1 teratoma cells using Seurat label transfer with the 7 H1 teratomas as reference and collapsed developmentally similar cell types (**Figure 3.2D, 3.3 Materials and Methods**). To determine the total effect of each knockout, we measured the difference in cell type composition between cells in each gene knockout with all cells belonging to the non-targeting control (NTC) separately for each screen using Earth Mover's Distance (EMD) [261] (**Figure 3.1A, 3.3 Materials and Methods**). For both the original and replicate screen, we ran a ridge regression model to assess effects of each gene knockout on cell type enrichment/depletion [253] (**Figure 3.1A**, **3.3 Materials and Methods**). For each gene, we plotted its EMD alongside the Pearson correlation of the regression coefficients for the both the original screen and the replicate screen, giving us a sense of both the effect size and reproducibility of each gene knockout (**Figure 3.1B**, **3.3 Materials and Methods**). We also

see that gene knockouts with strong effect sizes tend to be more reproducible (R = 0.59) (**Figure 3.1B**, **3.3 Materials and Methods**). We highlighted genes with a Pearson correlation of greater than 0.4 between the original and replicate screen for further analysis (**Figure 3.1B**).

For the highlighted genes, *TWIST1, RUNX1, CDX2, KLF6*, and *ASCL1,* we wanted to identify the gene knockout effects on cell types that were statistically signficant. Towards this we merged the cells from both screens and ran a combined ridge regression analysis, computing P-values using a permutation test and False Discovery Rates using the Benjamani-Hochberg correction (**3.3 Materials and Methods**). We then visualized all gene knockout effects with an FDR < 0.1 (**Figure 3.1C**, **3.3 Materials and Methods**).

CDX2 is known to be important for the development of the midgut and hindgut [264,265]. Our data shows that cells with a *CDX2* are enriched in the Foregut and depleted in the Mid/Hindgut, which lines up with past literature reports that *CDX2* knockout shifts the gut differentiation pathway away from intestine and towards gastric activation [266,267] (**Figure 3.1C, 3.1D**). *TWIST1* showed the largest effect size and is a known transcription factor for the epithelial-to-mesenchymal transition (EMT), which is important in development as well as metastatic cancers (**Figure 3.1B**) [268,269]. Our screen found that cells with a *TWIST1* knockout are depleted in mesodermal cell types (muscle, smooth muscle, pericytes, and mesenchymal stem cell/fibroblasts), and enriched in neuro-epithelium (retinal epithelium, neurons), confirming prior studies that have identified *TWIST1* as key to mesodermal specification [270] (**Figure 3.1C, 3.1D**). We see that *RUNX1* knockout results in a depletion of neurons and muscle cell types and an enrichment in mid/hindgut, which is consistent with previous mouse and stem cell studies that show *RUNX1* to be critical for neural crest formation, signaling in gut epithelium stem cells, and myoblast proliferation [271–275] (**Figure 3.1C, 3.1D**). *KLF6* knockout resulted in a depletion of pericytes, consistent with its role in promoting

endothelial activation during vascular repair [276] (**Figure 3.1C, 3.1D**). *ASCL1* interestingly resulted in an increase in the proportion of retinal epithelium and neural progenitors (**Figure 3.1C, 3.1D**). Since *ASCL1* is key to cell cycle exit and neuronal differentiation, knocking out *ASCL1* may slow down neurogenesis and result in a buildup of neural progenitors [277]. With this CRISPR knockout screen of key developmental regulators, we were able to assay the multi-lineage functions of these genes in a human-specific model, something that to our knowledge, no other human developmental model can currently accomplish.

3.4.2 Modeling Neural Disorders using Teratomas

While we were able to demonstrate the teratoma's unique ability to assess the multi-lineage function of embryonic lethal genes, we also wanted to see if the teratoma could model human neural disorders. Specifically, we looked into Pitt-Hopkins [278], Rett [279], and L1 [280] Syndromes. Pitt-Hopkins syndrome is a rare neurodevelopmental disorder most often caused by a *de novo* loss of function of one allele of the transcription factor 4 (*TCF4*) gene [281]. Rett Syndrome is a severe X-linked neurological disorder caused by a *de novo* mutation in the methyl-CpG-binding protein 2 (*MECP2*) gene. Finally, L1 syndrome is another X-linked syndrome with a mutation in the L1 cell adhesion molecule (*L1CAM*) gene important for neuron migration, adhesion, and neuronal differentiation [282]. To assess the downstream effects of perturbing these genes, we generated a CRISPR-KO library targeting *TCF4*, *MECP2,* and *L1CAM*, with 3 guides for each gene **(Table S7)**. We transduced PGP1-Cas9 cells with the neural disorder library, generated 2 teratomas, and then sequenced 2 scRNA-seq libraries for each teratoma using the 10X Genomics platform (**Table S1**) [63].

We integrated and clustered the teratomas using Seurat data integration and used Seurat's label transfer method to call cell types using the H1 teratomas as the reference. We then looked

for shifts in both cell type proportion and cell type specific gene expression as a result of the gene knockouts (**Figure 3.2E**). As one would expect, we found that the shift in cell type proportion (normalized EMD) was much smaller than for the embryonic lethal knockouts (**Figure 3.2F**). We thus looked at cell type specific shifts in gene expression from the neurological disorder knockouts instead. We merged our cell types into 7 broad cell types (Neurons, Neural Progenitors, Gut, Retinal Epithelium, Muscle, Immune, MSC/Fibroblast) and computed differential expression between each gene knockout and the NTCs (**3.3 Materials and Methods**). There was no significant gene expression shift due to the presence of a double stranded break (per AAVS control) (**Table S9**).

We then analyzed the effect of *L1CAM* in Neurons and the effect of *TCF4* and *MECP2* in Neural Progenitors and plotted the cell type specific log fold-changes for all DEGs with an FDR below 0.1 across both teratomas, showing that our hits are fairly reproducible (**Figures 3.2G – I**). Knocking out *L1CAM* in Neurons decreased the expression of clusterin (*CLU)*, an effect that has previously been shown in colorectal cancer cells [283], while also increasing the expression of *MAPT* (which produces the *tau* protein). *Tau* efflux via *L1CAM* exosomes is present in certain neurological diseases [284] (**Figure 3.2G, Table S9**). Knocking out *MECP2* in neural progenitors decreased the expression of transient receptor potential cation channel subfamily M member 3 (*TRPM3),* and previous literature has shown a similar decrease in expression and function of TRP channels in the hippocampus and several other brain regions of *MECP2* mutant mice contributing to Rett syndrome etiology [285–287] (**Figure 3.2H, Table S9**). Finally, knocking out *TCF4* in neural progenitors decreased the expression of *FOXO3* which is consistent with *TCF4* knockdown studies in the human neuroblastoma line SH-SY5Y showing a fold decrease in *FOXO3* which has been suggested to contribute to the molecular pathology of Pitt-Hopkins and other autism spectrum

disorders [288] (**Figure 3.2I, Table S9**). Overall, we were able to reproducibly discover cell type specific gene expression shifts that occurred when knocking out the genes underlying Rett, Pitt-Hopkins, and L1 syndromes, potentially building a resource for future in-depth study.

3.5 Discussion

The teratoma has the potential to be a fully vascularized, multi-lineage model for human development. Its major advantages are that it can grow to a large size due to its vascularization, and it can produce a wide array of relatively mature cell types from all major developmental lineages. We demonstrated with our CRISPR-Cas9 knockout screens, the teratoma's ability to generate cells from all lineages enables a comprehensive assessment of the effect of genetic perturbations on human development within a single integrated experiment. This experiment also shows the validity of the teratoma as a model for multi-lineage human development as the perturbation effects follow standard canonical pathways known to developmental biologists.

Any model system has its intrinsic strengths and weaknesses. One issue with the teratoma system (and organoids) is the intrinsic degree of heterogeneity [26,104,105,108]. In this regard, we found the use of internal controls when conducting perturbation experiments was important. For example, in our CRISPR-Cas9 screen, each teratoma contained both gene targeting guides and non-targeting controls, enabling us to compare cell type proportion shifts within each teratoma without having to worry about heterogeneity between teratomas.

Taken together, we believe the teratoma is a promising platform for modeling multi-lineage human development and pan-tissue functional genetic screening.

3.6 Acknowledgements

We thank members of the Mali lab for advice and help with experiments, Marianna Yusupova for help with initial studies, Alexander Militar for assistance in schematic generation,

in loving memory of Nakon Aroonsakool, and to the Moore's Cancer Center Histology Core, UC San Diego Microscopy Core, Sanford Consortium Flow Cytometry Core, and IGM Genomics Center for help with sample processing. This work was generously supported by UCSD Institutional Funds and NIH grants (R01HG009285, RO1CA222826, RO1GM123313).

Chapters 3 is in part reprints of the following materials of which the dissertation author was one of the primary investigators and authors of this paper:

Chapter 3, in part, is a reprint of the material as it appears in McDonald D*, Wu Y*, Dailamy A, Tat J, Parekh U, Zhao D, Hu M, Tipps A, Zhang K, Mali P. Defining the Teratoma as a Model for Multi-lineage Human Development. Cell. 2020 Nov 25;183(5):1402-1419.e18. doi: 10.1016/j.cell.2020.10.018.

*Both of these authors contributed equally

4 Engineering Teratomas via miRNA based Molecular Sculpting

4.1 Abstract

We propose that the teratoma, a recognized standard for validating pluripotency in stem cells, could be a promising platform for studying human developmental processes. We demonstrated teratomas can be molecularly sculpted via miRNA-regulated suicide gene expression to enrich for specific tissues. Specifically, we used miR-124 to enrich for neuro-ectoderm and validated through scRNA-seq analysis, histology, immunostaining, and RNA-FISH. The teratoma is a promising platform for modeling multi-lineage development and tissue engineering.

4.2 Introduction

We propose here the use of teratomas as a model for studying human development [42]. There has also been some progress in utilizing the inherent differentiation potential of teratomas to derive highly sought-after cell types. For instance, teratomas were recently utilized to derive skeletal myogenic progenitors by injecting PSCs into the *tibialis anterior* muscle of mice to enrich for muscle cell types in the teratomas that formed in those muscles [56]. Additionally, some groups have successfully enriched for hematopoietic stem cells (HSCs) from teratomas utilizing strategies such as human umbilical vein endothelial cell (HUVEC) pooling [57–60].

We have successfully characterized and validated 23 human cell types across all 3 germ layers and they are reproducible without significant cell type biasing, developmentally staged neuronal and gut tissue at gestational weeks 17 and 8 respectively, and utilized the teratoma as a tool to model human cell fate specification with reproducible results that are consistent with known biology in a single study. Since the teratoma is a multi-lineage system, we wanted to molecularly sculpt it to instead yield only specific lineages of interest. These lineages would be derived through

the normal processes of organogenesis and being vascularized the system would allow continuous maturation enable focused developmental biology and tissue engineering applications.

This next study performs molecular sculpting of teratomas with the assistance of endogenously expressed micro RNAs (miRNAs)[289–291]. MiRNAs are a class of regulatory non-coding RNA molecules that are approximately 21-24 nucleotides long. They form unique short hairpin structures and are present in plants, animals, and viruses. MiRNAs work through RNA-silencing and post-transcriptional regulation of gene expression. Their sequence is complementary to regions within specific messenger RNAs (mRNAs). When the miRNA (in association with argonaute proteins of the RNA-Induced Silencing Complex [RISC]) binds to its mRNA target, the mRNA is either cleaved, destabilized, or its translation efficiency is reduced[289–291]. This effect is critical in regulating gene expression in a cell-specific manner as many miRNAs are unique to explicit cell types, lineages, or disease states. The miRNA profile is often more precise and informative than the mRNA profile in characterizing developmental lineages[292,293]. To this end, we hijacked the miRNA capabilities to skew and molecularly sculpt teratomas down one lineage[294–296].

4.3 Materials and Methods

4.3.1 Experimental Model and Subject Details

4.3.1.1 Organoid Generation and Dissociation

Self-patterned whole brain organoids were generated following the Quadrato et al. 2017 protocol [104]. Briefly, H1 ESCs transduced with either miR-124-HSV-tk-GFP or HSV-tk-GFP were cultured as embryoid bodies for 5 days, transferred into Neural Induction (NI) media for 5 days, and finally embedded in Matrigel and cultured in Cortical Differentiation (CD) media for 25 days. Day 35 organoids were dissociated to single cell following a modified protocol using the

GentleMACS Human Tumor Dissociation Kit, but without use of the GentleMACS dissociator and instead cells were triturated post-37ºC 1-hr incubation with a 1000 µL pipetman prior to 70 µM filtration. Resulting single cell suspension was analyzed for GFP florescence via flow cytometry. Cells, embryoid bodies, and organoids were maintained under puromycin selection [0.75µg/µL] for the entirety of the experiment.

4.3.2 Method Details

4.3.2.1 Library Preparation

The lentiviral backbone plasmid for the miRNA-HSV-tk-GFP constructs was an EF1-alpha promoter, GFP, IRES domain, and puromycin-resistance gene (EGIP) backbone. The lentiviral backbone was digested with EcoRV-HF (New England Biolabs) at 37°C for 1 hour to excise out the GFP in a reaction consisting of: EGIP backbone, 5 µg, 1X Cutsmart Buffer (New England Biolabs) , 5 µl, EcoRV-HF, 5 µl, H20 up to 50 µl. After digestion, the vector was purified using a QIAquick PCR Purification Kit (Qiagen). We amplified a gBlock containing the Herpes Simplex Virus thymidine kinase (HSV-tk), 2A self-cleaving peptide, and GFP.

The primers used to amplify the gBlock contain unique miRNA binding sites (see below).

miR_Empty_F	TGGCTAGTTAAGCTTGATATCGAATTCCTGCAGCCCGGGGGATC CAGATCACACCGGTCGCCA
miR_Empty_R	GGGAGAGGGGGGGGGGGCGGAATTCCGCGGGCCCGTCGACGCG GTTAACGCCGCTTTACTTGTACAG
miR_21_F	TGGCTAGTTAAGCTTGATATCGAATTCCTGCAGCCCGGGGGATC CTCAACATCAGTCTGATAAGCTA AGATCACACCGGTCGCCA

miR_21_R	GGGAGAGGGGGGGGGGCGGAATTCCGCGGGCCCGTCGACGCG GTTTAGCTTATCAGACTGATGTTGAAACGCCGCTTTACTTGTAC AG
miR_122_F	TGGCTAGTTAAGCTTGATATCGAATTCCTGCAGCCCGGGGGATC CCAAACACCATTGTCACACTCCA AGATCACACCGGTCGCCA
miR_122_R	GGGAGAGGGGGGGGGGCGGAATTCCGCGGGCCCGTCGACGCG GTTTGGAGTGTGACAATGGTGTTTGAACGCCGCTTTACTTGTAC AG
miR_124_F	TGGCTAGTTAAGCTTGATATCGAATTCCTGCAGCCCGGGGGATC CGGCATTCACCGCGTGCCTTA AGATCACACCGGTCGCCA
miR_124_R	GGGAGAGGGGGGGGGGCGGAATTCCGCGGGCCCGTCGACGCG GTTTAAGGCACGCGGTGAATGCC AACGCCGCTTTACTTGTACAG
miR_126_F	TGGCTAGTTAAGCTTGATATCGAATTCCTGCAGCCCGGGGGATC CCGCATTATTACTCACGGTACGA AGATCACACCGGTCGCCA
miR_126_R	GGGAGAGGGGGGGGGGCGGAATTCCGCGGGCCCGTCGACGCG GTTTCGTACCGTGAGTAATAATGCGAACGCCGCTTTACTTGTAC AG
miR_302A_F	TGGCTAGTTAAGCTTGATATCGAATTCCTGCAGCCCGGGGGATC CAGCAAGTACATCCACGTTTAAGT AGATCACACCGGTCGCCA
miR_302A_R	GGGAGAGGGGGGGGGGCGGAATTCCGCGGGCCCGTCGACGCG GTTACTTAAACGTGGATGTACTTGCTAACGCCGCTTTACTTGTAC AG

We cloned this amplicon into our digested EGIP backbone using standard Gibson assembly.

The Gibson assembly reactions were set up as follows: 1:10 molar ratio of digested backbone to sgRNA insert, 2X Gibson assembly master mix (New England Biolabs), H20 up to 20 µl. After incubation at 50°C for 1 h, the product was transformed into One Shot Stbl3 chemically competent Escherichia coli (Invitrogen). A fraction (150 µL) of cultures was spread on carbenicillin (50 µg/ml) LB plates and incubated overnight at 37°C for 15-18hrs (miRNA constructs required longer incubation times). Individual colonies were picked, introduced into 5 ml of carbenicillin (50 µg/ml) LB medium and incubated overnight in a shaker at 37°C. The plasmid DNA was then extracted with a QIAprep Spin Miniprep Kit (Qiagen), and Sanger sequenced to verify correct assembly of the vector and to extract barcode sequences.

To assemble the library, individual sgRNA vectors were pooled together in an equal mass ratio along with 5 non-targeting control (NTC) sgRNAs which constituted 50% of the final pool.

4.3.2.2 Viral Transduction

For viral transduction, virus was added at a low MOI (ensuring a single barcode/cell or a single sgRNA/cell) to stem cells at 20% confluency alongside polybrene (5 µg/ml, Millipore) in fresh mTeSR medium. The following day, medium was replaced with fresh mTeSR. For miRNA-HSV-tk-GFP transduced cells puromycin selection did not begin until 5-7 days after transduction to allow for enough GFP positive cells.

4.3.2.3 GCV-HSV-tk Killing *in vitro*

Cells transduced with miRNA-HSV-tk-GFP construct and EGIP-transduced controls grew for a maximum of 5 days in standard medium conditions in the presence of Ganciclovir ([GCV, Sigma-Aldrich] 1µM, 10µM, or 100µM) with daily phase and fluorescent microscopy imaging.

GCV was resuspended and stored in 1 mL PBS (Gibco) aliquots at 3mg/mL in -20ºC. Cells were seeded at similar densities on Day 0 of experiment.

4.3.2.4 miRNA-HSV-tk-GFP Knockdown *in vitro*

Cells were transduced with miRNA-HSV-tk-GFP constructs and allowed to grow for a maximum of 5 days in standard medium conditions. After 5 days, cells were spun down and resuspended in PBS (Gibco) at $1x10^6$ cell / mL and ran on the Becton Dickinson FACScan flow cytometer gating for fluorescence (FL1-H [GFP positivity]) and forward scatter (FSC-H [shape and size]).

4.3.2.5 Molecular Sculpting of Teratomas

Standard teratoma formation protocol was followed using miRNA-HSV-tk-GFP transduced H1s. Once teratomas reach a size of at least 10mm in one axis, intratumoral (IT) or combined intraperitoneal intrautumoral (IPIT) administration of GCV begins at 80mg/kg/d or 100mg/kg/d (50mg/kg/d at each site) respectively, using standard needle and syringe injection. Teratoma was allowed to grow for a total of 10 weeks before extraction.

4.3.2.6 Immunostaining

For neuro-ectoderm staining, fresh frozen sections were rinsed onced with PBS before fixation at room temperature for 15 min with 4% paraformaldehyde. Three consecutive washes were then performed with PBS 5 min each before addition of blocking buffer (5% normal donkey serum, 0.2% triton x-100 in PBS) for 1 hr. Primary antibody (anti-PAX6 rabbit [Millipore Sigma] diluted 1:50 in blocking buffer) was added overnight (12 hrs) at 4C. Three consecutive washes were then performed with PBS 10 min each with gentle agitation before addition of secondary antibody (Anti-Rabbit Dylight 550 (Abcam) diluted 1:200 in blocking buffer) for 1 hr at 37C shielded from light. Three consecutive washes were then performed with PBS 5 min each with

gentle agitation before addition of DAPI (1:10,000 dilution in PBS) for 10 min. This was finally followed by three consecutive washes with PBS 10 min each with gentle agitation before imaging.

4.3.2.7 Microscopy

Following 24 hrs of incubation with RNAScope® probes in 4ºC, slides were imaged using Zeiss 880 Airyscan Confocal microscope with special thanks to Michael Hu for image processing utilizing the UC San Diego Microscopy Core. Raw images on the Leica DMi8 were obtained with 16bit bit-depth per color, and highlights and shadows were adjusted in the LASX software. Raw images on the Zeiss 880 were obtained with 16bit bit-depth per color, and highlights and shadows were adjusted in the ZEN software. RNAScope images were dilated using ImageJ's MorphoLib by splitting the image into the composite channels and dilating the dots in the appropriate channel. Dots were dilated to 3 pixels as disks.

4.3.3 Quantification and Statistical Analysis

4.3.3.1 Molecular Sculpting Analysis

To assess the enrichment or depletion of cell types in the miRNA-HSV-tk transduced H1 teratomas, we compared teratomas that had ganciclovir (GCV) added using intratumoral (IT) and both intratumoral and intraperitoneal (IPIT) injection methods, versus a control teratoma that had the construct miRNA-HSV-tk but no GCV. All teratomas were injected on the same date and extracted after 10 weeks of growth. To assign cell types, we again used Seurat's label transfer. We then collapsed cell types using the same merging strategy described in the **PGP1 Teratoma Screen Analysis (3.3 Materials and Methods)** section, and then computed the fraction of cell types present in each teratoma. Finally, we computed log2 fold-changes of cell type fractions by dividing the cell type fractions in the GCV+ IT/IPIT teratomas with the cell type fractions in the GCV- teratoma. To compute an estimated z-score, we subtracted the GCV- teratoma fractions

from the GCV+ IPIT/IT teratoma fractions and divided by the cell type fraction variance. The z-scores for IPIT and IT teratomas were computed separately, and the cell type fraction variance was computed by pooling the variance of the miRNA-HSV-tk teratomas and the variance of the plain H1 teratomas with Cohen's pooled standard deviation [297].

