iPS models for interrogating how genes and alleles shape cellular function

Tegtmeyer, Matthew

Awarding institution:
King's College London

iPS models for interrogating how genes and alleles shape cellular function

Matthew Thomas Tegtmeyer

Thesis submitted for the degree of Doctor of Philosophy

First supervisor: Dr Davide Danovi
Second supervisor: Prof Fiona Watt
Third supervisor (external): Dr Ralda Nehme

Centre for Gene Therapy & Regenerative Medicine
King's College London
2023

To mom and dad

Associated Manuscripts

Tegtmeyer M, Nehme R. Leveraging the Genetic Diversity of Human Stem Cells in Therapeutic Approaches. *J Mol Biol*. 2022 Feb 15;434(3):167221. doi: 10.1016/j.jmb.2021.167221. Epub 2021 Aug 30. PMID: 34474087.
This manuscript is incorporated in its entirety into Chapter 1 of this thesis.

Mitchell J*, Nemesh J*, **Tegtmeyer M***, Ghosh S, Handsaker R, Mello C, Meyer D, Gebre H, Raghunathan K, de Rivera H, Hawes D, Neumann A, Nehme R, Eggan K, McCarroll S. Census-seq: an experimental and computational toolkit for population scale cellular genetic studies. (For re-submission Summer 2023)
This manuscript is incorporated in its entirety into Chapter 3 of this thesis

Wells MF, Nemesh J, Ghosh S, Mitchell JM, Salick MR, Mello CJ, Meyer D, Pietilainen O, Piccioni F, Guss EJ, Raghunathan K, **Tegtmeyer M**, Hawes D, Neumann A, Worringer KA, Ho D, Kommineni S, Chan K, Peterson BK, Raymond JJ, Gold JT, Siekmann MT, Zuccaro E, Nehme R, Kaykas A, Eggan K, McCarroll SA. Natural variation in gene expression and viral susceptibility revealed by neural progenitor cell villages. Cell Stem Cell. 2023 Mar 2;30(3):312-332.e13. doi: 10.1016/j.stem.2023.01.010. Epub 2023 Feb 15. PMID: 36796362.
This manuscript is relevant for Materials and Methods, Chapters 3 and Chapters 5

Rapino F, Natoli T, Limone F, O'Connor E, Blank J, **Tegtmeyer M**, Chen W, Norabuena E, Narula J, Hazelbaker D, Angelini G, Barrett L, O'Neil A, Beattie UK, Thanos JM, de Rivera H, Sheridan SD, Perlis RH, McCarroll SA, Stevens B, Subramanian A, Nehme R, Rubin LL. Small-molecule screen reveals pathways that regulate C4 secretion in stem cell-derived astrocytes. Stem Cell Reports. 2023 Jan 10;18(1):237-253. doi: 10.1016/j.stemcr.2022.11.018. Epub 2022 Dec 22. PMID: 36563689.
This manuscript includes data generation which was compiled for Chapter 4

Berryer M*, **Tegtmeyer M***, Binan L, Valakh V, Nathanson A, Trendafilova D, Crouse E, Klein J, Meyer D, Pietilainen O, Rapino F, Farhi S, Rubin L, McCarroll S, Nehme R, Barrett L. Robust induction of functional astrocytes using NGN2 expression in human pluripotent stem cells. *iScience*. 2023. https://doi.org/10.1016/j.isci.2023.106995
This manuscript is incorporated in its entirety into Chapter 4 of this thesis.

Pietiläinen O, Trehan A, Meyer D, Mitchell J, **Tegtmeyer M**, Valakh V, Gebre H, Chen T, Vartiainen E, Farhi SL, Eggan K, McCarroll SA, Nehme R. Astrocytic cell adhesion genes linked to schizophrenia correlate with synaptic programs in neurons. Cell Rep. 2023 Jan 31;42(1):111988. doi: 10.1016/j.celrep.2022.111988. Epub 2023 Jan 12. PMID: 36640364.
This manuscript is relevant to Chapter 5 of this thesis.

Tegtmeyer M*, Arora J*, Asgari S*, Cimini B, Peirent E, Liyanage D, Way G, Weisbart E, Nathan A, Amariuta T, Eggan K, Haghighi M, McCarroll S, Carpenter A, Singh S, Nehme R, , Raychaudhuri S**.** High-dimensional phenotyping to define the genetic basis of cellular morphology. Biorxiv 2023 https://doi.org/10.1101/2023.01.09.522731 (*under revision, Nature Communications*)
This manuscript is incorporated in its entirety into Chapter 6 of this thesis

Nadig A***, Tegtmeyer M*,** Patil A, Collings C, Ling, E, Trehan A, Gu H, Weiner D, Erdin S, Yadav R, Bybjerg-Grauholm, Pietilainen O, Collins R, iPSYCH Consortium, ASD Working Group of the Psychiatric Genomics Consortium, Hougaard D, Børglum A, Talkowski M, O'Connor L, Lieberman-Aiden E, McCarroll S, Kadoch C, Nehme R, Robinson E. Regulatory architecture shapes the effect of genetic variation at chromosome 22q. (For submission June 2023)
Portions of this manuscript are incorporated into Chapter 7 of this thesis

Nehme R, Pietiläinen O, Artomov M, **Tegtmeyer M**, Valakh V, Lehtonen L, Bell C, Singh T, Trehan A, Sherwood J, Manning D, Peirent E, Malik R, Guss EJ, Hawes D, Beccard A, Bara AM, Hazelbaker DZ, Zuccaro E, Genovese G, Loboda AA, Neumann A, Lilliehook C, Kuismin O, Hamalainen E, Kurki M, Hultman CM, Kähler AK, Paulo JA, Ganna A, Madison J, Cohen B, McPhie D, Adolfsson R, Perlis R, Dolmetsch R, Farhi S, McCarroll S, Hyman S, Neale B, Barrett LE, Harper W, Palotie A, Daly M, Eggan K. The 22q11.2 region regulates presynaptic gene-products linked to schizophrenia. *Nat Commun*. 2022 Jun 27;13(1):3690. doi: 10.1038/s41467-022-31436-8. PMID: 35760976; PMCID: PMC9237031.
This publication is relevant to Chapter 7 and data generated from this work was utilized in Nadig & Tegtmeyer et al.

Acknowledgements

I would like to thank my supervisors, Dr Davide Danovi and Prof Fiona Watt for giving me the opportunity to be a part of their groups and to pursue my goal of a PhD with their advice. Along with King's College, their guidance and willingness to construct an unusual relationship to allow this for me will leave me eternally in their debt. The flexibility of the University, especially during the COVID-19 pandemic helped to foster a collaborative and support PhD experience. I am forever grateful for them.

I want to thank Dr. Ralda Nehme, who, for the last 6 years, has been an instrumental figure in my career and life. When I joined her lab as a technician following my master's degree, I couldn't have envisioned where things stand today and all of the great fun we have had along the way. To Ralda, I owe a tremendous amount of gratitude and look forward to lifelong collaborations.

I want to thank Prof Steven McCarroll, whose guidance over the past years has been instrumental in my growth as a young scientist.

I want to thank many funders including the Stanley Gift , the National Institutes of Health, and many others, for supporting my PhD, without their support I would not have been able to accomplish these goals.

I want to thank Prof Patrick Chinnery and Prof Siddharthan Chandran for agreeing to read and examine my thesis. I greatly appreciate your time and I am looking forward to discussing my project with you.

I want to provide a special thanks to everyone who has contributed to my projects over the course of my PhD. As will be made clear throughout my work, it has hardly been an effort of myself but rather a dedicated and concerted effort of so many people of whom I am tremendously grateful and will forever cherish my relationships with. Because their contributions are so important to me, I will highlight many individuals and their role in my work here.

Thank you to: Derek Hawes, for working together to establish our stem cell collection which formed the basis for the entirety of the work in my thesis. Jana Mitchell, for establishing early iterations of our cell village approach. Jim Nemesh, for endless support on computational projects and guidance on experimental design. Dia Ghosh, for endless support on computational projects and guidance on experimental design. Curtis Mello, for helping to establish Census-Seq related workflows. Daniel Meyer, for endless support on computational projects and guidance on experimental design. Hilena Gebre, for supporting work contained in the Census-Seq manuscript and other various projects under my umbrella. Kavya Raghunathan, for establishing early iterations of our cell village approach. Anna Neumann, for project management support across all of my projects. Kevin Eggan, for playing a key early role in forming my arrangement with King's College London. Ajay Nadig, for a wonderful new friendship and the endless stream projects we've been working on together. Ajinkya Patil, for his work on mapping BAF complexes in our Chr22 work. Clayton Collings, for his computational help on our Chr22 work. Emi Ling, for her herculean effort to generate snRNAseq data from nearly 200 individuals which was used in our Chr22 work. Aditi Trehan, who helped to generate samples used for Hi-C data contained in the Chr22 work. Huiya Gu, who helped generate Hi-C data contained in the Chr22 work. Daniel Weiner, whose work on 16p established the foundation for our work on Chr22. Dhara Liyanage, a wonderful undergraduate student who has made significant contributions to a range of my projects. Kathryn Boit, a wonderful undergraduate student who has made significant contributions to a range of my projects. Erez Aiden-Lieberman, whose lab generated and consulted on our Hi-C projects. Cigall Kadoch, whose world-class expertise on the BAF complex helped to solidify a mechanistic hypothesis in our work on Chr22. Elise Robinson, who has helped lead the Chr22 work and to inspire new ways of thinking about the genome. Jatin Arora, who worked on our rare variant analysis in the cmQTL work. Samira Asgari, who worked on our common variant analysis in the cmQTL work. Beth Cimini, who helped with image processing and analysis in the cmQTL work. Emily Peirent, who assisted with tissue culture in the cmQTL work. Dhara Liyanage, who worked on the CRISPRi experiments for the cmQTL work. Gregory Way, who helped establish computational pipelines related to the cmQTL work. Erin Weisbart, who provided meaningful comments on our cmQTL manuscript. Aparna Nathan, whose early contributions to the cmQTL project helped define the key parameters for our workflows. Tiffany Amariuta, whose early contributions to the cmQTL project helped define the key parameters for our workflows. Marzieh Haghighi, who contributed to data analyses for the cmQTL work. Anne E. Carpenter, who helped to lead and mentor imaging-based work for both the cmQTL project but also for Neuronal Cell Painting. Shantanu Singh, who helped to lead and mentor imaging-based work for both the cmQTL project but also for Neuronal Cell Painting. Soumya Raychaudhuri, who helped to lead and mentor the genomics work for the cmQTL project. Martin Berryer, for his teamwork in establishing our astrocyte protocol. Loic Binan, who led the calcium signaling experiments in our astrocyte method. Vera Valakh, who worked on calcium signaling experiments in our astrocyte method. Anna Nathanson, who contributed to biochemical validation of our astrocyte protocol. Darina Trendafilova, who worked on the down syndrome modeling contained in our astrocyte manuscript. Ethan Crouse, who worked on calcium signaling experiments in our astrocyte method. Olli Pietilainen, who contributed to analyses in our 22q11.2 work and astrocyte manuscript. Francesca Rapino, who shared her data used in our cell-type comparison in our astrocyte manuscript. Samouil Farhi, who oversaw the calcium signaling experiments in our astrocyte method. Lee L Rubin, who oversaw the development of alternative astrocyte methods highlighted in Rapino et al and incorporated into our astrocyte manuscript. Lindy Barrett, who helped lead and mentor the development of our astrocyte protocol.

I want to thank my family. Mom, dad, Liz, Chris, Kim, Kelsey, Mickey, Pepper, Linke, Emma, Logan, Olivia, and Kami. Your generous support and understanding while I figure out what to do with my life has been critical in making this an easy process. I love you all so very much.

COVID Impact Statement

The COVID-19 restrictions disrupted my PhD research in many ways.

- When I began my PhD in 2019, I had goals of splitting time between London and Boston, MA USA. The national and international lockdowns forced me to remain in Boston, MA USA for the bulk of my PhD due to an inability to return to the UK with the flight restrictions. When the restrictions were lifted, my supervisors and I determined that since I was close to wrapping up the work towards my PhD, it may have an overall negative impact on my work to completely change my physical location. Therefore, I remained in Boston, MA USA to complete my scientific work and made semi-frequent trips to London for meetings and presentations, in addition to maintaining a regular virtual presence in all possible relevant meetings and activities at King's. In order to remain tightly connected with King's, I participated in weekly and monthly group meetings and served as a Centre student representative for a large portion of my time in the U.S.

Lay Summary

Since the completion of the human genome project, rapid advances in sequencing technologies have provided unprecedented access to measure relationships between our DNA and human traits including diseases. My project explores several ways to help bridge the gap between human genetic variation and cellular phenotypes. I sought to develop and implement new experimental systems that allow us to measure the behavior of cells from many different people at once, which is key to link cell behavior to an individual's unique DNA. I show that one can capture and analyze a picture of cells from many people and uncover relationships between how cells look and the specific DNA sequence of their donor. Another key element is to find ways to make experiments larger so we can capture more people at a time. We accomplish this by mixing together cells from many different people together and use new sequencing techniques to link cells to their donors. I developed a method to generate a new brain cell type from stem cells, which allowed me to explore how various cell types within our brains interact with one another. Lastly, I found interesting relationships between rare mutations and general disease risk. My project provides a strong foundation for using human stem cells to understand how genetic variation impacts biology and contributes to risk for human illnesses.

Abstract

Tens of thousands of genetic variants shape human phenotypes, mostly by unknown cellular mechanisms. Human derived stem cells enable us to measure cellular phenotypes and their relationship to human genetic variation. Here, I provide a range of works focusing on leveraging these human stem cells to model functional genomics in an array of experimental paradigms. My efforts developing novel experimental systems now make population-scale (10s-100s of unique cell lines) experiments more feasible and accessible. I provide a new method for differentiating human stem cells into functional astrocytes, which are increasingly being implicated in a range of neuropsychiatric and neurological disorders. I then use these methods to understand astrocyte-neuron interactions and show that perturbations of this interaction may shed insights into mechanisms by which antipsychotic medications provide therapeutic benefit. I show state of the art imaging approaches are amenable to stem cell models and identify both common and rare genetic variants which mediate cell morphology. Last, I display functional and statistical convergence between rare and common genetic risk factors for neuropsychiatric phenotypes on Chromosome 22q. My work fills many gaps in the fields of stem cell biology, genomics, and human genetics to establish a strong foundation for other researchers to explore experimental biology at the scale of 100s to 1000s of genetically unique cell lines, across multiple levels of analysis, and develop new intellectual frameworks for investigating the functional consequences of human genetic variation.

Table of Contents

List of figures

Chapter 1

General Introduction

Leveraging the Genetic Diversity of Human Stem Cells in Therapeutic Approaches

1.1 – Introduction

Human pluripotent stem cells (hPSCs) provide unprecedented access to the tissues affected in disease, and to diverse cell types from individual patients with unique genetic backgrounds and diagnoses. The ability of these cells to proliferate indefinitely and differentiate into the three primary germ layers has begun to revolutionize in vitro biology. This is especially true in the context of neuroscience, where primary cultures of patient brain cells are mostly inaccessible, and the availability of post-mortem samples from the developing human brain is limited. Induced pluripotent stem cells (iPSCs), whereby somatic cells can be reprogrammed into a pluripotent state, were first generated in 2007 and continual developments since then have provided new insights into the mechanisms of human disease and paved the way for a new era in therapeutics and drug discovery (Takahashi et al, 2007; Yu et al, 2009).

hPSCs can be generated either from the collection of embryonic stem cells (ESCs) from the inner-cell mass of the blastocyst, an early-stage embryo, or through reprogramming of iPSCs from somatic tissues such as fibroblasts or peripheral blood mononuclear cells (PBMCs) from any given individual (Takahashi et al, 2007, Staerk et al, 2010). When grown in defined culture conditions, these hPSCs can be differentiated into a range of disease-relevant cell types (Li et al, 2011; Liu et al, 2012; Zhang et al, 2012; Du et al, 2015; Hu et al, 2009). Early efforts leveraged hPSC technology to investigate a spectrum of human diseases, from neurological conditions such as Parkinson's disease or amyotrophic lateral sclerosis (ALS) to blood disorders and cardiac and vascular disorders (Soldner et al, 2009; Dimos et al, 2008; Raya et al, 2009; Yazawa et al, 2011). Since their inception, hPSC-based disease models have undergone continuous evolutions in their quality and complexity to further investigate pathological mechanisms of diseases, especially in neuroscience, where many disorders are classified by both single-cell type and systems level pathology. Key innovations in relatively simple two-dimensional hPSC culture systems in combination with more complex three-dimensional models are critical to uncovering meaningful biological insights.

hPSC-based model systems, whether two- dimensional cultures or three-dimensional organoids, hold great promise for nominating potential therapeutic targets and identifying cell-type specific disease mechanisms. However, several shortcomings have limited the broader adoption of these tools, including incomplete understanding about the extent to which hPSC-derived cell types recapitulate living human tissues, and limited sample sizes which have hindered high-throughput validation of in vivo phenotypes using these models. Adding to the challenge is the necessity to navigate complex regulatory processes and protocols for accessing cell lines and datasets. Sample collections are frequently built upon diverse sets of compliance which, while essential to protect donor privacy and intent, can often complicate downstream experimental designs and may deter researchers from investing in these resources (Isai et al, 2014).

A unique, yet largely underexplored aspect of hPSC models is that they offer the ability to capture the vast complexity of human genetic diversity. This diversity not only plays a role in human disease but can also impact therapeutic targets and treatment response and efficacy (Fakunle et al, 2012; Yamasaki et al, 2017). The genetic contributions to complex human diseases have principally been explored through genome-wide association studies (GWAS), whereby large collections of genetic data compare the correlative relationships between single-nucleotide polymorphism (SNPs) and human traits. Within the context of neurological and neuropsychiatric conditions, GWAS have nominated hundreds of genomic loci associated with schizophrenia, Alzheimer's disease, autism spectrum disorder, Parkinson's disease, and many others (Chang et al, 2017; Ripke et al, 2020; Grove et al, 2019; Jansen et al 2019). HPSC models enable the investigation of how natural genetic variation, such as that detected in GWAS, contributes to functional biology. Understanding how these variants contribute to cellular behavior will allow researchers to uncover previously unknown mechanisms of complex disease, facilitating the development of new therapeutic avenues.

Below, we discuss the impact and future potential of human stem cell derived model systems in advancing therapeutic research and applications, with a focus on neurological disease. We review key innovations which gave rise to these technologies, main strategies for hPSC differentiation and their suitability for therapeutic discovery, along with avenues for scaling up these studies. We further highlight how genetic diversity, particularly with respect to ancestry and sex, can be leveraged to extend the applicability of these approaches to the majority of the global population.

1.2 - hPSCs as a powerful tool to capture genetic diversity of therapeutic populations

A primary benefit of the use of hPSC technologies in therapeutic discovery lies in their ability to capture biological interactions due to specific characteristics of individual genetic backgrounds such as single nucleotide, copy number or repeat polymorphisms. These differences are of particular importance within the context of sex and ancestry, both of which can affect treatment safety and efficacy. Leveraging iPSCs to understand the way ancestry and sex contribute to biological function can enable improved efficacy of treatment while minimizing negative side effects of current medicines.

1.2.3 - Ancestry

Patient-derived stem cells offer a valuable, yet underexplored path for examining many different genetic backgrounds and ancestries in functional studies. This diversification may provide significant value to not only societal equity, but importantly, for understanding the mechanisms by which ancestry-specific genetic variants increase an individual's susceptibility to disease and, additionally, their response to drug treatment.

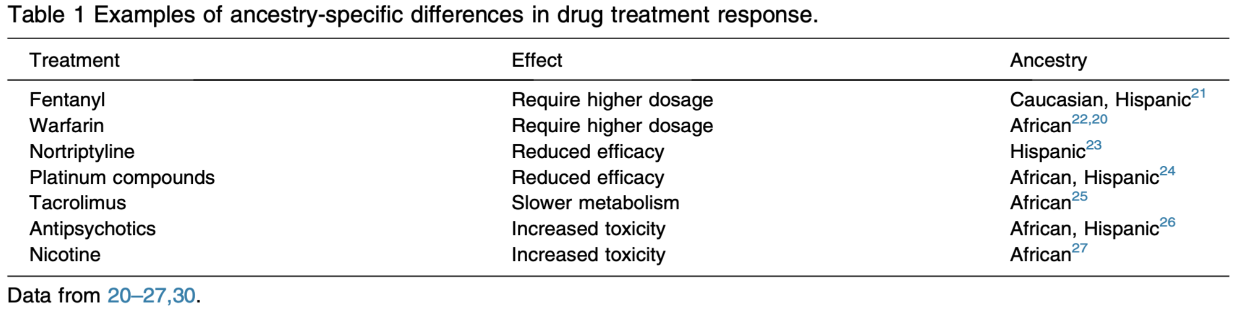


Table 1 Examples of ancestry-specific differences in drug treatment response.

Treatment	Effect	Ancestry
Fentanyl	Require higher dosage	Caucasian, Hispanic[21]
Warfarin	Require higher dosage	African[22,20]
Nortriptyline	Reduced efficacy	Hispanic[23]
Platinum compounds	Reduced efficacy	African, Hispanic[24]
Tacrolimus	Slower metabolism	African[25]
Antipsychotics	Increased toxicity	African, Hispanic[26]
Nicotine	Increased toxicity	African[27]

Data from [20–27,30].

[20, Mouton et al, 2016; 21, Lotsch et al, 2009; 22, Suarez-Kurtz et al, 2013; 23, Uher et al, 2010; 24, Bonifaz-Pena et al, 2014; 25, Burchard et al, 2003; 26, Tanaka et al, 2013; Luo et al, 2008]

Pharmacological treatment responses can have strong inter-ethnic differences, whereby persons from one ancestral lineage may have significantly different outcomes following treatment when compared to individuals from another ancestry. Examples of such differential responses have been documented in several previous clinical studies, summarized in **Table 1**. For example, nearly 8% of hospital admissions in South Africa are due to adverse reactions to common medications, which do not occur in European-ancestry individuals (Mouton et al, 2016). The same can be said for Brazilians and Caribbean Hispanics, where treatment efficacy of warfarin, an anticoagulant used to treat blood clots, drops precipitously when compared to individuals of European ancestry (Mouton et al, 2016). The fallibility of treatments across diverse ancestries provides an enormous challenge for the scientific and pharmaceutical industry, where experimental validation and patient recruitment for clinical trials are predominantly performed in the United States. The severity of this issue cannot be overstated. For example, within the U.S., two massive clinical trials funded by the National Institutes of Health (NIH) and the National

Cancer Institute (NCI) comprised less than 2% non-white subjects (Clinical Trials, SA 2018). Moreover, the NIH-funded trial was for a respiratory disorder, asthma, which predominantly affects the African-American population (NIH ASTHMA). This problem, however, is not specific to the United States. In a comprehensive analysis of clinical trials spanning 29 countries and greater than 150,000 participants, 86% were white (Clinical Trials, SA 2018).

While setting up clinical trials in different countries and continents can be difficult and expensive, stem cells can be generated from any given individual through the simple extraction of somatic tissues (i.e., hair, skin, blood, urine). This enables the collection of samples from donors across diverse geographical areas and as such paves the way for the subsequent investigation of pharmacogenetic effects on relevant hPSC derived cell types from diverse ancestral populations to better understand and optimize drug treatments across the globe. The availability of hPSC lines from diverse genetic backgrounds is thus a critical prerequisite to facilitate the survey of the ancestry-specific effects of drug treatments.

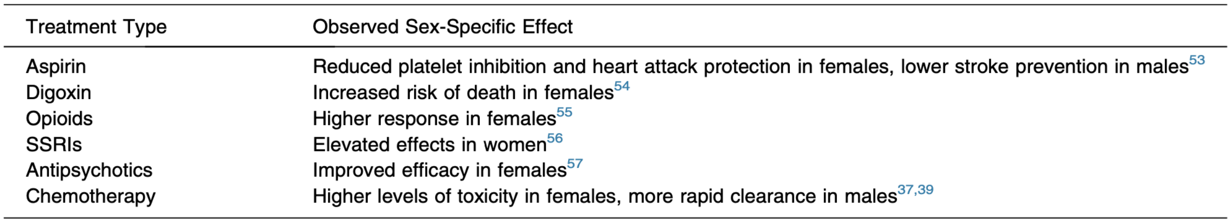
As with genetic studies, many current iPSC collections have a bias towards cell lines derived from individuals of Northern European ancestry (Nehme et al, 2020). However, efforts are underway to include ancestrally diverse populations in genetic studies, specifically populations from Africa, South America, and Southeast Asia, which have been significantly underrepresented when compared to their global population (Sirugo et al, 2019; Martin et al, 2019). In one such example, researchers investigating genetic variation within the context of HIV-1 infected individuals in a South Africa population, showed significant genetic variation in the APOBEC3 locus, which occurs at much higher frequencies than reported in the 1000 genomes data (Matume et al, 2019). This study illustrates how expanding genetic studies to specific populations may uncover previously unknown genetic associations to disease.

Thus, it is imperative to initiate parallel efforts for the generation and analysis of human pluripotent stem cells and their derivatives, to enable the functional investigation of genetic variants in non-European populations and ensure the efficacy and safety of future therapeutic strategies across different populations. A few initiatives are already underway to help achieve this aim of generating "ethnically-diverse" iPSCs (ED-iPSCs) from individuals of African, Hispanic, and Asian ancestry (Chang et al, 2015). Additionally, collections are currently ongoing in South Africa to generate a resource of iPSCs from a diverse population of Africans (UCT).

1.2.4 – Sex

Analogous to the disparities in treatment efficacy and risk of side-effects with respect to ancestry, sex-specific differences are another largely undervalued co-variable in common clinical prescriptions. For example, women often experience more common and more severe toxicity and side effects as a result of chemotherapy when compared to males undergoing the same treatment (Tsiouda et al, 2020). These sex-specific effects may be driven by biological factors such as hormones, liver and kidney function, other comorbidities, and both germline and somatic variation. For example, differences in androgenic and estrogenic receptors between males and females impact renal tubule structure and function (Sabolic et al, 2007). These differences may drive disease onset and pathogenesis, as well as treatment response whereby recent research shows that males are able to eliminate toxic byproducts of common cancer therapeutics more quickly than females (Schmetzer et al, 2012). Several examples of sex-differences in treatment response are documented in previous studies and illustrated in **Table 2**.

Table 2 Examples of differences in sex-specific responses to common drug treatments.



Treatment Type	Observed Sex-Specific Effect
Aspirin	Reduced platelet inhibition and heart attack protection in females, lower stroke prevention in males[53]
Digoxin	Increased risk of death in females[54]
Opioids	Higher response in females[55]
SSRIs	Elevated effects in women[56]
Antipsychotics	Improved efficacy in females[57]
Chemotherapy	Higher levels of toxicity in females, more rapid clearance in males[37,39]

[37, Tsiouda et al, 2020; 39, Schmetzer et al, 2012; 51, Li et al, 2017; 52, Genolet et al, 2021; 53, Becker et al, 2006; 54, Rathorpe et al, 2002; 55, Bijur et al, 2008]

Though many clinical trials typically enroll equal numbers of participants from both sexes, understanding the underlying biological mechanisms which give rise to these sex differences, and ultimately providing ways in which to overcome them, is a significant downstream challenge. Even with legislation in the U.S., Canada, Europe, Australia, and Japan requiring the inclusion of females into clinical trials, women continue to be underrepresented when compared to males, leaving them more susceptible to adverse drug reactions (Raz et al, 2012). This inequity has proved harmful to the female population, who on average suffers from greater numbers of side effects from common medications compared to men (Zucker et al, 2020). The application of iPSCs, derived from both sexes, to investigate pharmacogenomic effects of drug treatment may provide a novel approach to understanding, and ameliorating, sex-specific differences in therapeutic response.

Further, multiple studies have implicated genes or genetic variation which contribute to sex-biases in disease risk. Kamitaki et al showed within the locus containing C4, there are allele-specific contributions to an inverse proportion of risk for schizophrenia (SCZ) and systemic lupus erythematosus (SLE) in males and females, respectively (Kamitaki et al, 2020). Genetic studies have also suggested, based on inherited genetic variation, a female protective effect against autism spectrum disorder (ASD) (Wigdor et al, 2021). Similar, and yet stronger, sex-specific biology drives biases in neurodegenerative disorders such as Alzheimer's disease, for which women comprise nearly 2/3 of cases (Guo et al, 2021). For example, aging women have significantly higher rates of chromosome X aneuploidy compared to men. Chromosome X aneuploidy within the hippocampus of patients with AD is increased two-fold (Yurov et al, 2014).