4.3.3.2 Figure Generation

All figures were generated using original artwork or open source with InkScape, Adobe Illustrator®, and ImageJ.

4.4 Results

Since the teratoma is vascularized and has the potential to yield mature tissue, we sought to sculpt the teratoma towards specific lineages, which could allow for focused developmental modeling and tissue engineering. We used endogenously expressed micro RNAs (miRNAs) [289–291], which are often unique to specific cell types, lineages, or disease states [292,293]. Specifically, we appended tissue specific miRNA target sequences to the 5' and 3' UTR of a GFP fluorescent suicide gene (HSV-tk-GFP), thereby suppressing its expression in a miRNA specific lineage of interest (**Figure 4.1A**, **Table S10**)[294–296]. This design ensures that cell types that do not express the miRNA are killed by the suicide gene in the presence of gancliclovir (GCV), thus selecting for our desired lineage (**Figure 4.1B**).

We first tested the functionality our miRNA-HSV-tk-GFP constructs in H1 ESCs by showing that cells transduced with our miRNA-HSV-tk-GFP construct die in the presence of 10μM GCV after 5 days of culture, while cells transduced with a GFP control continue growing (**Figure 4.2A**). We then assessed the cell type specificity of the miRNA construct using miR-21 expressing HeLa cells[291,298–300]. HEK293T cells are show little to no expression of miR-21 and can serve as a control [301–303]. After transduction of both cell lines with our miR-21-HSV-tk-GFP

construct, we cultured the cells for 5 days and then performed flow cytometry analysis where we saw a decrease in GFP expression in the HeLA cells, but not in the HEK293T cells (**Figure 4.2B**). This would indicate that the GFP expression was silenced by the miR-21 expressed by HeLa cells. We used an HSV-tk-GFP construct without any miRNA binding sites as a control (**Figure 4.2B**). We repeated this experiment with a miR-126-HSV-tk-GFP construct (endothelial cell-specific) [304] and observed GFP repression ina decrease in GFP signal in HUVEC cells as compared to the HEK293T control (**Figure 4.2C**). With this we were able to validate both the HSV-tk killing with GCV, and the ability of our miRNA constructs to specifically repress GFP in target cell lines.

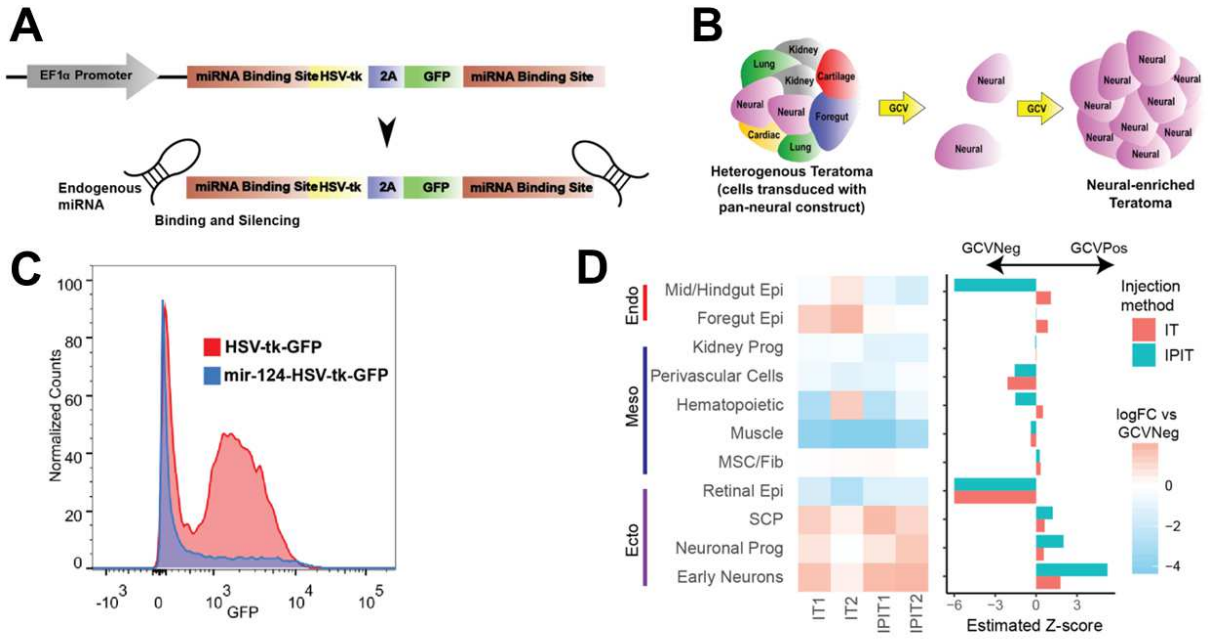
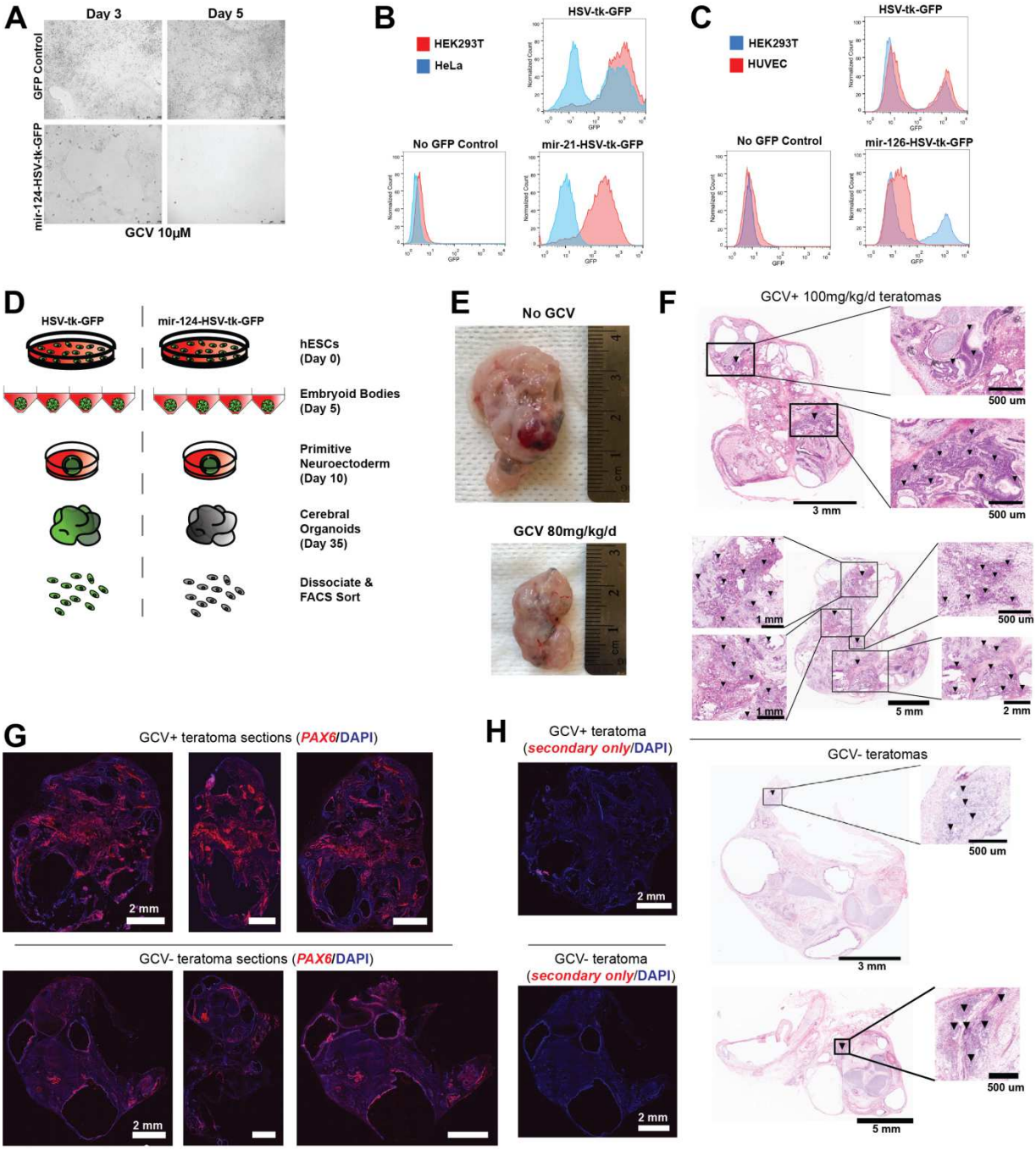


Figure 4.1. Engineering teratomas via miRNA based molecular sculpting. **(A)** Schematic of miRNA-HSV-tk-GFP construct. 2A encodes for a self-cleaving peptide. Upon transcription, the expression will be diminished if corresponding endogenously expressed miRNA is present in the cell. **(B)** Schematic of how a developing teratoma would form in the presence of Ganciclovir (GCV, 80mg/kg/d, **4.3 Materials and Methods)** if cells were transduced with a neural-specific miRNA-HSV-tk construct. **(C)** Quantification using flow cytometry and gating based on the presence or absence of GFP in 35-day self-patterned whole brain organoid single cells transduced with either HSV-tk-GFP control or miR-124-HSV-tk-GFP. **(D)** *In vivo* studies of miR-124-HSV-tk-GFP teratomas in the presence of GCV administration (80mg/kg/d, **4.3 Materials and Methods**) using both intratumoral (IT) and intratumoral and intraperitoneal injection methods. A heatmap showing cell type fraction log fold-change for each teratoma replicate compared to a control miR-124-HSV-tk-GFP teratoma in the absence of GCV. Z-scores for each cell type fraction change are plotted as well, with standard deviations calculated using a pooled variance (**4.3 Materials and Methods**).

Figure 4.2. Engineering teratomas via molecular sculpting. Related to Figure 4.1. (A) Phase images from light microscopy showing H1 cell survival after 3 and 5 days in the presence of GCV (10μM). H1 ESC line was either transduced with GFP control (EGIP backbone) or miR-124-HSV-tk-GFP. **(B)-(C)** Quantification using flow cytometry and gating based on the presence or absence of GFP in HEK293T and HeLa/HUVEC cells **(B)/(C)** transduced with either No GFP control, HSV-tk-GFP, or miR-21-HSV-tk-GFP/miR-126-HSV-tk-GFP for 5 days (**4.3 Materials and Methods**). **(D)** Schematic of generating self-patterned whole brain organoids (**4.3 Materials and Methods**). **(E)** Images of teratomas grown in the absence and presence of GCV administration (80mg/kg/d, **4.3 Materials and Methods**) for 10 weeks. **(F)** H&E stains of teratomas grown in the absence (left) and presence (right) of GCV administration. Arrowheads highlight regions of neuro-ecotoderm. Scalebars are directly labeled. **(G)** anti-*PAX6* (red) and DAPI (blue) immunostaining in GCV+ and GCV- control sections across 3 different regions of the corresponding teratoma. Scalebar = 2 mm. **(H)** Secondary antibody staining only (Dylight 550, red) and DAPI (blue) for a GCV+ and GCV- negative teratoma. Scalebar = 2 mm. **(I)** RNA FISH analysis of *HES5* (red) and DAPI (blue) in a GCV+ and GCV- teratoma. Scalebar = 2 mm, 200 μm (magnified insert).


A
Day 3
Day 5
GFP Control
mir-124-HSV-tk-GFP
GCV 10μM
B
HEK293T
HeLa
HSV-tk-GFP
No GFP Control
mir-21-HSV-tk-GFP
C
HEK293T
HUVEC
HSV-tk-GFP
No GFP Control
mir-126-HSV-tk-GFP
D
HSV-tk-GFP
mir-124-HSV-tk-GFP
hESCs (Day 0)
Embryoid Bodies (Day 5)
Primitive Neuroectoderm (Day 10)
Cerebral Organoids (Day 35)
Dissociate & FACS Sort
E
No GCV
GCV 80mg/kg/d
F
GCV+ 100mg/kg/d teratomas
500 um
3 mm
1 mm
5 mm
2 mm
GCV- teratomas
G
GCV+ teratoma sections (PAX6/DAPI)
2 mm
GCV- teratoma sections (PAX6/DAPI)
H
GCV+ teratoma (secondary only/DAPI)
GCV- teratoma (secondary only/DAPI)
I
GCV+ teratoma (HES5/DAPI)
200 um
GCV- teratoma (HES5/DAPI)

We further validated our construct in whole brain organoids. Following a standard self-patterned whole brain organoid protocol (**Figure 4.2D, 4.3 Materials and Methods**) [104], we created organoids using H1 ESCs transduced with either the miR-124-HSV-tk-GFP construct or the HSV-tk-GFP construct (lacking any miRNA binding sites). We used miR-124 since it is a pan-neural miRNA [305–307]. Day 35 organoids from both groups (miR-124-HSV-tk-GFP and HSV-tk-GFP) were dissociated down to single cell level and analyzed via flow cytometry for GFP fluorescence (Methods). As expected, HSV-tk-GFP organoid single cells maintained their GFP fluorescence while miR-124-HSV-tk-GFP organoids showed GFP repression (**Figure 4.1C**).

We then tested the miRNA-HSV-tk-GFP constructs *in vivo* using the miR-124-HSV-tk construct to generate teratomas enriched for the neural lineage. After the H1 ESC line was successfully transduced with the miR-124-HSV-tk-GFP construct, we formed teratomas as described in our previous studies (Methods). Once teratomas reached a minimum of 1cm in diameter, we began either intratumoral (IT) injections with GCV (80mg/kg/d, **4.3 Materials and Methods**) or two-site intraperitoneal and intratumoral (IPIT) injections (50/mg/kg/d for each site, **4.3 Materials and Methods**) all compared to a control miR-124-HSV-tk-GFP teratoma with no GCV (**4.3 Materials and Methods**). There were 2 teratomas for each injection condition for a total of 4 teratomas + 1 control teratoma and all teratomas were grown for up to 70 days. Post-extraction, teratomas were observed for external heterogeneity. The teratomas that received GCV injections were of smaller size (approx. 2cm compared to 4cm) and weight (approx. 1-2 gm compared to 5+ gm) than the control teratoma without GCV injections (**Figure 4.2E**).

We ran the 10X scRNA-seq protocol on each teratoma and classified cells using Seurat label transfer **(Table S1)** [68]. A comparison of the GCV+ teratomas cell type composition with the GCV- teratoma revealed enrichment in Early Neurons, Neuronal Progenitors, and Schwann cells

(Figure 4.1D). In addition, we saw depletion in muscle, retinal pigmented epithelium (lacks miR-124 expression), and other cell types (**Figure 4.1D**). The teratomas with the IPIT injection strategy showed a stronger enrichment for the neuro-ectoderm cell types, suggesting that the addition of an intraperitoneal injection site helps with GCV selection (**Figure 4.1D**). We also visualized the neuro-ectoderm enrichment in GCV+ teratomas with H&E staining of a GCV+ and GCV- teratoma (**Figure 4.2F**). The IPIT teratomas had a stronger enrichment for Early Neurons (Z-score > 3) than for Neuronal Progenitors or Schwann cells, possibly since the expression of miR-124 increases as the neuro-ectoderm cell types mature (**Figure 4.1D, Figure 4.2F**).

We further validated the enrichment of neuro-ectoderm in IPIT teratomas by immunostaining for *PAX6*, a key marker of neuronal fate determination (**Figure 4.2G**). The three GCV+ teratoma sections with IPIT injections showed higher levels of *PAX6* protein expression than the three GCV- teratoma sections, validating that our miR-124 circuit enriches for neuro-ectoderm (**Figure 4.2G**). We used a secondary antibody (Dylight 550) only staining to confirm that there was no non-specific secondary antibody binding (**Figure 4.2H**). Additionally, we validated that the GCV+ teratoma has higher expression of HES5, a key Radial Glia marker, using RNA FISH (**Figure 4.2I**).

In summary, we developed a miRNA circuit that enables us to engineer the teratoma towards a desired lineage. We demonstrated this circuit *in vitro* using miR-126 (endothelial lineage) and miR-21 (cancer), and *in vivo* using miR-124 (neuro-ectoderm lineage). Our *in vivo* results showed that administering GCV through multiple sites resulted in improved neuro-ectoderm enrichment. Our miRNA circuit can be extended to any cell-type specific miRNA, and could have applications in studying developmental biology and human disease, as well as in tissue engineering.

4.5 Discussion

The teratoma has the potential to be a fully vascularized, multi-lineage model for human development. Its major advantages are that it can grow to a large size due to its vascularization, and it can produce a wide array of relatively mature cell types from all major developmental lineages. Additionally, we show the teratoma can be engineered using miRNA circuits to grow/enrich specific tissues of interest *in vivo.*

Further optimization is necessary on the miRNA molecular sculpting technology. We anticipate there will be a considerable degree of silencing that occurs in the miRNA-suicide gene constructs due to the use of lentiviral vectors. Future studies could explore incorporating these in genomic regions such as the AAVS1 locus that would enable constitutive expression across all cell types. Safety switches based on suicide genes will also be critical for eliminating potential residual undifferentiated cells, and mouse cells within the teratoma, to mitigate impact on safety and utility in tissue engineering applications. This study could also still be enhanced by optimizing the timing, dosing, and route for GCV administration. Additionally, the use of multiple miRNA-regulated switches together may be beneficial to enrich for multiple lineages (i.e. miR-122 liver and miR-126 endothelial) in the same tumor tissue to assess multi-lineage interaction.

We have validated a proof-of-concept for molecular sculpting with our miRNA circuit. However, different lineages may have more effective miRNAs that are also more translationally relevant. A future study of conducting a miRNA circuit screen would be beneficial in assessing which miRNAs are most effective at translational lineage enrichment for downstream focused developmental biology and tissue engineering applications. Taken together, we believe the teratoma is a promising platform for modeling multi-lineage human development and cellular engineering.

4.6 Acknowledgements

We thank members of the Mali lab for advice and help with experiments, Marianna Yusupova for help with initial studies, Alexander Militar for assistance in schematic generation, in loving memory of Nakon Aroonsakool, and to the Moore's Cancer Center Histology Core, UC San Diego Microscopy Core, Sanford Consortium Flow Cytometry Core, and IGM Genomics Center for help with sample processing. This work was generously supported by UCSD Institutional Funds and NIH grants (R01HG009285, RO1CA222826, RO1GM123313).

Chapters 4 is in part reprints of the following materials of which the dissertation author was one of the primary investigators and authors of this paper:

Chapter 4, in part, is a reprint of the material as it appears in McDonald D*, Wu Y*, Dailamy A, Tat J, Parekh U, Zhao D, Hu M, Tipps A, Zhang K, Mali P. Defining the Teratoma as a Model for Multi-lineage Human Development. Cell. 2020 Nov 25;183(5):1402-1419.e18. doi: 10.1016/j.cell.2020.10.018.

*Both of these authors contributed equally

5 Engineering Teratomas via Material Microenvironment

5.1 Abstract

In this final chapter we engineer the teratoma microenvironment to aid in controlling PSC differentiation and ultimate tissue composition. We assayed materials both natural and synthetic across ranges of stiffness: collagen, fibrin, gelatin methacryloyl, polyethylene glycol, blended mixtures of fibrin/hyaluronic acid, blended mixtures of fibrin/Matrigel, blended mixtures of fibrin/gelatin/Matrigel, and finally Matrigel/mTeSR control. We find that the addition of hyaluronic acid allows for greater neural differentiation and fibrin (40mg/mL) allows for greater cardiac muscle differentiation. This validates a proof-of-concept of the importance of the PSC microenvironment and matrix composition for downstream differentiation. Taken together, the teratoma is a promising platform for modeling multi-lineage development and tissue engineering with many parameters that can be uniquely modified by the researcher for their desired teratoma outcome.

5.2 Introduction

Developmental biologists have long wanted to understand the key parameters that influence stem cell differentiation especially in a 3D context[308,309]. Understanding these parameters is vital for organotypic tissue engineering and developmental biology research. Researchers have modulated biomaterials previously to assess regulatory effects on stem cell fate[310,311], but never in a developing three dimensional growing tissue context.

We have previously shown that the teratoma, a recognized standard for validating pluripotency in stem cells, could be a promising platform for studying human developmental processes in 3D as we identified approximately 20 cell types across all 3 germ layers[312]. The inter-teratoma cell type heterogeneity was comparable to organoid systems and the teratoma gut and

brain cell types corresponded well to similar fetal cell types. Additionally, we demonstrated teratomas can be molecularly sculpted to enrich for specific tissues.

After utilizing a genetics approach to sculpt the teratoma, we wish to utilize a materials approach to prime for desired lineages. We assayed teratomas under multiple unique matrix conditions to assess cellular heterogeneity outcomes. We engineered microenvironments by varying matrix composition from our standard 1:1 Growth Factor Reduced Matrigel / mTeSR condition. Assessing an assortment of materials will allow us to understand the impact of a naturally derived environment compared to a fully synthetic environment, as well as the potential impact of certain materials known to be present at high levels in specific organs, such as hyaluronic acid in the brain[313]. Furthermore, each material is used over a range of concentrations to vary the elastic modulus, another factor known to impact stem cell differentiation[314,315]. Fibrin in particular exists as a natural material with a highly tunable stiffness[316,317], and as such, is included in a large portion of the compositions. The matrix composition upon PSC injection for teratoma formation is a unique parameter we can tune to enhance desired tissue types. Taken together, the teratoma is a promising platform for modeling multi-lineage development and tissue engineering.