Additionally, within the context of brain disorders, specifically psychiatric conditions, there exist sex-biases in diagnosis and treatment beyond biological differences. For example, though males are diagnosed with severe mental illness more frequently than females, women on average are prescribed psychiatric medication more often than men (Terilizzi et al, 2020). The same is true for self-diagnosed depression, leading to over-treatment of females when compared to males who report experiencing symptoms of depression (Thunander et al, 2017). Social and gender-based biases like these can have severe impacts on the health of individuals, who may be wrongfully or incorrectly treated for their disorders. It will be important to leverage iPSCs from both sexes to provide meaningful insights into divergent biological mechanisms which can elucidate potential biomarkers to improve psychiatric diagnostics in the future and overcome bias in treatment decisions and stigma surrounding these conditions.

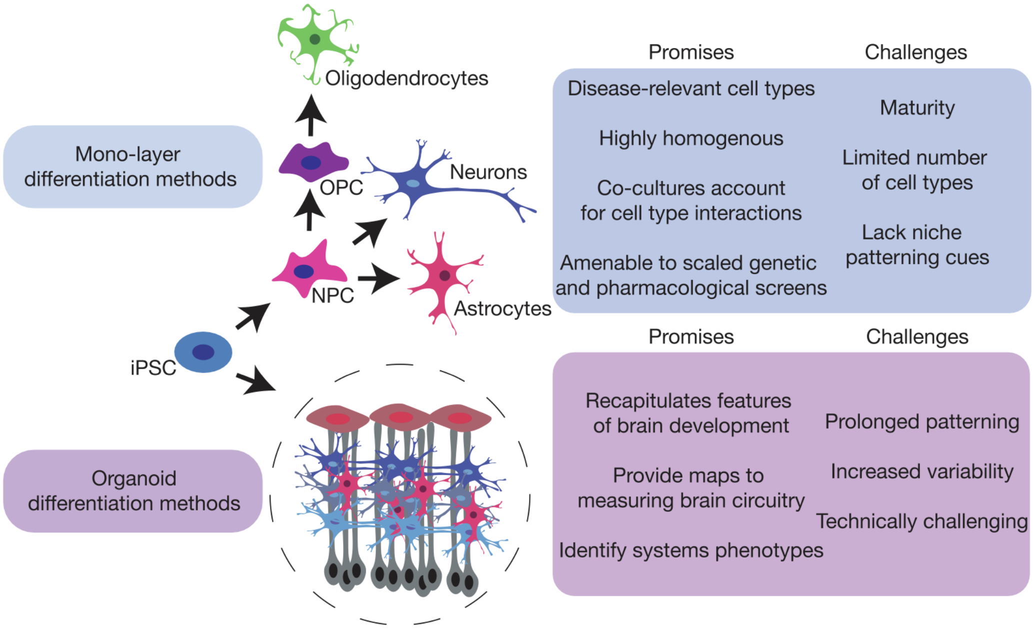
Utilizing both male and female cell lines is key for uncovering the biological mechanisms underlying disease in each sex. The massive epigenomic reorganization which occurs during reprogramming to generate iPSCs brings into question the stability of the X chromosomes, in which one is inactivated in adult females (Tchieu et al, 2010). For X-linked disorders, erosion of the inactivation of the wrong X chromosome may mask or produce unwanted biological signal which is key for uncovering the biological mechanisms of disorders such as mental retardation, where 11% of implicated genes are X-linked (Gecz et al, 2009). Several recent studies using hPSC-based models to study autism spectrum disorder (ASD) have excluded XX hPSCs lines due to erosion of X chromosome inactivation in many XX iPSC lines (reviewed in Nehme et al, 2020), limiting the insights gained from these functional studies to XY genetic backgrounds. However, many other studies have included XX cell lines to successfully identify sex-specific iPSC-based biology (Ronen et al, 2014; Li et al, 2017; Genolet et al, 2021). To understand how sex differences influence cellular programs in health and disease, along with response to drug treatment, it is thus important that future studies strive to utilize cohorts incorporating both XX and XY cell lines. Appropriately designed hPSC studies could provide insight into the biological mechanisms, including sex-specific cellular programs, by which genome structure may confer functional changes in cells derived from both sexes.

1.3 - Differentiation of human pluripotent stem cells into therapeutically relevant cell types

hPSCs can be differentiated into disease-relevant neuronal cell models via several strategies: two-dimensional (2D) monolayer formats, or more complex 3D (three-dimensional) systems (**Figure 1.1**). There are two main mechanisms for generating two-dimensional neuronal cultures; directed differentiation, in which many diverse cell types are generated from hPSC; and transcription-factor mediated differentiation, which often generate somewhat homogenous populations of single cell types. The generation of three-dimensional organoid models follows a similar trajectory whereby scientists may use undirected or patterned protocols to generate specific cellular structures. Here, we outline current hPSC derived cellular systems being used in therapeutic approaches, with an emphasis on cell types used to study neurological conditions. From neuropsychiatric to neurodegenerative disorders, brain disease is among the most challenging to research given the complex nature of the human brain and the lack of accessibility to cell types from the developing brain. While we focus on brain-derived cell types, the approaches, and limitations we describe could be extended to different tissues and cell types.

1.3.2 - Two-dimensional (monolayer) neuronal cell models

Directed differentiation: Establishing the culture conditions for embryonic stem cells (ESCs) provided a major breakthrough in the advancement of in vitro research (Evans et al, 2007; Martin et al, 1981). First attempts to differentiate these ESCs into neural cell types involved aggregates of cells known as embryoid bodies (EBs), which were chemically patterned with retinoic acid (RA) and dissociated to yield small populations of neural crest cells (Bain et al, 1995). Interestingly, in cultures not treated with RA, similar neuronal-like cell types were observed, suggesting neuronal differentiation may be a default mechanism. This differentiation can be achieved through endogenous and exogenous presence of both induction and inhibition of particular factors.



Oligodendrocytes
Mono-layer differentiation methods
OPC
Neurons
NPC
Astrocytes
iPSC
Organoid differentiation methods
Promises
Challenges
Disease-relevant cell types
Highly homogenous
Co-cultures account for cell type interactions
Amenable to scaled genetic and pharmacological screens
Maturity
Limited number of cell types
Lack niche patterning cues
Promises
Challenges
Recapitulates features of brain development
Provide maps to measuring brain circuitry
Identify systems phenotypes
Prolonged patterning
Increased variability
Technically challenging

Figure 1.1. Promises and limitations of the main approaches for hPSC differentiation.

Monolayer and organoid (3D) based methods for generating disease relevant neural cell types provide insights into disease mechanisms and potential therapeutic targets for neurological disorders.

In vitro differentiation into neuronal cell types recapitulates the multi-step process which occurs within the developing brain. First, ESCs are differentiated into neural rosettes, neural tube-like structures composed of neuroepithelial cells with polarized apical and basal proteins. Following, radial glial cells which surround the lumen show characteristics of interkinetic nuclear migration, then giving rise to neural precursor cells. As these neuronal precursor cells differentiate, they can be pushed towards specific cell fates such as glutamatergic and GABAergic neurons, in addition to diverse types of glia (astrocytes and oligodendrocytes) contained within the brain. These innovations enabled a first glance at biological mechanisms underpinning brain development with living human tissues and have been addressed in several recent reviews (Sidhaye et al, 2006; Velasco et al, 2021; Chiaradia et al, 2020).

Transcription factor-mediated differentiation: Emerging genetic and transcriptional findings implicate specific, disease-relevant cell types across the spectrum of human illnesses. For example, although neuropsychiatric disorders may have systems-level deficits which may involve dysfunction in many diverse cell types, recent genetic studies have implicated cortical excitatory and inhibitory neurons in the pathogenesis of schizophrenia and autism (Finucane et al, 2018; Satterstrom et al, 2020). Other disorders, such as bipolar disorder or Parkinson's disease can be characterized by aberrant function of inhibitory neurons and dopaminergic neurons, respectively (Lee et al, 2018; Yao et al, 2021). Cellular differentiation programs which utilize transcription factor (TF) mediating overexpression enable researchers to generate highly homogenous neuronal cell cultures for these specific cell types of interest. Several methods have been established for the generation of cortical-like excitatory neurons, first through overexpression of NEUROG2, which was later iterated on and combined with patterning factors to dorsalize the cell fate (Zhang et al, 2013, Nehme et al, 2018). Complementary to excitatory glutamatergic neurons are GABAergic inhibitory cells, which can be generated through transient expression of the TFs ASCL1 and DLX2 (Yang et al, 2017). More specific cells, such as dopaminergic neurons for studying Parkin- son's disease are derived by driving expression of rLmx1a, rNurr1 or rPitx3 during differentiation (Mahajani et al, 2019). There are continuous adaptations to these approaches to make more mature and pure populations of these cell types, as well as define novel ways to generate previously unidentified neuronal subtypes within the brain. In many instances, disease-specific signatures require a combination of discreet cell type characterization as well as larger, more complex cell-type to cell-type interaction studies to fully elucidate underlying mechanisms of disease. The advancement of methods to generate neuronal support cells, such as astrocytes and oligodendrocytes are rapidly changing how scientists approach establishing co-culture systems to begin studying these cell–cell interactions in vitro (Ehrlich et al, 2017; McPhie et al, 2018; Canals et al, 2018; Quist et al, 2021). However, there are instances where moving beyond two-dimensional models is fundamental to understanding how brain development and structure may be perturbed in human brain disorders.

Monolayer differentiation approaches are often simpler than more complex three-dimensional (3D) organoids, and yield more homogenous cultures, which facilitates the implementation of genetic and pharmacological screens across large sample sizes. In recent efforts, researchers have utilized such monolayer methods for nominating in vitro phenotypes and drug targets for a spectrum of brain disorders, including Parkinson's Disease, epilepsy, bipolar disorder, and autism (Mertens et al, 2015; Pasca et al, 2011; Tang et al, 2016; Surmeier et al, 2017). Additionally, genetic, and viral screens have revealed cell-type specific interactions within the context of viral infections, such as Zika virus (Wells et al, 2023).

Neuronal cells generated via many two-dimensional models express neuronal markers, are functionally active and display key synaptic contacts and neuronal processes (Nehme et al, 2018; Zhang et al, 2013; Mahajani et al, 2019). Yet the absence of a higher-order cortex-like architecture places limits on the investigation of neuronal activity and circuitry. Co-culture systems, whereby two or more cell types are pooled together in vitro, as well as 3D differentiation methods help address some of these limitations and further allow insights into cell autonomous and cell non-autonomous effects in cell–cell interactions. There currently exist many methods for co-culturing neural tissues, including a model of the blood–brain barrier, neuron-microglia systems to understand immune function, and neuron-glioma models to investigate tumor migration (Miranda-Azpiazu et al, 2018; Gava-Junior et al, 2020; Nadadhur et al, 2019). This is of high importance to the study of many disorders, such as neuropsychiatric conditions, which often are characterized by imbalances in synaptic transmission between neighboring types of neural cells (Fan et al, 2018; Lisman et al, 2008). In a recent cross- platform study, whereby experiments were repeated and validated across separate institutions, researchers showed in NXRN-1 mutant neurons derived from patients with schizophrenia, impaired neurotransmitter release, altering electrophysiological activity (Pak et al, 2021). Further, methods to generate homogenous astrocytes from iPSC have displayed altered calcium signaling in cells generated from schizophrenic patients (Szabo et al, 2020). Similar findings have been made in the context of neuromuscular disorders, such as ALS, where electrochemical signaling is disrupted between adjoining neural and muscular cells (Hawrot et al, 2020; Birger et al, 2019; Kiskinis et al, 2018).

1.3.3 - 3D systems: cerebral organoids

More complex models such as cerebral organoids promise to overcome some of the limitations of two-dimensional models. These organoids recapitulate many features represented within the developing fetal brain (Quadrato et al, 2017). During early differentiation, they often exhibit typical layers of the embryonic cortex, with ventricular zone-like bands which can be distinguished from the overlying subventricular zone region (Benito-Kwiecinsky et al, 2021). Analogous to the early cortex, radial-glia precursor cells maintain their polarized structure and display the same nuclear migration kinetics. Consistent with the fetal human brain, in which neurons are generated in time and location specific manners, prolonged organoid cultures generate greater diversity and maturity of specific neural cell types (Giandomenico et al, 2021).

Though the most dominating cell type in current organoid protocols have been glutamatergic excitatory neurons, populations of inhibitory neurons, oligodendrocytes, astrocytes, and their precursors have all shown a capacity to differentiate alongside one another in these more complex systems (Quadrato et al, 2017). The vast diversity of neural cell types generated by current organoid methods have been extensively characterized through single-cell RNA-sequencing studies (Velasco et al, 2019; He et al, 2020; Smits et al, 2020). Additionally, organoids have also been used to explore mechanisms of migration of inhibitory neurons to the cerebral cortex, as often the mature development of these diverse cell types is dependent on the time and development of additional cell types or secreted compounds which drive their cell fate selection (Giandomenico et al, 2021). Further efforts are needed to identify the optimal complexity and maturity of these systems as key features of the adult human brain, such as myelination of axons, have yet to be seen in long-term cultures of cerebral organoids (Giandomenico et al, 2021). Moreover, some dis- ease relevant cell types, such as microglia, which may play a role in the pathogenesis of schizophrenia, are not usually observed (Quadrato et al, 2017). This is likely due to cerebral organoids being derived from the neuro-ectoderm, whereby microglial cells are mesoderm derived. Thus, most organoids generate varied cell compositions which mimic the extent of variation within the human brain yet still lack key features and disease specific cell types (Marx et al, 2020). For example, neural organoids do not develop distinctive cellular subtypes which underlie key brain circuits found in humans, likely driven by active stress responses during organoid development (Bhaduri et al, 2020). One approach to address this is to introduce lacking cell types later on once the organoid has reached a particular level of maturity. For example, injecting microglia cells, typically absent in organoid differentiation protocols, mimics in vivo characteristics such as immune response and synaptic clearing (Abud et al, 2017). Additionally, a key limitation to increased complexity of modern cerebral organoids is a lack of vascularization to provide nutrients to the tissue, furthering its development. Recent evidence has shown that supplementation with mesodermal progenitor cells generates vascular networks throughout cortical organoids, providing a potential path to increased maturity amongst these models (Worsdorfer et al, 2019).

The adoption of organoid approaches for studying brain development and disease, while very promising, still presents some challenges within their ability to be used for genetic and drug perturbations, as well as limitations in reproducibility and variability across cell lines and replicates. These models often take many months to develop and can be difficult and expensive to produce at scale, thereby limiting their utility in therapeutic perturbations (Centeno et al, 2018). It is also a significant challenge to identify which of the cell types generated are being preferentially perturbed by a given intervention in the absence of more technically difficult and expensive downstream assays, which may not be easily accessible compared to more succinct methods applicable to conventional two-dimensional models. Though the protocols for generating these complex structures are constantly evolving and improving, there is still variability between different cell lines and replicates of the same cell line, which make comparisons between organoids a challenge (Velasco et al, 2021). In many cases, phenotypic variation between organoids could be driven by the varied abundance of cell types generated and may not be driven by a specific genetic or pharmacological effect of interest to the researcher.

Choosing the appropriate model to investigate brain disorders is paramount for answering important biological questions and several considerations should be deliberated in addition to those covered above. These include the timeline for model generation, where many two-dimensional models can be produced within a few weeks and many current organoid methods may take several months to fully mature (Centeno et al, 2018). Organoids may also require additional equipment which may not be accessible in each research environment such as bio-reactors and hydrogel scaffolding to support the structure and formation of three-dimensional systems (Centeno et al, 2018).

1.4 - Strategies for scaled hPSC based translational studies

For the majority of individuals suffering from complex diseases, there are comparatively few known genetic variants which can be identified as having casual impacts. In some cases, as in mendelian disorders, rare genetic mutations can be identified and often have a severe impact on the individual's health. For example, neurological conditions such as Huntington's Disease or ALS are linked to mutations in the HTT gene and C9orF72/SOD1 genes, respectively (Gusella et al, 1983; Brown et al, 2021). Alternatively, through genome-wide association studies (GWAS), researchers have implicated regions across the genome which may play a role in complex genetic disorders, whereby no single gene can be attributed to their pathophysiology (Chang et al, 2017; Ripke et al, 2020; Grove et al, 2019; Jansen et al, 2019). Functional in vitro investigations of these GWAS loci have been plagued by current barriers in cellular systems, such as limited sample sizes, technical variation, and the cost to scale up experiments so that researchers may be sufficiently powered to detect biologically significant signals (Nehme et al, 2020; Brennand et al, 2011). Additionally, run-to-run variation and batch effects across experimental parameters often introduce high levels of 'noise' into the system which may often be indistinguishable from the relevant disease or variant signatures being studied. Efforts to overcome these limitations are underway, and several studies have begun to establish frameworks for studying complex disease genetics in vitro (Brennand et al, 2011; Novikova et al, 2021; Hoffman et al, 2017).

Historically, cellular assays have been performed in an arrayed format, whereby cells derived from each donor are cultured and characterized separately. Advances in genomics have enabled the adaptation of novel methods of single-cell transcriptomics and cellular phenotyping, leveraging genetic sequencing technologies, to enable the analysis of a mixture of cells derived from different donors in the same culture environment. One such approach is 'pooled' or 'cell village' model systems (Chan et al, 2018; Mitchell et al, 2020; Jerber et al, 2021). Within this context, cells from several donors are cultured together in the same culture dish (**Figure 1.2**). These cells then undergo desired perturbations, or phenotyping and on the back end are genetically characterized to identify individuals based on their phenotype of interest. These tools enable the mapping of genetic influences, such as the presence of rare or common disease variants, on cellular phenotypes. Currently, these tools have been leveraged for both protein and RNA-based phenotyping. In one approach, researchers 'pooled' together 215 iPSCs lines across 17 different pools (or villages) and performed single-cell RNA-sequencing (Jerber et al, 2021). In this study, they identified over 1200 new disease-eQTL (expression Quantitative Trait Loci) colocalizations not present in the GTEx (Genotype-Tissue Expression) consortium as well as revealed intrinsic mechanisms which contribute to differentiation capacity. A second utility of village-based cellular models is the ability to map alternative phenotypes, beyond gene expression, to genetic background (Mitchell et al, 2020). Here, after combining cells from many donors in the same pool, cells were separated based on a protein phenotype using FACS (Fluorescence Activated Cell Sorting). Leveraging low-pass DNA sequencing, individual fractions of cells sorted for high or low expression were tested for their donor composition relative to their pre-sort composition. Researchers used this method to identify that gene copies within the SMN locus dictate treatment response to an anti-sense oligonucleotide for treatment of spinal muscular atrophy (SMA) (Mitchell et al, 2020).

These 'village' approaches increase the scale at which a single scientist can perform in vitro based studies and provide a powerful approach to reducing both the cost and variance which often plague current systems. The capacity to phenotype and analyze many cells at once, from multiple donors, limits the need for high numbers of replicates both within and across cell lines by ensuring their environment and technical conditions are identical throughout the experimental procedures. Further, the 'cell village' approach can be leveraged to scale the interrogation of the effects of diverse cell lines following exposure to identical pharmacological and genetic perturbations (Mitchell et al, 2020). Parallel advances in genetic screens, such as those utilizing CRISPR-based manipulations, are amenable to village-based experimental approaches. For example, in combining CRISPRa and CRISPRi perturbations on cell lines from different donors, researchers could investigate the degree to which changes in gene expression interact within the context of distinct genetic backgrounds.

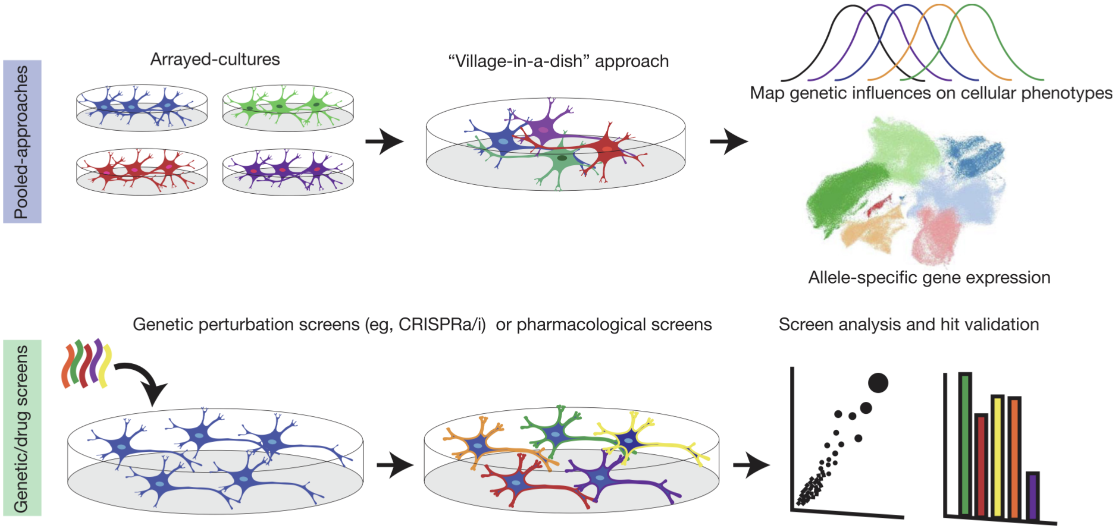

Pooled-approaches
Arrayed-cultures
"Village-in-a-dish" approach
Map genetic influences on cellular phenotypes
Allele-specific gene expression
Genetic/drug screens
Genetic perturbation screens (eg, CRISPRa/i) or pharmacological screens
Screen analysis and hit validation

Figure 1.2. Approaches for scaling hPSC-based studies.

Top: Cellular village methods enable scientists to infer how genetic variation influences cellular phenotypes and gene expression. Bottom: Genetic and pharmacological screens allow investigation of the functional impact of genes and drug treatments.

Though these tools provide great promise for scaling hPSC based translational studies, there remain several limitations to their utilization. While having large collections of iPSCs can be very useful for disease modeling, they can be very time consuming and expensive to generate. For an individual laboratory, producing a collection of stem cells for sufficiently powered studies can be very expensive, and take several years to complete, with many hurdles related to proper use and consenting of individual cell donors (Odeh et al, 2015). To overcome this, significant efforts have established large iPSC repositories across the globe, which together harbor more than 5,200 diverse iPSC lines (Huang et al, 2019). These collections make such cell lines more accessible to the scientific community. Beyond infrastructure hurdles, the biological consequences of village-in-a-dish approaches remain unclear. Sufficient efforts have not yet investigated the degree to which culturing cells from different donors and conditions may impact the function of any given cell. For example, culturing disease cells with control cells may mask biological signals from either condition, which may negate true effects of interest. Thus, while cell village approaches might be well poised for scaled measurements of cell-autonomous phenotypes, it is unlikely that cell non-autonomous effects could be currently measured using these approaches. Until more research is done, it will be important to complement village-based approaches with arrayed experiments to uncover mechanisms by which cells impact one another. Alternatively, understanding how disease cells and control cells interact within the same culture dish could uncover novel disease mechanisms or drug targets.

1.5 - Current therapeutic applications and considerations for hPSC models

The field of iPSC-based drug discovery is still in its infancy. Yet, its broader adoption amongst the therapeutic industry over the last decade has facilitated several ongoing advances in our efforts to develop novel medicines. We highlight key applications here as this has been the subject of several reviews (Rowe et al, 2019; Pintacuda et al, 2021; Villa et al, 2021).

1.5.1 - hPSC-based drug screening

A rapidly growing sector of iPSC-based modeling is that of pharmacogenomics screens, whereby the identification of drug targets and their efficacy in the context of diverse genetic backgrounds provides a path forward for improving and understanding how therapeutic treatments impact specific cell types of interest. Many currently approved treatments for brain disorders act through previously unknown mechanisms of action (Davis et al, 2020). This is due, in part, to the lack of access to disease-relevant tissues. One approach has been to use animal models, yet these don't fully recapitulate the disease and treatment responses often observed in human patients. However, the ability to differentiate iPSCs into human brain specific cell types offers a new opportunity to determine the mechanisms of action of treatments for neurological conditions, and to potentially suggest novel drug targets within the context of pre-existing treatments.

Over the past decade, hPSCs have been adopted in many drug screens to identify drug targets, mechanisms of action, and for repurposing already FDA (food and drug administration) approved compounds. The first to leverage this approach was McNeish et al, in which researchers tested >2 million compounds on hESC derived neurons to identify novel AMPA receptor potentiators (McNeish et al, 2010). High-throughput screens have subsequently been adopted to identify compounds which rescue phenotypes in the iPSC-derived neurons from patients with psychiatric diagnosis through modulation of WNT and BMP signaling (Zhao et al, 2012). In another study, researchers used iPSCs to nominate repurposing compounds for the treatment of Anti-RNA viruses. Here, they showed selective estrogen receptor modulators (SERMS) blocked entry of SARS-Cov2 into host cells, suggesting already FDA-approved drugs may be able to modulate host cell susceptibility of viral infections (Imamura et al, 2021). These efforts highlight how using hPSCs can help nominate biological targets for drugs through the use of large-scale compound screens.

Previously, we addressed the need for expanding the genetic diversity of iPSC collections for understanding how genetic variation, including in the form of ancestry and sex, may drive biology. Large scale drug screens on patients of diverse genetic backgrounds will be key in resolving these ancestry-specific responses to medicines, with a focus on improving efficacy and reducing harmful side-effects. The importance of exploring how genetic diversity impacts pharmacological screens has been illustrated in several studies (Wadhawan et al, 2001; Goncalves et al, 2020).

One example of this approach leveraged the 'cell village' system discussed above. In this study, genetic differences in the form of structural variation were shown to mediate treatment response to an antisense oligonucleotide (ASO). Here, copy number variation within the SMN locus, which regulates expression of the SMN protein, deficient in Spinal Muscular Atrophy (SMA), was targeted by the ASO dictated treatment response. Cell lines with higher levels of gene copies displayed increased treatment response when compared to lines with lower copy numbers. The results of this study also validated the mechanism of action for the ASO which has previously only been hypothesized (Mitchell et al, 2020). This example highlights the utility of iPSC-based models for mapping genetic influences on cellular phenotypes while also enabling the investigation of therapeutic mechanisms.

Each approach for drug screening, whether using a single or small number of donors, or larger scaled studies with donors from diverse ancestries, have their strengths and weaknesses. When using few donors, downstream analyses can control for biological signals which may be driven by the donor's intrinsic genetic variation. In these cases, this intrinsic biology may alter or mask the phenotypes as a result of the drug compound, providing unclear answers about the drug mechanism of action and treatment response. However, as we have highlighted above, incorporating an array of diverse genetic variation can also bring about true effects of treatment with drug compounds. It will be imperative when using hPSCs for drug-based screens to investigate both aspects, the target and mechanism of action for a compound on a stable, highly characterized hPSC background, and then to follow up exploring the ways in which these compounds may have varied effects on a range of cell lines from diverse ancestries.

1.5.2 - iPSCs in the clinic

The availability of hPSCs has presented tremendous opportunities for improving drug discovery and development. These tools can be leveraged for population scale safety and efficacy screens, often referred to as "clinical trials in a dish" (CTiD). In a measure to reduce development and cost, CTiD enables researchers to lead selection of drug candidates with much higher levels of success. There are several ongoing clinical trials utilizing drugs discovered using iPSC-based models. A compound developed by Roche, RG7800, demonstrated increased survival of iPSC-derived motor-neurons in a model for Spinal Muscular Atrophy (SMA) (Naryshkin et al, 2014). An FDA-approved treatment for epilepsy, Erzogabine, has shown efficacy in an iPSC model of ALS, and current clinical trials supported by GlaxoSmithKline are currently underway to repurpose Ezogabine (McNeish et al, 2015). More recently, scientists at CiRA, an iPSC research institute in Japan, screened thousands of small molecule compounds on a panel of patient-derived iPSCs and uncovered a novel treatment for fibrodysplasia ossificans progressiva (FOP), a devastating disease where soft tissue becomes ossified and turns into bone (Hino et al, 2018). Clinical trials have utilized iPSCs since their initial discovery in 2007 and the spectrum of disease classifications using these tools, as well as the number of donor sample sizes continue to increase (**Figure 1.3**) (Clinicaltrials.gov).

In addition to the utility of iPSCs for drug discovery and repurposing, they provide a perfectly suited system for modeling drug toxicity, which often plagues clinical trials and is a main reason for why only 16% of new medicines make it to market (Wouters et al, 2020). Within the brain, iPSC-derived neural cells provide a model way to perform toxicity screens (Liu et al, 2013; McGivern et al, 2014). Recently, studies have demonstrated neurotoxic effects of ketamine exposure, first in rodents and primates, then recapitulated using iPSC-derived neurons under high concentrations (Skiller et al, 2015; Ito et al, 2015). The negative effects of ketamine use may become a problem of increasing concern as ketamine becomes a widely used treatment for depression, a disorder which affects roughly 20% of the population at some point in their lives (Parikh et al, 2021). In another study, a compound screen across multiple neural lineage cell types derived from iPSCs showed that greater than half of the small molecule compounds displayed cytotoxic activity in multiple cell types within the brain, suggesting a strong need to further utilize these tools to investigate the long-term effects of potential new treatments and their effects in diverse cell types (Pei et al, 2016).