5.3 Methods

5.3.1 Method Details

5.3.1.1 Hydrogel Formation

Teratoma cells were encapsulated in matrices of 15 different compositions. Specifically, these included: Matrigel (5 mg/mL), collagen (5 mg/mL), fibrin (3, 20, or 40 mg/mL), a blended mixture of fibrin (3 or 20 mg/mL) and hyaluronic acid (2 mg/mL), a blended mixture of fibrin (3 or 20 mg/mL) and Matrigel (4 mg/mL), a blended mixture of fibrin (3 or 20 mg/mL), gelatin (10

mg/mL) and Matrigel (4 mg/mL), gelatin methacryloyl (5% or 10%), and polyethylene glycol (5% or 10%).

Matrigel and collagen matrices were allowed to incubate at 37°C to gelate with no additional components save mTeSR media. Blended matrices of fibrin, gelatin, and Matrigel were formulated with Matrigel (4 mg/mL), fibrinogen (3 or 20 mg/mL), gelatin (10 mg/mL), transglutaminase (2 mg/mL), $CaCl_2$ (2.5 mM), and thrombin (2 U/mL)[318]. Briefly, stock solutions of gelatin, $CaCl_2$, and thrombin were prepared prior to formulation. Type A porcine skin gelatin (Sigma-Aldrich) was dissolved overnight in water (15 wt/vol %) at 70 °C, buffered to pH 7.4 using 1 M NaOH, passed through a 0.22 mm filter (Millipore), and stored at 4 °C. $CaCl_2$ was dissolved at 250 mM in Dulbecco's phosphate buffered saline (dPBS), and Thrombin (Sigma-Aldrich) was prepared at 500 U/mL, aliquoted, and stored at -20 °C. Solutions of both bovine plasma fibrinogen (Millipore) and transglutaminase (MooGloo) were dissolved in dPBS at 37 °C immediately prior to use, and at respective concentrations of 100 mg/mL and 50 mg/mL. During formulation, all components except Matrigel and thrombin were mixed and incubated at 37 °C for 20 minutes, after which Matrigel and thrombin were rapidly added. Blended matrices of fibrin and Matrigel were formulated identically, but with the absence of gelatin, transglutaminase, and $CaCl_2$. Matrices of fibrin alone were formulated identically, but in the additional absence of Matrigel, and with the addition of the high-concentration 40 mg/mL fibrin formulation. Blended matrices of fibrin and hyaluronic acid were formulated from fibrinogen (3 or 20 mg/mL), hyaluronic acid (2 mg/mL), and thrombin (2 U/mL). All components except hyaluronic acid were prepared as previously described. Hyaluronic acid (LifeCore) was prepared at a stock concentration of 10 mg/mL by stirring overnight in PBS at 4 °C. Matrices were prepared by directly mixing all components.

Gelatin methacryloyl 300-bloom (Millipore-Sigma) and polyethylene glycol diacrylate (Millipore-Sigma) were both prepared by dissolving respective components in dPBS to form stock 20% solutions, diluting with mTeSR to the appropriate concentrations, encapsulating cells, and exposing to a UV light source in the presence of Irgacure 2959 (Millipore-Sigma) initiator to induce radical polymerization.

5.3.1.2 Hydrogel Implantation

Mice were anesthetized using intraperitoneal administration of ketamine (75mg/kg) / xylazine (15mg/kg). Once the mouse was fully anesthetized, the right flank was shaved and sterilized with alternating swabs of 7.5% povidone-iodine and 10% USP povidone-iodine respectively (PDI PVP #S141125) followed by 70% isopropyl alcohol. A small incision was made on the right flank subcutaneously with small animal surgical scissors. Subcutaneous connective tissue was released via blunt dissection to create a small pocket for the hydrogel. Upon hydrogel placement, the incision was closed with standard 4-0 silk sutures (UNIFY® #S-S418R13). Sutures were removed 10-14 days post-op.

5.3.1.3 Teratoma Processing

After growth for 8-10 weeks on average mice were euthanized by slow release of CO2 followed by secondary means via cervical dislocation. Tumor area was shaved, sprayed with 70% ethanol, and then extracted via surgical excision using scissors and forceps. Tumor was rinsed with PBS, weighed, and photographed. Tumors were cut in a semi-random fashion in ≤ 22 mm diameter pieces and frozen in OCT for sectioning and H&E staining courtesy of the Moore's Cancer Center Histology Core. Remaining tumor was cut into small pieces and flash frozen in LN2 for downstream bulk RNA extraction.

5.3.1.4 Histology

Sectioning and H&E staining was performed by the Moore's Cancer Center Histology Core. In brief, Optimal Cutting Temperature (O.C.T.) blocks were sectioned with a cryostat into 10 micron sections onto a positively charged glass slide. The slide was then stained with Harris hematoxylin and then rinsed in tap water and treated with an alkaline solution. The slide was then de-stained to remove non-specific background staining with a weak acid alcohol. The section was then stained with an aqueous solution of eosin and passed through several changes of alcohol, then rinsed in several baths of xylene. A thin layer of polystyrene mountant was applied, followed by a glass cover slip.

5.3.1.5 Bulk RNA Extraction

Teratoma samples were frozen via LN2, then pulverized with a pestle and mortar until fully powderized. Powderized samples were resuspended in Qiazol (Qiagen) at a ratio of 900 uL Qiazol per 100 mg of original tissue, mixed with an 18-gauge syringe, and allowed to incubate on a shaker for approximately 30 minutes. Samples were centrifuged at 12000g at 4 °C for 10 minutes, and supernatant was collected. Chloroform (Fisher Scientific) was added to samples at a ratio of 180 uL chloroform per 900 uL supernatant. Samples were mixed and allowed to incubate for 10 minutes, then centrifuged at 12000g at 4 °C for 15 minutes to separate into aqueous and organic phases. The aqueous phase was collected, while the organic phase was discarded. Samples were diluted at a 1:1 ratio with 70% ethanol, and the remainder of the extraction was performed using the RNeasy Kit (Qiagen).

Following RNA extraction, approximately 1 ug of RNA from each condition was used to synthesize cDNA and construct a transcriptomic library using the NEBNext poly(A) mRNA Magnetic Isolation Module (New England Biosystems) and NEBNext Ultra II RNA Library Prep

Kit (New England Biosystems). Multiplex indexing was performed with NEB Multiplex Primers. The final product was purified using Ampure XP beads, pooled in equal ratios, and sequenced using the NovaSeq with paired end 100 bp reads.

5.3.2 Quantification and Statistical Analysis

5.3.2.1 Cibersort

Reads were aligned to both human reference genome HG38 and mouse reference genome MM10 using STAR[70]. Resulting BAM files were processed using the XenofilteR analytical tool[319] to remove reads from transcripts from cells of the host mice. Read counts were then generated by mapping to reference transcriptome GenCode v33 using FeatureCounts. Prior single-cell sequencing data across seven teratomas was used to construct a signature matrix containing gene profiles associated with twenty-three different cell types identified within the teratomas. The CIBERSORTx analytical tool (https://cibersort.stanford.edu/)[320] was used to estimate abundances of each of the different cell types based on the read counts obtained from the transcriptomic libraries, and the gene profiles in the signature matrix.

To decompose the proportion of each cell type in a given sample from bulk RNA-seq data with CIBERSORTx, we created a reference matrix for input to CIBERSORTx. We obtained a single cell expression matrix by randomly sampling up to 200 cells from each cell type present in 7 wild type teratomas[312].

5.3.2.2 Gene Ontology

Read counts were normalized in DESeq2[260] both across all samples for a given gene using the geometric mean, and within each sample using the median. Relative expression profiles and differentially expressed gene lists were subsequently generated using the DESeq2 pipeline. Enriched and depleted pathways were identified using Metascape[321]. Differentially expressed

genes were classified as genes with both |Zscore| > 2 and FDR < 0.1, were input into Metascape, and pathway lists were restricted to terms within the Gene Ontology Biological Process domain.

5.3.2.3 Figure Generation

All figures were generated using original artwork or open source with InkScape, Adobe Illustrator®, and ImageJ.

5.4 Results

We have previously characterized the teratoma in terms of cell type presence and proportion with the standard Matrigel:mTeSR (1:1) matrix condition. We wanted to better understand changes in cell type heterogeneity when these matrix conditions are manipulated. We thus, encapsulated H1 ESCs in the following 15 matrix conditions for downstream analysis: Matrigel (5 mg/mL), collagen (5 mg/mL), fibrin (3, 20, or 40 mg/mL), a blended mixture of fibrin (3 or 20 mg/mL) and hyaluronic acid (2 mg/mL), a blended mixture of fibrin (3 or 20 mg/mL) and Matrigel (4 mg/mL), a blended mixture of fibrin (3 or 20 mg/mL), gelatin (10 mg/mL) and Matrigel (4 mg/mL), gelatin methacryloyl (5% or 10%), and polyethylene glycol (5% or 10%) (**Table 5.1**).

Table 5.1. Matrix Conditions Analyzed

Condition	Matrigel	Collagen	Fibrin	Hyaluronic Acid	Gelatin	GelMA	PEGDA
1	5 mg/mL						
2		5 mg/mL					
3			3 mg/mL				
4			20 mg/mL				
5			40 mg/mL				
6			3 mg/mL	2 mg/mL			
7			20 mg/mL	2 mg/mL			
8	4 mg/mL		3 mg/mL				
9	4 mg/mL		20 mg/mL				
10	4 mg/mL		3 mg/mL		10 mg/mL		
11	4 mg/mL		20 mg/mL		10 mg/mL		
12						5%	
13						20%	
14							5%
15							20%

After formation and implantation of hydrogels subcutaneously in the right flank of Rag2$^{-/-}$;γc$^{-/-}$ immunodeficient mice (**5.3 Materials and Methods**) teratomas were allowed to grow for up to 8 weeks until the tumors were of a sufficient size for extraction and downstream analyses. Post-extraction, tumors were photographed (**Figure 5.1A**), weighed, inspected, and sectioned in a semi-random fashion. Half of tissue was utilized for bulk RNA extraction and sequencing and other half was utilized for H&E staining (**5.3 Materials and Methods**). Of note, the collagen matrix condition did not form any appreciable tumor mass throughout the study. Additionally, all synthetic conditions (gelatin methacryloyl and polyethylene glycol) failed to form tumor tissue as well and upon extraction (8 weeks) only the initial matrix was present with a surrounding fibrous capsule. Upon initial inspection, the blended mixture of fibrin, gelatin, and Matrigel (FibGelMa)

condition formed tumors that were excessively black in color, the fibrin and hyaluronic acid (FibHA) condition formed softer tumors than the matrigel control, and the fibrin (40mg/mL) condition formed tumors with little to no black coloration at all. All other tumors seemed to be visually similar to matrigel control teratomas macroscopically (**Figure 1.1C**, **Figure 5.1A**).

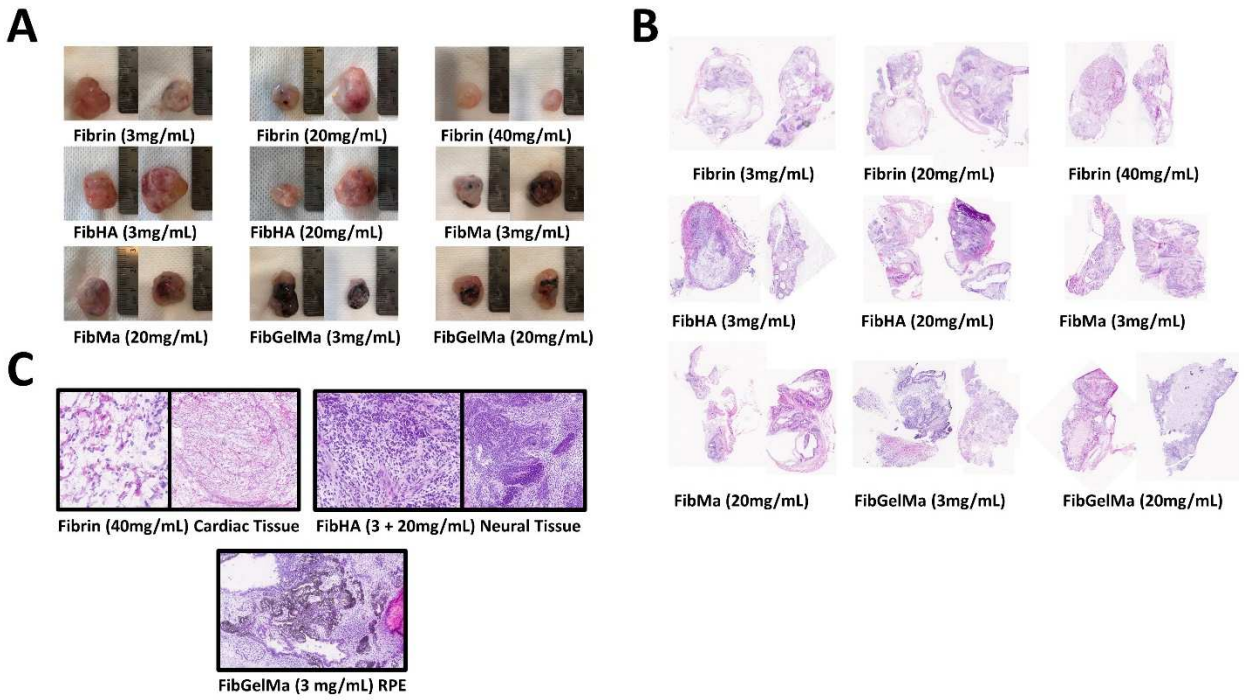


Figure 5.1. Images and Histology of Teratomas with Different Matrix Conditions. (A) Images of 18 teratomas generated from H1 cells encapsulated in varying matrix conditions. **(B)** H&E stains of the 18 teratoma histology sections. **(C)** Highlighted regions of enriched tissue types for Fibrin (40mg/mL), FibHA (3 + 20mg/mL), and FibGelMa (3mg/mL).

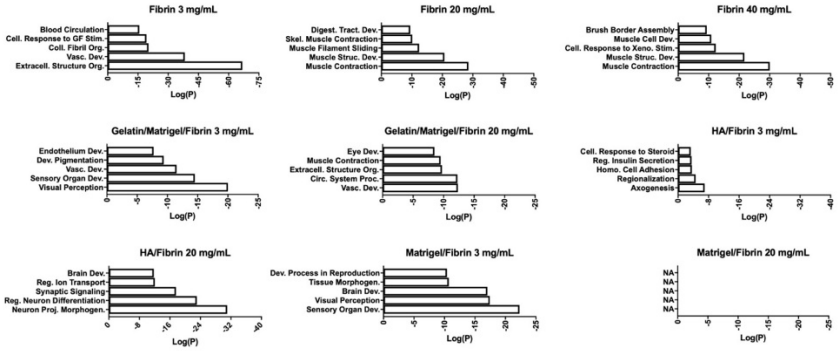
Upon deeper histological analysis we saw in general fibrin (3mg/mL) and fibrin (20mg/mL) showed unremarkable sections comparable to matrigel control tumors. However, fibrin (40mg/mL) showed an excess of muscle cell types, in particular cardiac muscle in both replicates (**Figure 5.1B,C**). The FibHA conditions (3mg/mL and 20mg/mL) showed greater neural architecture compared to matrigel controls histologically with FibHA (3mg/mL) showing more

differentiating neuroblasts interlaced with fetal skeletal muscle and FibHA (20mg/mL) showing greater primitive neuroectoderm (**Figure 5.1B,C**). The fibrin (3 or 20 mg/mL) and Matrigel (4 mg/mL) blended conditions histologically appeared somewhat unremarkable with some presence of retinal pigmented epithelium (**Figure 5.1B**). However, the FibGelMa (3 and 20mg/mL) was notable for a large retinal pigmented epithelium (RPE) presence in particular with the FibGelMa (3mg/mL) condition throughout the section (**Figure 5.1B,C**). The FibGelMa (20mg/mL) condition although had regions of RPE was also notable for a relatively intact matrix upon sectioning with little tissue and cell development in the interior of the matrix likely due to its high stiffness levels (**Figure 5.1B**).

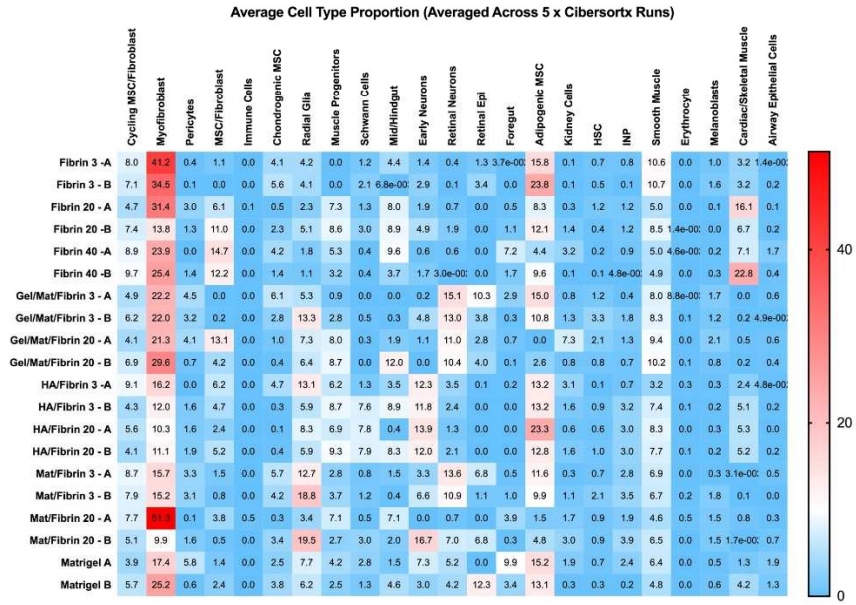
These results were all confirmed via gene ontology analysis from respective tumor tissue bulk RNA compared to matrigel control (**Figure 5.2A**). In particular, the fibrin (40mg/mL) conditions showed its most abundant gene pathways in muscle contraction, muscle structure development, muscle cell development, with some additional pathways related to intestinal development (**Figure 5.2A**). This result would be consistent with known biology of increased matrix stiffness (~9-11 kPa) leading to more muscular development[314]. We assessed fibrin (40mg/mL) via Cibersort to determine cell type distribution from bulk RNA data by projecting the data onto the original H1 teratoma dataset (**5.3 Materials and Methods**)[312]. The Cibersort dataset shows consistent upregulation in cardiac/skeletal muscle cell types compared to matrigel control and in addition, increased MSC populations (also mesodermal) (**Figure 5.2B,C**). Fibrin (40mg/mL) has the greatest enrichment in muscle cell types in comparison to all other tested matrix conditions (**Figure 5.2B,C**).

Figure 5.2. Gene Ontology and Heatmap of Teratomas with Different Matrix Conditions. (A) Gene ontology of teratomas with different matrix conditions compared to matrigel control **(B)** Average cell type proportion heat map of teratomas with different matrix conditions across 5 Cibersortx runs **(C)** Log Fold Change of Cell Type Proportions heat map relative to Matrigel control across 5 Cibersortx runs

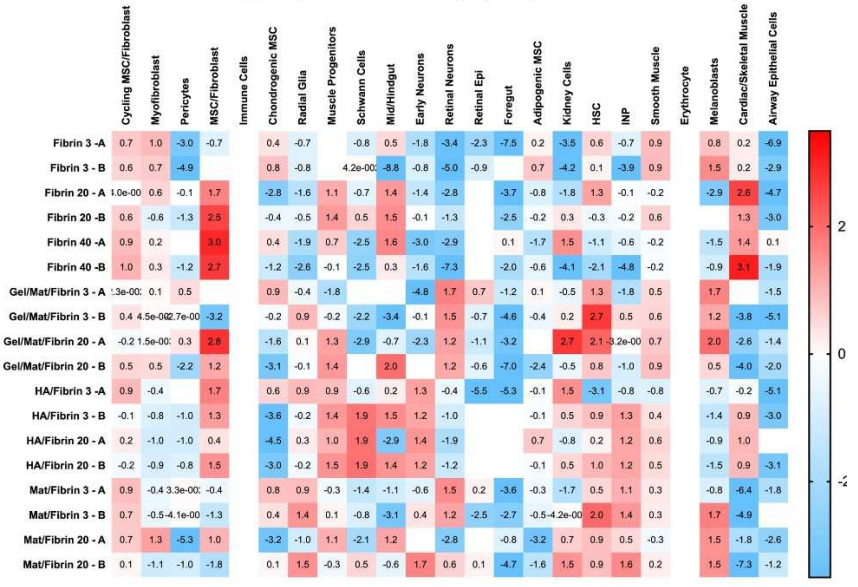
A


Fibrin 3 mg/mL
Blood Circulation
Cell. Response to GF Stim.
Coll. Fibril Org.
Vasc. Dev.
Extracell. Structure Org.
Log(P)
Fibrin 20 mg/mL
Digest. Tract. Dev.
Skel. Muscle Contraction
Muscle Filament Sliding
Muscle Struc. Dev.
Muscle Contraction
Log(P)
Fibrin 40 mg/mL
Brush Border Assembly
Muscle Cell Dev.
Cell. Response to Xeno. Stim.
Muscle Struc. Dev.
Muscle Contraction
Log(P)
Gelatin/Matrigel/Fibrin 3 mg/mL
Endothelium Dev.
Dev. Pigmentation
Vasc. Dev.
Sensory Organ Dev.
Visual Perception
Log(P)
Gelatin/Matrigel/Fibrin 20 mg/mL
Eye Dev.
Muscle Contraction
Extracell. Structure Org.
Circ. System Proc.
Vasc. Dev.
Log(P)
HA/Fibrin 3 mg/mL
Cell. Response to Steroid
Reg. Insulin Secretion
Homo. Cell Adhesion
Regionalization
Axogenesis
Log(P)
HA/Fibrin 20 mg/mL
Brain Dev.
Reg. Ion Transport
Synaptic Signaling
Reg. Neuron Differentiation
Neuron Proj. Morphogen.
Log(P)
Matrigel/Fibrin 3 mg/mL
Dev. Process in Reproduction
Tissue Morphogen.
Brain Dev.
Visual Perception
Sensory Organ Dev.
Log(P)
Matrigel/Fibrin 20 mg/mL
NA
NA
NA
NA
NA
Log(P)

B


Average Cell Type Proportion (Averaged Across 5 x Cibersortx Runs)
Cycling MSC/Fibroblast
Myofibroblast
Pericytes
MSC/Fibroblast
Immune Cells
Chondrogenic MSC
Radial Glia
Muscle Progenitors
Schwann Cells
Mid/Hindgut
Early Neurons
Retinal Neurons
Retinal Epi
Foregut
Adipogenic MSC
Kidney Cells
HSC
INP
Smooth Muscle
Erythrocyte
Melanoblasts
Cardiac/Skeletal Muscle
Airway Epithelial Cells
Fibrin 3 -A 8.0 41.2 0.4 1.1 0.0 4.1 4.2 0.0 1.2 4.4 1.4 0.4 1.3 3.7e-00 15.8 0.1 0.7 0.8 10.6 0.0 1.0 3.2 1.4e-00
Fibrin 3 - B 7.1 34.5 0.1 0.0 0.0 5.6 4.1 0.0 2.1 6.8e-00 2.9 0.1 3.4 0.0 23.8 0.1 0.5 0.1 10.7 0.0 1.6 3.2 0.2
Fibrin 20 - A 4.7 31.4 3.0 6.1 0.1 0.5 2.3 7.3 1.3 8.0 1.9 0.7 0.0 0.5 8.3 0.3 1.2 1.2 5.0 0.0 0.1 16.1 0.1
Fibrin 20 -B 7.4 13.8 1.3 11.0 0.0 2.3 5.1 8.6 3.0 8.9 4.9 1.9 0.0 1.1 12.1 1.4 0.4 1.2 8.5 1.4e-00 0.0 6.7 0.2
Fibrin 40 -A 8.9 23.9 0.0 14.7 0.0 4.2 1.8 5.3 0.4 9.6 0.6 0.6 0.0 7.2 4.4 3.2 0.2 0.9 5.0 4.6e-00 0.2 7.1 1.7
Fibrin 40 -B 9.7 25.4 1.4 12.2 0.0 1.4 1.1 3.2 0.4 3.7 1.7 3.0e-00 0.0 1.7 9.6 0.1 0.1 4.8e-00 4.9 0.0 0.3 22.8 0.4
Gel/Mat/Fibrin 3 - A 4.9 22.2 4.5 0.0 0.0 6.1 5.3 0.9 0.0 0.0 0.2 15.1 10.3 2.9 15.0 0.8 1.2 0.4 8.0 8.8e-00 1.7 0.0 0.6
Gel/Mat/Fibrin 3 - B 6.2 22.0 3.2 0.2 0.0 2.8 13.3 2.8 0.5 0.3 4.8 13.0 3.8 0.3 10.8 1.3 3.3 1.8 8.3 0.1 1.2 0.2 4.9e-00
Gel/Mat/Fibrin 20 - A 4.1 21.3 4.1 13.1 0.0 1.0 7.3 8.0 0.3 1.9 1.1 11.0 2.8 0.7 0.0 7.3 2.1 1.3 9.4 0.0 2.1 0.5 0.6
Gel/Mat/Fibrin 20 - B 6.9 29.6 0.7 4.2 0.0 0.4 6.4 8.7 0.0 12.0 0.0 10.4 4.0 0.1 2.6 0.8 0.8 0.7 10.2 0.1 0.8 0.2 0.4
HA/Fibrin 3 -A 9.1 16.2 0.0 6.2 0.0 4.7 13.1 6.2 1.3 3.5 12.3 3.5 0.1 0.2 13.2 3.1 0.1 0.7 3.2 0.3 0.3 2.4 4.8e-00
HA/Fibrin 3 - B 4.3 12.0 1.6 4.7 0.0 0.3 5.9 8.7 7.6 8.9 11.8 2.4 0.0 0.0 13.2 1.6 0.9 3.2 7.4 0.1 0.2 5.1 0.2
HA/Fibrin 20 - A 5.6 10.3 1.6 2.4 0.0 0.1 8.3 6.9 7.8 0.4 13.9 1.3 0.0 0.0 23.3 0.6 0.6 3.0 8.3 0.0 0.3 5.3 0.0
HA/Fibrin 20 - B 4.1 11.1 1.9 5.2 0.0 0.4 5.9 9.3 7.9 8.3 12.0 2.1 0.0 0.0 12.8 1.6 1.0 3.0 7.7 0.1 0.2 5.2 0.2
Mat/Fibrin 3 - A 8.7 15.7 3.3 1.5 0.0 5.7 12.7 2.8 0.8 1.5 3.3 13.6 6.8 0.5 11.6 0.3 0.7 2.8 6.9 0.0 0.3 3.1e-00 0.5
Mat/Fibrin 3 - B 7.9 15.2 3.1 0.8 0.0 4.2 18.8 3.7 1.2 0.4 6.6 10.9 1.1 1.0 9.9 1.1 2.1 3.5 6.7 0.2 1.8 0.1 0.0
Mat/Fibrin 20 - A 7.7 51.3 0.1 3.8 0.5 0.3 3.4 7.1 0.5 7.1 0.0 0.7 0.0 3.9 1.5 1.7 0.9 1.9 4.6 0.5 1.5 0.8 0.3
Mat/Fibrin 20 - B 5.1 9.9 1.6 0.5 0.0 3.4 19.5 2.7 3.0 2.0 16.7 7.0 6.8 0.3 4.8 3.0 0.9 3.9 6.5 0.0 1.5 1.7e-00 0.7
Matrigel A 3.9 17.4 5.8 1.4 0.0 2.5 7.7 4.2 2.8 1.5 7.3 5.2 0.0 9.9 15.2 1.9 0.7 2.4 6.4 0.0 0.5 1.3 1.9
Matrigel B 5.7 25.2 0.6 2.4 0.0 3.8 6.2 2.5 1.3 4.6 3.0 4.2 12.3 3.4 13.1 0.3 0.3 0.2 4.8 0.0 0.6 4.2 1.3
40
20
0

C


LFC of Cell Type Proportions Relative to Matrigel (Averaged Across 5 x Cibersortx Runs)
Cycling MSC/Fibroblast
Myofibroblast
Pericytes
MSC/Fibroblast
Immune Cells
Chondrogenic MSC
Radial Glia
Muscle Progenitors
Schwann Cells
Mid/Hindgut
Early Neurons
Retinal Neurons
Retinal Epi
Foregut
Adipogenic MSC
Kidney Cells
HSC
INP
Smooth Muscle
Erythrocyte
Melanoblasts
Cardiac/Skeletal Muscle
Airway Epithelial Cells
Fibrin 3 -A 0.7 1.0 -3.0 -0.7 0.4 -0.7 -0.8 0.5 -1.8 -3.4 -2.3 -7.5 0.2 -3.5 0.6 -0.7 0.9 0.8 0.2 -6.9
Fibrin 3 - B 0.6 0.7 -4.9 0.8 -0.8 4.2e-00 -8.8 -0.8 -5.0 -0.9 0.7 -4.2 0.1 -3.9 0.9 1.5 0.2 -2.9
Fibrin 20 - A 1.0e-00 0.6 -0.1 1.7 -2.8 -1.6 1.1 -0.7 1.4 -1.4 -2.8 -3.7 -0.8 -1.8 1.3 -0.1 -0.2 -2.9 2.6 -4.7
Fibrin 20 -B 0.6 -0.6 -1.3 2.5 -0.4 -0.5 1.4 0.5 1.5 -0.1 -1.3 -2.5 -0.2 0.3 -0.3 -0.2 0.6 1.3 -3.0
Fibrin 40 -A 0.9 0.2 3.0 0.4 -1.9 0.7 -2.5 1.6 -3.0 -2.9 0.1 -1.7 1.5 -1.1 -0.6 -0.2 -1.5 1.4 0.1
Fibrin 40 -B 1.0 0.3 -1.2 2.7 -1.2 -2.6 -0.1 -2.5 0.3 -1.6 -7.3 -2.0 -0.6 -4.1 -2.1 -4.8 -0.2 -0.9 3.1 -1.9
Gel/Mat/Fibrin 3 - A 3e-00 0.1 0.5 0.9 -0.4 -1.8 -4.8 1.7 0.7 -1.2 0.1 -0.5 1.3 -1.8 0.5 1.7 -1.5
Gel/Mat/Fibrin 3 - B 0.4 4.5e-00 2.7e-00 -3.2 -0.2 0.9 -0.2 -2.2 -3.4 -0.1 1.5 -0.7 -4.6 -0.4 0.2 2.7 0.5 0.6 1.2 -3.8 -5.1
Gel/Mat/Fibrin 20 - A -0.2 1.5e-00 0.3 2.8 -1.6 0.1 1.3 -2.9 -0.7 -2.3 1.2 -1.1 -3.2 2.7 2.1 -3.2e-00 0.7 2.0 -2.6 -1.4
Gel/Mat/Fibrin 20 - B 0.5 0.5 -2.2 1.2 -3.1 -0.1 1.4 2.0 1.2 -0.6 -7.0 -2.4 -0.5 0.8 -1.0 0.9 0.5 -4.0 -2.0
HA/Fibrin 3 -A 0.9 -0.4 1.7 0.6 0.9 0.9 -0.6 0.2 1.3 -0.4 -5.5 -5.3 -0.1 1.5 -3.1 -0.8 -0.8 -0.7 -0.2 -5.1
HA/Fibrin 3 - B -0.1 -0.8 -1.0 1.3 -3.6 -0.2 1.4 1.9 1.5 1.2 -1.0 -0.1 0.5 0.9 1.3 0.4 -1.4 0.9 -3.0
HA/Fibrin 20 - A 0.2 -1.0 -1.0 0.4 -4.5 0.3 1.0 1.9 -2.9 1.4 -1.9 0.7 -0.8 0.2 1.2 0.6 -0.9 1.0
HA/Fibrin 20 - B -0.2 -0.9 -0.8 1.5 -3.0 -0.2 1.5 1.9 1.4 1.2 -1.2 -0.1 0.5 1.0 1.2 0.5 -1.5 0.9 -3.1
Mat/Fibrin 3 - A 0.9 -0.4 3.3e-00 -0.4 0.8 0.9 -0.3 -1.4 -1.1 -0.6 1.5 0.2 -3.6 -0.3 -1.7 0.5 1.1 0.3 -0.8 -6.4 -1.8
Mat/Fibrin 3 - B 0.7 -0.5 -4.1e-00 -1.3 0.4 1.4 0.1 -0.8 -3.1 0.4 1.2 -2.5 -2.7 -0.5 -4.2e-00 2.0 1.4 0.3 1.7 -4.9
Mat/Fibrin 20 - A 0.7 1.3 -5.3 1.0 -3.2 -1.0 1.1 -2.1 1.2 -2.8 -0.8 -3.2 0.7 0.9 0.5 -0.3 1.5 -1.8 -2.6
Mat/Fibrin 20 - B 0.1 -1.1 -1.0 -1.8 0.1 1.5 -0.3 0.5 -0.6 1.7 0.6 0.1 -4.7 -1.6 1.5 0.9 1.6 0.2 1.5 -7.3 -1.2
2
0
-2

The FibHA conditions (3mg/mL and 20mg/mL) were also consistent in their gene ontology results showing greater levels of neuronal pathways (especially FibHA 20mg/mL) compared to matrix control, in particular: axonogenesis, neuron projection morphogenesis, regulation of neuronal differentiation, synaptic signaling, and brain development. The Cibersort dataset shows the greatest enrichment in Schwann cells compared to all other matrix conditions. There is also consistent enrichment in early neurons and intermediate neuronal progenitors (INP) compared to matrigel control (**Figure 5.2B,C**). This result is reasonable with known biology as it is well known in the field that hyaluronic acid has a high water-retaining capacity allowing for a soft sponginess (previously noted) and highly prevalent in the brain[313,322]. In addition, there is increased muscle progenitors and cardiac/skeletal muscle with this condition (**Figure 5.2B,C**).

Finally, the FibGelMa conditions (3 and 20mg/mL) had consistently upregulated gene ontology pathways for eye development, in particular: visual perception, sensory organ development, and developmental pigmentation (RPE importance). This was consistent with histological analysis and Cibersort confirms this as well showing the highest levels of retinal neurons compared to all other matrix conditions tested. In addition, there were high levels of melanoblasts (also a pigmented cell population) and MSCs.

These data show that stiffer fibrin conditions (40mg/mL, ~9-11kPa) can influence a more myogenic phenotype, addition of hyaluronic acid to a more neuronal phenotype, and including gelatin into a matrigel fibrin blend to a more pigmented and retinal phenotype. Further validation will still be needed on this front with more rigorous scRNA-seq analysis and immunostaining/RNA-FISH studies. Taken together, unique matrix conditions can influence the stem cell microenvironment and differentiation/lineage commitment in a 3D context.

5.5 Discussion

This study has highlighted the importance of the cellular microenvironment and its influence on stem cell growth and differentiation in 3D space utilizing the teratoma. When the proper cues are given to PSCs they can seemingly be primed and fed information in order to differentiate down specific lineages. This has been shown before in many studies[314,315,323–325], but this study highlights the importance in a 3D context with a growing maturing vascularized tissue. This study may offer some novel information for developmental biologists and tissue engineers. Cerebral organoids have long been grown and encapsulated in a Matrigel condition[20,102], but it may be more beneficial to be grown in a hyaluronic acid-containing matrix to boost growth dynamics and maturity from standard current protocols. More research would be needed to assess the validity of this idea. In addition, muscle differentiation seems to benefit best in a matrix condition with matched stiffness to native muscle tissue and may prove to be beneficial for future muscle generation studies.

This study, like any study, has its limitations. We allowed for random assembly of tissue which gives semi-random architecture. Though influencing matrix conditions may push PSCs towards a lineage of interest, structured tissue organization is still a key component to tackle. In addition, the system is not perfect with additional unwanted tissue still present from all other germ layers regardless of the enrichment seen in our conditions. Finding ways to trim unwanted tissue will be essential for tissue engineering applications. Utilizing this method in combination with a miRNA-regulated suicide gene circuit (**Chapter 4**)[312] may prove to help alleviate some of these issues. Additionally, for proper tissue engineering and future transplantation studies, the mouse vasculature will need to be eliminated as a chimeric transplantation will be unreasonable in any clinical application. Utilizing *ex vivo* growth methods with bioreactors and roller cultures in

combination with 3D bioprinting may prove to be essential in these cases[41,318,326]. Finally, further validation will be necessary with rigorous scRNA-seq analysis and immunostaining/RNA-FISH studies.

The strengths of tuning the teratoma matrix are tremendous with seemingly unlimited options. Here we tested a few key conditions for a proof-of-concept, but this can be taken further with addition of growth factors/cytokines and testing novel matrices. The researcher may, in addition, tweak other parameters to shift cell type heterogeneity such as cell number upon injection or implantation/injection location for unique signaling cues from the surrounding microenvironment (i.e. muscular injection)[56]. Taken together, we believe the teratoma is a promising platform for modeling multi-lineage development and tissue engineering with multi-faceted tuning ability of many unique parameters.

5.6 Acknowledgements

We thank members of the Mali lab for advice and help with experiments, to the Moore's Cancer Center Histology Core, and IGM Genomics Center for help with sample processing. This work was generously supported by UCSD Institutional Funds and NIH grants (R01HG009285, RO1CA222826, RO1GM123313).

Chapter 5 is unpublished material in which the dissertation author was one of the primary investigators and authors:

Hu, Michael*; McDonald, Daniella*. Chapter 5, Engineering Teratomas via Material Microenvironment.

*Both of these authors contributed equally.

6 Conclusions and Outlook

6.1 Conclusions

The teratoma has the potential to be a fully vascularized, multi-lineage model for human development. Its major advantages are that it can grow to a large size due to its vascularization, and it can produce a wide array of relatively mature cell types from all major developmental lineages. Additionally, as we demonstrated with our CRISPR-Cas9 knockout screens, the teratoma's ability to generate cells from all lineages enables a comprehensive assessment of the effect of genetic perturbations on human development within a single integrated experiment. Furthermore, we show the teratoma can be engineered using miRNA circuits and/or novel matrix conditions to grow/enrich specific tissues of interest *in vivo*.

In Chapter 1, we rigorously characterized the teratoma determining an array of 20+ cell types present across all germ layers and assessed reproducibility and maturity of these cell types and abundance patterns across different stem cell lines. The teratoma's reproducibility was comparable to patterned brain organoids when formed within the same cell line. In terms of maturity, we determined asynchronous development with neural tissue being staged as more mature (week 17 gestational age) than the teratoma's gut tissue (week 8 gestational age).

In Chapter 2, we discussed the importance of functional genomics utilizing CRISPR-Cas9 and clarified the importance of 3D *in vivo* screens to produce the most biomimetic and rigorous datasets (i.e. organoids or the teratoma).

In Chapter 3, we utilized CRISPR-Cas9 to perform our own developmental and neural disease screens within the teratoma to assess the effects of perturbing key developmental genes or mimicking genetic neural diseases in a 3D developing vascularized multi-lineage tissue. We found many key developmental gene knockouts reproduce biological effects consistent with literature

(i.e. CDX2 knockout leading to depletion in mid/hindgut and enrichment in foregut). We realized the power of this model as we had some perturbation effects span all germ layers in a single experiment (i.e. RUNX1), an experiment that would be seemingly impossible until now and requiring organoid systems from multiple lineages. Furthermore, we recapitulated 3 neural diseases in the teratoma (Pitt-Hopkins, Rett Syndrome, and L1 Syndrome) all in a single study to assess changes in differential gene expression.

In Chapter 4, we modulated the teratoma to sculpt tissue types of interest. We utilized a miRNA-regulated suicide gene circuit to enrich for the neural lineage while trimming away unwanted tissue types.

In Chapter 5, we modulated the teratoma by tuning the PSCs matrix conditions to influence growth and differentiation. This study showed the importance of hyaluronic acid in neural development and proper matrix stiffness for muscle development.

Taken together, this study has characterized, validated, utilized, and engineered the teratoma to assess its power as a model for multi-lineage human development.

Any model system has its intrinsic strengths and weaknesses, and below we discuss some of the limitations of the teratoma system and also considerations towards improving it for enabling basic science and engineering studies. One issue with the teratoma system (and organoids) is the intrinsic degree of heterogeneity[26,104,105,108]. In this regard, we found the use of internal controls when conducting perturbation experiments was important. For example, in our CRISPR-Cas9 screen, each teratoma contained both gene targeting guides and non-targeting controls, enabling us to compare cell type proportion shifts within each teratoma without having to worry about heterogeneity between teratomas.

While the teratoma has regions of organization and maturity, these may develop in an asynchronous manner. This lack of synchronization may prove to be a barrier in accessing certain mature cell types that need a highly ordered cellular context to develop.

Also, since the teratoma contains cell types from all lineages, finding a single dissociation protocol that captures as many cell types as possible is a challenge. The choice of dissociation method can drastically change the cell types profiled in single cell RNA-seq, and it is likely that the set of cell types we see in our data is biased by our dissociation protocol [109]. It may be the case that no single dissociation method can capture all cell types, and it will be necessary to design specific dissociation protocols to capture specific tissues.

Additionally, our cell type annotations are still preliminary. While we validated key cell types by comparison to fetal human/mouse reference datasets and RNA FISH, we were not able to validate all cell types due to limited developmental human reference scRNA-seq datasets, as well as cost constraints. Thus, some cell types, such as the neuro-ectoderm cell types, have more validation than others, giving us greater confidence in their identity **(Table 1)**. We may also still be underpowered in detecting less abundant cell types and additional single cell RNA-seq could enable us to resolve some missing cell types, as under sampling could result in smaller cell types being collapsed into a larger cell type during analysis.

In regard to lineage engineering, we anticipate there will be a considerable degree of silencing that occurs in the miRNA-suicide gene constructs due to the use of lentiviral vectors. Future studies could explore incorporating these in genomic regions such as the AAVS1 locus that would enable constitutive expression across all cell types. Safety switches based on suicide genes will also be critical for eliminating potential residual undifferentiated cells, and mouse cells within the teratoma, to mitigate impact on safety and utility in tissue engineering applications.

Additionally, the teratoma is a chimera utilizing the host vasculature for growth and development. The mouse vasculature will need to be eliminated as a chimeric transplantation will be unreasonable in any clinical application. Utilizing *ex vivo* growth methods with bioreactors and roller cultures in combination with 3D bioprinting may prove to be essential in these cases[41,318,326].

The feasibility of utilizing the teratoma as a model in terms of cost is as follows. Overall, the cost of profiling a single teratoma with the 10X RNA-seq system runs at about $1,300, including sequencing costs for ~8,000 cells (the output of a single 10X RNA-seq run) at a sequencing depth of 50,000 reads per cell. Mouse husbandry and reagents related to teratoma formation (cells, Matrigel, media) are relatively cheap in comparison. During teratoma growth, the researcher needs to only monitor the mice for health concerns, weights, and tumor measurements if desired. The teratoma can be extracted at any time after 3 weeks of growth. It is also theoretically possible to inject both flanks of the mouse to generate 2 teratomas per animal. With the availability of easy to use analysis tools such as Seurat/PAGODA2, as well as methods for integrating datasets (such as CONOS), running a basic clustering and cell type annotation of scRNA-seq data is fairly straightforward.

Taken together, we believe the teratoma is a promising platform for modeling multi-lineage human development, pan-tissue functional genetic screening, and cellular/tissue engineering with multi-faceted tuning ability of many unique parameters.