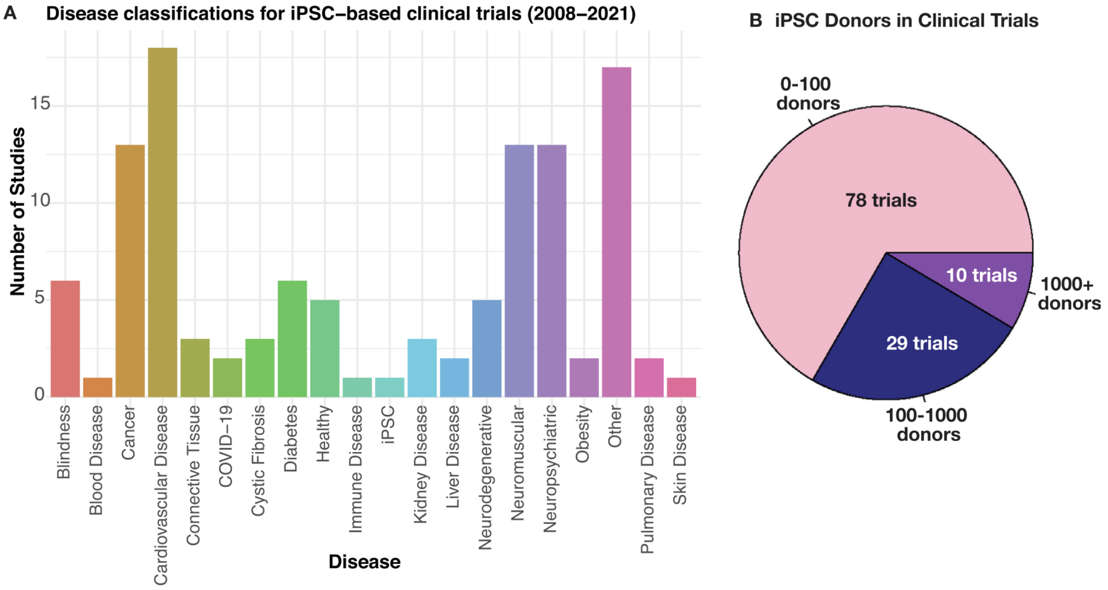


Figure 1.3. iPSC-based Clinical Trials.

A. Disease classifications for previous and ongoing clinical trials involving iPSCs for disease modeling and drug discovery. B. Number of donors per iPSC-based clinical trials. Data compiled from Clinicaltrials.gov.

The emerging application of iPSCs for disease modeling and cellular therapies is not without its challenges. A primary concern amongst the stem cell and therapeutic community is genomic integrity of iPSCs. Aberrations in genome integrity, such as copy number variants, often propagate tumorigenicity and immunogenicity amongst populations of iPSCs, potentially causing harm if transplanted into a patient or otherwise masking true experimental effects of a drug treatment. Several studies have identified factors which may drive tumorigenicity in iPSCs, such as the method of reprogramming, whereby addition of dominant mutant p53 or c-Myc have been shown to increase the presence of tumors in mice (Merkle et al, 2017; Yamanaka et al, 2020). Other mechanisms, such as karyotypic abnormalities, impact the activity of cancer-causing genes, such as TP53, and may promote cancer-like behaviors if transplanted into patients. Next generation sequencing technologies have enabled rapid and reliable methods for identification of these perturbations within iPSCs and to track their behavior in culture, as high passage number cell lines have been shown to be more susceptible to these alterations (Merkle et al, 2017). Fortunately, these concerns have been met with the establishment of increasingly rigorous cross-field standards to characterize genomic integrity and further to outline parameters for which stem cells harboring these aberrations should be dis- carded and not used for therapeutic purposes or downstream experimental applications.

1.6 - Perspectives

The use of hPSCs is revolutionizing how researchers model human disease and develop novel medicines to combat them. Surpassing the status quo of animal models for therapeutic development, hPSCs enable clinicians and scientists to leverage the significant genetic diversity of human populations to understand how it impacts drug treatment response. These approaches hold great, yet underexploited, promise to ameliorate many of the negative side effects of common medications on large swaths of the global population.

Here, we highlight key considerations for inclusion of hPSC-based systems in therapeutic discovery and development work. First, the necessity of using scalable and precise approaches to stem cell derived systems. It is increasingly important to adequately expand the sample sizes commonly used in in vitro cellular studies to ensure reliability and validity of the data. Second, a deep understanding and characterization of the cellular differentiation methods used is necessary. Cellular populations which are heterogenous may mask the impact of certain genetic variants or responses to particular treatments, resulting in uninterpretable data. Depending on the cellular phenotype being investigated, the use of either reductionist, yet scalable monolayer cultures or more complex models such as co-cultures or 3D organoids might be more appropriate. Third, it is critical to scrutinize genomic integrity in iPSCs. Though it is not often the case that cell lines acquire highly penetrant karyotype abnormalities, more subtle aberrations of the genome are common and may confound or construct experimental results and impact downstream therapeutic applications. It is thus imperative to carefully monitor cellular behaviors (including at the hPSC state) and perform routine low-level DNA sequencing or karyotyping to catalog any potential abnormalities. Fourth, examining key findings in the context of broader genetic studies across diverse ancestral populations, and both XX and XY genetic backgrounds, whenever possible. When following up with results centered around particular genetic variants or drug targets, it is critical to consider results within the context of diversity across genomic populations.

We strongly advocate for the continual adoption of hPSC models by the pharmaceutical industry. Whether through collaboration or partnerships, large collections of genetically diverse hPSCs will continue to become pivotal to address many of the challenges we have discussed here. Leading academic institutions and pharmaceutical companies have begun to unite to aggregate a collection of compound screens in order to nominate small molecule drug targets from several classes of compounds. Identifying how these compounds impact diverse ancestries using collections of iPSCs has the potential to revolutionize therapeutic discovery. We would further like to inspire organizations developing novel therapies to include iPSC-based toxicity and efficacy screens alongside early and late phase trials. The added expense of generating a few patient iPSC lines pales in the context of the cost for clinical drug development. Moreover, hPSC based studies could reveal mechanisms by which adverse effects can be identified, and drugs could be modified to potentially increase the success rate of clinical trials. hPSC-based models are not without limits. There remain several hurdles which require due diligence to facilitate the broader academic and industry adoption of these technologies to revolutionize how we understand and treat human diseases. As these advancements continue, we believe that hPSCs can provide a large step closer to a world in which precision medicine, at the personal and population level, could become more readily accessible.

1.7 - Aims & hypothesis

Risk for complex traits is mediated by a great number of both common and rare genetic variants which impact the regulation of nearly all genes within our genomes. Genetic studies have made tremendous progress in nominating thousands of variants associated with risk of illnesses to non-medical phenotypes. A critical gap remains - to complement efforts in human genetics with downstream functional and mechanistic studies to understand the underlying function of these risk alleles. Pioneering studies have shown that lymphoblastoid cell lines or pluripotent stem cells (PSCs) from human donors can be used to identify how common DNA variation shapes certain cellular phenotypes, especially RNA expression (Cheung et al., 2005, Morley et al., 2004, Stranger et al., 2007a, Stranger et al., 2007b, Kilpinen et al., 2017, Lo Sardo et al., 2017, McFarland et al., 2019, Pickrell et al., 2010, Jerber et al, 2021, Wells et al, 2023). There are many challenges to fully understanding the mechanisms which underly genetic risk for biological traits which include reaching the necessary scale for experiments, having experimental models which represent disease relevant tissues, and uncovering overlapping mechanisms between common and rare variant cases of similar clinical phenotypes. Overcoming these would help to facilitate progress in functional genomics, especially within the context of neuroscience.

In my project, I work to overcome these challenges. I will do so across four distinct, but complimentary, aims:

(1) To develop experimental systems which would enable scientists to systematically experiment on cells from many genetically unique donors simultaneously. This would facilitate the study of many people at once to scale *in vitro* experiments.

a. I hypothesize that we could combine next generation sequencing approaches map experimental phenotypes to individual cell lines which were cultured and assayed together.

(2) To create a new method for generating astrocytes from iPSCs.

a. I hypothesize that I could adapt existing methods for generating excitatory neurons to generate astrocyte like cells from iPSCs. As astrocytes and neurons often originate from a common progenitor cell in the developing cortex, perhaps there is a critical window during cellular differentiation which mimics this intermediary and therefore cells could be diverted to an astrocyte fate during this time.

(3) To implement novel image-based assays to explore whether genetic variation impacts cellular phenotypes

a. I hypothesize that unbiased cell morphology assays, will be a promising tools for functional genomics studies due to their reduced cost relative to transcriptional assays. If we are able to capture imaging data from enough genetically diverse iPSCs, we might be able to use morphology-based readouts to nominate the function of genetic variants.

(4) To discover whether there are shared biological mechanisms between common and rare variant cases of schizophrenia.

a. I hypothesize that the underlying cellular and molecular mechanisms of psychotic disorders may be shared between cases of psychosis which results from rare mutations and cases which occur irrespective of rare mutations. With similar or identical clinical manifestations, it is crucial to understand their similarity at the cellular level so that we can better understand how to treat these conditions. It stands that rare mutations with large effect are more likely to elicit a phenotype in an experimental system. If we can identify shared mechanisms, it will pave the way for future studies of rare variants whereby findings can be extrapolated back to common variant cases of those same conditions.

My project utilizes iPSCs for the investigation of complex human traits, to expand and adapt existing approaches in functional genomics to comprehensively map how human genetics influence living cells across multiple levels of analysis. My goal is to fill several critical gaps in the field and improve our understanding of genetic contributions to cellular phenotypes.

Chapter 2

Materials and methods

2.1 - Cell culture

2.1.1 - iPSC derivation and collection

iPSC lines were obtained from the Stanley Center for Psychiatric Research stem cell collection at the Broad Institute (https://sites.google.com/broadinstitute.org/sc-stem-cell-resource/). Cell lines were derived from donated skin fibroblasts or PBMCs from consented research volunteers using Sendai and mRNA vectors expressing OCT4, SOX2, KLF4 and c-MYC. Each cell line was given a project alias and collaborator ID in order to anonymize donor identification. Quality control, including pluripotency tests and genetic karyotyping, were performed in order to measure the quality of the cell lines used in this work. Cell lines with limited differentiation capacity and/or the presence of known deleterious copy number variants were excluded from the study.

2.1.2 - iPSC culture

Each iPSC line was cultured in feeder-free conditions on Geltrex (15mg/ml, ThermoFisher) in mTeSR1 media (StemcellTech). For routine maintenance, cultured cells underwent daily medium changes and were passaged when reaching 70-80% confluence. Here, new 6-well NUNC plates were coated with Geltrex (15mg/ml) for 1hr at 37C. iPSC colonies were dissociated with Accutase (StemcellTech) for 5-10 min at 37C. After incubation, cells were titruated to remove any excess cells from the plate bottom. Accutase-cell suspensions were added to mTeSR1 medium + 10uM Y-27632 in a 15mL Falcon tube. Cells were centrifuged at 300g x 5 min. The cell pellets were then resuspended in mTeSR1 medium + 10uM Y-27632 and plated across new plates at a desired split ratio (between 1:5 and 1:20). Cells were maintained in a humidified incubator at 37C and 5% CO2. iPSCs between passage 10 and 35 were used in this work.

2.1.3 - Cryopreservation of iPSCs

Cell stocks were generated for long term cryopreservation in liquid nitrogen. iPSC colonies were dissociated using Accutase (StemcellTech) for 5-10 min at 37C. After incubation, cells were titruated to remove any excess cells from the plate bottom. Accutase-cell suspension was transferred to a tube with mTeSR1 medium + 10uM Y-27632 in a 15mL Falcon tube. Cells were centrifuged at 300g x 5 min. The cell pellets were then resuspended is CryoStor 10 (StemcellTech) and placed into cryotubes. Cryotubes were placed in a Mr.Frosty in a -80C freezer for 24hrs before transferring to -180C liquid nitrogen tank.

2.1.4 - Differentiation of neuronal progenitor cells

Based on adapted protocol from Nehme et al, 2018. Once iNGN2-iPSCs reach 70-80% confluent, cells were dissociated into a single-cell suspension using Accutase and incubated at 37C for 5-10 min. Following incubation, cells are titruated to remove excess cells from the plate bottom. Accutase-cell suspension was transferred into a 15mL Falcon tube filled with mTeSR1 + 10uM Y-27632. Cells were centrifuged at 300g x 5 min. Cells were next resuspended in mTeSR1 + 10uM Y-27632 and counted using a Countess Fluorescent cell counter. Cells were next seeded at 1M cells per well in a geltrex coated 6-well NUNC plate. The following day (D1), cells are fed with Neural Induction Media (NIM) supplemented with XAV, LDN, SB4312, and 2 ug/ml doxycycline. On D2 and D3 the cells are fed with the same media as D1 but supplemented with

1 ug/ml Zeocin. In the morning on D4, cells are fed with Neurobasal Medium supplemented with GDNF, CNTF, BDNF and FUDR for at least 6hrs. D4 NPCs are then ready to be harvested for downstream assays.

2.1.5 - Differentiation of excitatory neurons

Based on adapted protocol from Nehme et al, 2018. Once iNGN2-iPSCs reach 70-80% confluent, cells were dissociated into a single-cell suspension using Accutase and incubated at 37C for 5-10 min. Following incubation, cells are titruated to remove excess cells from the plate bottom. Accutase-cell suspension was transferred into a 15mL Falcon tube filled with mTeSR1 + 10uM Y-27632. Cells were centrifuged at 300g x 5 min. Cells were next resuspended in mTeSR1 + 10uM Y-27632 and counted using a Countess Fluorescent cell counter. Cells were next seeded at 1M cells per well in a geltrex coated 6-well NUNC plate. The following day (D1), cells are fed with Neural Induction Media (NIM) supplemented with XAV, LDN, SB4312, and 2ug/ml doxycycline. On D2 and D3 the cells are fed with the same media as D1 but supplemented with 1 ug/ml Zeocin. On D4 cells are fed with Neurobasal Medium supplemented with GDNF, CNTF, BDNF and FUDR. Cells underwent half media changes every three days until the cells were terminally differentiated and placed into downstream experiments (D28).

2.1.6 - Differentiation of astrocytes

On day 0, hPSCs were differentiated in N2 medium [500 mL DMEM/F12 (1:1) (Gibco, 11320-033), 5 mL Glutamax (Gibco, 35050-061), 7.5 mL Sucrose (20%, SIGMA, S0389), 5 mL N2 supplement B (StemCell Technologies, 07156)] supplemented with SB431542 (10 µM, Tocris, 1614), XAV939 (2 µM, Stemgent, 04-00046) and LDN-193189 (100 nM, Stemgent, 04-0074) along with doxycycline hyclate (2 µg.mL^{-1}, Sigma, D9891) with Y27632 (5 mM, Stemgent 04-

0012). Day 1 was a step-down of small molecules, where N2 medium was supplemented with SB431542 (5 µM, Tocris, 1614), XAV939 (1 µM, Stemgent, 04-00046) and LDN-193189 (50 nM, Stemgent, 04-0074) with doxycycline hyclate (2 µg.mL^{-1}, Sigma, D9891) and Zeocin (1 µg.mL^{-1}, Invitrogen, 46-059). On day 2, N2 medium was supplemented with doxycycline hyclate (2 µg.mL^{-1}, Sigma, D9891) and Zeocin (1 µg.mL^{-1}, Invitrogen, 46-059). Starting on day 2 human induced neural progenitor-like cells were harvested with Accutase (Innovative Cell Technology, Inc., AT104-500) and re-plated at 15,000 cells.cm^{-2} in Astrocyte Medium (ScienCell, 1801) with Y27632 (5 mM, Stemgent, 04-0012) on geltrex coated plates. Cells were maintained for > 30 days in Astrocyte Medium (ScienCell, 1801).

2.1.7 - iNGN2 integration

iPSCs were cultured using standard methods from the aforementioned. When cells were 70-80% confluent, they were dissociated with Accutase and incubated for 5-10 min at 37C. Following incubation, cells were titruated to remove any excess cells from the well bottom. Accutase-cell suspension was then transferred to a 15mL Falcon tube filled with mTeSR1 + 10 uM Y-27632. Cells were centrifuged at 300g x 5 min. Cell pellets were resuspended in NEON buffer R containing 10ug pGEP116, 1.5ug pGEP43, and 1.5ug pGEP44. The cell and DNA suspension was next loaded into a NEON transfection pipette and electroporate using the following parameters: 1040mv, 30ms, 2 pulse. Following electroporation, cells were plated into Geltrex-coated 10cm dishes. Media was changed the following day and supplemented with 100ug/ml Geneticin for chemical selection of non-integrating cells.

2.1.8 - Lentiviral transduction

Cell lines were transduced with lentivirus cocktail to express tetracycline-inducible murine Neurogenin 2 (Ngn2) tagged with a puromycin resistance gene, in combination with tetracycline inducible GFP, as described (Ho et al., 2016; Nehme et al., 2018). PSCs were grown using StemFlex™ Medium (Gibco™, A3349401) on tissue culture dishes coated with Geltrex™ LDEV-Free Reduced Growth Factor Basement Membrane Matrix (Gibco™, A1413202) and maintained in 5% CO_2 incubators at 37 °C. iPSCs were dissociated to single cell suspension using Accutase (Innovative Cell Technologies, Inc., AT 104). The concentration of cells in suspension was estimated using the Scepter Automated Cell Counter (Millipore Sigma). Cells were plated at a density of 100 000 cells/cm^2 and incubated while in suspension with media containing 10 uM ROCK-Inhibitor (Sigma, Y27632), and lentivirus particles at a final MOI of 2. Three independent viruses were co-transduced to induce the expression of NGN2 using doxycycline (pTet-O-Ngn2-puro; pTet-O-EGFP; Ub-rtTA, gift from Marius Wernig; Lentivirus was produced by ALSTEM). Transduced cells were expanded to 10 cm plates and 10 vials (2M cells/vial) of each cell line was cryopreserved in CryoStor® (StemCell Technologies, 07930) for further use.

2.1.9 - Human primary astrocytes

Human primary cortical astrocytes (hpA) were obtained from ScienCell Research Laboratories (1800) and cultured according to the manufacturer's instructions.

2.1.10 - Village construction and experimental design

Cell lines were maintained as independent cultures for 1-2 passages. At passage 2-3, cell lines were dissociated with Accutase and centrifuged at 300 xg for 5 min. Following centrifugation, cells were suspended in mTeSR1 + 10uM ROCKi and counted using a Countess fluorescent cell counter. Once all cell lines were counted, they were aliquoted in equal proportions by cell number into a new 50mL conical Falcon tube. After all cell lines were added to the 50mL tube, the cell suspension was centrifuged at 300 x g for 5 min to pellet the mixture. Next, the cells can be either suspended in culture medium and plated into new dishes or suspended in freezing medium for cryopreservation.

2.1.11 - Cell seeding and staining

For each batch of imaging, cells were detached from 6-well NUNC plates using Accutase (StemcellTech; cat#07920) for generating single-cell suspensions. Following detachment, cells were centrifuged at 1000 rpm x 5:00 and re-suspended in StemFlex medium supplemented with ROCK inhibitor. After each cell line was counted to determine cell solution concentration and viability, the desired cell solution volume was aliquoted into a 96-deep well low attachment plate. In order to disperse a high number of cell lines across a 384-well plate in a semi-random fashion, we optimized the use of an Agilent Bravo liquid handling device. Here, using an 8-channel head, cell solutions were transferred from the 96-well low attachment plate and distributed into a geltrex-coated Perkin Elmer Cell Carrier 384-well plate for staining and imaging.

2.2 – High-content imaging

2.2.1 - Cell seeding and staining

For each batch of imaging, cells were detached from 6-well NUNC plates using Accutase (StemcellTech; cat#07920) for generating single-cell suspensions. Following detachment, cells were centrifuged at 1000 rpm x 5:00 and re-suspended in StemFlex medium supplemented with

10uM ROCK inhibitor. After each cell line was counted to determine cell solution concentration and viability, the desired cell solution volume was aliquoted into a 96-deep well low attachment plate following a specific plate map to ensure that wells from any given cell line were not predominantly on the edge wells or too close together. To disperse a high number of cell lines across a 384-well plate in a semi-random fashion, we optimized the use of an Agilent Bravo liquid handling device (Figure S1B). Here, using an 8-channel head, cell solutions were transferred from the 96-well low attachment plate and distributed into a geltrex-coated Perkin Elmer Cell Carrier 384-well plate at a density of 10,000 cells per well. Each cell line was plated into 8 distinct wells on the final screening plate in four-well quadrants (see Figure S2). These parameters were selected based on a pilot experiment with 6 cell lines across a range of densities and fixation conditions. We observed that we could maximize variability across cell lines using 10k cells per well fixed 6hrs after plating, when compared to 24hrs post-plating.

2.2.3 - Cell Painting and imaging

Cells were stained and imaged with minor adaptations to procedures described previously (Cimini et al, 2023, Bray et al, 2016). Six hours post seeding in 384-well plates, cells were treated for 30 min with 0.5 uM MitoTracker Deep Red FM - Special Packaging (Thermo Fisher cat#: M22426) dye at 37oC. Following the MitoTracker treatment, cells were fixed with 16% paraformaldehyde diluted to a final concentration of 4% (Thermo Fisher cat#: 043368.9M) for 20 minutes in the dark at RT. After three washes with 1X HBSS cells were permeabilized and stained using a solution of 1X HBSS (Thermo Fisher cat#: 14175095), 0.1% Triton-X-100 (Sigma Aldrich cat#: X100-5ML), 1% Bovine Serum Albumin, 8.25nM Alexa Fluor 568 Phalloidin (Thermo Fisher cat#: A12380), 0.005mg/ml Concanavalin A, Alexa Fluor 488 Conjugate (Thermo Fisher cat#: C11252), 1ug/ml Hoechst 33342, Trihydrochloride, Trihydrate (Thermo Fisher cat#: H3570), 6uM SYTO 14 Green Fluorescent Nucleic Acid Stain (Thermo Fisher cat#: S7576), and 1.5ug/ml Wheat Germ Agglutinin, Alexa Fluor 555 Conjugate (Thermo Fisher cat#: W32464) for 1hr at RT in the dark. Following the staining, plates were washed 3X with 1X HBSS and sealed until imaging. Cell Painted plates were imaged on a Perkin Elmer Phenix Automated Microscope under a standardized protocol (Cimini et al, 2023). Configuration files for imaging protocols can be found with their associated images at https://registry.opendata.aws/cellpainting-gallery/ under project ID "cpg0022-cmqtl." All 297 cell lines were dispersed across seven plates which were imaged in four separate batches.

2.2.4 - Quantification of cellular morphology traits and their quality control

The segmentation of individual cells in the image into its cellular compartments (whole cell, cytoplasm, and nuclei) and subsequently quantification of morphology traits for each cellular compartments was done using CellProfiler 3.1.8 (McQuinn et al, 2018); pipelines are available at https://github.com/broadinstitute/imaging-platform-pipelines/tree/master/cellpainting_ipsc_20x_phenix_with_bf_bin1. Analysis of CRISPR experiments was done in CellProfiler 4.2.4 with pipelines available at https://github.com/broadinstitute/imaging-platform-pipelines/tree/master/cellpainting_ipsc_20x_phenix_with_bf_bin1_cp4 (Stirling et al, 2021). Subsequently, cells missing measurement for more than 5% of traits were removed. Morphology traits a priori known to be problematic, not measured across all cells or non-variable across cells were removed using Caret v6.0-86 package. QC-ed cells were then segregated in two groups based on the number of neighbors: isolated cells having no neighbors and colony cells having one or more neighbors. Individual morphology traits were then summarized to well level measurement by averaging them across all cells per well, resulting in a well by trait matrix. Following this, each morphology trait was gaussianized across all 7 plates using inverse normal transformation (INT) method. For a list of morphological traits measured and for guidelines on how to interpret them, see https://carpenter-singh-lab.broadinstitute.org/blog/help-interpreting-image-based-profiles.

2.2.5 - Selection of traits for association analysis

A set of morphology traits for association analysis (with both common variants and rare variant burden) was selected by considering their pairwise correlation across colony and isolate cells in the following steps: Step 1. Calculate Pearson correlation matrix for colony and isolate cells at donor level (total 2 correlation matrices). Step 2. Identify that single trait having the Pearson r >= 0.9 with the largest number of other traits across both correlation matrices. We specifically chose Pearson r >= 0.9 as cutoff here because most traits (93.7% and 91.2% traits in colony and isolated cells, respectively) had a correlation Pearson >= 0.9 with at-least one other trait (Fig S7). Step 3. Include that individual trait for association analysis. Remove it and other traits having Pearson >= 0.9 with it from correlation matrices. Step 4. Repeat step 1 to 3 until there are no more traits to include in association analysis.

2.2.6 - Whole genome sequencing (WGS), variant calling and genes to test

DNA was obtained from cell line pellets with the Qiagen Quick-Start DNeasy Blood and Tissue Kit (cat. no. 69506). DNA samples were submitted to the Genomics Platform at the Broad Institute of MIT and Harvard. Whole genome sequencing (30x) was performed for all individuals (n=297) at the Broad Institute Genomics Platform using Illumina Nextera library preparation, quality control, and sequencing on the Illumina HiSeq 2500 platform. Raw sequences were QC-ed and sequencing reads (150 bp, paired-end) were aligned to the hg38 reference genome using the BWA alignment program. Variants were called and annotated (VQSLOD filter) using HapMap reference.

2.2.7 - WGS data quality control for common variant association analysis

The QC-ed WGS VCF file was processed using plink v1.90b3 to remove sex chromosomes, multi-allelic variants, variants with duplicated positions, and small insertions and deletions larger than 5bp. Of 38,239,223 variants loaded from the VCF file, 33,348,914 passed these filters. Donor-level genotype missingness rates were checked to exclude donors with genotype missingness rates > 10%. All 297 individuals passed this filter. Finally, variants with minor allele frequency (MAF) < 5%, missingness > 5%, and Hardy-Weinberg equilibrium p-value < 10-5 were excluded, following which, 7,020,633 remained for common variant association analysis.

2.2.7 - Principal components analysis (PCA)

Plink v1.90b3 was used on common (MAF > 5%) and post-QC variants to remove regions with known long-range linkage disequilibrium (LD) and variants in high LD (r2 > 0.1 in a window of 50 kb and a sliding window of 10 kb) (Price A. L. Am. J. Hum. Genetics 2008). The remaining 291,493 variants were loaded to GCTA v1.91.1 to generate a genetic relatedness matrix (GRM) using the --make-grm command with default options. The resulting GRM was used to generate 20 PCs using GCTA v1.91.1 --pca command with default options.

2.2.8 - Variance component analysis

Variance component of fixed (cell neighbor density and donor's age) and random effects (iPSC source tissue, cell line ID, plate and well of imaging, donor's sex, and disease status) was estimated for selected traits using linear mixed model (lmer function in lmertest package). We included the first 4 PCs derived from genetic variation, corresponding to the elbow in scree plot, for ancestry/population stratification. The p-value of each factor was Bonferroni corrected for the number of all tested traits (n=3,418).

Linear model question for variance component analysis:
Gaussianized trait ~ (1|Diseaseyes|no) + (1|iPSCfibroblast|Bcell) + (1|Sexmale|female) + Age + ∑PCi=1-4 + (1|Plate) + (1|Well) + (1|OnEdgeyes|no) + (1|Cell line ID) + Neighbor countif not isolets

2.2.9 - Common variant association analysis

The linear regression framework implemented in GCTA v1.91.1 (--fastGWA-lr command) was used to test the association of common (MAF > 5%), post-QC variants with 246 post-QC, INT traits that were described above. Like the rare variant association analysis, plate and sex were included as categorical and four genotyping PCs, number of cell neighbors (for cells in colony) and the edge variable were included as quantitative variables in the model. Associations were considered statistically significant if they passed the genome-wide significance threshold for 246 tests (P < 5x10-8/246).

Linear model equation for isolate cells:
Gaussianized trait ~ Variant + Age + Sex + ∑PC1-4 + (1|Plate)

Linear model equation for colony cells:
Gaussianized trait ~ Variant + Age + Sex + ∑PC1-4 + (1|Plate) + Neighbor count

2.2.10 - Rare variant burden test

To perform the rare variant burden test, the variants which were autosomal, passed the VQSLOD filter and called in >95% individuals and had maf< 1% were retained. These variants were annotated for their functional effect using SnpEff v5.0. After annotation, those variants were kept which resided in the protein-coding region and had high or moderate effects on encoded protein. For each gene, multiple rare variants were grouped and coded as present or absent. The association between individual morphology traits and the presence of rare variants in a gene was investigated using linear regression models. The p-values of associations were corrected for both the number of tested traits (n = 246) and the number of genes (n = 9105) using Bonferroni correction method.