6.2 Outlook

The power of the teratoma has just begun to be fully realized. What makes this system so potent is the tunability of many parameters to ultimately form a tissue of interest to the researcher. With the teratoma, the researcher can adjust which cells enter the teratoma upon injection. For example, the researcher can blend PSCs with other cell types to influence growth (i.e. HUVEC

pooling)[59] or change the cell density for the injection site. Additionally, cells may be pulsed with factors prior to injection or during growth such as SMAD inhibitors for greater neural populations[327]. The surrounding matrix that the cells reside in can be manipulated to influence PSC growth. Additionally, injection site can potentially enrich for desired cell types (i.e. muscle for muscle cells)[56]. Finally, time is a key factor to increase (or decrease) maturity. The allowance for larger host animals besides mice can increase the time these tumors are allowed to be ethically formed in a host such as the use of mini-pigs or larger farm animals. Tweaking all of these parameters may allow for the greatest control of tissue types the researcher is interested in. For example, for greatest neural enrichment perhaps the researcher can utilize PSCs containing the miR-124 suicide gene circuit pulsed with SMAD inhibitors prior to injection into the cerebral area of a mouse for growth while encapsulated in a hyaluronic acid-containing matrix. Thus PSCs are given the greatest chance for neural growth while also exogenously adding GCV to continuously trim away undesired lineages throughout growth. With this tissue, researchers may study the effects of drugs on human tissue, perform screens, or even span into transplantation studies (for endodermal or cardiac tissues) which may in the future aid in ameliorating the ongoing issue of donor deficiency and the extensive UNOS list[328]. The researcher may also use multiple miRNA-regulated circuits simultaneously to enrich for multiple lineages together to study unique questions regarding the brain-gut axis or neuromuscular junction for example[329,330]. The teratoma intrinsically has low endodermal levels (as is the case in human development) so tuning multiple parameters to push for endodermal tissue may be of highest interest for the future of these studies and for the most clinically translational studies.

One hindrance with the teratoma is how it is chimeric in nature utilizing the host vasculature. Finding *ex vivo* culturing methods is critical for understanding human vascular studies

in the teratoma in addition to allowing continued growth and maturation ethically without the need of an animal host[41,318].

Benchmarking with human patient-derived teratomas would be valuable, especially as many of these often can become quite mature. Additionally, expanding screens to patient-derived cell types (iPSCs) will better model human pathologies while providing a potential way to identify patient-specific disease vulnerabilities utilizing the teratoma. Another critical future study is assessing the impact of different dissociation methods on teratoma cell type proportion. The ability to achieve greater cell numbers with the most current single cell RNA sequencing protocols, such as SPLiT-seq [61] and sci-RNA-seq [62], will be vital for identifying additional cell types. A time series analysis of teratomas at multiple stages of maturity could help uncover developmental pathways that the cell types follow. Growing patient-specific teratomas could benefit disease research through isogenic iPSC lines aiding in understanding the disease state in various tissues that otherwise may be inaccessible with current technologies. Additionally, we have validated a proof-of-concept for molecular sculpting with our miRNA circuit, but different lineages may have more effective miRNAs that are also more translationally relevant. Thus, a future study conducting a miRNA circuit screen would be beneficial in assessing which miRNAs are most effective at translational lineage enrichment for downstream focused developmental biology and tissue engineering applications. Taken together, we believe the teratoma is a promising platform for modeling multi-lineage human development. From the earliest depictions of teratomas in 600 to 900 BCE on ancient tablets[43] to the initial thorough description made by Thürlbeck and Scully in 1960[47], we have taken the teratoma to greater heights in developmental biology and tissue engineering research in the 21st century.

REFERENCES

1. Vastag, L., Jorgensen, P., Peshkin, L., Wei, R., Rabinowitz, J. D. & Kirschner, M. W. Remodeling of the metabolome during early frog development. *PLoS One* **6**, (2011).

2. Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A. & Schier, A. F. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science (80-.).* **3131**, eaar3131 (2018).

3. Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-Nieto, F. J., Mulas, C., Ibarra-Soria, X., Tyser, R. C. V., Ho, D. L. L., Reik, W., Srinivas, S., Simons, B. D., Nichols, J., Marioni, J. C. & Göttgens, B. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* (2019) doi:10.1038/s41586-019-0933-9.

4. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C. & Shendure, J. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

5. Peter, I. S. & Davidson, E. H. Evolution of gene regulatory networks controlling body plan development. *Cell* **144**, 970–985 (2011).

6. Royo, J. L., Maeso, I., Irimia, M., Gao, F., Peter, I. S., Lopes, C. S., D'Aniello, S., Casares, F., Davidson, E. H., Garcia-Fernandez, J. & Gomez-Skarmeta, J. L. Transphyletic conservation of developmental regulatory state in animal evolution. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 14186–14191 (2011).

7. Lin, Y., Chen, D., Fan, Q. & Zhang, H. Characterization of SoxB2 and SoxC genes in amphioxus (Branchiostoma belcheri): implications for their evolutionary conservation. *Sci. China. Ser. C, Life Sci.* **52**, 813–822 (2009).

8. Richard, I., Abitbol, M., Wilson, D., Fougerousse, F., Beckmann, J. S., Suel, L., Durand, M., Herasse, M., Bullen, P., Robson, S., Lindsay, S. & Strachan, T. Human–mouse differences in the embryonic expression patterns of developmental control genes and disease genes. *Hum. Mol. Genet.* **9**, 165–173 (2000).

9. Richardson, M. K., Hanken, J., Gooneratne, M. L., Pieau, C., Raynaud, A., Selwood, L. & Wright, G. M. There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anat. Embryol. (Berl).* **196**, 91–106 (1997).

10. Raff, R. A. *The Shape of Life; Genes, Development and the Evolution of Animal Form.* (University of Chicago Press, 1996).

11. Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., Close, J. L., Long, B., Penn, O., Yao, Z., Eggermont, J., Hollt, T., Levi, B. P., Shehata, S. I., Aevermann, B., Beller, A., Bertagnolli, D., Brouner, K., Casper, T., Cobbs, C., Dalley, R., Dee, N., Ding, S.-L., Ellenbogen, R. G., Fong, O., Garren, E., Goldy, J., Gwinn, R. P.,

Hirschstein, D., Keene, C. D., Keshk, M., Ko, A. L., Lathia, K., Mahfouz, A., Maltzer, Z., McGraw, M., Nguyen, T. N., Nyhus, J., Ojemann, J. G., Oldre, A., Parry, S., Reynolds, S., Rimorin, C., Shapovalova, N. V, Somasundaram, S., Szafer, A., Thomsen, E. R., Tieu, M., Scheuermann, R. H., Yuste, R., Sunkin, S. M., Lelieveldt, B., Feng, D., Ng, L., Bernard, A., Hawrylycz, M., Phillips, J., Tasic, B., Zeng, H., Jones, A. R., Koch, C. & Lein, E. S. Conserved cell types with divergent features between human and mouse cortex. *Nature* 384826 (2019) doi:10.1101/384826.

12. Zhu, Y., Sousa, A. M. M., Gao, T., Skarica, M., Li, M., Santpere, G., Esteller-Cucala, P., Juan, D., Ferrández-Peral, L., Gulden, F. O., Yang, M., Miller, D. J., Marques-Bonet, T., Imamura Kawasawa, Y., Zhao, H. & Sestan, N. Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science (80-.).* **362**, eaat8077 (2018).

13. Miller, J. A., Ding, S. L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., Ebbert, A., Riley, Z. L., Royall, J. J., Aiona, K., Arnold, J. M., Bennet, C., Bertagnolli, D., Brouner, K., Butler, S., Caldejon, S., Carey, A., Cuhaciyan, C., Dalley, R. A., Dee, N., Dolbeare, T. A., Facer, B. A. C., Feng, D., Fliss, T. P., Gee, G., Goldy, J., Gourley, L., Gregor, B. W., Gu, G., Howard, R. E., Jochim, J. M., Kuan, C. L., Lau, C., Lee, C. K., Lee, F., Lemon, T. A., Lesnar, P., McMurray, B., Mastan, N., Mosqueda, N., Naluai-Cecchini, T., Ngo, N. K., Nyhus, J., Oldre, A., Olson, E., Parente, J., Parker, P. D., Parry, S. E., Stevens, A., Pletikos, M., Reding, M., Roll, K., Sandman, D., Sarreal, M., Shapouri, S., Shapovalova, N. V., Shen, E. H., Sjoquist, N., Slaughterbeck, C. R., Smith, M., Sodt, A. J., Williams, D., Zöllei, L., Fischl, B., Gerstein, M. B., Geschwind, D. H., Glass, I. A., Hawrylycz, M. J., Hevner, R. F., Huang, H., Jones, A. R., Knowles, J. A., Levitt, P., Phillips, J. W., Šestan, N., Wohnoutka, P., Dang, C., Bernard, A., Hohmann, J. G. & Lein, E. S. Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).

14. Yao, Z., Mich, J. K., Ku, S., Menon, V., Krostag, A. R., Martinez, R. A., Furchtgott, L., Mulholland, H., Bort, S., Fuqua, M. A., Gregor, B. W., Hodge, R. D., Jayabalu, A., May, R. C., Melton, S., Nelson, A. M., Ngo, N. K., Shapovalova, N. V., Shehata, S. I., Smith, M. W., Tait, L. J., Thompson, C. L., Thomsen, E. R., Ye, C., Glass, I. A., Kaykas, A., Yao, S., Phillips, J. W., Grimley, J. S., Levi, B. P., Wang, Y. & Ramanathan, S. A Single-Cell Roadmap of Lineage Bifurcation in Human ESC Models of Embryonic Brain Development. *Cell Stem Cell* **20**, 120–134 (2017).

15. Wang, J., Angarica, V. E., Bhinge, A., Jenjaroenpun, P., Del Sol, A., Kuznetsov, V. A., Nookaew, I. & Stanton, L. W. Single-cell gene expression analysis reveals regulators of distinct cell subpopulations among developing human neurons. *Genome Res.* **27**, 1783–1794 (2017).

16. Jang, S., Choubey, S., Furchtgott, L., Zou, L. N., Doyle, A., Menon, V., Loew, E. B., Krostag, A. R., Martinez, R. A., Madisen, L., Levi, B. P. & Ramanathan, S. Dynamics of embryonic stem cell differentiation inferred from single-cell transcriptomics show a series of transitions through discrete cell states. *Elife* **6**, 1–28 (2017).

17. Parekh, U., Wu, Y., Zhao, D., Worlikar, A., Shah, N., Zhang, K. & Mali, P. Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and

Single-Cell RNA-Sequencing Readout. *Cell Syst.* 1–8 (2018) doi:10.1016/j.cels.2018.10.008.

18. Tsunemoto, R., Lee, S., Szucs, A., Chubukov, P., Sokolova, I., Blanchard, J. W., Eade, K. T., Bruggemann, J., Wu, C., Torkamani, A., Sanna, P. P. & Baldwin, K. K. Diverse reprogramming codes for neuronal identity. *Nature* **557**, 375–380 (2018).

19. Liu, C., Oikonomopoulos, A., Sayed, N. & Wu, J. C. Modeling human diseases with induced pluripotent stem cells : from 2D to 3D and beyond. 1–6 (2018) doi:10.1242/dev.156166.

20. Brown, J., Quadrato, G. & Arlotta, P. *Studying the Brain in a Dish : 3D Cell Culture Models of Human Brain Development and Disease*. *Human Embryonic Stem Cells in Development* vol. 129 (Elsevier Inc., 2018).

21. Huch, M. & Koo, B.-K. Modeling mouse and human development using organoid cultures. *Development* **142**, 3113–3125 (2015).

22. Yin, X., Mead, B. E., Safaee, H., Langer, R., Karp, J. M. & Levy, O. Engineering Stem Cell Organoids. *Cell Stem Cell* **18**, 25–38 (2016).

23. Clevers, H. Review Modeling Development and Disease with Organoids. *Cell* **165**, 1586–1597 (2016).

24. Dutta, D., Heo, I. & Clevers, H. Disease Modeling in Stem Cell-Derived 3D Organoid Systems. *Trends Mol. Med.* **23**, 393–410 (2017).

25. Fligor, C. M., Langer, K. B., Sridhar, A., Ren, Y., Shields, P. K., Edler, M. C., Ohlemacher, S. K., Sluch, V. M., Zack, D. J., Zhang, C., Suter, D. M. & Meyer, J. S. Three-Dimensional Retinal Organoids Facilitate the Investigation of Retinal Ganglion Cell Development, Organization and Neurite Outgrowth from Human Pluripotent Stem Cells. *Sci. Rep.* **8**, 14520 (2018).

26. Capowski, E. E., Samimi, K., Mayerl, S. J., Phillips, M. J., Pinilla, I., Howden, S. E., Saha, J., Jansen, A. D., Edwards, K. L., Jager, L. D., Barlow, K., Valiauga, R., Erlichman, Z., Hagstrom, A., Sinha, D., Sluch, V. M., Chamling, X., Zack, D. J., Skala, M. C. & Gamm, D. M. Reproducibility and staging of 3D human retinal organoids across multiple pluripotent stem cell lines. *Development* **146**, dev171686 (2019).

27. Collin, J., Queen, R., Zerti, D., Dorgau, B., Hussain, R., Coxhead, J., Cockell, S. & Lako, M. Deconstructing Retinal Organoids: Single Cell RNA-Seq Reveals the Cellular Components of Human Pluripotent Stem Cell-Derived Retina. *Stem Cells* **37**, 593–598 (2019).

28. Bigorgne, A. E., Farin, H. F., Lemoine, R., Mahlaoui, N., Lambert, N., Gil, M., Schulz, A., Philippet, P., Schlesser, P., Abrahamsen, T. G., Oymar, K., Graham Davies, E., Ellingsen, C. L., Leteurtre, E., Moreau-Massart, B., Berrebi, D., Bole-Feysot, C., Nischke, P., Brousse, N., Fischer, A., Clevers, H. & De Saint Basile, G. TTC7A mutations disrupt

intestinal epithelial apicobasal polarity. *J. Clin. Invest.* **124**, 328–337 (2014).

29. Dekkers, J. F., Wiegerinck, C. L., De Jonge, H. R., Bronsveld, I., Janssens, H. M., De Winter-De Groot, K. M., Brandsma, A. M., De Jong, N. W. M., Bijvelds, M. J. C., Scholte, B. J., Nieuwenhuis, E. E. S., Van Den Brink, S., Clevers, H., Van Der Ent, C. K., Middendorp, S. & Beekman, J. M. A functional CFTR assay using primary cystic fibrosis intestinal organoids. *Nat. Med.* **19**, 939–945 (2013).

30. Gao, D., Vela, I., Sboner, A., Iaquinta, P. J., Karthaus, W. R., Gopalan, A., Dowling, C., Wanjala, J. N., Undvall, E. A., Arora, V. K., Wongvipat, J., Kossai, M., Ramazanoglu, S., Barboza, L. P., Di, W., Cao, Z., Zhang, Q. F., Sirota, I., Ran, L., MacDonald, T. Y., Beltran, H., Mosquera, J.-M., Touijer, K. A., Scardino, P. T., Laudone, V. P., Curtis, K. R., Rathkopf, D. E., Morris, M. J., Danila, D. C., Slovin, S. F., Solomon, S. B., Eastham, J. A., Chi, P., Carver, B., Rubin, M. A., Scher, H. I., Clevers, H., Sawyers, C. L. & Chen, Y. Organoid Cultures Derived from Patients with Advanced Prostate Cancer. *Cell* **159**, 176–187 (2014).

31. Bartfeld, S., Bayram, T., van de Wetering, M., Huch, M., Begthel, H., Kujala, P., Vries, R., Peters, P. J. & Clevers, H. In Vitro Expansion of Human Gastric Epithelial Stem Cells and Their Responses to Bacterial Infection. *Gastroenterology* **148**, 126-136.e6 (2015).

32. Boj, S. F., Hwang, C.-I., Baker, L. A., Chio, I. I. C., Engle, D. D., Corbo, V., Jager, M., Ponz-Sarvise, M., Tiriac, H., Spector, M. S., Gracanin, A., Oni, T., Yu, K. H., van Boxtel, R., Huch, M., Rivera, K. D., Wilson, J. P., Feigin, M. E., Öhlund, D., Handly-Santana, A., Ardito-Abraham, C. M., Ludwig, M., Elyada, E., Alagesan, B., Biffi, G., Yordanov, G. N., Delcuze, B., Creighton, B., Wright, K., Park, Y., Morsink, F. H. M., Molenaar, I. Q., Borel Rinkes, I. H., Cuppen, E., Hao, Y., Jin, Y., Nijman, I. J., Iacobuzio-Donahue, C., Leach, S. D., Pappin, D. J., Hammell, M., Klimstra, D. S., Basturk, O., Hruban, R. H., Offerhaus, G. J., Vries, R. G. J., Clevers, H. & Tuveson, D. A. Organoid Models of Human and Mouse Ductal Pancreatic Cancer. *Cell* **160**, 324–338 (2015).

33. van de Wetering, M., Francies, H. E., Francis, J. M., Bounova, G., Iorio, F., Pronk, A., van Houdt, W., van Gorp, J., Taylor-Weiner, A., Kester, L., McLaren-Douglas, A., Blokker, J., Jaksani, S., Bartfeld, S., Volckman, R., van Sluis, P., Li, V. S. W., Seepo, S., Sekhar Pedamallu, C., Cibulskis, K., Carter, S. L., McKenna, A., Lawrence, M. S., Lichtenstein, L., Stewart, C., Koster, J., Versteeg, R., van Oudenaarden, A., Saez-Rodriguez, J., Vries, R. G. J., Getz, G., Wessels, L., Stratton, M. R., McDermott, U., Meyerson, M., Garnett, M. J. & Clevers, H. Prospective Derivation of a Living Organoid Biobank of Colorectal Cancer Patients. *Cell* **161**, 933–945 (2015).

34. Aurora, M. & Spence, J. R. hPSC-derived lung and intestinal organoids as models of human fetal tissue. *Dev. Biol.* **420**, 230–238 (2016).

35. Chambers, S. M., Tchieu, J. & Studer, L. Build-a-brain. *Cell Stem Cell* **13**, 377–378 (2013).

36. Jabaudon, D. & Lancaster, M. Exploring landscapes of brain morphogenesis with organoids. 2016–2019 (2018) doi:10.1242/dev.172049.

37. Sato, T., Vries, R. G., Snippert, H. J., Van De Wetering, M., Barker, N., Stange, D. E., Van Es, J. H., Abo, A., Kujala, P., Peters, P. J. & Clevers, H. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature* **459**, 262–265 (2009).

38. Sato, T., Stange, D. E., Ferrante, M., Vries, R. G. J., Van Es, J. H., Van Den Brink, S., Van Houdt, W. J., Pronk, A., Van Gorp, J., Siersema, P. D. & Clevers, H. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* **141**, 1762–1772 (2011).

39. Jung, P., Sato, T., Merlos-Suárez, A., Barriga, F. M., Iglesias, M., Rossell, D., Auer, H., Gallardo, M., Blasco, M. A., Sancho, E., Clevers, H. & Batlle, E. Isolation and in vitro expansion of human colonic stem cells. *Nat. Med.* **17**, 1225–1227 (2011).

40. Moris, N., Anlas, K., Brink, S. Van Den & Alemany, A. An in vitro model for anteroposterior organisation during human development. *Nature* (2020) doi:10.1038/s41586-020-2383-9.

41. Aguilera-Castrejon, A., Oldak, B., Shani, T., Ghanem, N., Itzkovich, C., Slomovich, S., Tarazi, S., Bayerl, J., Chugaeva, V., Ayyash, M., Ashouokhi, S., Sheban, D., Livnat, N., Lasman, L., Viukov, S., Zerbib, M., Addadi, Y., Rais, Y., Cheng, S., Stelzer, Y., Keren-Shaul, H., Shlomo, R., Massarwa, R., Novershtern, N., Maza, I. & Hanna, J. H. Ex utero mouse embryogenesis from pre-gastrulation to late organogenesis. *Nature* **593**, (2021).

42. Lensch, M. W., Schlaeger, T. M., Zon, L. I. & Daley, G. Q. Teratoma Formation Assays with Human Embryonic Stem Cells: A Rationale for One Type of Human-Animal Chimera. *Cell Stem Cell* **1**, 253–258 (2007).

43. Wheeler, J. E. History of Teratomas. in *The Human Teratomas: Experimental and Clinical Biology* (eds. Damjanov, I., Knowles, B. B. & Solter, D.) 1–22 (Humana Press, 1983). doi:10.1007/978-1-4612-5628-1_1.

44. Stevens, L. THE BIOLOGY OF TERATOMAS. *Adv Morphog* **6**, 1–31 (1967).

45. Stevens, L. C. & Pierce., G. B. Teratomas: Definitions and Terminology. *Teratomas Differ.* 13–14 (1975).

46. Stevens, L. The biology of teratomas including evidence indicating their origin form primordial germ cells. *Annee Biol.* **1**, 585–610 (1962).

47. Thurlbeck, William M., R. E. S. Solid Teratoma of the Ovary: A Clinicopatlzological Analysis of 9 Cases. 2563–2571 (1973).

48. Nikolic, A., Volarevic, V., Armstrong, L., Lako, M. & Stojkovic, M. Primordial Germ Cells : Current Knowledge and Perspectives. **2016**, (2016).

49. Saffman, E. E. & Lasko, P. Germline development in vertebrates and invertebrates. *Cell. Mol. Life Sci.* **55**, 1141–1163 (1999).

50. Cibas, E. S. Cytology (Third Edition) Diagnostic Principles and Clinical Correlated: Chapter 15 - Ovary. 433–450 (2009).

51. Willis, R. A. The Structure of Teratoma. *J. Pathol. Bacteriol.* **XL**, (1934).

52. Willis, R. A. THE HISTOGENESIS OF NEURAL TISSUE IN TERATOMAS . (PLATES. *J. Pathol. Bacteriol.* (1935).

53. Bocker, W. WHO classification of breast tumors and tumors of the female genital organs: pathology and genetics. *Verh. Dtsch. Ges. Pathol.* **86**, 116–119 (2002).

54. Smith, K. P., Luong, M. X. & Stein, G. S. Pluripotency: Toward a gold standard for human ES and iPS cells. *J. Cell. Physiol.* **220**, 21–29 (2009).

55. Avior, Y., Biancotti, J. C. & Benvenisty, N. TeratoScore: Assessing the Differentiation Potential of Human Pluripotent Stem Cells by Quantitative Expression Analysis of Teratomas. *Stem Cell Reports* **4**, 967–974 (2015).

56. Chan, S. S. K., Arpke, R. W., Filareto, A., Xie, N., Pappas, M. P., Penaloza, J. S., Perlingeiro, R. C. R. & Kyba, M. Skeletal Muscle Stem Cells from PSC-Derived Teratomas Have Functional Regenerative Capacity. *Cell Stem Cell* **23**, 74-85.e6 (2018).

57. Suzuki, N., Yamazaki, S., Yamaguchi, T., Okabe, M., Masaki, H., Takaki, S., Otsu, M. & Nakauchi, H. Generation of Engraftable Hematopoietic Stem Cells From Induced Pluripotent Stem Cells by Way of Teratoma Formation. *Mol. Ther.* **21**, 1424–1431 (2013).

58. Tsukada, M., Ota, Y., Wilkinson, A. C., Becker, H. J., Osato, M., Nakauchi, H. & Yamazaki, S. In Vivo Generation of Engraftable Murine Hematopoietic Stem Cells by Gfi1b, c-Fos, and Gata2 Overexpression within Teratoma. *Stem Cell Reports* **9**, 1024–1033 (2017).

59. Philipp, F., Selich, A., Rothe, M., Hoffmann, D., Rittinghausen, S., Morgan, M. A., Klatt, D., Glage, S., Lienenklaus, S., Neuhaus, V., Sewald, K., Braun, A. & Schambach, A. Human Teratoma-Derived Hematopoiesis Is a Highly Polyclonal Process Supported by Human Umbilical Vein Endothelial Cells. *Stem Cell Reports* **11**, 1051–1060 (2018).

60. Amabile, G., Welner, R. S., Nombela-arrieta, C., Alise, A. M. D., Ruscio, A. Di, Ebralidze, A. K., Kraytsberg, Y., Ye, M., Kocher, O., Neuberg, D. S., Khrapko, K., Silberstein, L. E. & Tenen, D. G. In vivo generation of transplantable human hematopoietic cells from induced pluripotent stem cells. **121**, 1–3 (2019).

61. Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Gray, L., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B. & Seelig, G. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science (80-.).* **12**, eaam8999 (2018).

62. Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C. & Shendure, J.

Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).

63. 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., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. & Bielas, J. H. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 1–12 (2017).

64. Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Chen, W., Peeler, D. J., Yao, Z., Tasic, B., Sellers, D. L., Pun, H. & Seelig, G. Scaling single cell transcriptomics through split pool barcoding. *Bioarxiv* (2017) doi:10.1101/105163.

65. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A. & McCarroll, S. A. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

66. Klein, A. M., Mazutis, L., Weitz, D. A., Kirschner, M. W., Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V. & Peshkin, L. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells Resource Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187–1201 (2015).

67. Ding, J., Adiconis, X., Simmons, S. K., Kowalczyk, M. S., Hession, C. C., Marjanovic, N. D., Hughes, T. K., Wadsworth, M. H., Burks, T., Nguyen, L. T., Kwon, J. Y. H., Barak, B., Ge, W., Kedaigle, A. J., Carroll, S., Li, S., Hacohen, N., Rozenblatt-Rosen, O., Shalek, A. K., Villani, A.-C., Regev, A. & Levin, J. Z. Systematic comparative analysis of single cell RNA-sequencing methods. *bioRxiv* 632216 (2019) doi:10.1101/632216.

68. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M. 3rd, Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

69. Wu, Y., Tamayo, P. & Zhang, K. Visualizing and interpreting single-cell gene expression datasets with Similarity Weighted Nonnegative Embedding. *Cell Syst.* **7**, 656-666.e4 (2018).

70. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

71. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4096.

72. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

73. Abdi, H. & Williams, L. J. Principal component analysis. *Chemom. Intell. Lab. Syst.* **2**, 433–459 (2010).

74. Houle, M. E., Kriegel, H.-P., Kröger, P., Schubert, E. & Zimek, A. Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? in *Scientific and Statistical Database Management* (eds. Gertz, M. & Ludäscher, B.) 482–500 (Springer Berlin Heidelberg, 2010).

75. Tarlow, D., Swersky, K., Charlin, L., Sutskever, I. & Zemel, R. Stochastic k-Neighborhood Selection for Supervised and Unsupervised Learning. *Proc. 30th Int. Conf. Mach. Learn.* **28**, 199–207 (2013).

76. Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., Kim, H., Cho, A., Kim, E., Lee, T., Kim, H., Kim, K., Yang, S., Bae, D., Yun, A., Kim, S., Kim, C. Y., Cho, H. J., Kang, B., Shin, S. & Lee, I. TRRUST: a reference database of human transcriptional regulatory interactions. *Sci. Rep.* **5**, 11432 (2015).

77. Wilcoxon, F. Individual Comparisons of Grouped Data by Ranking Methods. *J. Econ. Entomol.* **39**, 269–270 (1946).

78. McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* 1–18 (2018).

79. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F. & Newell, E. W. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, (2018).

80. Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K. & Kharchenko, P. V. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).

81. Wu, Y., Tamayo, P. & Zhang, K. Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding. *Cell Syst.* **7**, 656-666.e4 (2018).

82. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

83. Sammon, J. W. A Nonlinear Mapping for Data Structure Analysis. *IEEE Trans. Comput.* **C**–**18**, 401–409 (1969).

84. Polioudakis, D., de la Torre-Ubieta, L., Langerman, J., Elkins, A. G., Shi, X., Stein, J. L., Vuong, C. K., Nichterwitz, S., Gevorgian, M., Opland, C. K., Lu, D., Connell, W., Ruzzo, E. K., Lowe, J. K., Hadzic, T., Hinz, F. I., Sabri, S., Lowry, W. E., Gerstein, M. B., Plath, K. & Geschwind, D. H. A Single-Cell Transcriptomic Atlas of Human Neocortical

Development during Mid-gestation. *Neuron* 1–17 (2019) doi:10.1016/j.neuron.2019.06.011.

85. Bansod, S., Kageyama, R. & Ohtsuka, T. Hes5 regulates the transition timing of neurogenesis and gliogenesis in mammalian neocortical development. *Dev.* **144**, 3156–3167 (2017).

86. Khalaf-Nazzal, R., Stouffer, M. A., Olaso, R., Muresan, L., Roumegous, A., Lavilla, V., Carpentier, W., Moutkine, I., Dumont, S., Albaud, B., Cagnard, N., Roest Crollius, H. & Francis, F. Early born neurons are abnormally positioned in the doublecortin knockout hippocampus. *Hum. Mol. Genet.* **26**, 90–108 (2017).

87. Gao, Z., Ure, K., Ables, J. L., Lagace, D. C., Nave, K.-A., Goebbels, S., Eisch, A. J. & Hsieh, J. Neurod1 is essential for the survival and maturation of adult-born neurons. *Nat. Neurosci.* **12**, 1090–2 (2009).

88. Gao, S., Yan, L., Wang, R., Li, J., Yong, J., Zhou, X., Wei, Y., Wu, X., Wang, X., Fan, X., Yan, J., Zhi, X., Gao, Y., Guo, H., Jin, X., Wang, W., Mao, Y., Wang, F., Wen, L., Fu, W., Ge, H., Qiao, J. & Tang, F. Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing. *Nat. Cell Biol.* **20**, 721–734 (2018).

89. Bort, R., Martinez-Barbera, J. P., Beddington, R. S. P. & Zaret, K. S. Hex homeobox gene-dependent tissue positioning is required for organogenesis of the ventral pancreas. *Development* **131**, 797–806 (2004).

90. Que, J., Okubo, T., Goldenring, J. R., Nam, K. T., Kurotani, R., Morrisey, E. E., Taranova, O., Pevny, L. H. & Hogan, B. L. M. Multiple dose-dependent roles for Sox2 in the patterning and differentiation of anterior foregut endoderm. *Development* **134**, 2521–2531 (2007).

91. Green, M. D., Chen, A., Nostro, M. C., D'Souza, S. L., Schaniel, C., Lemischka, I. R., Gouon-Evans, V., Keller, G. & Snoeck, H. W. Generation of anterior foregut endoderm from human embryonic and induced pluripotent stem cells. *Nat. Biotechnol.* **29**, 267–273 (2011).

92. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W., Stoeckius, M., Smibert, P. & Satija, R. Comprehensive integration of single cell data. *bioRxiv* 1–34 (2018) doi:10.1101/460147.

93. Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G.-C., Chen, M. & Guo, G. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e17 (2018).

94. An, Z., Sabalic, M., Bloomquist, R. F., Fowler, T. E., Streelman, T. & Sharpe, P. T. A quiescent cell population replenishes mesenchymal stem cells to drive accelerated growth

in mouse incisors. *Nat. Commun.* **9**, (2018).

95. Zovein, A. C., Hofmann, J. J., Lynch, M., French, W. J., Turlo, K. A., Yang, Y., Becker, M. S., Zanetta, L., Dejana, E., Gasson, J. C., Tallquist, M. D. & Iruela-Arispe, M. L. Fate tracing reveals the endothelial origin of hematopoietic stem cells. *Cell Stem Cell* **3**, 625–636 (2008).

96. Cathery, W., Faulkner, A., Maselli, D. & Madeddu, P. Concise Review: The Regenerative Journey of Pericytes Toward Clinical Translation. *Stem Cells* **36**, 1295–1310 (2018).

97. Wolburg, H., Wolburg-Buchholz, K., Mack, A. F. & Reichenbach, A. Ependymal cells. *Encycl. Neurosci.* 1133–1140 (2009) doi:10.1016/B978-008045046-9.01001-9.

98. Goding, C. R. Mitf from neural crest to melanoma: signal transduction and transcription in the melanocyte lineage. *Genes Dev.* **14**, 1712–1728 (2000).

99. Mort, R. L., Jackson, I. J., Patton, E. E., Mort, R. L., Jackson, I. J. & Patton, E. E. The melanocyte lineage in development and disease. *Dev.* **142**, 620–632 (2015).

100. Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G. & Klein, A. M. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo Daniel. *Science (80-.).* **25**, 289–313 (2018).

101. Kim, J. W., Botvinnik, O. B., Abudayyeh, O., Birger, C., Rosenbluh, J., Shrestha, Y., Abazeed, M. E., Hammerman, P. S., DiCara, D., Konieczkowski, D. J., Johannessen, C. M., Liberzon, A., Alizad-Rahvar, A. R., Alexe, G., Aguirre, A., Ghandi, M., Greulich, H., Vazquez, F., Weir, B. A., Van Allen, E. M., Tsherniak, A., Shao, D. D., Zack, T. I., Noble, M., Getz, G., Beroukhim, R., Garraway, L. A., Ardakani, M., Romualdi, C., Sales, G., Barbie, D. A., Boehm, J. S., Hahn, W. C., Mesirov, J. P. & Tamayo, P. Characterizing genomic alterations in cancer by complementary functional associations. *Nat. Biotechnol.* **34**, 3–5 (2016).

102. Velasco, S., Kedaigle, A. J., Simmons, S. K., Nash, A., Rocha, M., Quadrato, G., Paulsen, B., Nguyen, L., Adiconis, X., Regev, A., Levin, J. Z. & Arlotta, P. Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature* **570**, 523–527 (2019).

103. Guo, C., Biddy, B. A., Kamimoto, K., Kong, W. & Morris, S. A. CellTag Indexing: a genetic barcode-based multiplexing tool for single-cell technologies. *bioRxiv* 335547 (2018) doi:10.1101/335547.

104. Quadrato, G., Nguyen, T., Macosko, E. Z., Sherwood, J. L., Min Yang, S., Berger, D. R., Maria, N., Scholvin, J., Goldman, M., Kinney, J. P., Boyden, E. S., Lichtman, J. W., Williams, Z. M., McCarroll, S. A. & Arlotta, P. Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* **545**, 48 (2017).

105. de Souza, N. Organoid variability examined. *Nat. Methods* **14**, 655 (2017).

106. Ortmann, D. & Vallier, L. Variability of human pluripotent stem cell lines. *Curr. Opin. Genet. Dev.* **46**, 179–185 (2017).

107. Zhong, S., Zhang, S., Fan, X., Wu, Q., Yan, L., Dong, J., Zhang, H., Li, L., Sun, L., Pan, N., Xu, X., Tang, F., Zhang, J., Qiao, J. & Wang, X. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* (2018) doi:10.1038/nature25980.

108. Phipson, B., Er, P. X., Combes, A. N., Forbes, T. A., Howden, S. E., Zappia, L., Yen, H.-J., Lawlor, K. T., Hale, L. J., Sun, J., Wolvetang, E., Takasato, M., Oshlack, A. & Little, M. H. Evaluation of variability in human kidney organoids. *Nat. Methods* **16**, 79–87 (2019).

109. Denisenko, E., Guo, B. B., Jones, M., Hou, R., de Kock, L., Lassmann, T., Poppe, D., Clement, O., Simmons, R. K., Lister, R. & Forrest, A. R. R. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *bioRxiv* 832444 (2019) doi:10.1101/832444.

110. Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E. & Church, G. M. RNA-Guided Human Genome Engineering via Cas9. *Science (80-.).* **339**, 823–826 (2013).

111. Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Hsu, P. D., Wu, X., Jiang, W. & Marraffini, L. A. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (80-.).* **339**, 819–823 (2013).

112. Sanjana, N. E. Genome-scale CRISPR pooled screens. *Anal. Biochem.* **532**, 95–99 (2017).

113. Barrangou, R. & Gersbach, C. A. Expanding the CRISPR Toolbox: Targeting RNA with Cas13b. *Mol. Cell* **65**, 582–584 (2017).

114. Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr. Opin. Microbiol.* **37**, 67–78 (2017).

115. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).

116. Heyer, W.-D., Ehmsen, K. T. & Liu, J. Regulation of Homologous Recombination in Eukaryotes. *Annu. Rev. Genet.* **44**, 113–139 (2010).

117. Montalbano, A., Canver, M. C. & Sanjana, N. E. High-Throughput Approaches to Pinpoint Function within the Noncoding Genome. *Mol. Cell* **68**, 44–59 (2017).

118. Doench, J. G. Am I ready for CRISPR? A user's guide to genetic screens. *Nat. Rev. Genet.* (2017) doi:10.1038/nrg.2017.97.

119. Dominguez, A. A., Lim, W. A. & Qi, L. S. Beyond editing: Repurposing CRISPR-Cas9 for precision genome regulation and interrogation. *Nat. Rev. Mol. Cell Biol.* **17**, 5–15

(2016).

120. Shalem, O., Sanjana, N. E., Zhang, F. & Sciences, C. High-throughput functional genomics using CRISPR-Cas9. **16**, 299–311 (2015).

121. Wang, T. Genetic Screens in Human Cells Using. **80**, 80–85 (2014).

122. Polstein, L. R., Gersbach, C. A., Carolina, N., States, U., Biology, C., Carolina, N. & Carolina, N. HHS Public Access. **11**, 198–200 (2015).

123. Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., Nureki, O. & Zhang, F. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).

124. Kampmann, M. CRISPRi and CRISPRa screens in mammalian cells for precision biology and medicine. *ACS Chem. Biol.* acschembio.7b00657 (2017) doi:10.1021/acschembio.7b00657.

125. Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., Qi, L. S., Kampmann, M. & Weissman, J. S. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647–661 (2014).

126. Hess, G. T., Tycko, J., Yao, D. & Bassik, M. C. Methods and Applications of CRISPR-Mediated Base Editing in Eukaryotic Genomes. *Mol. Cell* **68**, 26–43 (2017).

127. Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I. & Liu, D. R. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).

128. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

129. Hess, G. T., Frésard, L., Han, K., Lee, C. H., Li, A., Cimprich, K. A., Montgomery, S. B. & Bassik, M. C. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods* **13**, 1036–1042 (2016).

130. Ma, Y., Zhang, J., Yin, W., Zhang, Z., Song, Y. & Chang, X. Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. *Nat. Methods* **13**, 1029–1035 (2016).

131. Kuscu, C., Parlak, M., Tufan, T., Yang, J., Szlachta, K., Wei, X., Mammadov, R. & Adli, M. CRISPR-STOP: Gene silencing through base-editing-induced nonsense mutations. *Nat. Methods* **14**, 710–712 (2017).

132. Vojta, A., Dobrinic, P., Tadic, V., Bockor, L., Korac, P., Julg, B., Klasic, M. & Zoldos, V.

Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Res.* **44**, 5615–5628 (2016).

133. Liu, X. S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R. A. & Jaenisch, R. Editing DNA Methylation in the Mammalian Genome. *Cell* **167**, 233-247.e17 (2016).

134. Laufer, B. I. & Singh, S. M. Strategies for precision modulation of gene expression by epigenome editing: An overview. *Epigenetics and Chromatin* **8**, 1–12 (2015).

135. Hilton, I. B., D'Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E. & Gersbach, C. A. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33**, 510–517 (2015).

136. Kearns, N. A., Pham, H., Tabak, B., Genga, R. M., Silverstein, N. J., Garber, M. & Maehr, R. Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat. Methods* **12**, 401–403 (2015).

137. Kwon, D. Y., Zhao, Y. T., Lamonica, J. M. & Zhou, Z. Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC. *Nat. Commun.* **8**, 1–8 (2017).

138. Oren, M. & Rotter, V. Mutant p53 gain-of-function in cancer. *Cold Spring Harb. Perspect. Biol.* **2**, 1–15 (2010).

139. Albert, P. R., Le François, B. & Millar, A. M. Transcriptional dysregulation of 5-HT1A autoreceptors in mental illness. *Mol. Brain* **4**, 1–14 (2011).

140. Gonda, T. J. & Ramsay, R. G. Directly targeting transcriptional dysregulation in cancer. *Nat. Rev. Cancer* **15**, 686–694 (2015).

141. Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., Lim, W. A., Weissman, J. S. & Qi, L. S. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).

142. Polstein, L. R., Perez-pinera, P., Kocak, D. D., Vockley, M., Bledsoe, P., Song, L., Safi, A., Crawford, G. E., Reddy, T. E., Gersbach, C. a, Carolina, N. & Surgery, O. Genome-wide specificity of DNA binding , gene regulation , and chromatin remodeling by TALE- and CRISPR / Cas9-based transcriptional activators. *Genome Res.* **25**, 1158–1169 (2015).

143. Boettcher, M., Tian, R., Blau, J. A., Markegard, E., Wagner, R. T., Wu, D., Mo, X., Biton, A., Zaitlen, N., Fu, H., McCormick, F., Kampmann, M. & McManus, M. T. Dual gene activation and knockout screen reveals directional dependencies in genetic networks. *Nat. Biotechnol.* **36**, 170–178 (2018).

144. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).

145. Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A. & Mundlos, S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

146. Nowak, C. M., Lawson, S., Zerez, M. & Bleris, L. Guide RNA engineering for versatile Cas9 functionality. *Nucleic Acids Res.* **44**, 9555–9564 (2016).

147. Zalatan, J. G., Lee, M. E., Almeida, R., Gilbert, L. A., Whitehead, E. H., La Russa, M., Tsai, J. C., Weissman, J. S., Dueber, J. E., Qi, L. S. & Lim, W. A. Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell* **160**, 339–350 (2015).

148. Ma, H., Tu, L. C., Naseri, A., Huisman, M., Zhang, S., Grunwald, D. & Pederson, T. Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. *Nat. Biotechnol.* **34**, 528–530 (2016).

149. Agrotis, A. & Ketteler, R. A new age in functional genomics using CRISPR/Cas9 in arrayed library screening. *Front. Genet.* **6**, 1–15 (2015).

150. Boutros, M., Heigwer, F. & Laufer, C. Microscopy-Based High-Content Screening. *Cell* **163**, 1314–1325 (2015).

151. Neumann, B., Held, M., Liebel, U., Erfle, H., Rogers, P., Pepperkok, R. & Ellenberg, J. High-throughput RNAi screening by time-lapse imaging of live human cells. *Nat. Methods* **3**, 385–390 (2006).

152. Moffat, J., Grueneberg, D. A., Yang, X., Kim, S. Y., Kloepfer, A. M., Hinkle, G., Piqani, B., Eisenhaure, T. M., Luo, B., Grenier, J. K., Carpenter, A. E., Foo, S. Y., Stewart, S. A., Stockwell, B. R., Hacohen, N., Hahn, W. C., Lander, E. S., Sabatini, D. M. & Root, D. E. A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen. *Cell* **124**, 1283–1298 (2006).

153. Schneeberger, K. Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nat. Rev. Genet.* **15**, 662–676 (2014).

154. Canver, M. C., Haeussler, M., Bauer, D. E., Orkin, S. H., Sanjana, N. E., Shalem, O., Yuan, G.-C., Zhang, F., Concordet, J.-P. & Pinello, L. Integrated design, execution, and analysis of arrayed and pooled CRISPR genome editing experiments. *Doi.Org* 125245 (2017) doi:10.1101/125245.

155. Joung, J., Konermann, S., Gootenberg, J. S., Abudayyeh, O. O., Platt, R. J., Brigham, M. D., Sanjana, N. E. & Zhang, F. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nat. Protoc.* **12**, 828–863 (2017).