2.2.11 - CRISPRi sgRNA design, cloning, and virus production

To functionally validate the rare-variant burden associations, we designed sgRNAs targeting the transcriptional start site (TSS) for each gene using CRISPick software (Doench, 2016, Sanson, 2018). sgRNA oligonucleotides were cloned into the CROPseq vector using a Golden Gate cloning protocol (Addgene: #106280, Juong, 2017). To validate sequence insertion, DNA plasmids were sequenced by a 3rd party provider. Plasmids with successful insertion were packaged for lentivirus generation using TransIT-293 reagent (Mirus Bio cat#: MIR 2704) and packaging plasmids VSV-G and DVPR (Addgene: #12259 and #12259).). HEK239T (ATCC cat#: CRL-3216) cells were transfected with sgRNA packaging plasmid and incubated for 48hrs. HEK239T media supernatant was collected, and lentivirus was concentrated using LENTI-X concentrator (Takara) per manufacturer's instructions. Virus supernatant was then aliquoted and stored at -80C.

2.2.12 - sgRNA transduction in dCas9-iPSCs

An iPSC line, WTC11_TO-NGN2_dCas9-BFP-KRAB (gift from Michael Ward), was seeded at 250k cells per well in a 12 well plate and 50ul of sgRNA lentivirus was added to each designated well. This iPSC line was cultured using mTeSR1 medium according to source recommendations (Stemcell Technologies, cat#: 85850). The following day, 1mL of mTeSR1 complete media was added on top of the existing media. 48hrs post transduction, cells underwent a full media change with the addition of 1 ug/ml puromycin (Sigma Aldrich cat#: P8833) for chemical selection of cells which did not uptake the sgRNAs. Puromycin is supplemented in the feeding media for the duration the cell line is in culture to avoid uninfected cells from populating the dish.

2.2.13 - qPCR analysis

RNA isolation was performed with the Direct-Zol RNA miniprep kit (ZYMO: cat# R2051) according to the manufacturer's instructions. To prevent DNA contamination, RNA was treated with DNase I (ZYMO: cat# R2051). The yield of RNA was determined with a Denovix DS-11 Series Spectrophotometer (Denovix). 200 ng of RNA was reverse transcribed with the iScript cDNA Synthesis Kit (Bio-Rad, cat# 1708890). For all analyses, RT–qPCR was carried out with iQ SYBR Green Supermix (Bio-Rad, cat# 1708880) and specific primers for each gene (listed below) with a CFX384 Touch Real-Time PCR Detection System (Bio-Rad). Target genes were normalized to the geometric mean of control genes, RPL10 and GAPDH, and relative expression compared to the mean Ct values for non-targeting control sgRNAs and gene targeting sgRNAs, respectively.

The following primers were used:

WASF2_forward 5′-TAGTAACGAGGAACATCGAGCC-3′

WASF2_reverse 5′-AAGGGAGCTTACCCGAGAGG-3′

PRLR_forward 5′-TCTCCACCTACCCTGATTGAC-3′

PRLR_reverse 5′-CGAACCTGGACAAGGTATTTCTG-3′

TSPAN15_forward 5′-TCCCTCCGTGACAACCTGTA-3′

TSPAN15_reverse 5′-CCGCCACAGCACTTGAACT-3′

RPL10_forward 5′-GCCGTACCCAAAGTCTCGC-3′

RPL10_reverse 5′-CACAAAGCGGAAACTCATCCA-3′

GAPDH_forward 5′-GGAGCGAGATCCCTCCAAAAT-3′

GAPDH_reverse 5′-GGCTGTTGTCATACTTCTCATGG-3′

2.2.14 - Modeling cmQTL effect size distributions with FMR-noLD

We used FMR-noLD (O'Connor, 2021) to model the effect size distribution for both common and rare variant summary statistics from our analyses. FMR-noLD is a simplified version of the main FMR method that does not model linkage disequilibrium (LD) between variants. We used FMR-noLD rather than FMR for this analysis as 1) the mixed ancestry of our sample complicates LD-score style estimators such as FMR, and 2) FMR-noLD is the appropriate choice for rare variants, which have very little LD.

For the common-variant analysis, we used the PLINK2 (Chang et al, 2015) –indep-pairwise (with parameters: variant count 50, variant count shift 5, threshold 0.2) utility to find a set of approximately 350,000 variants in approximate linkage equilibrium. We then submitted the concatenated set of summary statistics across all traits for FMR-noLD. For the rare-variant analysis, we used the same set of summary statistics used in the main burden test analysis (i.e. with no need for variant pruning), concatenated across all traits.

For power analysis for rare variant association, we first predicted effect size distributions at varying sample sizes by adding sampling variance 1/N to our inferred distribution of true effect sizes. We then computed the cumulative distribution function of these predicted distributions at our significance threshold for the main rare variant burden analysis, p = 2.2e-8, which represents the proportion of tests that are expected to be significant at each sample size.

2.2.15 - Immunocytochemistry

Immunofluorescence was performed using an automatic liquid handling dispenser (ApricotDesigns, Personal Pipettor). Cells were washed abundantly in 1x PBS, fixed for 20 minutes in PFA (4%, Electron Microscopy Sciences, 15714-S) plus Sucrose (4%, SIGMA, S0389), washed abundantly in 1x PBS, permeabilized and blocked for 20 minutes in Horse serum (4%, ThermoFisher, 16050114), Triton X-100 (0.3%, SIGMA, T9284) and Glycine (0.1M, SIGMA, G7126) in 1x PBS. Primary antibodies were then applied at 4°C overnight in 1x PBS supplemented with Horse serum (4%, ThermoFisher, 16050114). The following primary antibodies were used: Rabbit anti-human Aquaporin 4 (1:100, Millipore, AB3594), Rat anti-human CD44 (1:500, ThermoFisher, 14-0441-82), Rabbit anti-human GFAP (1:400, Dako, Z0334), Rabbit anti-human SLC1A3 (1:500, Boster, PA2185), Rabbit anti-human S100b (1:200, Abcam, ab52642) and Rabbit anti-human Vimentin (1:100, Cell Signaling, 3932S).

2.2.16 - Calcium imaging and analysis

Cells were incubated in fura-4AM dye at 2μM for 30 min at 37C. Cells were then washed and imaged in 200 μL recording solution (125mM NaCl, 2.5mM KCl, 15mM HEPES, 30mM glucose, 1mM $MgCl_2$, 3mM $CaCl_2$ in water, pH 7.3, mOsm 305). Time lapse videos were acquired at 4X on a Nikon Ti2-E microscope at 2 Hz for 5 mins. Cells were stimulated after 1 min by addition of 200 μL of ATP at 500μ M in recording solution, generating a final concentration of 250 μM in the well. Analysis of calcium videos was done using a custom MATLAB script. First, cells were segmented using a watershed algorithm. A table of mean fluorescence per cell across time was generated. Traces were smoothed with a Savitzky-Golay filter with a span of 50 images, aligned on the X-axis by subtracting the minimal value, then the peaks of a minimum height and local prominence of 10, with a minimal distance of 3 seconds between peaks were detected. We then extracted features for each cell: the number of peaks, peak height, interval and duration, rising time and falling time (defined as the time between the peak and the previous/next change of sign in the derivative). Data from all samples (4 wells of primary cells, and 8 wells of induced astrocytes from 2 parental cells, 4 wells each) were Z-scored then pooled together before k-means clustering with k=5. Data were then re-split into 12 tables according to the origin of each cell for statistics. For each cluster, example traces for 60 cells randomly picked from all 12 samples of origin were generated by normalizing all traces between 0 and 255 and generating an 8-bit image (Supplementary fig 2). Scripts are available here https://github.com/lbinan/astrocyteInduction.

2.3 - Next Generation Sequencing

2.3.1 - Generation and sequencing of libraries

Sequencing libraries were generated from isolated DNA using either TruSeq Nano DNA Library Prep Kit (Illumina, NP-101-1001) or Nextera DNA Flex Library Prep Kit (Illumina FC-121-1030) in combination with the NeoPrep Library Prep System (Illumina, SE-601-1001). Libraries were sequenced using the NextSeq 500 Sequencing System (Illumina, SY-415-1001) with the NextSeq 500 High Output v2 Kit (75 cycles, FC-404-2005). Runs were set up as a single 85 bp reads and included an index read when libraries were pooled. For scaled analyses, 16 Census-Seq samples are pooled into one NextSeq run with a minimum requirement of 16-32 million reads per library (about 1X coverage).

In the cases where we had only a handful of libraries, it was more cost-effective to use fee-for-service DNA sequencing providers such as Azenta. Under those circumstances, we would ship out either extracted genomic-DNA using the same method as above or send and entirely frozen cell pellet.

2.3.2 – Sequence Alignment Protocol

Raw sequence data were demultiplexed using the Picard tools ExtractIlluminaBarcodes and IlluminaBasecallsToSam. The resulting demultiplexed libraries were validated for both relative library size (library balance) as well as absolute size, to flag potential bioinformatic issues with demultiplexing, as well as benchside library generation issues. The demultiplexed libraries were then aligned to a human reference genome with BWA. The reference genome used was selected to match the same build used in the VCF file that contained the reference genotypes for the experiment's donor pool. For experiments for which human cells were grown on a bed of mouse glia, alignment was performed using a multi-organism reference and the reads were competitively aligned to both genomes. Sequencing reads were then filtered to reads that mapped at high quality (MQ>=10) to the human genome.

2.3.3 - Variant Call Format (VCF) pre-analysis Processing

Prior to running Census-seq, VCF files were processed to filter variants and add additional site-level information. Variants were first normalized to their appropriate reference sequence using BCFTools; this splits multiallelic SNPs into multiple biallelic SNPs and sets the reference allele to be the reference base of the genome at that position. Variants that were monomorphic were dropped, as well as those without a PASS filter, where the site was flagged as problematic during VCF generation. Sites without rsID annotations were updated using information from dbSNP when possible, and otherwise site names were changed to chromosome:position:ref_allele:alt_allele. Allele frequencies calculated from the 1000 Genomes Project were annotated at all available sites.

2.3.4 - scRNA-sequencing and donor assignment

For single-cell analyses, cells were harvested and prepared with 10X Chromium Single Cell 3' Reagents V3 and sequenced on a NovaSeq 6000 (Illumina) using a S2 flow cell at 2 x 100bp. Raw sequence files were then aligned and prepared following the Drop-seq workflow (Macosko et al., 2015). Human reads were aligned to GRCh18 and filtered for high quality mapped reads (MQ 10). In order to identify donor identity of each droplet, variants were filtered through several quality controls as described previously to be included in the VCF files (Wells, 2021), with the goal of only using sites that unambiguously and unequivocally can be detected as A/T or G/C. Once both the sequenced single-cell libraries and VCF reference files are filtered and QC'ed, the Dropulation algorithm is run. Dropulation analyzes each droplet, or cell, independently and for each cell generates a number representing the likely provenance of each droplet from one donor. Each variant site is assigned a probability score for a given allele in the sequenced unique molecular identifier (UMI) calculated as the probability of the base observed compared to expected based, and 1 – probability that those reads disagree with the base sequenced. Donor identity is then assigned as the computed diploid likelihood at each UMI summed up across all sites (Wells et al, 2023).

2.3.5 - scRNAseq analysis of villages and integrated datasets

Gene by cell matrices from hiA villages were built from separate runs of 10X Chromium Single Cell 3’ Reagents V3 as described above. Cells with less than 200 genes and more than 15% mitochondrial RNA were trimmed away from downstream analyses. SNN graphs were computed using batch-balanced k-nearest neighbors (BBKNN) to remove batch effects across 10X reactions (Polanski et al., 2020). Leiden clustering was performed across resolutions in BBKNN space (0.2,0.4,0.6) and then visualized to determine the dimensionality of the data. LEIDEN_BBKNN_0.2 was used for the downstream analysis. For the metagene analysis, summed expression of gene sets for astrocyte precursor markers and mature astrocyte markers were divided by a random control set of 500 genes. For the analyses where we integrated existing iPSC-astrocyte data as well as human brain data, raw matrices were loaded in Seurat v4.0.1. The same parameters were applied as above, excluding cells with fewer than 200 genes and greater than 15% mitochondrial gene expression. Given the technical variability across datasets, and cell sources, we first computed a new count matrix using SCTransform (Hafemeister and Satija, 2019). Next, the transformed data was integrated using linked inference of genomic experimental relationships (LIGER) (Welch et al., 2019). Downstream analytical steps were performed using Seurat v4.0.1 basic functions.

2.3.6 - High-depth Hi-C in neural cell lines

In Situ Hi-C Protocol All cell lines were cultured following the manufacturer’s recommendations. Two to five million cells were crosslinked with 1% formaldehyde for 10 min at room temperature. Nuclei were permeabilized. DNA was digested with 100 units of MboI, and the ends of restriction fragments were labeled using biotinylated nucleotides and ligated in a small volume. After reversal of crosslinks, ligated DNA was purified and sheared to a length of

400 bp, at which point ligation junctions were pulled down with streptavidin beads and prepped for Illumina sequencing. Dilution Hi-C was performed as in Lieberman-Aiden et al. (2009).

2.3.7 - Hi-C Data Pipeline

The raw Hi-C sequencing data was processed using Juicer54 and aligned against the hg19. reference genome. Both contact matrices used for downstream analysis were KR-normalized with Juicer. We visualized the contact map in hic format using Juicebox.

2.3.8 - Single-nucleus RNA sequencing of postmortem human brain

We analyzed a previously generated resource of paired genotypic and single-nucleus RNA-seq data from the dorsolateral prefrontal cortex (DLPFC) of post-mortem human brain samples from the HBTRC/NIH NeuroBioBank (Ling et al, under revision). The generation of these data will be described in detail in a forthcoming primary manuscript. Briefly, we developed a workflow for generating and analyzing pools of nuclei from DLPC from 20 donors per pool. We started by dissecting tissue from each donor, ensuring that we obtained a similar mass of tissue from each specimen while sampling all cortical layers. The frozen tissue samples were then pooled for simultaneous isolation of nuclei, and subsequent steps (nuclear isolation, encapsulation in droplets and preparation of snRNA-seq libraries) used all donors pooled together. This "dropulation" workflow minimizes experimental variability, including effects of messenger RNA ascertainment and cell-free ambient RNA. We reassigned nuclei to donor-of-origin using combinations of hundreds of transcribed SNPs that disambiguated individual genotypes. We performed global clustering and identification of marker genes to assign nuclei to seven major cell classes (glutamatergic neurons, GABA-ergic neurons, astrocytes, oligodendrocytes, polydendrocytes, microglia, and endothelia). Median cell-type proportions were: 48% glutamatergic neurons, 19% GABA-ergic neurons, 14% astrocytes, 8%

oligodendrocytes, 5% polydendrocytes, 2% microglia, and 1% endothelia. The cell type-specific gene-by-donor expression matrices were normalized and variance-stabilized using the scTransform utility (Hafemeister & Satija, 2019, Genome Biol). As previously described (Weiner et al, 2022, NG), we used PCA to identify European ancestry samples and exclude samples with mean expression >3 s.d. From the cohort mean, yielding a final sample of 122 individuals. Notably, this cohort includes individuals with schizophrenia (n = 58)

2.3.9 - Bulk-RNA sequencing of postmortem human brain

We analyzed paired a resource of paired genotypic and bulk RNA-seq from dorsolateral prefrontal cortex (DLPFC) samples from the CommonMind consortium. Generation of the expression count matrices is described extensively in the primary publication for this resource (Hoffman et al, 2019, Sci Data). Within this dataset, we restricted analysis to donors from the NIMH Human Brain Collection Core (HBCC) and the University of Pittsburgh (PITT) biobanks due to previous analysis suggesting high concordance with the snRNA-seq resource described above. We processed these data with scTransform with similar parameters as our snRNA-seq analysis.

As previously described (Weiner et al, 2022, NG), we used PCA to identify European- (N = 302) and African-ancestry (N = 206) samples and analyzed each population separately. Notably, this cohort includes individuals with diagnoses of schizophrenia (n = 69) and bipolar disorder (n = 89)

2.4 – Census-seq Computational analysis methods

Methods to detect the presence of individuals' DNA within DNA mixtures have been a lively area of computational investigation (Egeland et al., 2003, Hu and Fung, 2003, Balding, 2003, Clayton et al., 1998). Our goal with Census-seq, Roll Call and CSI was to develop a suite of algorithms with which to detect and precisely quantify individual donors' contributions to cell/DNA mixtures and to detect the presence of contaminating DNA/cells of known or unknown genotypes.

2.4.1 - Census-seq algorithm (precise quantification of donor contribution to cell/DNA mixtures)

The goal of Census-seq is to measure the contribution of each donor to a
cell mixture – both to monitor population dynamics, and for quantitative phenotyping. We do this systematically, routinely, and inexpensively, without the need for single-cell analysis, simply by lightly sequencing genomic DNA from the cells. The donor mixture determines what ratio of alleles are present at every SNP. We developed a gradient-descent algorithm to find the donor-mixing coefficients that maximize the likelihood of any observed sequence data.

For Census-seq to perform accurately, input sequencing and VCF data is filtered on a per-run basis. Sequence reads are filtered to high quality mappings (MQ>=10) on the autosomes that have not been flagged as PCR duplicates. VCF sites are considered if they meet all of the following criteria: each site has GQ score of at least 30, is a diploid site, is polymorphic in the subset of donors in the population, and at least 90% of donors have a genotype quality score >=30. In addition, for genotype array-based data where site quality scores may not be available, sites where the reference base is ambiguous [A/T, C/G] are not considered. Variant sites are also rejected if they are not common in the population – only sites with >5% allele frequency are included in analysis. In training data, we found that private SNPs provide more error than gain. This could be driven by genotype error in genetic data and their exclusion slightly improved the accuracy of our results. Therefore, common alleles were more than sufficient for inclusion.

Given these filtered inputs, a matrix of donor genotypes and the counts of the reference and alternate allele at each variant are generated. Census-seq then uses these to find a vector of donor-specific contributions (to the mixture) that best explains the observed counts of alleles at each site in the sequence data. The algorithm initializes with the donor proportions set to equal values (1/number of donors), then runs through an estimation maximization (EM) procedure. During each step, the allele frequency of each site is calculated from the genotypes of the donors and their relative proportion in the pool. The initial likelihood of the sequencing data given the starting donor ratios is calculated at each SNP by the likelihood function (see below) and the results summed across all sites. To determine how to change the donor ratios to explain the data, an adjustment term is calculated for every donor/site, and the results are summed across sites for each donor. This adjustment factor is then scaled by an additional parameter and added to each donor's representation. To determine this scaling value the algorithm employs a univariate optimizer to maximize the donor likelihood. The adjustment is then applied to the data, and the algorithm repeats the adjustment/ likelihood optimization loop until convergence.

Log-likelihood Function:

For any set of donor-mixing coefficients (and resulting allele frequencies in the DNA mixture), the log-likelihood of the Census-seq sequencing data is calculated as:

$$\sum_{i=1}^{n} \log_{10}\left({f_{(a,i)}}^{A_i} \times {f_{(b,i)}}^{B_i} \right)$$

where *i* (= 1, 2, …, n) indexes the full set of SNPs used in analysis; $f_{(a,i)}$ and $f_{(b,i)}$ are the frequencies of the two alleles for SNP *i*; and A_i and B_i are the numbers of observations of these alleles in the Census-seq sequencing data.

The derivative of the above likelihood function with respect to the donor-specific mixing coefficients is used to calculate a gradient ascent direction in the form of an adjustment factor for each donor; this adjustment term reflects the extent to which an increase (or decrease) in that donor's mixing coefficient improved the likelihood of the Census-seq data. During each loop of the EM, each donor's mixing coefficient (contribution estimate) is adjusted by a factor that is proportional to the result of the following formula:

$$\sum_{i=1}^{n} \log_{10}\left(\left(\frac{A_i}{f_{(a,i)}} - \frac{B_i}{f_{(b,i)}}\right) \times \left(g_i - f_a\right)\right)$$

Where *i* (= 1, 2, …, n) indexes the full set of SNPs used in analysis; g_i is the genotype of the donor at SNP *i* (= 1 if homozygous for the reference allele, 0 if homozygous for the alternate allele, and 0.5 if heterozygous); and $f_{(a,i)}$ and $f_{(b,i)}$ are the frequencies of the two alleles for SNP *i*.

2.4.2 - Roll Call algorithm (to establish which donors have contributed to a cell/DNA mixture)

Census-seq requires a complete and accurate list of donors in order to estimate each donor's contribution accurately. However, cell lines can become contaminated with cells from other donors; tubes can be mislabeled; and sample swaps can occur even despite best efforts. The probability of such errors increases with the number of cell lines in an experiment – thus, population-scale experiments are inherently more vulnerable to error. We thus developed an algorithm ("Roll Call") to create a complete list of the donors who have contributed to a cell/DNA mixture (from a larger set of genomically characterized candidates, for example, all of the cell lines in a lab). Roll Call uses each donor's private (singleton) alleles to find evidence that a donor is present in a mixture in which s/he doesn't belong. To do this, we measure (for all of the sequence reads that touch these sites) the fraction of the sequence reads that individual's IRVs. Since this donor is in principle the only source of alternate alleles at the IRV sites, the fraction of these alleles observed can be directly related to the presence of the donor to the mixture. In the equation below, we count the observations (sequence reads) that (in the absence of sequence errors) could only arise from a given individual, divided by the total number of reads at those sites. This is less precise than Census-seq at quantifying donor representation, because it draws upon a small fraction of all variable sites; the utility of Roll Call is to search through a very large set of potential donors to determine presence/absence and thereby authenticate a village prior to Census-seq analysis.

Roll Call uses the same VCF and sequencing read filtering as Census-seq, with one exception - instead of retaining common sites, Roll Call leverages IRVs (rare identifying variants) - sites that are private to a single donor in the VCF. Since these IRVs are the only source of alternate alleles in the sequencing data, the fraction of alleles in the sequencing data observed in a sequencing pool can directly be related to the proportion of donors in the pool.

To generate the counts of IRVs, the algorithm filters the VCF to a set of heterozygous sites that are private to donors in the VCF. The pileup of reference and alternate alleles is generated at those sites in the sequencing data. For each donor, those pileups are then aggregated into a single result, and the Roll Call score is calculated:

$$\frac{2a}{a+r}$$

where *a* is the number of alternate alleles (and r is the number of reference alleles) observed for that donor's IRVs at heterozygous sites.

Note that PCR and sequencing error cause the Roll Call score to be slightly positive even for donors who have not contributed to a mixture. For any experiment, a background null distribution of Roll Call scores can be estimated by including (in the input VCF) many individuals not expected to have contributed to the cell/DNA mixture; this distribution can then inform the selection of an experiment-specific threshold for the Roll Call score.

We routinely use Roll Call to validate and if necessary refine the donor list prior to Census-seq analysis.

2.4.3 - CSI - Contaminating Sample Identifier (detection of cells/DNA from genomically unknown donors)

What if a pool were visited by cells from a genomically "unknown" donor? If we don't have genotypes for the contaminating donor(s) *a priori*, we need another way to detect the presence of "unknown unexpected" visitors. We developed the CSI algorithm to do this. CSI utilizes observations of alternate alleles that could not have arisen from the expected donors' genomes and must therefore represent sequencing errors or contaminating cells. We first identify all genomic sites and alleles that are absent among the donors we expect in the pool but present at some minimum frequency in the wider population (as estimated from the Thousand Genomes Project or gnomAD data). We then look for evidence of such alleles in the village DNA sequencing data. By correcting for sequencing error rate, we distinguish between two models: sequencing errors and an unwelcome visitor.

CSI utilizes observations of alternate alleles that could not have arisen from the expected donors' genomes and must therefore represent sequencing errors or contaminating cells. We first identify all alleles that are absent among the donors we expect in the pool but present at some minimum frequency in the wider population.

CSI uses the same VCF and sequencing read filtering as Census-Seq with one exception - the variant sites selected from the VCF are those for which all donors in the experiment have the reference genotype, and the minor allele frequency (MAF) of these sites in a wider population is at least 2.5% This population allele frequency can be computed from a variety of potential sources, including (i) all donors in the VCF file not expected to be present in the experimental mixture; or, (ii) an external reference population such as those provided by the 1000 Genomes Project or gnomAD.

To calculate the CSI intrusion score, we count observations of reference and alternate alleles across all such sites and aggregate the results. We then calculate the CSI intrusion score by considering the sequencing error rate and average minor allele frequency of the sites queried.

CSI intrusion score:

Where Fa is the fraction of sequencing reads (at these sites) observed to contain the alleles that are absent among the candidate donors; s is the sequencing error rate; and m is the mean minor allele frequency (of these alleles) in the population from which the potential donors are sampled.

Note that while the CSI intrusion score is proportional to the degree of contamination from unexpected donor(s), two possibilities can cause it to deviate: (i) if the unexpected donor is related to an expected/candidate donor; or (ii) if the unexpected donor comes from a different population than the population used to estimate the mean MAF of the absent alleles. (However, *changes* in the CSI score across time are likely to represent changes in the contribution of a contaminating donor.) For this reason, we generally use the CSI score to authenticate or flag cell villages rather than to precisely measure contamination; a cell village contaminated by an unknown donor would not be suitable for Census-seq analysis anyway.

2.4.4 - Quantification and Statistical Analysis

In analyzing the SMN-expression phenotype, we aggregated results from two villages of iPSCs, one of 113 donors, which we called CIRM1, and another with 38 donors, which we called SMN6.

For each village, we calculate an SMN-expression phenotype measure for each donor as the log10-fold change in the representation of that donor in the SMN-high fraction of the village over their representation in the SMN-low fraction of the village: $SMNexpression = log_{10}\,(rep_{high}/rep_{low})$.

Census-seq has reduced precision in quantifying donors who have contributed less than 0.2% of the cells/DNA in a mixture. To account for this, as a quality control measure, we exclude from association analysis any donors who have less than 0.3% representation in any of the relevant derived villages (in this case, SMN-high or SMN-low). 72 of the donors from the village of 113 and 37 donors from the village of 38 donors passed this filtering step and were used in downstream genetic analysis.

20 of the 28 donors shared between the two villages were above threshold in both villages. To construct an aggregate table of the two villages, we averaged the SMN-expression phenotype measurements from both villages for these 20 donors. Then, for the remaining 76 unique donors, we used the SMN-expression phenotype calculated from the one village in which it passed QC.

This resulted in a set of 96 donors with SMN-expression phenotype measurements suitable for genetic analysis. We performed linear regression of these 96 measurements against the number of copies of the *SMN* genes (**Figure 4C**).

2.4.5 - Measurement of quantitative phenotype and of quantitative drug response

We had collected two additional fractions of SMN-high and SMN-low for each village after treatment with the drug LMI070. Using these two LMI070-treated fractions, we calculated $SMNexpression$for the donors in each village when treated with LMI070. We then calculated the difference of $SMNexpression$ of the donor when treated with LMI070 and $SMNexpression$of the donor when not treated with LMI070, calling this difference $LMI070DrugResponse$. $LMI070DrugResponse = SMNexpression_{LMI070treated} - SMNexpression_{control}$

Since we used two new fractions to calculate drug response, our quality control of the data had to adjust to take these into account. In particular, we filtered the data from each village to only include donors that had a representation of at least 0.3% in all 4 fractions. 67 donors from the village of 113 and 36 of the village of 38 passed QCl, with 11 of the 28 common donors passing in both villages. We averaged values for the 11 common donors and created an aggregate table to use for downstream analysis that had measurements for 92 unique donors. Finally, we ran a linear regression of the effect of *SMN2* copy number on LMI070 drug response and found that there was a significant linear correlation between *SMN2* copy number and LMI070 drug response (**Figure 6D**) that was even stronger when we discount any copies of the *SMNdel* variant of the gene in the regression model (**Figure 6E**).

2.5 - Statistical analyses

2.5.1 - Partition-level Hi-C Analysis

To compare the regulatory architecture of chr22q to other similarly sized regions, we constructed 61 33-Mb partitions of the human genome, one of which represented chr22q (Supplementary Table 1) were constructed to avoid centromeres and telomeres. To estimate the degree of intrachromosomal contect in each partition in lymphoblastoid cell lines, we used previously described 1-Mb resolution HiC data from GM06990 (Lieberman-Aiden et al, 2009, Science). For each partition, we estimated the degree of intra-partition contact as the mean of all elements in the 33 Mb x 33 Mb count matrix for that partition. To estimate the degree of intrachromosomal contact in developing brain samples, we carried out an identical analysis in a previously described 100kb-resolution HiC sequencing dataset in midgestational cortical plate (Won et al, 2013, Nature).

Degree of physical contact has been demonstrated to vary as a function of gene density and segmental duplication content (Weiner et al, 2022, NG). In order to assess variation in physical content conditional on these genomic features, we regressed gene count and segmental duplication out of degree of physical contact. The scaled residuals from this model are used in the primary analysis in Figure 1A. To estimate segmental duplication content, we calculated the fraction of nucleotides for each partition that overlapped at least one segmental duplication in the UCSC Genome Browser (Kent et al, 2002, Genome Res.). We used BEDTools v2.30.0 to calculate segmental duplication coverage.