156. LeProust, E. M., Peck, B. J., Spirin, K., McCuen, H. B., Moore, B., Namsaraev, E. & Caruthers, M. H. Synthesis of high-quality libraries of long (150mer) oligonucleotides by

a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522–2540 (2010).

157. Kosuri, S., Eroshenko, N., Leproust, E., Super, M., Way, J., Li, J. B. & Church, G. M. A Scalable Gene Synthesis Platform Using High-Fidelity DNA Microchips. *Nat Biotechnol* **28**, 1295–1299 (2011).

158. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: Technologies and applications. *Nat. Methods* **11**, 499–507 (2014).

159. Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G. & Zhang, F. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science (80-.).* **343**, 84–88 (2014).

160. Pattanayak, V., Lin, S., Guilinger, J. P., Ma, E., Doudna, J. A. & Liu, D. R. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).

161. Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F. P., Rissland, O. S., Durocher, D., Angers, S. & Moffat, J. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515–1526 (2015).

162. Doerflinger, M., Forsyth, W., Ebert, G., Pellegrini, M. & Herold, M. J. CRISPR/Cas9—The ultimate weapon to battle infectious diseases? *Cell. Microbiol.* **19**, 1–10 (2017).

163. Puschnik, A. S., Majzoub, K., Ooi, Y. S. & Carette, J. E. A CRISPR toolbox to study virus-host interactions. *Nat. Rev. Microbiol.* **15**, 351–364 (2017).

164. Park, R. J., Wang, T., Koundakjian, D., Hultquist, J. F., Monel, B., Schumann, K., Yu, H., Kevin, M., Garcia-beltran, W., Piechocka-trocha, A., Krogan, N. J., Marson, A., Sabatini, D. M., Lander, E. S., Walker, B. D., Sciences, H., Francisco, S., Institutes, J. D. G., Francisco, S., Hospital, M. G. & Medical, H. A genome-wide CRISPR screen identifies a restricted set of HIV host dependency factors. **49**, 193–203 (2017).

165. Egan, E. S. Beyond Hemoglobin: Screening for Malaria Host Factors. *Trends Genet.* **34**, 133–141 (2017).

166. Singh, A. K., Carette, X., Potluri, L.-P., Sharp, J. D., Xu, R., Prisic, S. & Husson, R. N. Investigating essential gene function in *Mycobacterium tuberculosis* using an efficient CRISPR interference system. *Nucleic Acids Res.* **44**, e143–e143 (2016).

167. Kleinstiver, B. P., Prew, M. S., Tsai, S. Q., Topkar, V. V., Nguyen, N. T., Zheng, Z., Gonzales, A. P. W., Li, Z., Peterson, R. T., Yeh, J. R. J., Aryee, M. J. & Joung, J. K. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).

168. Chuai, G. hui, Wang, Q. L. & Liu, Q. In Silico Meets In Vivo: Towards Computational

CRISPR-Based sgRNA Design. *Trends Biotechnol.* **35**, 12–21 (2017).

169. Wu, N., Matand, K., Kebede, B., Acquaah, G. & Williams, S. Enhancing DNA electrotransformation efficiency in Escherichia coli DH10B electrocompetent cells. *Electron. J. Biotechnol.* **13**, 1–9 (2010).

170. Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J. & Root, D. E. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

171. McDade, J. R., Waxmonsky, N. C., Swanson, L. E. & Fan, M. Practical considerations for using pooled lentiviral CRISPR libraries. *Curr. Protoc. Mol. Biol.* **2016**, 31.5.1-31.5.13 (2016).

172. Koike-Yusa, H., Li, Y., Tan, E. P., Velasco-Herrera, M. D. C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* **32**, 267–273 (2014).

173. Wang, T., Lander, E. S. & Sabatini, D. M. Viral packaging and cell culture for CRISPR-based screens. *Cold Spring Harb. Protoc.* **2016**, 289–296 (2016).

174. Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y. & Wei, W. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* **509**, 487–491 (2014).

175. Song, M. The CRISPR/Cas9 system: Their delivery, in vivo and ex vivo applications and clinical development by startups. *Biotechnol. Prog.* **33**, 1035–1045 (2017).

176. Escors, D. & Breckpot, K. Lentiviral vectors in gene therapy: Their current status and future potential. *Arch. Immunol. Ther. Exp. (Warsz).* **58**, 107–119 (2011).

177. Chow, R. D., Guzman, C. D., Wang, G., Schmidt, F., Youngblood, M. W., Ye, L., Errami, Y., Dong, M. B., Martinez, M. A., Zhang, S., Renauer, P., Bilguvar, K., Gunel, M., Sharp, P. A., Zhang, F., Platt, R. J. & Chen, S. AAV-mediated direct in vivo CRISPR screen identifies functional suppressors in glioblastoma. *Nat. Neurosci.* **20**, 1329–1341 (2017).

178. Grieger, J. C. & Samulski, R. J. Packaging Capacity of Adeno-Associated Virus Serotypes: Impact of Larger Genomes on Infectivity and Postentry Steps. *J. Virol.* **79**, 9933–9944 (2005).

179. Liu, C., Zhang, L., Liu, H. & Cheng, K. Delivery strategies of the CRISPR-Cas9 gene-editing system for therapeutic applications. *J. Control. Release* **266**, 17–26 (2017).

180. Vargas, J. E., Chicaybam, L., Stein, R. T., Tanuri, A., Delgado-Cañedo, A. & Bonamino, M. H. Retroviral vectors and transposons for stable gene therapy: Advances, current challenges and perspectives. *J. Transl. Med.* **14**, 1–15 (2016).

181. Xu, C., Qi, X., Du, X., Zou, H., Gao, F., Feng, T., Lu, H., Li, S., An, X., Zhang, L., Wu, Y., Liu, Y., Li, N., Capecchi, M. R. & Wu, S. *piggyBac* mediates efficient in vivo CRISPR library screening for tumorigenesis in mice. *Proc. Natl. Acad. Sci.* **114**, 722–727 (2017).

182. Wang, P., Zhang, L., Xie, Y., Wang, N., Tang, R., Zheng, W. & Jiang, X. Genome Editing for Cancer Therapy: Delivery of Cas9 Protein/sgRNA Plasmid via a Gold Nanocluster/Lipid Core–Shell Nanocarrier. *Adv. Sci.* **4**, (2017).

183. Mout, R., Ray, M., Yesilbag Tonga, G., Lee, Y. W., Tay, T., Sasaki, K. & Rotello, V. M. Direct Cytosolic Delivery of CRISPR/Cas9-Ribonucleoprotein for Efficient Gene Editing. *ACS Nano* **11**, 2452–2458 (2017).

184. Manguso, R. T., Pope, H. W., Zimmer, M. D., Brown, F. D., Yates, K. B., Miller, B. C., Collins, N. B., Bi, K., La Fleur, M. W., Juneja, V. R., Weiss, S. A., Lo, J., Fisher, D. E., Miao, D., Van Allen, E., Root, D. E., Sharpe, A. H., Doench, J. G. & Haining, W. N. In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. *Nature* **547**, 413–418 (2017).

185. Naldini, L., Blömer, U., Gallay, P., Ory, D., Mulligan, R., Gage, F. H., Verma, I. M., Trono, D., Naldini, L., Blomer, U., Gallay, P., Ory, D., Mulligan, R., Gage, F. H., Verma, I. M. & Trono, D. In Vivo Gene Delivery and Stable Transduction of Nondividing Cells by a Lentiviral Vector. *Science (80-.).* **272**, 263–267 (1996).

186. Shen, J. P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A. N., Sanchez, K. S., Thomas, A., Kuo, C. C., Du, D., Roguev, A., Lewis, N. E., Chang, A. N., Kreisberg, J. F., Krogan, N., Qi, L., Ideker, T. & Mali, P. Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. *Nat. Methods* **14**, 573–576 (2017).

187. Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A. & Weissman, J. S. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882.e21 (2016).

188. Parnas, O., Jovanovic, M., Eisenhaure, T. M., Herbst, R. H., Dixit, A., Ye, C. J., Przybylski, D., Platt, R. J., Tirosh, I., Sanjana, N. E., Shalem, O., Satija, R., Raychowdhury, R., Mertins, P., Carr, S. A., Zhang, F., Hacohen, N. & Regev, A. A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* **162**, 675–686 (2015).

189. Wu, J., Platero Luengo, A., Gil, M. A., Suzuki, K., Cuello, C., Morales Valencia, M., Parrilla, I., Martinez, C. A., Nohalez, A., Roca, J., Martinez, E. A. & Izpisua Belmonte, J. C. Generation of human organs in pigs via interspecies blastocyst complementation. *Reprod. Domest. Anim.* **51**, 18–24 (2016).

190. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D.,

Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N. & Regev, A. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).

191. Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T. M., Tanay, A., van Oudenaarden, A. & Amit, I. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e15 (2016).

192. Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D. & Bock, C. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).

193. Hardcastle, T. & Kelly, K. Empirical Bayesian methods for differential expression in count data. (2009).

194. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).

195. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

196. Di, Y., Schafer, D. W., Cumbie, J. S. & Chang, J. H. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.* **10**, (2011).

197. Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M. & Liu, X. S. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).

198. Winter, J., Breinig, M., Heigwer, F., Brügemann, D., Leible, S., Pelz, O., Zhan, T. & Boutros, M. CaRpools: An R package for exploratory data analysis and documentation of pooled CRISPR/Cas9 screens. *Bioinformatics* **32**, 632–634 (2015).

199. Jeong, H. H., Kim, S. Y., Rousseaux, M. W. C., Zoghbi, H. Y. & Liu, Z. CRISPRcloud: A secure cloud-based pipeline for CRISPR pooled screen deconvolution. *Bioinformatics* **33**, 2963–2965 (2017).

200. Li, W., Köster, J., Xu, H., Chen, C. H., Xiao, T., Liu, J. S., Brown, M. & Liu, X. S. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biol.* **16**, 1–13 (2015).

201. Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., Vasilevich, A. S., Barry, J. D., Bansal, H. S., Kraus, O., Wawer, M., Paavolainen, L., Herrmann, M. D., Rohban, M., Hung, J., Hennig, H., Concannon, J., Smith, I., Clemons, P. A., Singh, S., Rees, P., Horvath, P., Linington, R. G. & Carpenter, A. E. Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017).

202. Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., Guertin, D. A., Chang, J. H., Lindquist, R. A., Moffat, J., Golland, P. & Sabatini, D. M. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* (2006) doi:10.1186/gb-2006-7-10-r100.

203. Pau, G., Fuchs, F., Sklyar, O., Boutros, M. & Huber, W. EBImage-an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979–981 (2010).

204. Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C. W. V, Bussey, H., Andrews, B., Tyers, M. & Boone, C. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science (80-.).* **294**, 2364–2368 (2001).

205. Zhao, D., Badur, M. G., Luebeck, J., Magaña, J. H., Birmingham, A., Sasik, R., Ahn, C. S., Ideker, T., Metallo, C. M. & Mali, P. Combinatorial CRISPR-Cas9 Metabolic Screens Reveal Critical Redox Control Points Dependent on the KEAP1-NRF2 Regulatory Axis. *Mol. Cell* **69**, 648-663.e7 (2018).

206. Zhao, D., Shen, J. P., Sasik, R., Ideker, T. & Mali, P. Combinatorial CRISPR-Cas9 Knockout Screen. *Protoc. Exch.* 3–8 (2017) doi:10.1038/protex.2017.063.

207. Muellner, M. K., Duernberger, G., Ganglberger, F., Kerzendorfer, C., Uras, I. Z., Schoenegger, A., Bagienski, K., Colinge, J. & Nijman, S. M. B. TOPS: A versatile software tool for statistical analysis and visualization of combinatorial gene-gene and gene-drug interaction screens. *BMC Bioinformatics* **15**, 1–12 (2014).

208. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, (2018).

209. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).

210. Miles, L. A., Garippa, R. J. & Poirier, J. T. Design, execution, and analysis of pooled in vitro CRISPR/Cas9 screens. *FEBS J.* **283**, 3170–3180 (2016).

211. Rodenburg, R. J. The functional genomics laboratory : functional validation of genetic variants. 297–307 (2018).

212. Chuai, G., Yang, F., Yan, J., Chen, Y., Ma, Q., Zhou, C., Zhu, C., Gu, F. & Liu, Q. Deciphering relationship between microhomology and in-frame mutation occurrence in human CRISPR-based gene knockout. *Mol. Ther. - Nucleic Acids* **5**, e323 (2016).

213. Ipsaro, J. J., Shen, C., Arai, E., Xu, Y., Kinney, J. B., Joshua-Tor, L., Vakoc, C. R. & Shi, J. Rapid generation of drug-resistance alleles at endogenous loci using CRISPR-Cas9 indel mutagenesis. *PLoS One* **12**, 1–16 (2017).

214. Sheel, A. & Xue, W. Genomic amplifications cause false positives in CRISPR screens. *Cancer Discov.* **6**, 824–826 (2016).

215. Munoz, D. M., Cassiani, P. J., Li, L., Billy, E., Korn, J. M., Jones, M. D., Golji, J., Ruddy, D. A., Yu, K., McAllister, G., Deweck, A., Abramowski, D., Wan, J., Shirley, M. D., Neshat, S. Y., Rakiec, D., De Beaumont, R., Weber, O., Kauffmann, A., Robert McDonald, E., Keen, N., Hofmann, F., Sellers, W. R., Schmelzle, T., Stegmeier, F. & Schlabach, M. R. CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions. *Cancer Discov.* **6**, 900–913 (2016).

216. Aguirre, A. J., Meyers, R. M., Weir, B. A., Vazquez, F., Zhang, C. Z., Ben-David, U., Cook, A., Ha, G., Harrington, W. F., Doshi, M. B., Kost-Alimova, M., Gill, S., Xu, H., Ali, L. D., Jiang, G., Pantel, S., Lee, Y., Goodale, A., Cherniack, A. D., Oh, C., Kryukov, G., Cowley, G. S., Garraway, L. A., Stegmaier, K., Roberts, C. W., Golub, T. R., Meyerson, M., Root, D. E., Tsherniak, A. & Hahn, W. C. Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discov.* **6**, 914–929 (2016).

217. Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Yorick, PostWei, J. J., Lander, E. S. & Sabatini, D. M. Identification and characterization of essential genes in the human genome. *Science (80-.).* **350**, 1096–1101 (2015).

218. Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., Dharia, N. V., Montgomery, P. G., Cowley, G. S., Pantel, S., Goodale, A., Lee, Y., Ali, L. D., Jiang, G., Lubonja, R., Harrington, W. F., Strickland, M., Wu, T., Hawes, D. C., Zhivich, V. A., Wyatt, M. R., Kalani, Z., Chang, J. J., Okamoto, M., Stegmaier, K., Golub, T. R., Boehm, J. S., Vazquez, F., Root, D. E., Hahn, W. C. & Tsherniak, A. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).

219. Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J. & Root, D. E. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).

220. Mohr, S. E., Hu, Y., Ewen-Campen, B., Housden, B. E., Viswanatha, R. & Perrimon, N. CRISPR guide RNA design for research applications. *FEBS J.* **283**, 3232–3238 (2016).

221. Cross, B. C. S., Lawo, S., Archer, C. R., Hunt, J. R., Yarker, J. L., Riccombeni, A., Little, A. S., Mccarthy, N. J. & Moore, J. D. Increasing the performance of pooled CRISPR-Cas9 drop-out screening. *Sci. Rep.* **6**, 1–8 (2016).

222. Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G. & Zhang, F. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).

223. Zhang, X. H., Tee, L. Y., Wang, X. G., Huang, Q. S. & Yang, S. H. Off-target effects in

CRISPR/Cas9-mediated genome engineering. *Mol. Ther. - Nucleic Acids* **4**, (2015).

224. Lin, Y., Cradick, T. J., Brown, M. T., Deshmukh, H., Ranjan, P., Sarode, N., Wile, B. M., Vertino, P. M., Stewart, F. J. & Bao, G. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42**, 7473–7485 (2014).

225. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).

226. Tycko, J., Myer, V. E. & Hsu, P. D. Methods for Optimizing CRISPR-Cas9 Genome Editing Specificity. *Mol. Cell* **63**, 355–370 (2016).

227. Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z. & Joung, J. K. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).

228. Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. A., Yan, W. X. & Zhang, F. Rationally engineered Cas9 nucleases with improved specificity. *Science (80-.).* **351**, 84–88 (2016).

229. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. 1012–1019 (2014) doi:10.1101/gr.171322.113.

230. Canver, M. C., Lessard, S., Pinello, L., Wu, Y., Ilboudo, Y., Stern, E. N., Needleman, A. J., Galactéros, F., Brugnara, C., Kutlar, A., McKenzie, C., Reid, M., Chen, D. D., Das, P. P., A Cole, M., Zeng, J., Kurita, R., Nakamura, Y., Yuan, G. C., Lettre, G., Bauer, D. E. & Orkin, S. H. Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nat. Genet.* **49**, 625–634 (2017).

231. Paquet, D., Kwart, D., Chen, A., Sproul, A., Jacob, S., Teo, S., Olsen, K. M., Gregg, A., Noggle, S. & Tessier-Lavigne, M. Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533**, 125–129 (2016).

232. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**, 120–123 (2014).

233. Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., Syed, T., Emons, B. J. M., Gifford, D. K. & Sherwood, R. I. High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167–174 (2016).

234. Roy, K. R., Smith, J. D., Vonesch, S. C., Lin, G., Tu, C. S., Lederer, A. R., Chu, A., Suresh, S., Nguyen, M., Horecka, J., Tripathi, A., Burnett, W. T., Morgan, M. A., Schulz, J., Orsley, K. M., Wei, W., Aiyar, R. S., Davis, R. W., Bankaitis, V. A., Haber, J. E., Salit, M. L., Onge, R. P. S. & Steinmetz, L. M. Multiplexed precision genome editing with trackable genomic barcodes in yeast. *Nat. Publ. Gr.* (2018) doi:10.1038/nbt.4137.

235. Jacobson, E. F. & Tzanakakis, E. S. Human pluripotent stem cell differentiation to functional pancreatic cells for diabetes therapies: Innovations, challenges and future directions. *J. Biol. Eng.* **11**, 21 (2017).

236. Pharoah, P. D. P., Antoniou, A., Bobrow, M., Zimmern, R. L., Easton, D. F. & Ponder, B. A. J. Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* **31**, 33–36 (2002).

237. Lvovs, D., Favorova, O. O. & Favorov, A. V. A Polygenic Approach to the Study of Polygenic Diseases. *Acta Naturae* **4**, 59–71 (2012).

238. Wong, A. S. L., Choi, G. C. G., Cui, C. H., Pregernig, G., Milani, P., Adam, M., Perli, S. D., Kazer, S. W., Gaillard, A., Hermann, M., Shalek, A. K., Fraenkel, E. & Lu, T. K. Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM. *Proc. Natl. Acad. Sci.* **113**, 2544–2549 (2016).

239. Najm, F. J., Strand, C., Donovan, K. F., Hegde, M., Sanson, K. R., Vaimberg, E. W., Sullender, M. E., Hartenian, E., Kalani, Z., Fusi, N., Listgarten, J., Younger, S. T., Bernstein, B. E., Root, D. E. & Doench, J. G. Orthologous CRISPR–Cas9 enzymes for combinatorial genetic screens. *Nat. Biotechnol.* **36**, 179–189 (2018).

240. Aubrey, B. J., Kelly, G. L., Kueh, A. J., Brennan, M. S., O'Connor, L., Milla, L., Wilcox, S., Tai, L., Strasser, A. & Herold, M. J. An Inducible Lentiviral Guide RNA Platform Enables the Identification of Tumor-Essential Genes and Tumor-Promoting Mutations InVivo. *Cell Rep.* **10**, 1422–1432 (2015).

241. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015).

242. Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K. C., Huang, H., Liu, T., Marina, R. J., Jung, I., Shen, Y., Guan, K. L. & Ren, B. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* **14**, 629–635 (2017).

243. Diao, Y., Li, B., Meng, Z., Jung, I., Lee, A. Y. & Dixon, J. A New Class of Temporarily Phenotypic Enhancers Identified by CRISPR / Cas9 Mediated Genetic Screening. *Genome Res.* 1–9 (2016) doi:10.1101/gr.197152.115.

244. Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G. C., Zhang, F., Orkin, S. H. & Bauer, D. E. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).

245. Chu, V. T., Graf, R., Wirtz, T., Weber, T., Favret, J., Li, X., Petsch, K., Tran, N. T., Sieweke, M. H., Berek, C., Kühn, R. & Rajewsky, K. Efficient CRISPR-mediated mutagenesis in primary immune cells using CrispRGold and a C57BL/6 Cas9 transgenic mouse line. *Proc. Natl. Acad. Sci.* **113**, 12514–12519 (2016).

246. Sharma, A., Toepfer, C. N., Ward, T., Wasson, L., Agarwal, R., Conner, D. A., Hu, J. H. & Seidman, C. E. CRISPR/Cas9□Mediated Fluorescent Tagging of Endogenous Proteins in Human Pluripotent Stem Cells. 21.11.1-21.11.20 (2018) doi:10.1002/CPHG.52.

247. Chen, S., Sanjana, N. E., Zheng, K., Shalem, O., Lee, K., Shi, X., Scott, D. A., Song, J., Pan, J. Q., Weissleder, R., Lee, H., Zhang, F. & Sharp, P. A. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* **160**, 1246–1260 (2015).

248. Gao, X., Bali, A. S., Randell, S. H. & Hogan, B. L. M. GRHL2 coordinates regeneration of a polarized mucociliary epithelium from basal stem cells. *J. Cell Biol.* **211**, 669–682 (2015).

249. Budnik, B., Levy, E., Harmange, G. & Slavov, N. Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *bioRxiv* 102681 (2018) doi:10.1101/102681.