2.5.2 - Partition-level Enhancer Density Analysis

To estimate the proportion of each 33-Mb partition composed by active enhancers, we used ChromHMM predictions for dorsolateral prefrontal cortex samples from the Roadmap Epigenomics Project (Kundaje et al, 2015, Nature). Briefly, this resource assigns each 200-bp window in the genome to one of 18 chromatin states. For this analysis, we considered four states to represent enhancers: EnhG1/EnhG2 (genic enhancers) and EnhA1 and EnhA2 (active enhancers). For each 33-Mb partition, we computed the fraction of 200-bp windows within that partition that were assigned to one of these four states.

2.5.3 - Differential expression analyses in isogenic cell lines

We reanalyzed previously described bulk-RNA sequencing data in two 22q11.2del-carrying cell lines and isogenic controls, in four different in vitro cell types: human pluripotent stem cell (hPSC-Day0), induced neural progenitor cell (iNPC-Day4), and two induced excitatory neuron stages (iExN-Day28 and iExN-Day49). Briefly, these cell lines were generated from H1 hESCs with CRISPR-Cas9 sgRNAs targeted towards LCR-A and LCR-D, the most common breakpoints for recurrent 22q11.2del. The protocol for generation and differentiation of these cell lines is described in detail in Nehme, Pietlanen et al, 2022, Nat Comms.

For each cell-type, we had deletion and isogenic control data from two cell lines, with 3 replicates each (e.g. four cell lines * 3 replicates each = 12 total samples). We performed differential expression analysis with a negative binomial generalized linear model, implemented in DESeq2 (Love et al, 2014, Bioinformatics). Briefly, DESeq2 models the mean-variance relationship across genes to regularize overdispersion estimates and uses these regularized estimates to estimate log2FCs. For normalization, we used the default DESeq2 median-of-ratios approach, which estimates a common size factor across all genes. We used surrogate variable analysis (SVA; Leek et al, 2007, PLoS Genetics) to control for unobserved confounding variables.

For each cell type, we used the following regression equation:

Count = Deletion + Surrogate Variables

We also integrated all of our data in one large model to boost power. For that analysis of all 48 samples (e.g. 4 cell types * 12 samples per cell type), we used the following regression equation:

Count = Deletion + CellType + Surrogate Variables

2.5.4 - Differential expression in case/control cohort

We used a similar analytic strategy to assess differential expression in a case-control sample collection from the same previous study (Nehme, Pietlanen et al, 2022, Nat Comms). For this analysis, we had access to data from three cell types: hPSC, iNPC, and iExN (e.g. without a second iExN stage, as in the isogenic analysis). For each cell type, we had data from 28 control individuals and 20 individuals carrying 22q11.2del, for a total of 144 samples. For this analysis, we also used DESeq2 and SVA to perform differential expression analysis, with identical regression equations to those for the isogenic analysis.

2.5.5 - Correspondence between differential expression and physical contact

For each gene on chr22q outside the canonical 22q11.2 deletion interval (chr22:18637094-21809133), we calculated the median degree of contact with the 22q11.2 deletion interval on the observed/expected scale. First, we assigned each gene to a 100kb bin along chr22q. We then queried the intrachromosomal contact between each gene's bin and the 100kb bins residing in the deletion interval and took the median of these observed/expected values to avoid excessive influence of outliers. We used this relatively coarse bin-size (100kb) because at lower bin sizes, we did not have sufficient depth to accurately estimate observed/expected for these very distal contacts. Specifically, using the guideline from Rao et al, 2014, Cell to use a minimum of 400 reads to calculate expected contact, we were unable to achieve this minimum value at a finer 10kb bin size.

2.5.6 - Generation of local polygenic scores

We generated local polygenic risk scores for 8 complex traits and common diseases. For seven of these traits, we used publicly available summary statistics: schizophrenia (Trubetskoy et al, 2022, Nature), ADHD (Demontis et al, 2019, NG), intelligence (Savage-Jansen et al, 2018, NG), height (Neale Lab, see URLs), body mass index (Neale Lab, see URLs), low-density lipoprotein (Neale Lab, see URLs), and red blood cell count (Neale Lab, see URLs). For autism, we used an unpublished set of summary statistics from an in-progress update to the iPSYCH/Psychiatric Genomics Consortium autism GWAS (Grove et al, 2019, NG). Briefly, this sample comprises 26,067 cases and 46,455 controls drawn from the Danish iPSYCH resource and the Psychiatric Genomics Consortium, which includes several cohorts from around the world.

We first used LD Score Regression (Bulik-Sullivan et al, 2015, NG) to quality-control summary statistics, and estimate SNP-heritability. Then, to estimate polygenic score weights, we used the best linear unbiased predictor (BLUP) approach as implemented in LDPred2-Inf (Prive et al, 2020, Bioinformatics). To calculate PGS22q, we then filtered these weights to SNPs residing on chr22q. For both the snRNA-seq and bulk RNA-seq studies described below, we used PLINK2 (Chang et al, 2015, Gigascience) to compute polygenic scores from genotypes with the above-described weights.

2.5.7 - Differential expression in Phelan-McDermid Syndrome

We downloaded gene expression count matrices from Breen et al, 2020, Mol Autism. Briefly, this project sought to generate pluripotent stem cell lines from 7 individuals with Phelan-McDermid Syndrome and their unaffected siblings. Within this collection, some individuals had single-gene deletions (e.g. of SHANK3) and some had larger deletions at 22q13.3. A subset of these individuals had successful generation of cell lines and differentiation into neural cells; we analyzed data from the two families in this collection with large genomic deletions, each of which had multiple replicates. We performed differential expression analysis in an identical manner to the above-described analyses, using DESeq2 and SVA. We used the following regression equation:

Count = Deletion + Family + Surrogate Variables

2.5.8 - Differential expression in isogenic 16p11.2del and 15q13.3del cell lines

We analyzed data from a previous report of CRISPR-mediated deletions at 16p11.2 and 15q13.3 via Single-guide-CRISPR/Cas-targeting-Of-Repetitive-Elements (SCORE) (Tai et al, 2016, Nat Neurosci; Tai et al, 2022, Am J Hum Gen). Briefly, we used sgRNAs targeted towards the segmental duplications that flank these recurrent CNVs, to mimic the process of non-allelic homologous recombination by which they are created in vivo. For 16p11.2del iNSCs (induced neural stem cells), we analyzed n = 7 heterozygous deletion lines and n=12 controls exposed to the CRISPR construct but not to the guide RNA; for 16p11.2del iNs (Ngn2-induced neurons), we analyzed n = 7 heterozygous deletion lines and n = 6 controls. For 15q13.3 iNSCs and iNs, we analyzed n = 11 heterozygous deletion lines and n = 6 controls exposed to CRISPR construct but not guide RNAs. Similarly, to the above-described analyses, we used DESeq2 and SVA to perform differential expression analysis, with the regression equation:

Count = Deletion + Surrogate Variables

2.6 – PCR

2.6.1 – Quantitative analysis of CRISPRi knockdown by qPCR

RNA isolation was performed with the Direct-Zol RNA miniprep kit (ZYMO: cat# R2051) according to the manufacturer’s instructions. To prevent DNA contamination, RNA was treated with DNase I (ZYMO: cat# R2051). The yield of RNA was determined with a Denovix DS-11 Series Spectrophotometer (Denovix). 200 ng of RNA was reverse-transcribed with the iScript cDNA Synthesis Kit (Bio-Rad, cat# 1708890). For all analyses, RT–qPCR was carried out with iQ SYBR Green Supermix (Bio-Rad, cat# 1708880) and specific primers for each gene (listed below) with a CFX384 Touch Real-Time PCR Detection System (Bio-Rad). Target genes were normalized to the geometric mean of control genes, RPL10 and GAPDH, and relative expression compared to the mean Ct values for control and wild-type isogenic samples, respectively.

The following primers were used:
SMARCB1_forward 5’-GCGAGTTCTACATGATCGGCT-3’
SMARCB1_reverse 5’-CACAGTGGCTAGTCGCCTC-3’

RPL10_forward 5′-GCCGTACCCAAAGTCTCGC-3′

RPL10_reverse 5′-CACAAAGCGGAAACTCATCCA-3′

GAPDH_forward 5′-GGAGCGAGATCCCTCCAAAAT-3′

GAPDH_reverse 5′-GGCTGTTGTCATACTTCTCATGG-3′

2.6.2 - Quantitative analysis of SMN1 and SMN2 transcript abundance by qPCR

SMN transcript abundance was quantified in cell lines of varying *SMN1* and *SMN2* copy number as follows: Genea52 (copy number *SMN1*:*SMN2*; 2/2), RUES1 (2/0), and iPSC322A (0/2). Cells were treated in the presence or absence of LMI070 as described. Total RNA was isolated from cell pellets using the RNeasy Mini Kit (Qiagen, 74104). cDNA was synthesized using iScript cDNA Synthesis Kit (Bio-Rad, 1708890). Two primer sets were used to detect *SMN* transcripts as described previously (Integrated DNA Technologies (Sumner et al., 2006): *SMNFL* spanning exons 6, 7 and 8 (forward, 5′-CAAAAAGAAGGAAGGTGCTCACATT-3′; reverse, 5′-GTGTCATTTAGTGCTGCTCTATGC-3′; probe, 5′-6FAM-CAGCATTTCTCCTTAATTTA-MGBNFQ-3′), and *SMNdelta7* spanning the exon 6-8 junction (forward, 5′-CATGGTACATGAGTGGCTATCATACTG-3′; reverse, 5′- TGGTGTCATTTAGTGCTGCTCTATG-3′; probe, 5′-6FAM- CCAGCATTTCCATATAATAGC-MGBNFQ-3′). mRNA levels were normalized to an internal control (Human Beta Glucuronidase (GUSB), Life Technologies, 4333767T). For samples treated with LMI070, *SMN1* and *SMN2* levels were compared to a DMSO-treated control. qRT-PCR was performed using TaqMan™ Fast Advanced Master Mix (ThermoFisher Scientific, 4444963).

2.7 - Miscellaneous

2.7.1 - Immunoblotting

For collection, neurons grown on glia were washed with DPBS and lysed with RIPA buffer and 1x protease inhibitor cocktail. Lysates were boiled, sonicated, and centrifuged at 16,000xg for 5 minutes. The soluble fraction was separated on SDS-PAGE using Bolt system (Novex). The proteins were transferred onto nitrocellulose membrane using iBlot2 Gel Transfer Device and immunostained using SMARCB1/BAF47 antibody (Cell signaling technology 91735S, 1:1,000) and GAPDH (Proteintech, 60004-1-Ig, 1:5000). and detected via HRP-conjugated secondary antibodies on the Chemidoc system.

2.7.2 - TNFα stimulation

Human primary and induced astrocytes (> day 30 of differentiation) were seeded at 15,000 cells.cm^{-2} in 96-well plates and incubated 24 hours later with human Tumor Necrosis Factor alpha (100ng.mL^{-1}) or 0.1% Bovine Serum Albumin (Sigma Aldrich, A0281-10G) in Astrocyte medium (ScienCell, 1800) for 7 days. The media was replaced with fresh treatment after 4 days of incubation. Media were collected and stored at -80°C until further processing. IL-6 concentration was measured by ELISA (Abcam, ab229334) in supernatants diluted 1:50 according to the manufacturer's protocol (Abcam).

2.7.3 – Treatment with LMI070

To enhance splicing of the *SMN2* transcript to include exon 7, the small molecule LMI070 (also referred to as NVS-SM1 or Branaplam (Novartis; MedChemExpress, HY-19620)) was added to cell culture media for 24 hr under standard culture conditions. To determine the optimal dose of LMI070 in our cellular systems, a dose response was performed using a range of concentrations from 0.01 to 1 uM (**Figure S4**). Inclusion of exon 7 in the *SMN2* transcript was used as an indicator of effectivity. 0.1 uM was found to be optimal and used in all future experiments.

2.7.4 - Antibody-mediated pluripotency selection and precise sorting of cell lines using flow cytometry

To generate villages of pluripotent stem cells used to characterize growth/proliferation phenotypes (**Figure 2**), flow cytometry using antibodies selecting for markers of pluripotency was used to select for cells of the highest quality combined with precision counting. Prior to immunostaining, cells were dissociated using Accutase® and cell densities were measured using the Countess™ Hemocytometer. 5 x 10^6 cells from each cell line were resuspended in FACS buffer (1 x PBS, 7.5% BSA, 1 mM EDTA, 10 uM Y27632, and Nomrocin™) and cells were immunostained with fluorescently labeled antibodies directed against SSEA4 and TRA-1-60 (BD Biosciences, Alexa Fluor® 647 Mouse anti-SSEA-4 Clone MC813-70 (RUO), 560219; Alexa Fluor® 488 Mouse anti-Human TRA-1-60 Antigen Clone TRA-1-60 (RUO), 560173) at a concentration of 1:500 for 30 min at RT. Cells were washed with FACS buffer repeatedly and stained with DAPI (1 ug/ml). 0.5 x 10^6 DAPI-/SSEA4+/TRA-1-60+ cells were collected for each cell line using the BD FACSAria™ II flow cytometer (BD Biosciences). Collected cells were pooled and plated at a density of 30 000 /cm^2.

2.7.5 - Phenotype-based sorting of SMN-high and SMN-low cellular populations

To allow for genotype-phenotype correlations based on SMN protein expression, villages of iPSCs or NGN2-induced neurons (at 14 days of differentiation) were immunostained with a monoclonal antibody directed against SMN (BD Biosciences, 610647) and segregated into new 'sub-villages' comprising populations of the lowest and highest SMN expressing cells. Following dissociation with Accutase®, 10 x 10^6 cells from villages of iPSCs or neurons were fixed and permeabilized using the Fixation/Permeabilization Solution Kit (BD Biosciences, 554714). Cells were resuspended in FACS buffer (1 x PBS, 7.5% BSA, 1 mM EDTA, 10 uM Y27632, and Nomrocin™) and incubated in the presence of anti-SMN at a concentration of 1: 500 for 30 min at RT. Cells were washed repeatedly in FACS buffer and incubated with donkey-anti-mouse-Alexa-647 secondary antibody (Life Technologies, A-31571) at a concentration of 1:5000 in FACS buffer supplemented with DAPI (1 ug/ml) for 30 min at RT. After repeated washing, cells were sorted using the BD FACSAria™ II flow cytometer (BD Biosciences). Gating strategies were developed to select for single, live, intact cells in the absence of autofluorescence (using an empty 555 nm channel as control) (Figure S5). For neurons, an additional GFP (488 nm) selection gate was applied to separate Human neurons (GFP+) from co-cultured mouse glia. Cells determined to be positively stained for SMN (647 nm) were compared to unstained or secondary antibody-only controls, and gates were set using the patient cell line iPSC3222A as an indicator of low levels of SMN expression. To generate sub-villages, fractions encompassing the bottom (SMN-low) and top (SMN-high) 20% of cells were independently collected (500 000 cells each). An aliquot of unsorted cells representing donor representation of the original village was collected to serve as a control. Collected cells were pelleted and stored at -20°C for DNA isolation.

Chapter 3

Experimental approaches for scaling *in vitro* genetic studies

Mapping genetic effects on cellular phenotypes with Census-seq

Manuscript for re-submission Fall 2023
(This is an updated version from a 2020 biorxiv preprint
https://doi.org/10.1101/2020.06.29.174383)

3.1 - Introduction

Human populations harbor vast numbers of common and rare alleles; such alleles affect the protein-coding sequence or regulation of almost all human genes. Human genetic studies have associated tens of thousands of alleles to risk of illnesses and other quantitative traits. A core goal of human genetics is to help identify cellular processes that underlie disease. And yet we know little today about how human alleles affect cells and their biology. We understand even less about how combinations of alleles – whether from one or many genes – converge upon cell-biological processes that might mediate normal variation and vulnerabilities.

Pioneering studies have shown that lymphoblastoid cell lines or pluripotent stem cells (PSCs) from human donors can be used to identify how common DNA variation shapes certain cellular phenotypes, especially RNA expression (Cheung et al., 2005, Morley et al., 2004, Stranger et al., 2007a, Stranger et al., 2007b, Kilpinen et al., 2017, Lo Sardo et al., 2017, McFarland et al., 2019, Pickrell et al., 2010, Jerber et al, 2021, Wells et al, 2023). The first challenge involves reaching the necessary scale by culturing and assaying the large number of cell lines necessary to associate phenotype with genotype. Challenges of scale have largely limited genome-wide genetic studies to a few labs or consortia and a few phenotypes. The second challenge involves control and rigor: how to accurately measure and compare phenotypes across many cell lines cultured separately. Without such control, it is often feared that biology can learn only from "alleles of large effect" – alleles that cause dramatic phenotypes in deterministic ways and thereby overwhelm noise in experimental measurement.

Here we describe an experimental system and computational methods ("Census-seq") that we developed to perform population-scale cellular experiments that enable insights from genetic influences of all kinds and frequencies upon diverse cellular phenotypes. Our approach involves what we call "cell village" experiments, in which cells from many genetically unique donors are mixed together. Cells undergo standard culture and experimental procedures in a shared environment, before being scored for a given phenotype all together. Census-seq analysis relates cells' phenotypes to the individual cell donors by analyzing the donors' DNA contributions to cell mixtures: after sorting or selecting the cell village for the phenotype of interest, we sequence the genomic DNA in the resulting, derived villages; computational analysis reveals the proportion of cells from each donor before and after sorting or selection. This approach allows many different kinds of cellular phenotypes to be analyzed for association to donor genotypes or other donor characteristics.

Census-seq addresses many challenges that have limited cellular genetic studies. First, we show that cell villages make it facile and inexpensive to do population-scale phenotype readouts and genetic analyses with cells. Second, we are able to measure phenotypes across hundreds of unique cell lines in a rigorous, well-controlled way. And third, we show Census-seq facilitates genetic analysis of phenotypes beyond RNA expression, which many previous methods have focused on. These experimental approaches provide a novel tool kit for the study of human genetics in living systems.

We sought here to establish such systems, understand their practical execution, apply them to a model phenotype, and enable other scientists to adopt similar approaches. We chose as a model phenotype the expression of the SMN protein, for which deficiency underlies Spinal Muscular Atrophy (SMA), a common congenital disorder engaged by emerging therapeutics (Chen, 2020, Lefebvre et al., 1995).

3.2 - Census-seq: determining the donor composition of a cellular 'village'

In order to perform cell village-based experiments, whereby we mix together cells from many donors, we first needed a computational tool that would enable us to leverage next generation sequencing (NGS) to measure the composition of each individual donor in the experiment. To do so, we developed Census-seq. Analogous to pooled CRISPR screening, with a key difference: Census-seq interrogates natural genetic variation rather than synthesized libraries of guide RNAs. In pooled CRISPR screens, cells are administered a library of gene-perturbing guide RNAs (each tagged with a DNA barcode) and then sorted or selected for a phenotype of interest; the relative frequencies of barcodes are compared between the initial population of cells and the population created by selection or sorting, and the effects of each guide on the phenotype are inferred from the change in that guide's representation (Adelmann et al., 2019, Canver et al., 2015, Gasperini et al., 2019, Hsu et al., 2018, Shalem et al., 2014, Wang et al., 2014). For Census-seq, we begin by constructing villages of cell lines derived from many donors; the cells in the village may be cultured or stimulated together over some period of time. Then, as in pooled CRISPR screening, we fractionate the cell village by sorting (or performing selections) to obtain cells with a phenotype of interest. By sequencing genomic DNA derived from the cell village, and applying a computational method we describe below, we then determine the relative contribution of each donor's genomic DNA to the cell mixture – and therefore the fraction of all cells that come from each donor – comparing the initial population to the population that results from fractionation or comparing fractionated populations to one another. Recent efforts have progressed these ideas using ddPCR to measure allele frequency in a mixture of hPSCs (Cederquist et al. 2020). However, this necessitates the design of custom primers for every individual contained within the experiment, which greatly reduces the long-term scalability of such experiments. Further, this method is vulnerable to common issues which may arise when performing population-scale studies such as sample swaps and DNA contamination.

Census-seq uses natural genetic variation (rather than synthetic barcodes (Yu et al., 2016)) to measure each individual's contribution to cell mixtures. The use of inherited variation as a natural barcode makes it possible to use human cells without the further perturbations (e.g. viral transfection and cloning) that alter cells' biology or contribute to the acquisition of mutations. To infer each donor's cellular contribution to a mixture, we isolate genomic DNA from the village, then perform low-coverage whole genome sequencing on that genomic DNA (generally about 1X average genomic coverage, at a cost of about $100 per cell village). We routinely analyze as few as 1x10^5 cells of a village, enabling many experiments to be performed in relatively small reaction chambers such as those on a 12-well plate.

We developed a computational approach to estimate each donor's contribution to this DNA mixture (**Figure 3.1A**). Making this estimate requires *a priori* genetic information on the individual donors, which can come from SNP arrays, exome sequencing, or whole genome sequencing (WGS) (**Figure S3.1E**). Sequencing the village's genomic DNA generates millions of sequence reads; these reads sample the donors' genomes in proportions that reflect the representation of each donor's cells in the village. A large minority of the sequence reads are "allelically informative" in the sense that they contain a genomic site that we know to vary among the potential donors; for example, in an analysis of a village of 40 donors whose genetic variation has been ascertained by WGS, about 42% of all 151-bp reads contain a genomic site for which the donors have varying genotypes (6,828,488 SNPs tested). The allele present on any such read offers partial information about the composition of the mixture, as only a subset of the potential donors' genomes can be sources of that sequence read.

In the mathematical analysis underlying Census-seq, we find the set of mixing coefficients (one coefficient for each donor, summing to 1.0) that make the observed sequence data – the millions of allelically informative reads, considered together – maximally likely to have been generated by random sampling from the donors' genomes (**Figure 3.1B**). The mixing coefficients are inferred via an expectation-maximization algorithm which works in the following way. At every variable site in the human genome, any hypothetical mixture of the donors involves an implicit "village allele frequency" for every allele in the DNA mixture (**Figure 3.1B**). (A default initial condition for analysis can be that each donor has contributed equal numbers of cells; in this case the village allele frequency for each allele is simply that allele's frequency among the donors, without weighting). We measure the likelihood of the village's sequence data by multiplying the village allele frequencies of all of the alleles from allelically informative sequence reads, making small adjustments to account for the possibility of sequencing error. To refine the mixing coefficients, we calculate the partial derivative of the likelihood of the data with respect to each donor's mixing coefficient; this calculation yields a set of adjustment factors by which we increase or decrease each individual donor's mixing coefficient to improve the data likelihood (**Figure 3.1B, C, D, Methods**). This "gradient ascent" process is repeated; as the mixing coefficients are adjusted, the likelihood of the observed sequence data increases toward an asymptote (**Figure 3.1D**). The computational analysis converges quickly – typically requiring just 10-30 iterations – to a set of mixing coefficients under which the observed sequence data are as likely as possible.

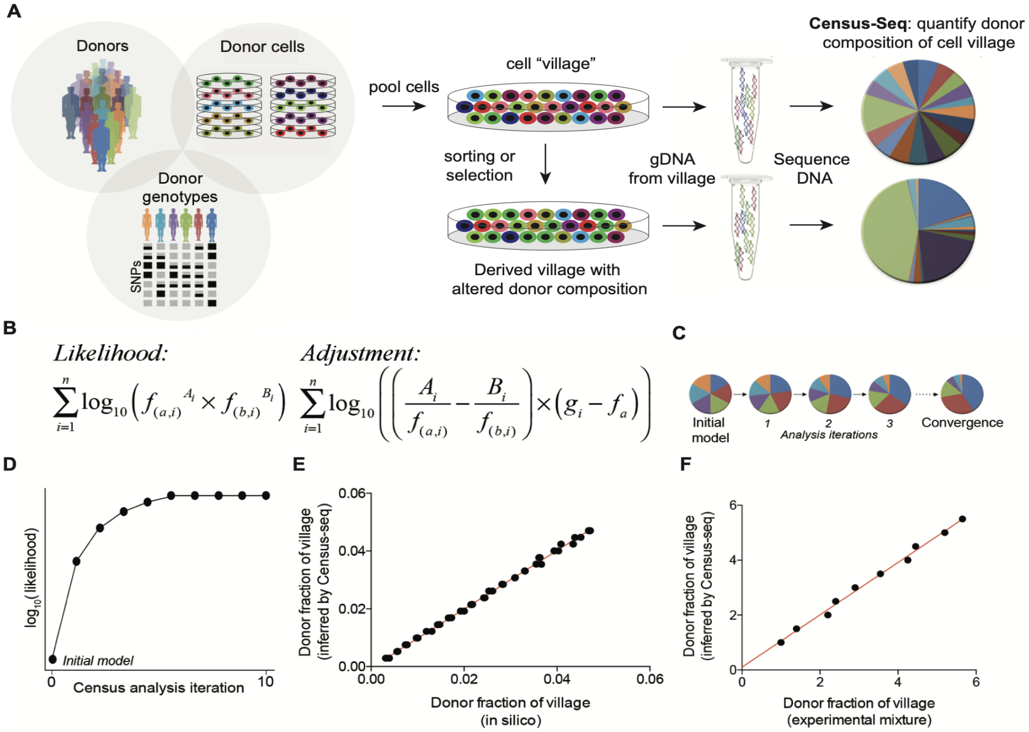

A
Donors
Donor cells
Donor genotypes
SNPs
pool cells
cell "village"
sorting or selection
Derived village with altered donor composition
gDNA from village
Sequence DNA
Census-Seq: quantify donor composition of cell village
B
Likelihood:
$\sum_{i=1}^{n} \log_{10}\left(f_{(a,i)}^{A_i} \times f_{(b,i)}^{B_i}\right)$
Adjustment:
$\sum_{i=1}^{n} \log_{10}\left(\left(\frac{A_i}{f_{(a,i)}} - \frac{B_i}{f_{(b,i)}}\right) \times (g_i - f_a)\right)$
C
Initial model
1
2
3
Analysis iterations
Convergence
D
$\log_{10}$(likelihood)
Initial model
0
10
Census analysis iteration
E
0.06
0.04
0.02
0.00
Donor fraction of village (inferred by Census-seq)
0.00
0.02
0.04
0.06
Donor fraction of village (in silico)
F
6
4
2
0
Donor fraction of village (inferred by Census-seq)
0
2
4
6
Donor fraction of village (experimental mixture)

Figure 3.1. Village-in-a-dish experimental systems.

A) Cells from many donors are cultured together as a "village" to enable scalability and minimize technical sources of variation. Sorting or selection enriches cells with phenotypes of interest, creating a derived village. Genomic DNA is extracted from each village and sequenced. Census-seq analysis measures each donor's contribution to the village's genomic DNA, and thus (indirectly) the relative number of cells from each donor in each village. **B**) Census-seq analysis is based on an expectation-maximization (EM) algorithm. The algorithm seeks the set of donor-mixing coefficients (summing to 100%) that maximize the likelihood of the observed sequence data. For any set of coefficients, this likelihood is measured by multiplying the modeled allele frequencies (in the village mixture) of all of the alleles observed on all allelically informative sequence reads. The donor-mixing coefficients are then adjusted in the direction that most strongly increases the likelihood of the observed data; the adjustments are derived by taking the derivative of the likelihood with respect to each of the donor-specific mixing coefficients (Methods) (fa = the frequency of the reference allele, A = the counts of the reference allele in the sequence data, fb = the frequency of the alternate allele, B = the counts of the alternate allele in the sequence data, gi = the genotype of the donor at the site, formatted as 1 for the reference genotype, 0.5 for the heterozygous genotype, and 0 for the alternate allele). **C**) Iterated rounds of adjustment optimize the estimates of donor mixing coefficients to fit the sequencing data, converging asymptotically to a final estimate. **D**) Convergence is typically reached after just a few iterations of the EM algorithm. **E**) Simulated "village DNA" data sets were made by mixing whole genome sequence data from 40 unrelated donors in such a way that each donor contributed a different proportion of the overall data. Census-seq was used to estimate the quantitative contribution of each donor to this data mixture. The plot compares the known, in silico mixing coefficients to the estimates from Census-seq. **F**) Genomic DNA from ten donors was mixed in such a way that each donor contributed a different proportion of the total DNA. The DNA mixture was sequenced and analyzed by Census-seq. The plot compares the aliquoted donor-contribution proportions to the Census-seq estimates from sequencing the DNA mixture.

To evaluate whether the donor composition of villages inferred by Census-seq corresponded to their known, actual composition, we performed many control analyses and experiments, including (i) analyzing *in silico* simulations in which we mixed DNA sequencing data from many individuals in known proportions; (ii) mixing genomic DNA from different individuals in known concentrations; and (iii) mixing cells from individuals in known proportions. In each case, the Census-seq estimates of donor representation in the mixture corresponded closely to the ratios in which sequence data, DNA or cells from different donors had been mixed (**Figure 3.1E, F, 3.2**). Overall, we found that a donor's DNA representation in a village could be measured accurately down to a limit at which a donor contributes about 0.2% of the cells in a mixture (**Figure 3.2F**), a limit that is related to the sequencing error rate and thus is not addressed by simply sequencing Census-seq libraries more deeply. This limit places an upper bound (of a few hundred) on the number of unique donors that can be accurately quantified in one village. This bound is still far above the scale of experiments that can be accomplished comfortably in traditional formats; still-larger experiments are possible by meta-analyzing many villages with overlapping membership for calibration.

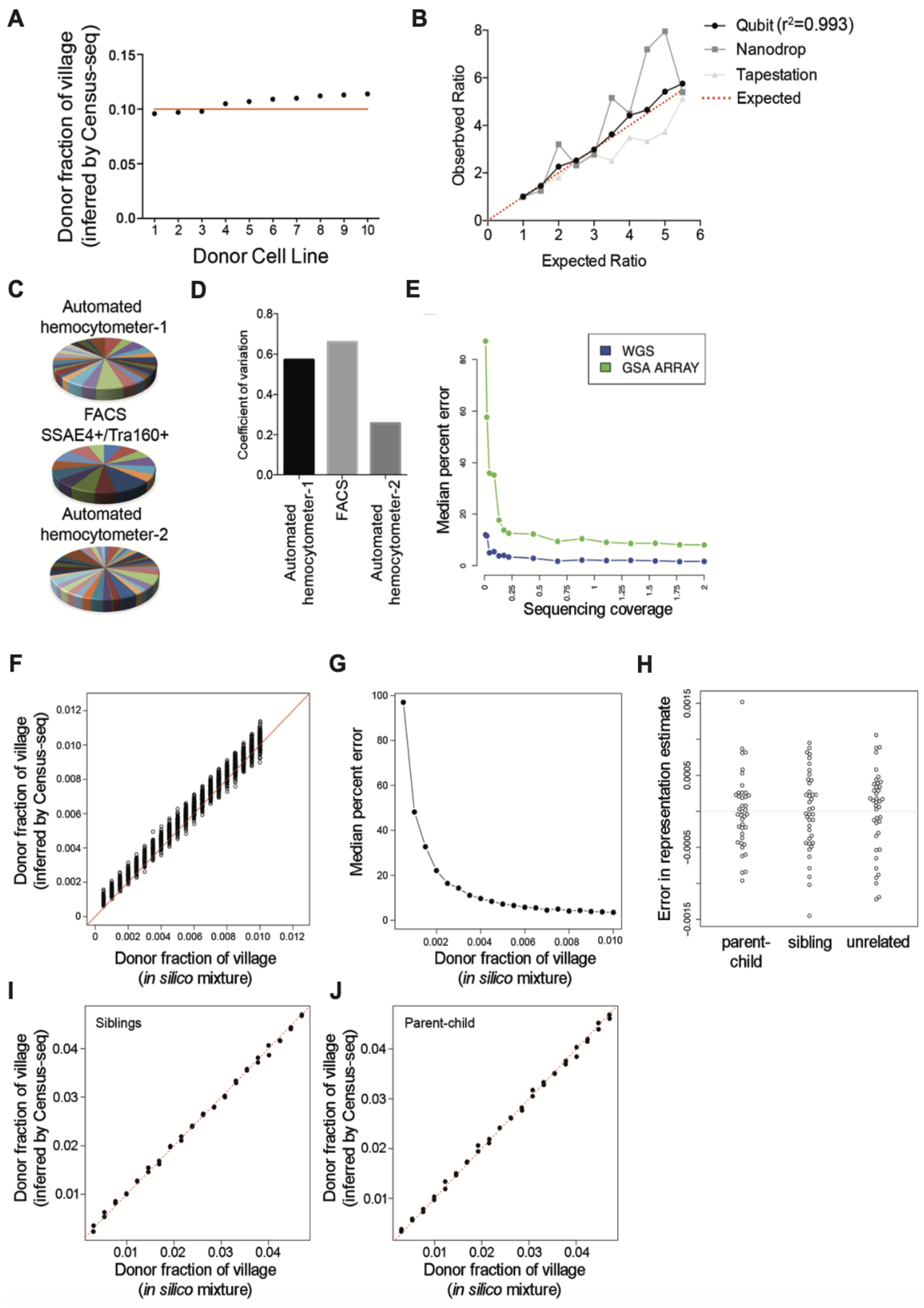


Figure 3.2. Construction and quantification of cell villages.

A) Equal quantities of genomic DNA from 10 donors were mixed, sequenced and analyzed by Census-seq. The plot compares the aliquoted donor-contribution proportions of the DNA mixture to the Census-seq estimates from the sequencing data. **B**) Genomic DNA from six donors was mixed in an arithmetic progression of quantities as estimated by three DNA-quantitation technologies (Qubit, Nanodrop, Tapestation). Donor contribution to the resulting DNA mixtures was then estimated by Census- seq. Census-seq estimates of donor-specific contributions to the mixture agreed most strongly with input estimates from the Qubit (r2 = 0.993). **C**) Cell lines were mixed using three different methods of cell quantification, with the goal of making equimolar mixtures; pie charts show the resulting donor composition of each mixture as estimated by Census-seq. The three cell-counting methods were: a hemocytometer (Countess, for live/dead estimation; flow cytometry with selection for SSEA4 and Tra-1-60 positive and pluripotent) cells; and an automated cell counter (Scepter) based on cell size. **D**) Using the data from (D), the coefficient of variation (standard deviation divided by the mean) of the donor contributions was estimated for each of the methods of cell quantification. **E**) Precision of the Census-seq algorithm as a function of (i) the type of a priori genome-variation data available for each donor (line colors) and (ii) the depth to which the village genomic DNA is sequenced (y-axis). A key relationship is that deeper a priori genetic analysis (e.g., WGS, blue curve) of the individual donors' genomes makes it possible to infer the donor composition of mixtures even from sequence data that are relatively low-coverage. Analysis is based on WGS data mixed from 40 donors. **F**) Results of in silico data-mixing experiments to estimate the bias and variance of Census-seq inferences for donors who have contributed small proportions (0.05% to 1%) to a mixture. WGS data from 40 unrelated donors was mixed in silico, with 30 donors at an arithmetic series of representations from 0.05% to 1.00% in 0.05% increments, and 10 donors (for whom data not shown) at higher representations, such that the 40 donors' representations summed to 1. This was repeated for 10 simulations, in each of which donors were permuted between the low and high representation groups at each iteration to generate a total of 300 observations of each bin. **G**) Bias of Census-Seq estimates (as a fraction of the estimate) for donors who have contributed small fractions (< 1%) of a cell or DNA mixture. Using the data shown in S1F, the median absolute error in representation was calculated as exp(median (abs(log(donor % representation in silico / donor % representation inferred by Census-Seq)))). Bias in donor representation was substantially greater for donors contributing <0.3% to a mixture. Bias was <15% at a representation of 0.3%, and <10% at a representation of 0.4%. We believe that this bias arises from PCR and sequencing errors, which result in reads that (as a group) tend to bias upward the estimated representation of very-low-contribution donors. For this reason, we exclude from some genetic analyses those donors with contributions of <0.3% to a mixture. **H**) To assess the potential effect of having genetically related donors in a mixture, three in silico mixtures of 40 donors were constructed for: 40 unrelated individuals; 20 parent/child pairs; and 20 sibling pairs. In each analysis, WGS data from these donors was mixed uniformly at a representation of 0.025; the data mixture was then analyzed by Census-seq. Error in representation was calculated as the difference between the known and Census-seq-inferred donor-contribution estimates. 95% of inference were within an absolute error of 0.001. The median absolute error in representation estimates were similar: unrelated 3.4x10-4, sibling= 4.1x10-4, parent-child= 2.6x10-4. **I**) To evaluate the robustness of Census-seq inference to the inclusion of genetically related individuals in a village, in silico data mixing was used to simulate a village of 20 sibling pairs, with the same distribution of representations as in Figure 3.1E. **J**) To evaluate the robustness of Census-seq inference to the inclusion of genetically related individuals in a village, in silico data mixing was used to simulate a village of 20 parent-child pairs, with the same distribution of representations as in Figure 3.1E.

We found that composition of 40-donor villages could be inferred accurately with modest amounts of sequencing that corresponded to less than 1X coverage of a single human genome (**Figure 3.2E**). The required depth of sequencing depended on the complexity of the village (number of donors) and the amount of available genome information on these donors (**Figure 3.2E**) (Methods), with deeper *a priori* genetic characterization (e.g., WGS) causing more sites to be allelically informative and thus allowing lighter sequencing of the village's DNA at the time of phenotype analysis. For example, if a donor has contributed 2.0% of the DNA in a mixture, sequencing the village genomic DNA to a sequencing depth of about 1X (16 million 150-bp reads) yields estimates of 2.0 +/- 0.1%. We routinely analyze 16 villages in each run of a desktop sequencer (Illumina NextSeq) (Methods) or send samples out to fee-for-service providers at a cost of about $100 per village.

These results encouraged us to use this approach to analyze a great many cell villages of experimental composition and to study the population dynamics of villages of PSCs as they grew in culture together. While the initial results were sobering, with significant variance across donor cell line distributions overtime, continued optimization of cell culture workflows greatly increased the feasibility of performing these kinds of experiments. Further, we tested whether our cell village-based approach was amenable to common *in vitro* practices including cryopreservation and cellular differentiation, which greatly reduces the technical challenges of cell village experiments and increases their application to cell type specific levels of analysis.

3.3 - Cell villages dramatically increase the feasibility of population-scale studies

To better understand population dynamics of PSC villages, we generated more than 30 villages, each consisting of 10-100 unique donors, and measured their donor dynamics across several passages; in total, we measured 3,705 Census-seq growth phenotypes in cell lines from 247 unique individuals (**Figure 3.3A-C, Table S3.1, S3.2**). To do this, cells from many donors are collected and thawed in their own, separate culture vessel. The cell lines then expand until they reach relative levels of confluency across each cell line selected. Next, cells are dissociated and counted and mixed together in equal proportions before being replated into a 'village' (**Methods, Figure 3.4A**). For initial experiments we utilized cell lines drawn from a large collection of human embryonic stem cells we had assembled and recently genetically characterized by whole genome sequencing (Merkle et al., 2022, Merkle et al., 2017). In most experiments, we found that one or a few cell lines progressively took over the village (**Figure 3.3B, C**). Analyses of the whole genome sequences of these cell lines indicated that a majority of the hyperproliferative cell lines had growth-promoting mutations previously identified (**Figure 3.3B**), some of which – such as mutations in the *TP53* gene and the gene encoding the p53 inhibitor MDM4 – recurred in multiple cell lines (Loh et al., 2018, Merkle et al., 2017). These experiments demonstrated the utility of Census-seq to confirm the cell biological impact of these growth-promoting mutations which we had identified in earlier work (that would ideally be excluded from translational efforts) (Merkle et al., 2017). These results also indicated that variation in growth rates among proliferative cell types such as stem cell lines must be managed to maintain large villages. We noted that such donor dynamics could be ameliorated when generating villages from non-proliferative cell types, such as glutamatergic neurons (**Figure 3.3D**). However, this still required differentiation of each cell line individually before being mixed together.

We next sought to ameliorate some of the population dynamics observed in our early experiments in an effort to continue to scale cell village methods and extend their applications to proliferative cell types. We implemented a rigorous quality control process and excluded cell lines which harbored deleterious mutations as identified by our *a priori* genetic analysis. In this way, we approached each cell village experiment with a more informed donor cell line selection. We observed that donor dynamics shifted to the greatest degree during the passaging of cells (**Figure 3.4B**). This suggested that in some cases, the variance in donor distributions may be attributed, not only to the varying growth rates across cell lines, but to differences in adherence and survival following mechanical manipulation. To test these assumptions, we generated several new cell villages containing 45-48 induced pluripotent stem cell (iPSC) lines derived from unique donors for whom we identified no known growth-promoting mutations. These villages were generated and plated for expansion (without being passaged), and their donor composition was ascertained by Census-seq every few days. When we incorporated these more intentional village designs, and minimized the number of passages, we observed that donor dynamics of the iPSC villages remained stable over time in culture (**Figure 3.3E**). These results suggested that with well-informed experimental practices, we could perform cell village experiments at scale in proliferative cell types and maintain stable compositions of different donor cell lines, thereby extending the utility of cell villages beyond post-mitotic cell types.

Another hurdle to scaling up the number of cell lines in a given cell village is the challenge of growing dozens to hundreds of cell lines simultaneously to create a village. We wondered whether cell villages were amenable to cryopreservation in the same way a standard single cell line culture was. If cell villages could be generated ahead of time and frozen down, this would greatly facilitate the execution of population-scale experiments, possibly starting from a single cryovial. To test this, we took the same three villages from above and cryobanked them at the time of village generation. Several months later, the villages were thawed, recovered, and collected for Census-seq. We used Census-seq to measure the donor distributions at the time of banking (before freezing) and following 24hr recovery post-thaw (to ensure we only measure cells which attach to the culture dish substrate). We found that donor distributions remained stable and were not significantly disrupted by the freeze-thaw cycler (**Figure 3.3F**).

Having demonstrated that we could successfully generate cell villages which remained stable overtime in culture, and could be cryopreserved for use later on, we sought to move beyond PSCs into differentiated cell types. We had shown previously that we could generate a village from post-mitotic neurons (**Figure 3.3D**), however this limits our ability to explore phenotypes at critical windows in cellular differentiation and development.

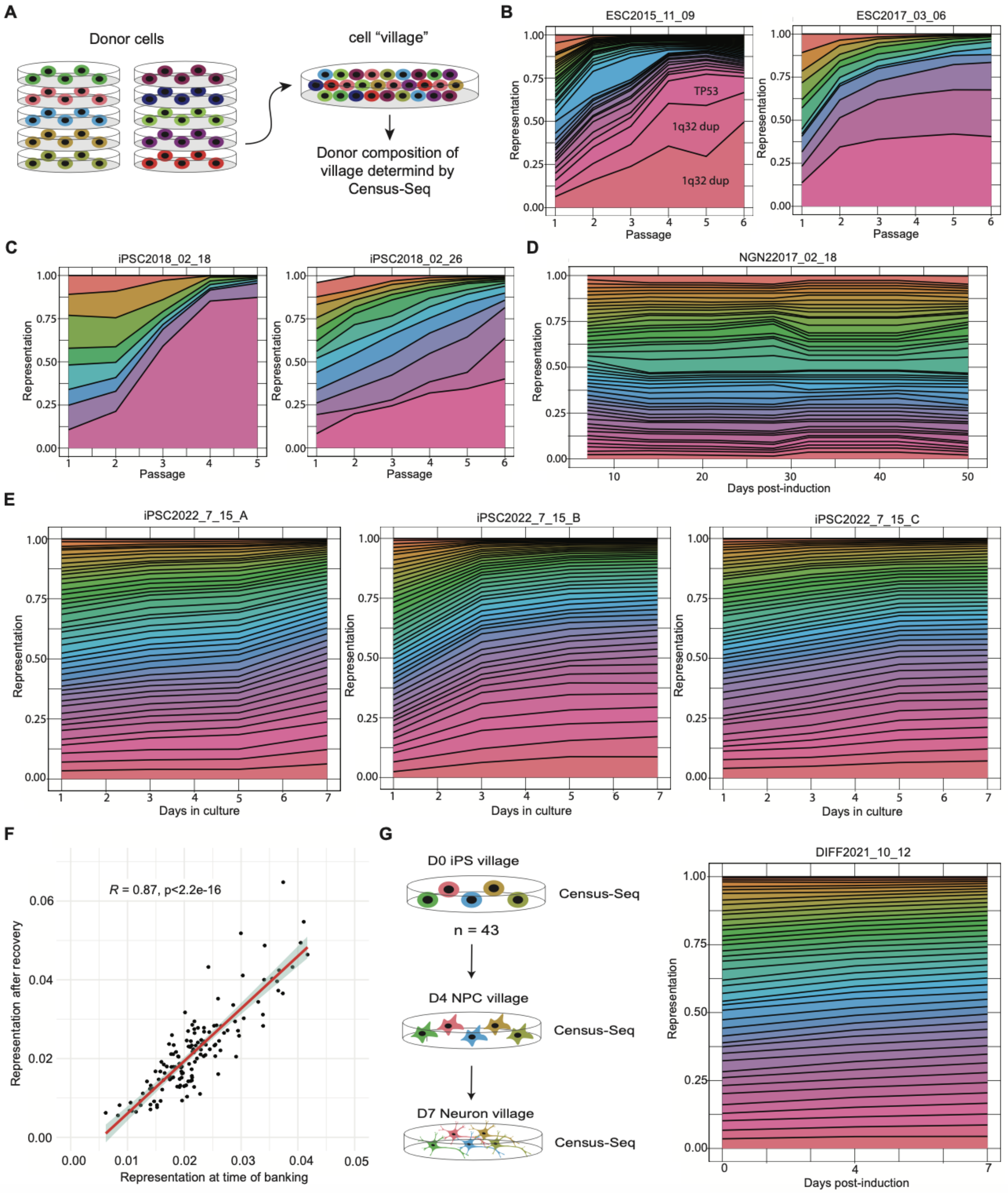
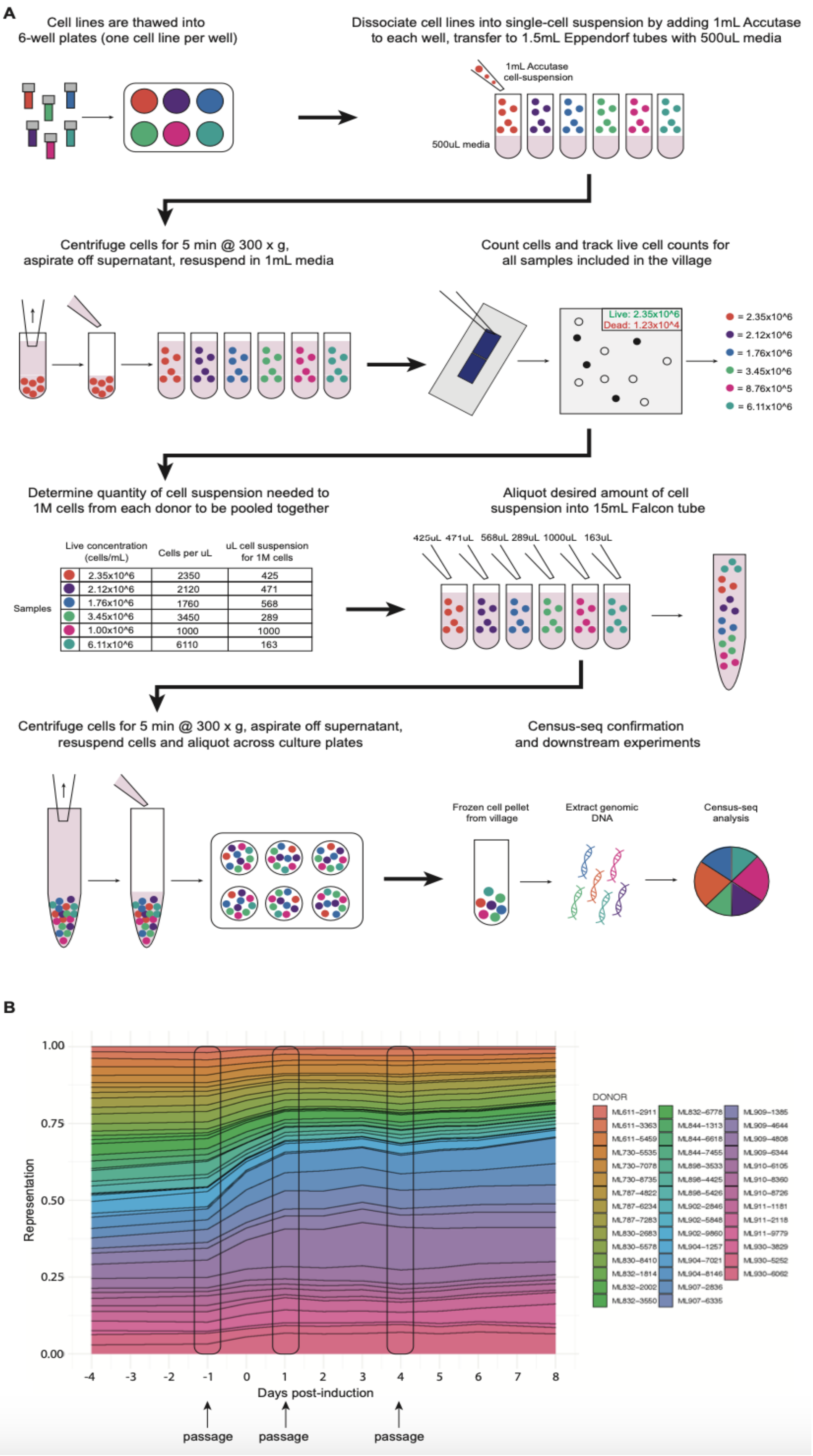


Figure 3.3. Cell villages dramatically increase the accessibility of population-scale studies.

A) Cell villages were generated from hPSCs, and donor compositions are monitored using Census-Seq. **B**) Early experiments with ESCs identified several hyperproliferative lines which harbored growth-promoting mutations. **C**) In iPSC villages, hyper-proliferative lines can quickly distort the donor composition of villages, rendering them unsuitable for many kinds of experiments. **D**) Dynamics donor distributions could be ameliorated when using non-proliferative cell types. **E**) Workflow optimizations and informed village selection vastly reduces donor cell line variability overtime in culture. F) Comparison of donor representation between cell villages prior to cryopreservation and following thaw and recovery of villages. **G**) Analysis of donor representation dynamics throughout neuronal differentiation from cell villages of iPSCs.

To explore whether our optimized workflows could help overcome this limitation, we constructed a village of 43 iPSC lines from unique donors and differentiated them to glutamatergic neurons by inducing the expression of Neurogenin 2 (Ngn2) combined with small-molecule patterning (Nehme et al., 2018, Zhang et al., 2013). We ascertained the donor composition of D0 iPSCs, D4 neuronal progenitor cells (NPCs), and D7 post-mitotic neurons by Census-seq. We found that donor distributions remained consistent across neuronal differentiation, confirming that cell lines could be mixed together prior to the induction of differentiation, and highlighting the utility of cell villages for cell-type specific experiments (**Figure 3.3G**).

While we had reliably established this novel experimental system, we were aware of various issues which may arise when performing large scale cellular studies, such as sample mix-ups. This may lead to cells from unintended donors being included in a given experiment. Commonly the result of human error, we needed to develop tools which would allow us to determine if unwelcome cell lines had intruded our experiments, and to correctly identify their donor.



Cell lines are thawed into 6-well plates (one cell line per well)

Dissociate cell lines into single-cell suspension by adding 1mL Accutase to each well, transfer to 1.5mL Eppendorf tubes with 500uL media

Centrifuge cells for 5 min @ 300 x g, aspirate off supernatant, resuspend in 1mL media

Count cells and track live cell counts for all samples included in the village

Determine quantity of cell suspension needed to 1M cells from each donor to be pooled together

		Live concentration (cells/mL)	Cells per uL	uL cell suspension for 1M cells
Samples	●	2.35x10^6	2350	425
	●	2.12x10^6	2120	471
	●	1.76x10^6	1760	568
	●	3.45x10^6	3450	289
	●	1.00x10^6	1000	1000
	●	6.11x10^6	6110	163

Aliquot desired amount of cell suspension into 15mL Falcon tube

Centrifuge cells for 5 min @ 300 x g, aspirate off supernatant, resuspend cells and aliquot across culture plates

Census-seq confirmation and downstream experiments

Figure 3.4. Workflow for generating cell villages.

A) hPSCs are thawed into 1w/6w plate and grown until each line reaches ~50% confluency. Next, each cell line is dissociated using Accutase to form a single cell suspension. Once cells are detached, they are titruated and transferred to a 1.5mL Eppendorf tube filled with 500uL of mTeSR1 medium supplemented with 10uM ROCK inhibitor. The Eppendorf tubes are centrifuged at 300 x g for 5 min to pellet the cells. Once cells have finished spinning, aspirate media from each individual tube such that a few uL of media are left to avoid aspirating the cell pellet. All cell pellets are then resuspended using 1mL of mTeSR1 medium supplemented with 10uM ROCK inhibitor to produce a single cell suspension. Concentration of cell suspensions are quantified using a Countess Fluorescence Cell Counter. Once all cell lines are counted, determine the amount of each cell suspension desired to generate a cell village. As a general rule of thumb, we aim for 1 million cells per individual line but one can reliably generate villages with many fewer cells. Aliquot the desired amount from each cell line into a new 50mL Falcon tube until all cell lines have been added. Centrifuge the 50mL Falcon tube containing all cell lines at 300 x g for 5 min to pellet the cell village. Following the centrifugation, resuspend cells in mTeSR1 medium supplemented with ROCK inhibitor and plate across new geltrex coated tissue culture plates. The following day, collect one well from the new plate to measure donor distribution using Census-Seq. **B)** Cell villages of iPSCs show donor distributions shift most dramatically following passaging rather than intrinsic cell growth behaviors.

3.4 - A computational toolkit for village experiments: “Roll Call” and “CSI”

We developed computational approaches to quickly authenticate a village (or other DNA mixture) and identify and diagnose the presence of cells with unexpected genomes, which we refer to here as contamination (**Figure 3.5, Methods**). These methods have the added benefit of being generally useful for validating the provenance of large numbers of cell lines (Nelson-Rees et al., 1981).

Contamination of a village could in principle arise from a donor of known genotype, or a donor of unknown genotype; we developed two computational approaches to address these cases. To illustrate these approaches, we have drawn on the example of a village whose 12 donors initially appeared to have maintained a stable balance through six passages (**Figure 3.5E**).

To detect the presence of donors with known genome sequences, we developed “Roll Call”, which utilizes variants that distinguish individual donors from all other donors for whom cells might be present in a given lab or project (**Figure 3.5A, 3.6A-D**); we call such variants Identifying Rare Variants (IRVs). The presence of a sufficient number of sequence reads with an individual donor’s IRVs confirms the presence of that donor’s cells in a village. In the experiment in question, Roll Call identified that the village DNA contained a great many IRVs from an unexpected cell line (CSES15) – a cell line whose genome we had previously analyzed but was not meant to be included in the village – and suggested that this cell line had become more abundant in the village with each passage (**Figure 3.5C**). Identifying the contaminating donor made it possible to correct the Census-seq analysis to account for this unexpected donor by including their genotypes among those of the other candidate donors in Census-seq analysis; this analysis revealed that cells from the CSES15 line had in fact taken over the village (**Figure 3.5F**). Examination of whole genome sequence data from CSES15 revealed that it harbored an acquired mutation in *MDM4.*

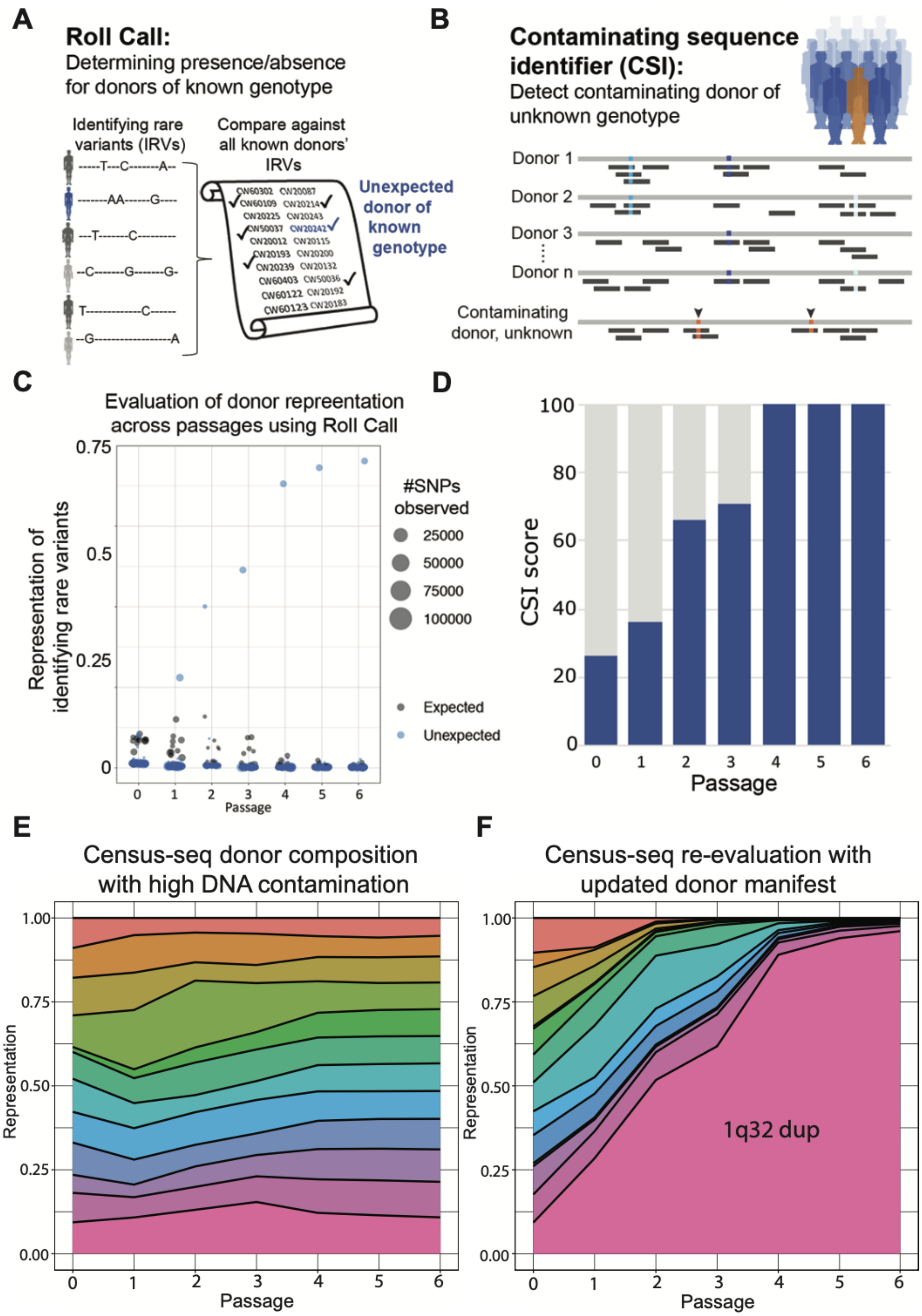

A
Roll Call:
Determining presence/absence for donors of known genotype
Identifying rare variants (IRVs)
-----T---C-------A--
-------AA------G----
---T-------C---------
--C-------G-------G-
T------------C------
--G-----------------A
Compare against all known donors' IRVs
CW60302 CW20087
CW60109 CW20214
CW20225 CW20243
CW50037 CW20242
CW20012 CW20115
CW20193 CW20200
CW20239 CW20132
CW60403 CW50036
CW60122 CW20192
CW60123 CW20183
Unexpected donor of known genotype
B
Contaminating sequence identifier (CSI):
Detect contaminating donor of unknown genotype
Donor 1
Donor 2
Donor 3
Donor n
Contaminating donor, unknown
C
Evaluation of donor repreentation across passages using Roll Call
Representation of identifying rare variants
0.75
0.5
0.25
0
0
1
2
3
4
5
6
Passage
#SNPs observed
25000
50000
75000
100000
Expected
Unexpected
D
CSI score
100
80
60
40
20
0
0
1
2
3
4
5
6
Passage
E
Census-seq donor composition with high DNA contamination
Representation
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
5
6
Passage
F
Census-seq re-evaluation with updated donor manifest
Representation
1.00
0.75
0.50
0.25
0.00
1q32 dup
0
1
2
3
4
5
6
Passage

Figure 3.5. Using Roll Call and CSI to inspect donor composition of villages.