250. Boggio, K. J., Obasuyi, E., Sugino, K., Nelson, S. B., Agar, N. Y. R. & Agar, J. N. Recent advances in single-cell MALDI mass spectrometry imaging and potential clinical impact. *Expert Rev. Proteomics* **8**, 591–604 (2011).

251. Liang, X., Potter, J., Kumar, S., Ravinder, N. & Chesnut, J. D. Enhanced CRISPR/Cas9-mediated precise genome editing by improved design and delivery of gRNA, Cas9 nuclease, and donor DNA. *J. Biotechnol.* **241**, 136–146 (2017).

252. Chu, V. T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K. & Kühn, R. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.* **33**, 543–548 (2015).

253. Dixit, A., Parnas, O., Li, B., Weissman, J. S., Friedman, N., Regev, A., Org, C. A., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M. & Lander, E. S. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1857.e17 (2016).

254. Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., Lim, W. A., Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., Horvath, P., Beerli, R. R., Barbas, C. F., Campbell, R. E., Tour, O., Palmer, A. E., Steinbach, P. A., Baird, G. S., Zacharias, D. A., Tsien, R. Y., Cho, S. W., Kim, S., Kim, J. M., Kim, J.-S., Churchman, L. S., Weissman, J. S., Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., Zhang, F., Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J., Charpentier, E., Gasiunas, G., Barrangou, R., Horvath, P., Siksnys, V., Hannon, G. J., Hwang, W. Y., Fu, Y., Reyon, D., Maeder, M. L., Tsai, S. Q., Sander, J. D., Peterson, R. T., Yeh, J.-R. J., Joung, J. K., Jiang, W., Bikard, D., Cox, D., Zhang, F., Marraffini, L. A., Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., Charpentier, E., Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., Doudna, J., Klug, A., Lewis, M., Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., Arkin, A. P., Lutz, R., Bujard, H., Makarova, K. S., Haft, D.

H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J. M., Wolf, Y. I., Yakunin, A. F., al., et, Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., Dicarlo, J. E., Norville, J. E., Church, G. M., Marraffini, L. A., Sontheimer, E. J., Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., Wold, B., Nudler, E., Goldfarb, A., Kashlev, M., Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C., Waldo, G. S., Qi, L., Haurwitz, R. E., Shao, W., Doudna, J. A., Arkin, A. P., Wang, H. H., Isaacs, F. J., Carr, P. A., Sun, Z. Z., Xu, G., Forest, C. R., Church, G. M., Wiedenheft, B., Lander, G. C., Zhou, K., Jore, M. M., Brouns, S. J. J., Oost, J. van der, Doudna, J. A., Nogales, E., Wiedenheft, B., Sternberg, S. H., Doudna, J. A., Zamore, P. D., Tuschl, T., Sharp, P. A., Bartel, D. P., Zhang, F., Cong, L., Lodato, S., Kosuri, S., Church, G. M. & Arlotta, P. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152**, 1173–1183 (2013).

255. Chen, M. & Qi, L. S. Repurposing CRISPR System for Transcriptional Activation. in *RNA Activation* (ed. Li, L.-C.) 147–157 (Springer Singapore, 2017). doi:10.1007/978-981-10-4310-9_10.

256. Akcakaya, P., Bobbin, M. L., Guo, J. A., Lopez, J. M., Clement, M. K., Garcia, S. P., Fellows, M. D., Porritt, M. J., Firth, M. A., Carreras, A., Baccega, T., Seeliger, F., Bjursell, M., Tsai, S. Q., Nguyen, N. T., Nitsch, R., Mayr, L. M., Pinello, L., Bohlooly-Y, M., Aryee, M. J., Maresca, M. & Joung, J. K. In vivo CRISPR-Cas gene editing with no detectable genome-wide off-target mutations. *bioRxiv* 272724 (2018) doi:10.1101/272724.

257. Black, J. B., Adler, A. F., Wang, H. G., D'Ippolito, A. M., Hutchinson, H. A., Reddy, T. E., Pitt, G. S., Leong, K. W. & Gersbach, C. A. Targeted Epigenetic Remodeling of Endogenous Loci by CRISPR/Cas9-Based Transcriptional Activators Directly Converts Fibroblasts to Neuronal Cells. *Cell Stem Cell* **19**, 406–414 (2016).

258. Dijk, D. Van, Sharma, R., Nainys, J., Wolf, G., Krishnaswamy, S., Pe, D., Dijk, D. Van, Sharma, R., Nainys, J., Yim, K., Kathail, P. & Carr, A. J. Recovering Gene Interactions from Single-Cell Data Resource Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729.e27 (2018).

259. Pinello, L., Canver, M. C., Hoban, M. D., Orkin, S. H., Kohn, D. B., Bauer, D. E. & Yuan, G. C. Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat. Biotechnol.* **34**, 695–697 (2016).

260. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

261. Chen, W. S., Zivanovic, N., Dijk, D. Van, Wolf, G., Bodenmiller, B. & Krishnaswamy, S. Uncovering axes of variation among single-cell cancer specimens. *Nat. Methods* (2020) doi:10.1038/s41592-019-0689-z.

262. Friedman, A. J., Hastie, T., Simon, N., Tibshirani, R. & Hastie, M. T. Lasso and Elastic-Net Regularized Generalized Linear Models. *Available online https//cran.r-project.org/web/packages/glmnet/glmnet.pdf. (Verified 29 July. 2015)*. (2015).

263. Thissen, D., Steinberg, L. & Kuang, D. Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons. *J. Educ. Behav. Stat.* **27**, 77–83 (2002).

264. Gao, N., White, P. & Kaestner, K. H. Establishment of Intestinal Identity and Epithelial-Mesenchymal Signaling by Cdx2. *Dev. Cell* **16**, 588–599 (2009).

265. Silberg, D. G., Sullivan, J., Kang, E., Swain, G. P., Moffett, J., Sund, N. J., Sackett, S. D. & Kaestner, K. H. Cdx2 Ectopic Expression Induces Gastric Intestinal Metaplasia. 689–696 (2002) doi:10.1053/gast.2002.31902.

266. Kim, T. & Shivdasani, R. A. Stomach development , stem cells and disease. 554–565 (2016) doi:10.1242/dev.124891.

267. Simmini, S., Bialecka, M., Huch, M., Kester, L., Wetering, M. Van De, Sato, T., Beck, F., Oudenaarden, A. Van, Clevers, H. & Deschamps, J. Transformation of intestinal stem cells into gastric stem cells on loss of transcription factor Cdx2. *Nat. Commun.* **5**, 1–10 (2014).

268. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. **119**, (2009).

269. Yang, J., Mani, S. A., Donaher, J. L., Ramaswamy, S., Itzykson, R. A., Come, C., Savagner, P., Gitelman, I., Richardson, A., Weinberg, R. A., Val, C. & Lamarque, A. Twist , a Master Regulator of Morphogenesis , Plays an Essential Role in Tumor Metastasis. **117**, 927–939 (2004).

270. Qin, Q., Xu, Y., He, T., Qin, C. & Xu, J. Normal and disease-related biological functions of Twist1 and underlying molecular mechanisms. *Nat. Publ. Gr.* **22**, 90–106 (2011).

271. Sarper, S. E., Inubushi, T., Kurosaka, H., Ono Minagi, H., Kuremoto, K. ichi, Sakai, T., Taniuchi, I. & Yamashiro, T. Runx1-Stat3 signaling regulates the epithelial stem cells in continuously growing incisors. *Sci. Rep.* **8**, 1–12 (2018).

272. Scheitz, C. J. F. & Tumbar, T. New insights into the role of Runx1 in epithelial stem cell biology and pathology. *J. Cell. Biochem.* **114**, 985–993 (2013).

273. Umansky, K. B., Gruenbaum-Cohen, Y., Tsoory, M., Feldmesser, E., Goldenberg, D., Brenner, O. & Groner, Y. Runx1 Transcription Factor Is Required for Myoblasts Proliferation during Muscle Regeneration. *PLoS Genet.* **11**, 1–31 (2015).

274. Marmigère, F., Montelius, A., Wegner, M., Groner, Y., Reichardt, L. F. & Ernfors, P. The Runx1/AML1 transcription factor selectively regulates development and survival of TrkA nociceptive sensory neurons. *Nat. Neurosci.* **9**, 180–187 (2006).

275. Fijneman, R. J. A., Anderson, R. A., Richards, E., Liu, J., Tijssen, M., Meijer, G. A., Anderson, J., Rod, A., O'Sullivan, M. G., Scott, P. M. & Cormier, R. T. Runx1 is a tumor suppressor gene in the mouse gastrointestinal tract. *Cancer Sci.* **103**, 593–599 (2012).

276. Garrido-Martín, E. M., Blanco, F. J., Roquè, M., Novensà, L., Tarocchi, M., Lang, U. E., Suzuki, T., Friedman, S. L., Botella, L. M. & Bernabéu, C. Vascular injury triggers Krüppel-like factor 6 mobilization and cooperation with specificity protein 1 to promote endothelial activation through upregulation of the activin receptor-like kinase 1 gene. *Circ. Res.* **112**, 113–127 (2012).

277. Castro, D. S., Martynoga, B., Parras, C., Ramesh, V., Pacary, E., Johnston, C., Drechsel, D., Lebel-Potter, M., Garcia, L. G., Hunt, C., Dolle, D., Bithell, A., Ettwiller, L., Buckley, N. & Guillemot, F. A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. *Genes Dev.* **25**, 930–945 (2011).

278. Dean, L. Pitt-Hopkins Syndrome. in (eds. Pratt, V. M., McLeod, H. L., Rubinstein, W. S., Scott, S. A., Dean, L. C., Kattman, B. L. & Malheiro, A. J.) (2012).

279. Ehinger, Y., Matagne, V., Villard, L. & Roux, J.-C. Rett syndrome from bench to bedside: recent advances. *F1000Research* **7**, 398 (2018).

280. Stumpel, C. & Vos, Y. J. L1 Syndrome. in (eds. Adam, M. P., Ardinger, H. H., Pagon, R. A., Wallace, S. E., Bean, L. J. H., Stephens, K. & Amemiya, A.) (1993).

281. Forrest, M. P., Hill, M. J., Quantock, A. J., Martin-Rendon, E. & Blake, D. J. The emerging roles of TCF4 in disease and development. *Trends Mol. Med.* **20**, 322–331 (2014).

282. Samatov, T. R., Wicklein, D. & Tonevitsky, A. G. L1CAM: Cell adhesion and more. *Prog. Histochem. Cytochem.* **51**, 25–32 (2016).

283. Shapiro, B., Tocci, P., Haase, G., Gavert, N. & Ben-Ze'ev, A. Clusterin, a gene enriched in intestinal stem cells, is required for L1-mediated colon cancer metastasis. *Oncotarget* **6**, 34389–34401 (2015).

284. Shi, M., Kovac, A., Korff, A., Cook, T. J., Ginghina, C., Bullock, K. M., Yang, L., Stewart, T., Zheng, D., Aro, P., Atik, A., Kerr, K. F., Zabetian, C. P., Peskind, E. R., Hu, S. C., Quinn, J. F., Galasko, D. R., Montine, T. J., Banks, W. A. & Zhang, J. CNS tau efflux via exosomes is likely increased in Parkinson's disease but not in Alzheimer's disease. *Alzheimer's Dement.* **12**, 1125–1131 (2016).

285. Li, W. & Pozzo-Miller, L. BDNF deregulation in Rett syndrome. *Neuropharmacology* **76 Pt C**, 737–746 (2014).

286. Chapleau, C. A., Lane, J., Pozzo-Miller, L. & Percy, A. K. Evaluation of current pharmacological treatment options in the management of Rett syndrome: from the present to future therapeutic alternatives. *Curr. Clin. Pharmacol.* **8**, 358–369 (2013).

287. Suzuki, A., Shinoda, M., Honda, K., Shirakawa, T. & Iwata, K. Regulation of transient receptor potential vanilloid 1 expression in trigeminal ganglion neurons via methyl-CpG binding protein 2 signaling contributes tongue heat sensitivity and inflammatory

hyperalgesia in mice. *Mol. Pain* **12**, 1–11 (2016).

288. Forrest, M. P., Waite, A. J., Martin-Rendon, E. & Blake, D. J. Knockdown of Human TCF4 Affects Multiple Signaling Pathways Involved in Cell Survival, Epithelial to Mesenchymal Transition and Neuronal Differentiation. *PLoS One* **8**, (2013).

289. Ambros, V. The functions of animal microRNAs. **431**, (2004).

290. Bartel, D. P., Lee, R. & Feinbaum, R. MicroRNAs : Genomics , Biogenesis , Mechanism , and Function Genomics : The miRNA Genes. **116**, 281–297 (2004).

291. Bartel, D. P. Review Metazoan MicroRNAs. *Cell* **173**, 20–51 (2018).

292. Shivdasani, R. A. Review in translational hematology MicroRNAs : regulators of gene expression and cell differentiation. **108**, 3646–3654 (2006).

293. Lu, J., Getz, G., Miska, E. A., Alvarez-saavedra, E., Lamb, J., Peck, D., Sweet-cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., Downing, J. R., Jacks, T., Horvitz, H. R. & Golub, T. R. MicroRNA expression profiles classify human cancers. **435**, (2005).

294. Miki, K., Endo, K., Takahashi, S., Funakoshi, S., Takei, I., Katayama, S., Toyoda, T., Kotaka, M., Takaki, T., Umeda, M., Okubo, C., Nishikawa, M., Oishi, A., Narita, M., Miyashita, I., Asano, K., Hayashi, K., Osafune, K., Yamanaka, S., Saito, H. & Yoshida, Y. Efficient Detection and Purification of Cell Populations Using Synthetic MicroRNA Switches. *Cell Stem Cell* **16**, 699–711 (2015).

295. Nissim, L., Wu, M. R., Pery, E., Binder-Nissim, A., Suzuki, H. I., Stupp, D., Wehrspaun, C., Tabach, Y., Sharp, P. A. & Lu, T. K. Synthetic RNA-Based Immunomodulatory Gene Circuits for Cancer Immunotherapy. *Cell* 1–13 (2017) doi:10.1016/j.cell.2017.09.049.

296. Hirosawa, M., Fujita, Y., Parr, C. J. C., Hayashi, K., Kashida, S., Hotta, A., Woltjen, K. & Saito, H. Cell-type-specific genome editing with a microRNA-responsive CRISPR-Cas9 switch. *Nucleic Acids Res.* **45**, e118 (2017).

297. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Routledge, 1988).

298. Yao, Q., Xu, H., Zhang, Q., Zhou, H. & Qu, L. MicroRNA-21 promotes cell proliferation and down-regulates the expression of programmed cell death 4 (PDCD4) in HeLa cervical carcinoma cells. *Biochem. Biophys. Res. Commun.* **388**, 539–542 (2009).

299. Medina, P. P. & Slack, F. J. MicroRNAs and cancer : An overview. **4101**, (2008).

300. Lu, Z., Liu, M., Stribinskis, V., Klinge, C. M., Ramos, K. S., Colburn, N. H. & Li, Y. MicroRNA-21 promotes cell transformation by targeting the programmed cell death 4 gene. *Oncogene* **27**, 4373 (2008).

301. Chak, K., Roy-Chaudhuri, B., Kim, H. K., Kemp, K. C., Porter, B. E. & Kay, M. A. Increased precursor microRNA-21 following status epilepticus can compete with mature

microRNA-21 to alter translation. *Exp. Neurol.* **286**, 137–146 (2016).

302. Li, J., Huang, H., Sun, L., Yang, M., Pan, C., Chen, W., Wu, D., Lin, Z., Zeng, C., Yao, Y., Zhang, P. & Song, E. MiR-21 indicates poor prognosis in tongue squamous cell carcinomas as an apoptosis inhibitor. *Clin. Cancer Res.* **15**, 3998–4008 (2009).

303. Zhu, S., Wu, H., Wu, F., Nie, D., Sheng, S. & Mo, Y.-Y. MicroRNA-21 targets tumor suppressor genes in invasion and metastasis. *Cell Res.* **18**, 350 (2008).

304. Wang, S., Aurora, A. B., Johnson, B. A., Qi, X., Mcanally, J., Hill, J. A., Richardson, J. A., Bassel-duby, R. & Olson, E. N. The Endothelial-Specific MicroRNA miR-126 Governs Vascular Integrity and Angiogenesis. 261–271 (2008) doi:10.1016/j.devcel.2008.07.002.

305. Sun, Y., Luo, Z.-M., Guo, X.-M., Su, D.-F. & Liu, X. An updated role of microRNA-124 in central nervous system disorders: a review. *Front. Cell. Neurosci.* **9**, 193 (2015).

306. Seiler, A., Schumacher, S., Nitsch, R., Wulczyn, F. G., Smirnova, L. & Gra, A. Regulation of miRNA expression during neural cell specification. **21**, 1469–1477 (2005).

307. Lagos-quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W. & Tuschl, T. Identification of Tissue-Specific MicroRNAs from Mouse. **12**, 735–739 (2002).

308. Wang, X. Stem cells in tissues, organoids, and cancers. *Cell. Mol. Life Sci.* **76**, 4043–4070 (2019).

309. Liu, G., David, B. T., Trawczynski, M. & Fessler, R. G. Advances in Pluripotent Stem Cells: History, Mechanisms, Technologies, and Applications. *Stem cell Rev. reports* **16**, 3–32 (2020).

310. Tsou, Y.-H., Khoneisser, J., Huang, P.-C. & Xu, X. Hydrogel as a bioactive material to regulate stem cell fate. *Bioact. Mater.* **1**, 39–55 (2016).

311. Bertucci, T. B. & Dai, G. Biomaterial Engineering for Controlling Pluripotent Stem Cell Fate. *Stem Cells Int.* **2018**, 9068203 (2018).

312. McDonald, D., Wu, Y., Dailamy, A., Tat, J., Parekh, U., Zhao, D., Hu, M., Tipps, A., Zhang, K. & Mali, P. Defining the Teratoma as a Model for Multi-lineage Human Development. *Cell* **183**, 1402-1419.e18 (2020).

313. Ruoslahti, E. Brain extracellular matrix. *Glycobiology* **6**, 489–492 (1996).

314. Even-Ram, S., Artym, V. & Yamada, K. M. Matrix Control of Stem Cell Fate. *Cell* **126**, 645–647 (2006).

315. Engler, A. J., Sen, S., Sweeney, H. L. & Discher, D. E. Matrix elasticity directs stem cell lineage specification. *Cell* **126**, 677–689 (2006).

316. Ozasa, Y., Gingery, A. & Amadio, P. C. Muscle-derived stem cell seeded fibrin gel interposition produces greater tendon strength and stiffness than collagen gel in vitro. *The Journal of hand surgery, European volume* vol. 40 747–749 (2015).

317. Litvinov, R. I. & Weisel, J. W. Fibrin mechanical properties and their structural origins. *Matrix Biol.* **60–61**, 110–123 (2017).

318. Hu, M., Dailamy, A., Lei, X. Y., Parekh, U., McDonald, D., Kumar, A. & Mali, P. Facile Engineering of Long-Term Culturable Ex Vivo Vascularized Tissues Using Biologically Derived Matrices. *Adv. Healthc. Mater.* **7**, 1800845 (2018).

319. Kluin, R. J. C., Kemper, K., Kuilman, T., de Ruiter, J. R., Iyer, V., Forment, J. V, Cornelissen-Steijger, P., de Rink, I., ter Brugge, P., Song, J.-Y., Klarenbeek, S., McDermott, U., Jonkers, J., Velds, A., Adams, D. J., Peeper, D. S. & Krijgsman, O. XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data. *BMC Bioinformatics* **19**, 366 (2018).

320. Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., Diehn, M. & Alizadeh, A. A. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

321. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C. & Chanda, S. K. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).

322. Dong, Q., Guo, X., Li, L., Yu, C., Nie, L., Tian, W., Zhang, H., Huang, S. & Zang, H. Understanding hyaluronic acid induced variation of water structure by near-infrared spectroscopy. *Sci. Rep.* **10**, 1387 (2020).

323. Birbrair, A. Stem Cell Microenvironments and Beyond. *Advances in experimental medicine and biology* vol. 1041 1–3 (2017).

324. Kong, L. Signaling Pathways Involved in Stem Cell Differentiation and Relevant Therapy. *Current stem cell research & therapy* vol. 14 213 (2019).

325. Chen, L., Huang, T., Qiao, Y., Jiang, F., Lan, J., Zhou, Y., Yang, C., Yan, S., Luo, K., Su, L. & Li, J. Perspective into the regulation of cell-generated forces toward stem cell migration and differentiation. *J. Cell. Biochem.* **120**, 8884–8890 (2019).

326. Lee, V. K., Lanzi, A. M., Haygan, N., Yoo, S.-S., Vincent, P. A. & Dai, G. Generation of Multi-Scale Vascular Network System within 3D Hydrogel using 3D Bio-Printing Technology. *Cell. Mol. Bioeng.* **7**, 460–472 (2014).

327. Chambers, S. M., Fasano, C. A., Papapetrou, E. P., Tomishima, M., Sadelain, M. & Studer, L. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat. Biotechnol.* **27**, 275–280 (2009).

328. OPTN/SRTR 2019 Annual Data Report: Introduction. *Am. J. Transplant. Off. J. Am. Soc. Transplant. Am. Soc. Transpl. Surg.* **21 Suppl 2**, 11–20 (2021).

329. Mehrian-Shai, R., Reichardt, J. K. V, Harris, C. C. & Toren, A. The Gut-Brain Axis, Paving the Way to Brain Cancer. *Trends in cancer* **5**, 200–207 (2019).

330. Verschuuren, J., Strijbos, E. & Vincent, A. Neuromuscular junction disorders. *Handb. Clin. Neurol.* **133**, 447–466 (2016).