A) The Roll Call algorithm is used to confirm the presence or absence of individual donors (with known, a priori genetic variation information). Roll Call utilizes Identifying Rare Variants (IRVs) – rare variants that distinguish individual candidate donors from all other candidates. **B**) Roll Call identified the intrusion of a familiar (but unexpected) cell line into a village, indicating that this line (light blue) dominated donor representation within a few passages. **C**) The Contaminating Sequence Identifier (CSI) algorithm detects the presence of DNA from unexpected donors of unknown genotype (orange). **D**) In an in-silico analysis blinded to any genetic information about the contaminating cell line, CSI predicted that an unfamiliar donor was present and had increased representation with each passage. **E**) Census-Seq donor composition over time in culture shows a balanced donor distribution. However, this village displayed a high degree of DNA contamination (**C, D**), suggesting a donor of unknown origin was included. **F**) Re-evaluating the donor composition of this village using Census-seq with an updated roster of donor cell lines confirmed the presence on a dominant intruding cell line. This line was determined to have acquired multiple growth-promoting mutations, including within MDM4, which encodes a regulator of the P53 tumor suppressor.

A more challenging analytical problem can arise if contaminating cells come from a donor whose genotypes are unknown. We developed the Contaminating Sample Identifier (CSI) algorithm to detect the presence of contaminating cells from genetically unknown donors (**Figure 3.5B, D, 3.6E-G**). CSI utilizes sequence reads that suggest the presence of alleles that are segregating in human populations yet (by chance) absent among the candidate members of a village; CSI determines whether such reads are sufficiently numerous that sequencing error is unlikely to explain them. To evaluate CSI, we asked whether it could identify the presence of contaminating cells in the village described above but do so in the absence of any *a priori* genetic data for the contaminating CSES15 line. Indeed, CSI analysis suggested the presence of an unexpected donor, at increasing frequency with each passage (**Figure 3.5D, F**). Follow-up CSI and Roll Call analyses of individual cell lines (from which the village was made) allowed us to identify the tube that had initially become contaminated with CSES15.

Roll Call and CSI can be used to authenticate cellular reagents – for individual cell lines as well as villages – in a wide variety of laboratory contexts. With our experimental and computational toolkit now in hand, we explored the use of Census-seq to simultaneously perform cellular phenotyping and genetics analyses on cell villages.

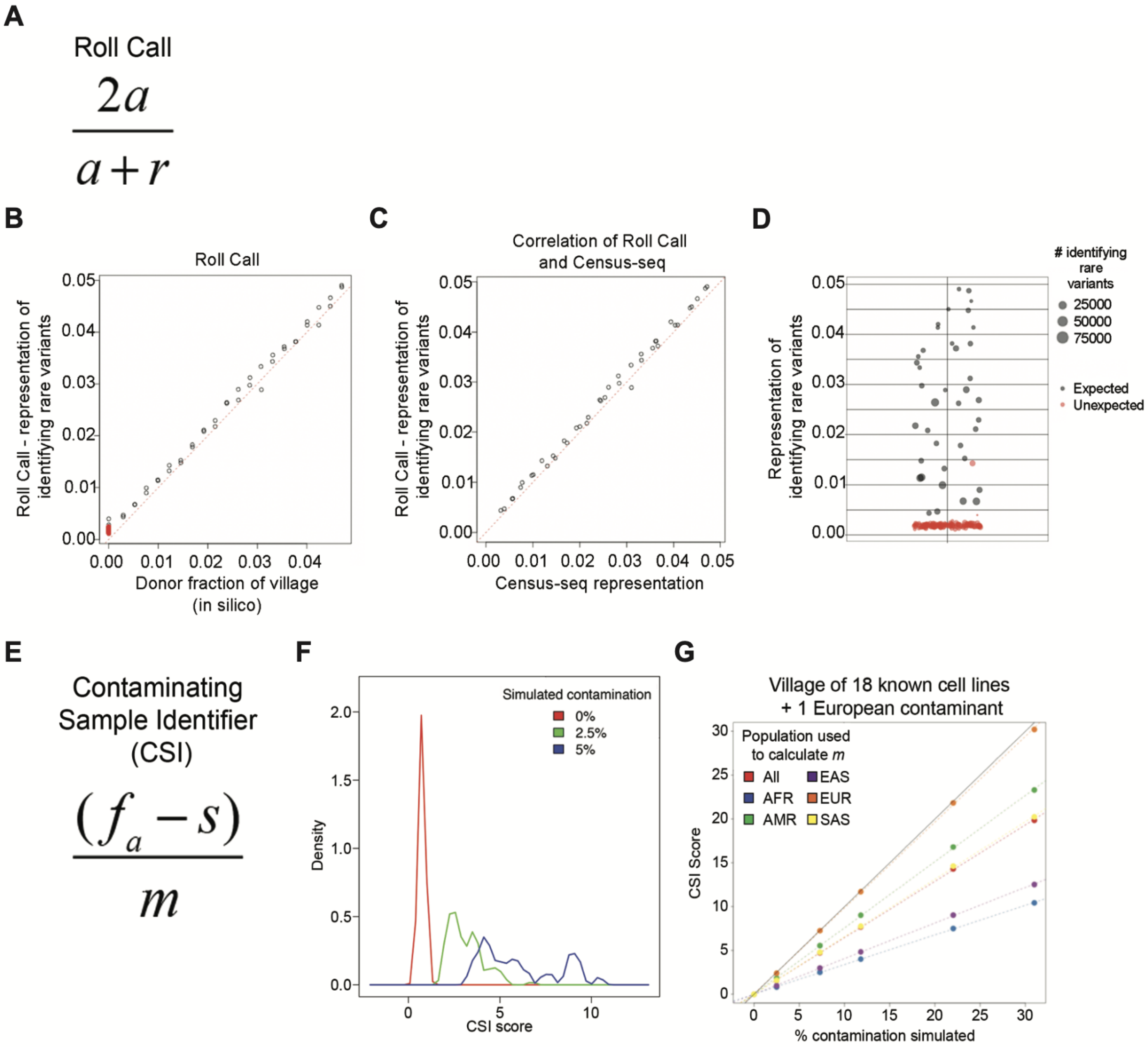


Figure 3.6. Algorithms and validation analyses for the Roll Call and CSI tools for authenticating cell villages.

A) The Roll Call algorithm leverages Identifying Rare Variants (IRVs) – rare alleles that are present in the genome of only one of the candidate donors – to determine which candidate donors have contributed cells/DNA to a mixture. (Census-seq offers more- accurate quantification of donors' quantitative contributions than Roll Call does, but only when starting with a complete and accurate list of donors.) The formula shows the calculation of the Roll Call score (for an individual donor, using that donor's IRVs), with which the algorithm evaluates whether a donor has contributed to a DNA mixture. In this formula, a and r refer to the numbers of observations of the reference and alternate alleles at that donor's IRVs in the village sequencing data. **B**) DNA sequence data from 40 unrelated donors was mixed at known proportions. Roll Call scores for these donors (x-axis) are compared to the proportions in which DNA data have been mixed (y-axis). An additional 147 donors (shown in red), whose DNA sequence data was not included in the data mixture, received scores < 0.005 [0.002-0.004]. **C**) Analysis by Roll Call and Census-seq of a synthetic mixture of WGS data from 24 unrelated individuals, mixed at an arithmetic of contributions across donors. Donors that are known to be in the mixture are observed at close to their correct mixtures. Donors not in the mixture have scores < 0.005 [0.003-0.004]. **D**) Roll Call analysis of an example cell village confirmed the presence of 45 expected donors (grey circles); confirmed the absence of all but one of the unexpected donors absent (red circles); and flagged the presence of a donor not intended to have been included in that village (red circle at y=0.014); follow-up analysis confirmed contamination by cells from this unexpected donor. **E**) The CSI (Contaminating Sample Identifier) algorithm calculates an intrusion score from sites that are monomorphic among the expected donors (but variable in the wider population from which the donors are sampled) – i.e., from alleles that should not have arisen from the expected donors' genomes and must therefore represent contaminating cells or sequencing errors. CSI utilizes genomic sites known to vary in the wider population (from which the donors are sampled) that happen to be monomorphic among the candidate donors. fa is the fraction of allelically informative reads (at these sites) that contain the alternative allele; s is the sequencing error rate; m is the mean minor allele frequency of these alleles. **F**) Distributions of CSI scores for in silico villages created by WGS data mixing to have 0%, 2.5%, or 5% contamination from an "unknown" donor to whose genetic data the CSI analysis was blinded. The in-silico mixing experiments each involved a mixture of 39 known, unrelated donors at varying concentrations and (in the 2.5% and 5% cases) an additional randomly selected unrelated donor, to whose genetic data the CSI analysis was blinded. 100 simulated villages per contamination level were analyzed. The 100 null (uncontaminated) simulations yielded CSI scores of 0.0074 +/- 0.0011; CSI scores in all 100 of the 2.5% contamination simulations exceeded any result from this null distribution. **G**) Results of CSI analyses of synthetic mixtures of WGS data from 18 known donors, plus an additional unexpected donor for whom WGS data was spiked in (to the mixture) at several proportions ("% contamination simulated", x-axis). Note that because the CSI formula utilizes the mean minor allele frequency of "unexpected" alleles in the sampled population, the CSI intrusion score estimates %contamination correctly only if the intruding cell line is indeed from the population used to estimate this – in this case, 1000 Genomes European-ancestry sample (EUR, orange), since the simulated contaminating line was of European ancestry. When other populations are used, or when the unexpected donor is related to an expected donor (not shown), the intrusion score mis- estimates %contamination. Since the identity of a genomically unknown contaminating line is by definition not knowable a priori, the intrusion score is primarily intended to be used as a diagnostic for contamination rather than as a precise measurement of %contamination; however, its change within a village (e.g., over cell passages) in principle reflects change in the proportion of contaminating cells within the village.

3.5 - Cellular phenotyping and genetic analysis in villages

To analyze how a cellular phenotype of interest varies among donors, we analyze how sorting or selection for that phenotype changes the representation of each donor's DNA in the cell mixture (**Figure 3.7**). Although a donor's individual cells may also vary in phenotype, donor-to-donor variation in the mean and/or variance of such distributions will alter the representation of each donor's DNA in the resulting villages (**Figure 3.7A**). For example, to analyze the expression level of a specific protein, we stain about 10^6-10^7 cells from a cell village with an antibody to that protein, then sort the cells based on the immunofluorescence signal (**Figure 3.7A, 7B**). Each donor's quantitative phenotype is then the ratio of their DNA representation in two different selected villages (**Figure 3.7C**).

The ability to analyze DNA recovered from fixed, sorted cell mixtures enables many kinds of analyses. Intracellular as well as cell-surface proteins can be analyzed. Flow cytometry conditions and gating thresholds can be optimized, for example by using mutant cells, strongly perturbed cells, or secondary-antibody-only conditions to define reference distributions of phenotypic values (**Figure 3.7C, 3.8**). Aliquots of a cellular village, once prepared and fixed, can be used to analyze many different proteins.

As a model cellular phenotype, we analyzed genetic influences on the Survival of Motor Neuron (SMN) protein, which in humans is encoded by the paralogous *SMN1* and *SMN2* genes on chr5q13 (Lefebvre et al., 1995, Lorson et al., 1999, Monani et al., 1999). SMN deficiency results in widespread splicing defects and causes Spinal Muscular Atrophy (SMA). Though the coding-sequence differences that distinguish *SMN1* from *SMN2* are all synonymous changes in codon usage, *SMN2* lacks a key splicing enhancer, with the result that the majority of *SMN2* mRNAs produce a shorter protein (*SMNdelta7*) whose inability to rescue *SMN1* deficiency has been attributed to protein instability and perhaps to nonsense-mediated decay (Burnett et al., 2009, Hua et al., 2008). SMN deficiency is primarily caused by mutations in *SMN1,* which are under these

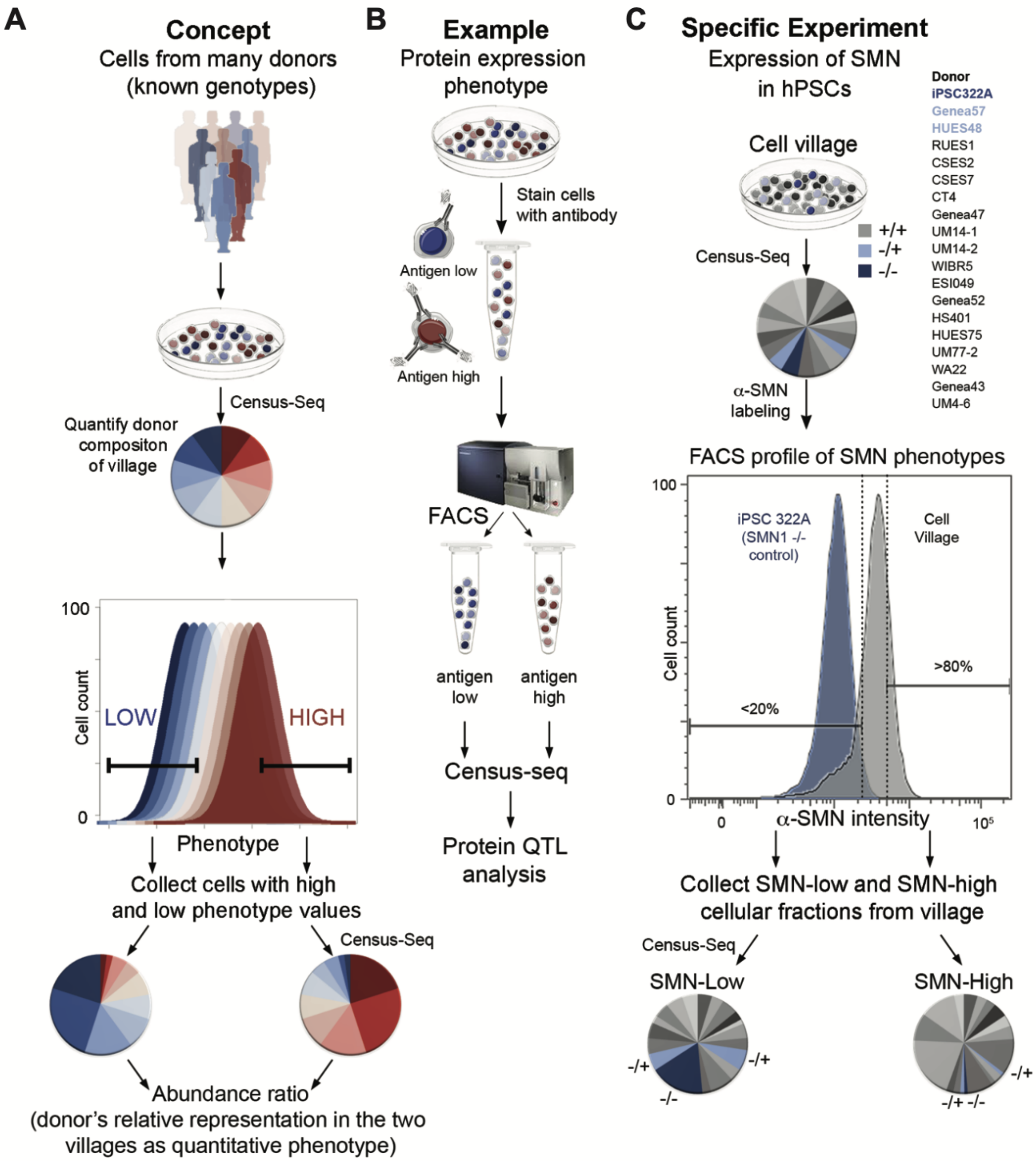


Figure 3.7. Using cell villages and Census-seq to analyze cellular phenotypes at population scale.

A) A cellular phenotype may be affected by both inter-individual biological variation and single-cell variation (in biology or measurement). Sorting or selecting the cells in the village based on this phenotype creates derived villages. If inter-individual biological variation shapes this cellular phenotype, then the derived villages will have different donor compositions. A donor's change in representation between such derived villages is a quantitative phenotype that can then be analyzed genetically or in relationship to other variables, such as donor age or health status. **B**) Ascertainment of inter-individual variation in protein expression from cell villages using Census-seq. A cell village is fixed and stained with an antibody to a protein or post-translational modification of interest. The cell village is FACS-sorted for level of immunoreactivity. The donor composition of each cell fraction is analyzed by Census-seq. **C**) A cell village was analyzed for expression levels of the SMN protein. The pilot village, consisting of PSCs from 19 donors, included cells from a donor with spinal muscular atrophy (SMA, a recessive genetic disorder caused by SMN deficiency) and two carriers of recessive SMA mutations. Comparisons of the FACS-derived SMN-high and SMN-low cell villages by Census-seq indicated that cells from all three donors were more abundant in the SMN-low than the SMN-high fraction. This effect was strongest for cells from the SMA patient, and also detectable in the two carriers.

circumstances not rescued by *SMN2*. An emerging therapeutic strategy for SMN deficiency is to cause the *SMN2* pre-mRNA to splice in an *SMN1*-like manner, producing a protein that can rescue *SMN1* deficiency (Finkel et al., 2016, Groen et al., 2018, Hua et al., 2010, Meyer et al., 2009, Palacino et al., 2015, Ramdas and Servais, 2020).

To first see whether Census-seq could be used to recognize an individual with a strong SMN-protein-expression phenotype, we created a 19-donor PSC village that included iPSCs derived from an SMA patient (iPSC322A). The cell village was fixed, stained with a monoclonal antibody recognizing the SMN protein (produced by both the *SMN1* and *SMN2* genes), and sorted based on anti-SMN immunoreactivity (**Figure 3.7B, C**). We then separately collected cells whose immunoreactivity was in the upper and lower quintiles relative to the rest of the village (**Figure 3.7C, 3.8**). Census-seq comparison of these "SMN-immunoreactivity-high" and "SMN-immunoreactivity-low" villages revealed that, as expected, cells from the SMA patient were greatly over-represented (3.1-fold) in the SMN-low fraction relative to the SMN-high fraction (**Figure 3.7C**). Interestingly, cells from two additional cell lines were also over-represented in the SMN-low fraction (1.9-fold and 1.4-fold, **Figure 3.7C**). Examination of WGS data from the 19 cell lines in the village revealed that these two cell lines came from heterozygous carriers of an *SMN1* deletion.

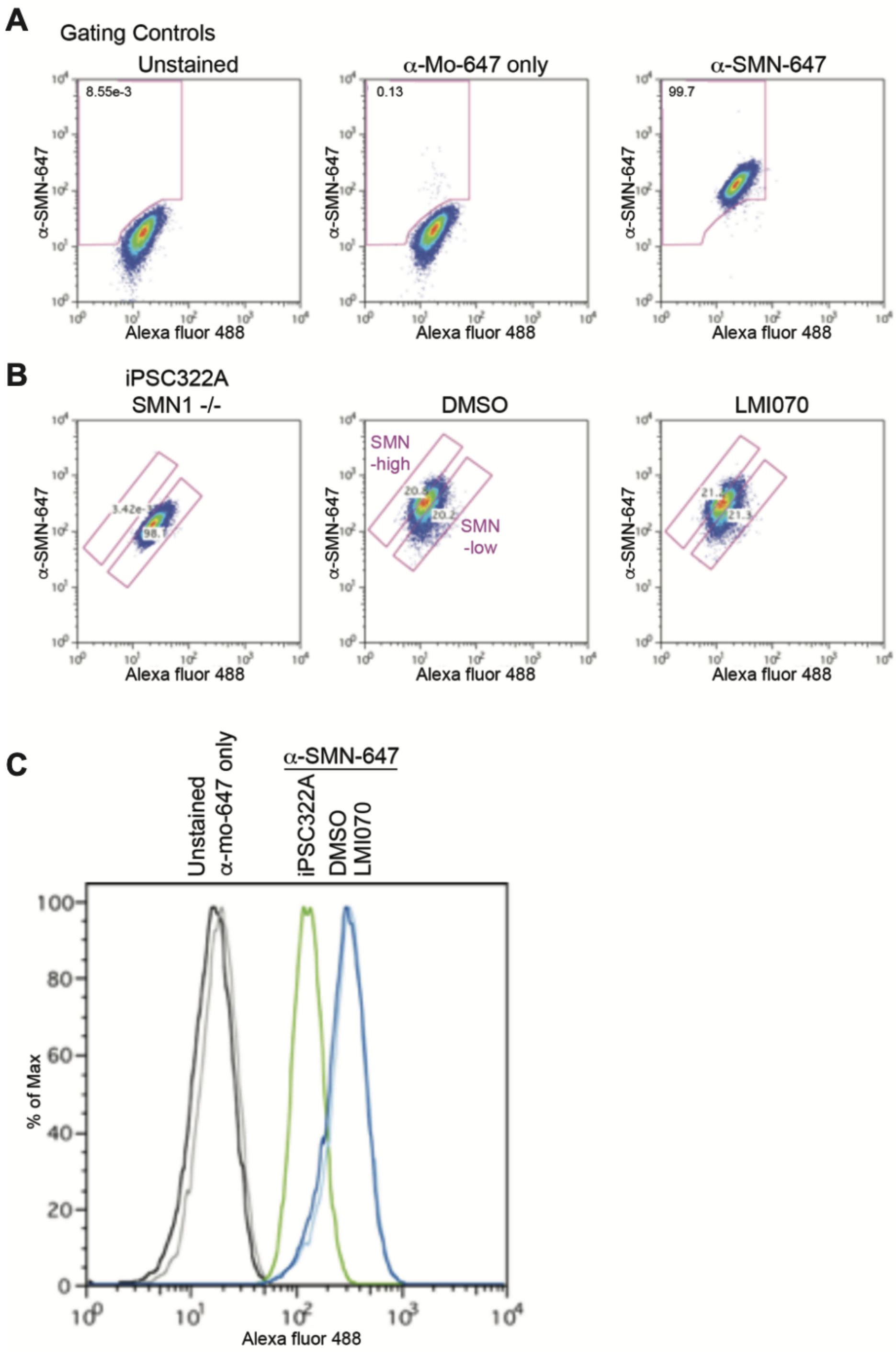


Figure 3.8. Use of flow cytometry to enrich cell villages for SMN-high and SMN-low cells.

A) Cells from a village were dissociated, fixed and immunostained with a monoclonal antibody directed against the protein encoded by SMN1 and SMN2. Gating controls were set using unstained cells (left panel) and cells stained with an anti-mouse Alexa fluor- 647 conjugated secondary antibody alone (middle panel). 99.7% of cells stained with the anti-SMN antibody were captured using these gates (right panel). The Alexa fluor-488 channel served as an internal control for autofluorescence. **B**) Gates were established to capture the top 20% of SMN-stained cells (SMN-high) and the bottom 20% (SMN-low) of cells based on SMN immunoreactivity. Cells from the iPSC322A patient donor line were also analyzed to characterize antibody staining and inform gating (left panel); >98% of cells from the patient cell line. Then, cell fractions were collected from villages treated with a DMSO vehicle control (middle panel) or LMI070 (right panel). **C**) Histogram representation of the anti-SMN staining data from panels S4A and S4B, to facilitate comparison of these distributions across experiments and conditions.

In addition to the SMA patient and two *SMN1* deletion carriers, the other 16 PSC donors also varied in their DNA contribution to the SMN-low and SMN-high cell villages (**Figure 3.7C**). To see whether such variability was driven by genetic variation, we analyzed this phenomenon in a larger village of 113 iPSC lines in the Stanley Center Stem Cell collection (**Figure 3.9A**), whose genomes we also analyzed by WGS and make available as part of this work (Lin et al., 2020). Census-seq revealed abundant variation in the SMN-expression phenotype (**Figure 3.9A**) even among donors whose WGS data indicated that they did not harbor heterozygous or homozygous deletions in *SMN1.* To determine the genetic source of this phenotypic variation, we used the Census-seq measurements to perform a genome-wide association study (GWAS) of the SMN-expression phenotype – which we scored as the log-ratio of each donor's representation in the SMN-high and SMN-low fractions – focusing on 96 iPSC lines that were sufficiently well-represented in the villages to yield high-precision measurements of their representations (**Figure 3.9B, 3.10D, blue symbols; Methods**). This analysis yielded a genome-wide significant (p = 2.91×10^{-14}) association to the locus on chromosome 5 containing the *SMN* genes. The phenotype mapped most strongly to common copy-number variation of the *SMN1* and *SMN2* genes. We found most individuals had inherited 2 to 6 such genes (total), which we measured in each donor by applying the Genome STRiP algorithm to the individual donors' WGS data (**Figure 3.9C, 3.11B**) (Handsaker et al., 2011, Handsaker et al., 2015). The strong statistical significance of this relationship – which resulted from both the number of donors in the analysis (96) and the high correlation of gene copy number with the Census-seq SMN phenotype measurements (r^2 = 0.59) – meant that the genetic basis of this phenotype could in principle have been mapped in an unbiased genome-wide search.

Each copy of *SMN2* can only partially rescue loss of an *SMN1* allele – despite being expressed in the same tissues and cell types – a failure that could in principle be due to differences in the stability or activity of the proteins generated by *SMN1* and *SMN2* (**Figure 3.9D**). To separately quantify the contributions of *SMN1* and *SMN2* to SMN protein abundance, we used the fact that *SMN1* and *SMN2* each exhibit common variation in copy number; we inferred each donor's gene copy number for *SMN1* and *SMN2* by utilizing paralogous sequence variants that distinguish between the genes (**Methods**). Linear regression of the donors' Census-seq SMN-expression phenotypes against their *SMN1* and *SMN2* gene copy numbers revealed that both *SMN1* and *SMN2* copy number contributed positively to SMN protein expression, with *SMN1* making a greater contribution (**Figure 3.9E**). This result confirms that *SMN1* generates somewhat more or longer-enduring SMN protein than does *SMN2*. However, the modest difference between the per-copy effects of *SMN1* and *SMN2* (**Figure 3.9E**) suggests that protein instability is not on its own a sufficient explanation for the inability of *SMN2* to rescue *SMN1* deficiency and is consistent with the hypothesis that *SMNdelta7* also has reduced activity.

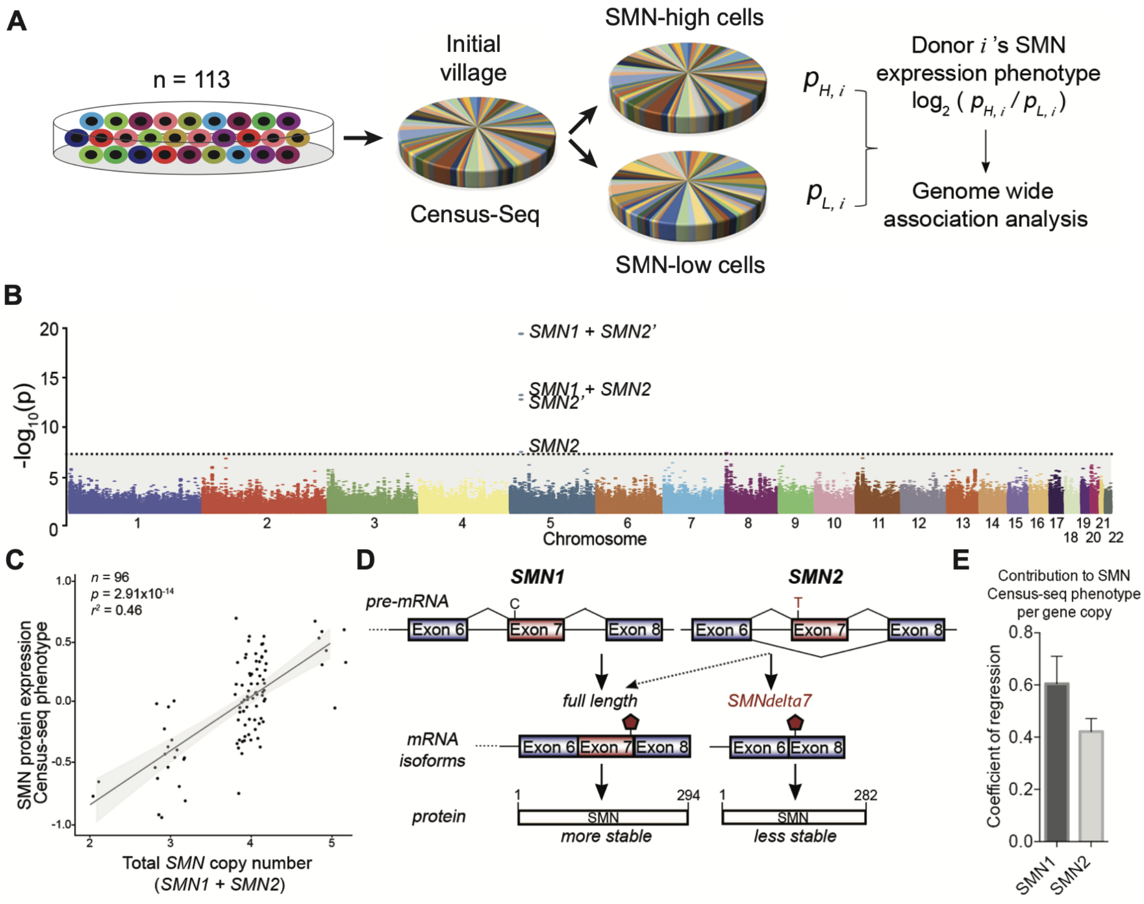


Figure 3.9. Genetic basis of an SMN protein expression phenotype.

A) A village of iPSCs from 113 donors was assembled. The cells in the village were fixed, immunostained for SMN protein, and sorted into SMN-high and SMN-low fractions. Genomic DNA from the two fractions was analyzed by Census-seq. For genetic analysis, each donor's SMN-expression phenotype was quantified from the relative abundance of his/her genomic DNA (cells) in the SMN-high and SMN-low fractions. **B**) Manhattan plot showing genome-wide association analysis of this SMN protein-expression phenotype. This analysis revealed genome-wide-significant association at the locus containing the SMN1 and SMN2 genes, which exhibit common variation in gene copy number. Genome-wide-significant associations involved SMN2 gene copy number and (more strongly) the combined copy number of SMN1 and SMN2. SMN2' refers to a calculation of SMN2 gene copy number that excludes a potential null allele characterized in Figure. 7. **C**) Correlation of the Census-seq SMN protein-expression phenotype with the summed gene copy number of SMN1 and SMN2 genes. **D**) Though SMN1 and SMN2 encode identical amino acid sequences, a sequence variant in a splice enhancer causes many SMN2 mRNAs to splice in a way that excludes exon 7, resulting in a protein that is less stable and potentially less functional. **E**) SMN1 and SMN2 contribute unequally to the Census-seq SMN protein-expression phenotype. Bars indicate coefficients of a linear regression of the SMN protein-expression phenotype against SMN1 and SMN2 gene copy numbers. Error bars indicate standard error.

Now that we had a systematic framework for mapping genetic influences on cellular phenotypes using cell villages, we wanted to apply this approach to understand pharmacogenomic interactions. Specifically, we sought to understand if therapeutic treatments for SMA would have *SMN1/SMN2* genotype dependent responses.

3.6 - Pharmacogenomics in cell villages

An important therapeutic approach is to coax *SMN2* to splice in an *SMN1*-like manner, generating a more stable and/or more effective protein. The clinical candidate LMI070 (Branaplam) was identified in a screen for modulators of *SMN2* splicing (Cheung et al., 2018, Singh and Singh, 2018). LMI070 appears to interact with the splicing enhancer in exon 7 and to increase levels of SMN protein in cells in a concentration-dependent manner (**Figure 3.10A**). The efficacy of LMI070 for treating SMA is currently being evaluated in phase I/II clinical trials in multiple countries (NCT02268552).

As humans often vary in drug responses, a key need in biomedical research is to be able to anticipate individuals' response to drugs and predict who might have an optimal or non-optimal response. We sought to understand whether cell villages could be used to identify variation in drug response and uncover genetic contributions to drug response. We first found concentrations of LMI070 that could cause changes in SMN protein expression (full-length and *SMNdelta7*) across three cell lines with varying gene copy number of *SMN1* and *SMN2* (**Figure 3.11**). Villages of iPSCs were then exposed to either LMI070 (0.1uM) or a vehicle control (DMSO) for 24 hours (**Figure 3.10B**) before being sorted into SMN-high and SMN-low fractions. We measured each donor's relative-drug-response phenotype by calculating how LMI070 treatment changed the distribution of that donor's DNA into the SMN-high and SMN-low cell fractions, relative to the vehicle control (DMSO) (**Figure 3.10C, Methods**); this drug-response phenotype variable was calculated from the results of four Census-seq analyses (**Figure 3.10B, C**).

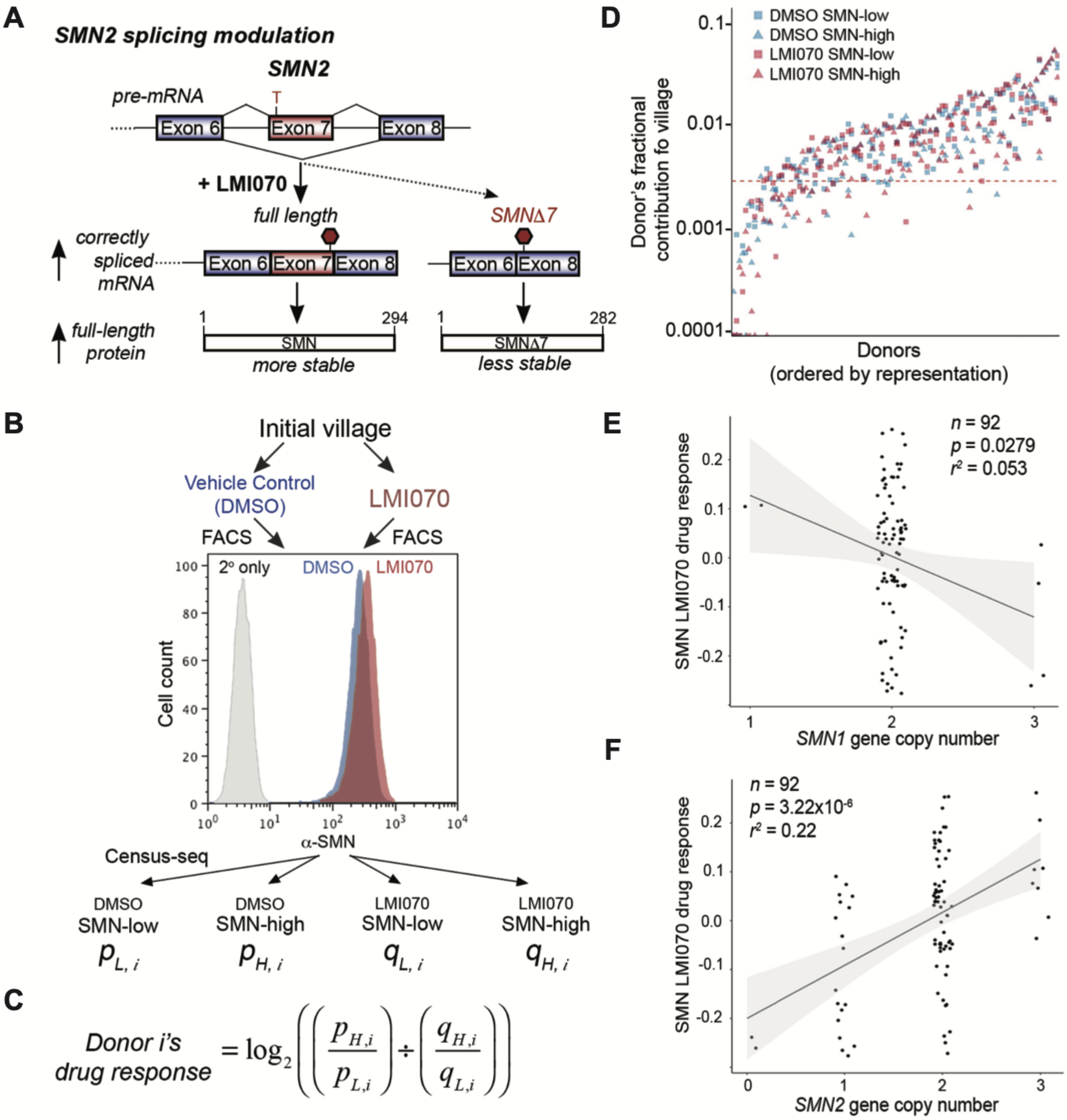


Figure 3.10. Pharmacogenetic analysis of response to SMN therapy.

A) An emerging therapeutic approach for SMA is to cause SMN2 to splice in an SMN1-like manner. The drug LMI070 was developed to do this. **B**) An iPSC village of 113 donors was split into two villages, which were then treated with either LMI070 or a vehicle control (DMSO). Both the LMI070-treated and the vehicle-treated villages were then fixed, immunostained and sorted into SMN-high and SMN-low fractions. **C**) For genetic analysis, a donor's LMI070-response phenotype was calculated from her/his relative cellular contributions to these four fractions. **D**) Estimated contribution of each of the 113 donors to each of the four derived villages. **E**) Correlation of the Census-seq LMI070-response phenotype with SMN1 gene copy number. **F**) Correlation of the Census-seq LMI070-response phenotype with SMN2 gene copy number.

Cells' LMI070-response phenotype correlated strongly with gene copy number of *SMN2* (p = 3.22x10^{-6}) but not *SMN1* (p > 0.01), consistent with the hypothesis that LMI070 affects SMN protein levels by acting specifically upon *SMN2* (**Figure 3.10E, F**). These results replicated in a distinct village of hESCs (**Figure 3.12**). These results contrasted strongly with the baseline SMN-expression phenotype, which was affected more strongly by *SMN1* than *SMN2* variation (**Figure 3.9E**).

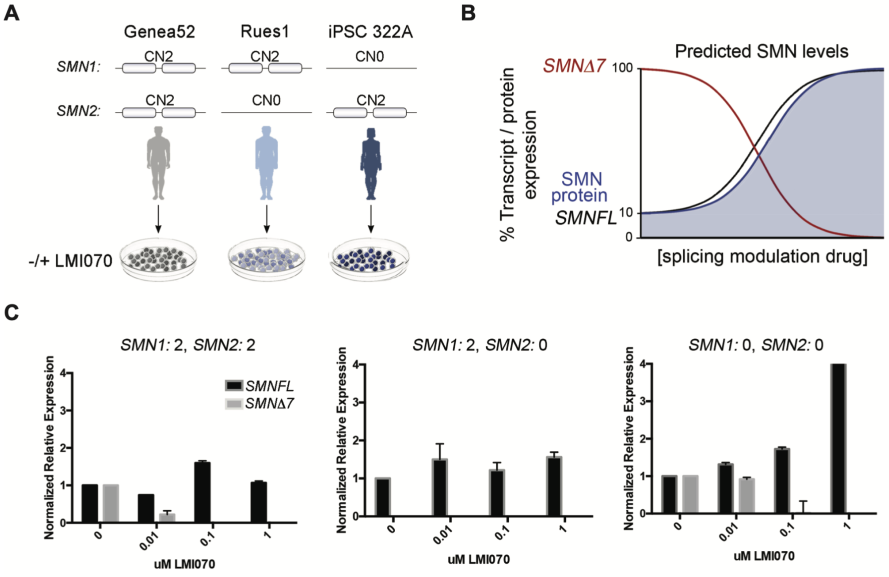


Figure 3.11. Characterization of LMI070 effects on SMN RNA transcript expression in PSCs in vitro.

A) Cell lines from individual donors of varying SMN1 and SMN2 gene copy numbers were chosen to find effective doses of LMI070 treatment in culture, for subsequent cell-village experiments. **B**) In this model, as levels of SMNdelta7 transcripts are decreased by the activity of LMI070, we expect a corresponding increase in levels of SMN full-length transcript and SMN protein. **C**) Measurements (by RT-PCR) of the levels of SMN full length (black) and SMNdelta7 (grey) RNA transcripts in donor cell lines of varying SMN copy number. Note that SMNdelta7 transcripts are produced only in the two donors with SMN2 genes (left and right panels), and that the expression of such transcripts is reduced by treatment with LMI070.

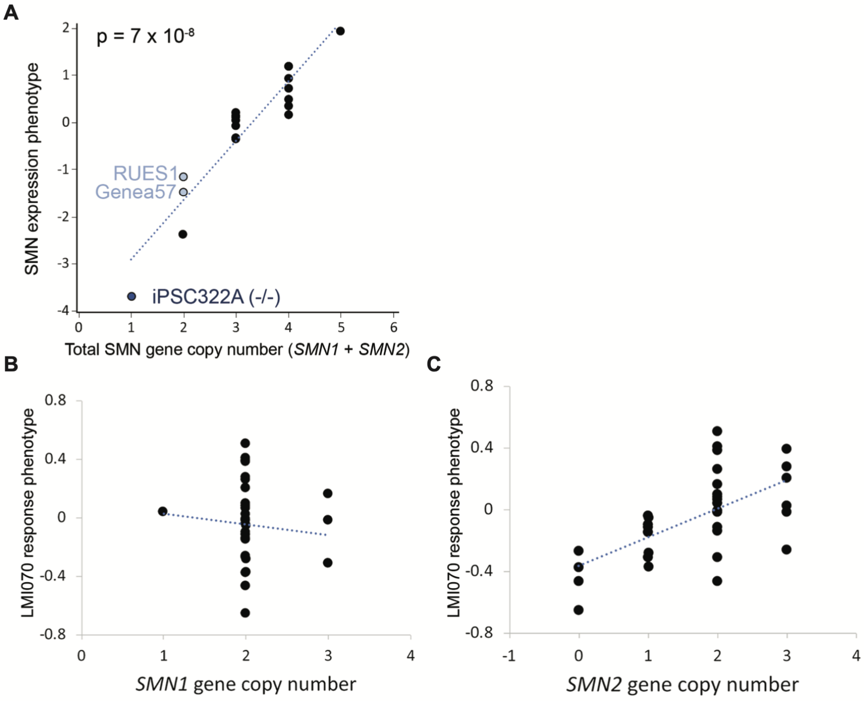


Figure 3.12. Using Census-seq to analyze genetic contributions to SMN protein phenotypes in a pilot village of hESCs.

Companion replication data to Figures 5C and 6EF, drawing upon an additional cell village of hESC lines. **A**) Correlation of the Census-seq SMN protein-expression phenotype with common variation in the total (summed) copy number of SMN1 and SMN2 genes. **B**) Correlation of the Census-seq LMI070-response phenotype with SMN1 gene copy number. **C**) Correlation of the Census-seq LMI070-response phenotype with SMN2 gene copy number.

Because *SMN1* deficiency strongly affects neurons, we also characterized the LMI070 response phenotype in villages of neural cells. A village of neural cells (from 50 donors) was generated using a lentiviral-based delivery system to induce the expression of *Ngn2* (**Figure 3.14**) and treated with LMI070 for 24hr before Census-seq analysis. Both the SMN-expression phenotype and LMI070 drug-response phenotype associated with *SMN2* copy number (**Figure 3.14C**, D; $p = 2.36 \times 10^{-4}$, $p = 1.15 \times 10^{-3}$), replicating the results from the iPSC village.

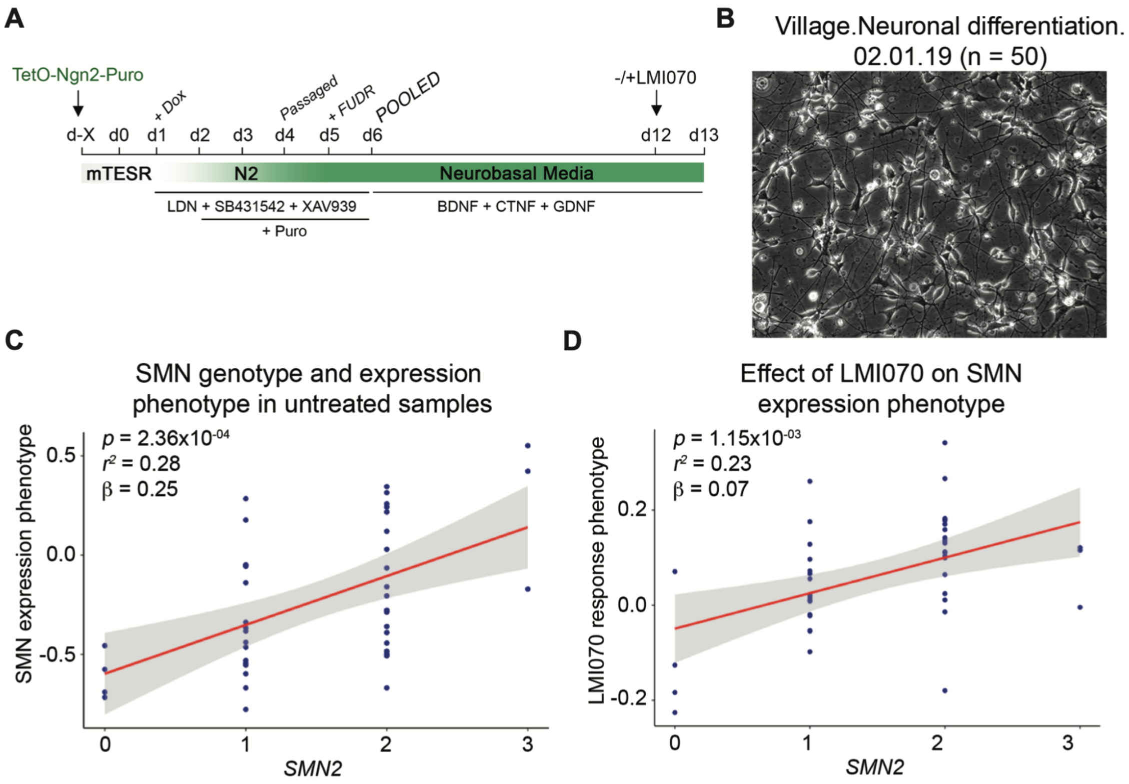


Figure 3.13. Pharmacogenetic analysis of response to LMI070 treatment in a village of differentiated neurons.

A) Differentiation protocol used to generate a village of neurons derived in vitro from 50 iPSC donors. Cell lines transduced with lentivirus to drive Ngn2 expression were induced with doxycycline, selected using puromycin, and pooled at day 6 of differentiation. The village was treated with LMI070 on day 12 and harvested for flow cytometry 24 hours later. **B**) Micrograph of the neuronal village at day 13 of differentiation. **C**) Correlation of the Census-seq SMN protein-expression phenotype with common variation in SMN2 gene copy number. **D**) Correlation of the Census-seq LMI070 drug-response phenotype with common variation in SMN2 gene copy number.

Despite the large apparent effect of *SMN2* gene copy number on response to LMI070, we found that cells from donors with the same number of *SMN2* gene copies often exhibited quite different responses to LMI070 (**Figure 3.10F**), potentially reflecting additional genetic effects. To better ascertain the full spectrum of DNA variation at the *SMN1/SMN2* locus, we analyzed WGS data from 767 individuals (**Methods**). We found that many donors carried an apparent deletion of *SMN2* exons 7 and 8, including the LMI070 binding site (**Figure 3.14A, B, C**; we refer to this allele as "*SMNdel*" below). Although the deletion was present in the genomes of 10% of the sampled individuals with European ancestry, it has only recently been described (Vijzelaar et al., 2019).

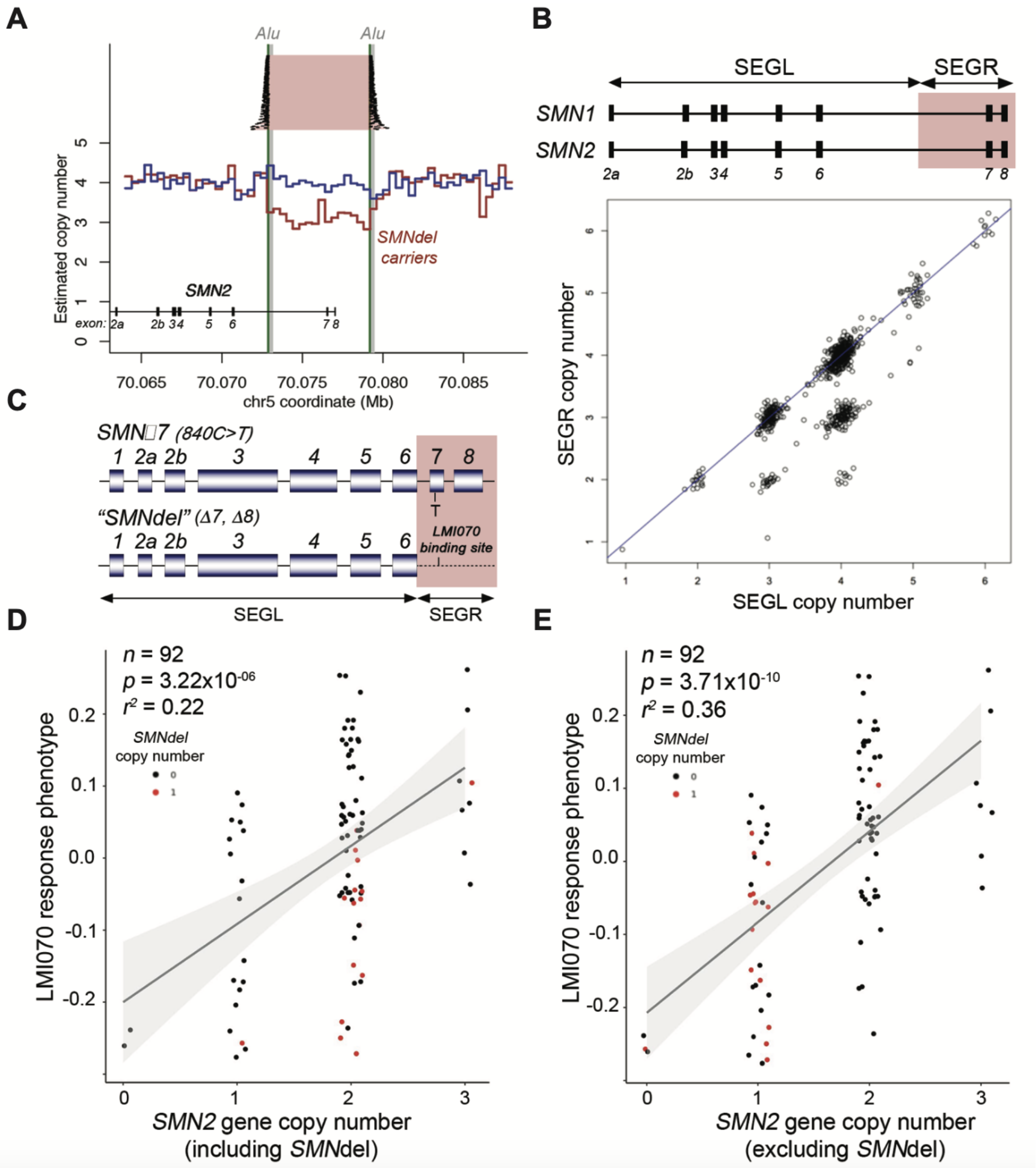


Figure 3.14. Pharmacogenetic analysis of response to SMN therapy.

A) In analysis of whole genome sequence generated from genomic DNA from the individual iPSC donors, 22 of the 113 donors were found to have sequence reads or mate-paired reads that appeared to jump over a 6-kb genomic segment (flanked by two Alu repeats) containing two exons of the SMN2 gene. Those same individuals tended to have reduced copy number of the 6-kb segment (red) relative to the other individuals (blue), as estimated from read depth of coverage across the genomic locus. **B**) Across 767 genomes analyzed by whole genome sequencing, copy number of this 6-kb segment (SEGR) was in many individuals less than copy number of the rest of the SMN1/SMN2 gene (SEGL). (Note that the >99% sequence identity between SMN1 and SMN2 requires that the paralogous sequences from these two genes be counted together for this analysis.) This population-level pattern confirms that SEGR is affected by a cryptic, common deletion allele. **C**) This cryptic, common deletion allele ("SMNdel") removes two exons, including the exon that encodes the putative binding site for LMI070. **D**) Cells from individuals with the SMNdel allele in their genomes tend to have a smaller response to LMI070 (as measured by the Census-seq LMI070-response phenotype) relative to cells from other individuals with the same SMN2 gene copy number. Red points: SMNdel carriers. Black points: Other iPSC donors. **E**) The Census-seq LMI070-response measurements more strongly fit a model in which SMNdel is treated as a null allele of SMN2.

Re-analysis of the Census-seq data to account for the *SMNdel* allele showed that carriers of *SMNdel* were indeed the donors whose cells exhibited a weak response of SMN expression to LMI070 treatment (**Figure 3.14D**, red). The Census-seq LMI070-response phenotype correlated much more strongly with a measure of "intact" *SMN2* gene copy number that we obtained by treating SMNdel as a null allele (r^2 = 0.36, p = 3.71^{-10}), suggesting that SMNdel carriers under-respond to LMI070 (**Figure 3.14E**). In fact, we found evidence that *SMNdel* is either not translated, or encodes a much less stable protein than *SMN2* does: excluding *SMNdel* copies from genetic measurements of *SMN2* gene copy number greatly strengthened the association with the SMN baseline protein-expression phenotype (**Figure 3.9B**). Simultaneous linear regression of the Census-seq SMN-expression and LMI070-response phenotypes against gene copy numbers of *SMN1*, *SMN2* (excluding *SMNdel*), and *SMNdel* (**Figure 3.15**) indicated the *SMNdel* allele, in addition to encoding an LMI070-unresponsive RNA, also encodes a much less stable isoform of the SMN protein than canonical *SMN2* gene copies do.

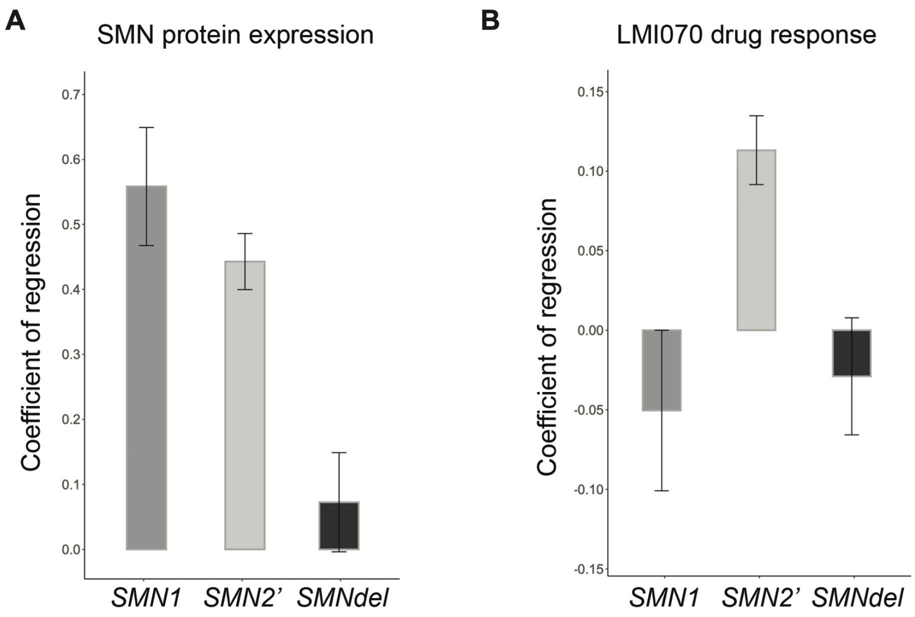


Figure 3.15. Contributions of SMN1, intact SMN2 genes (excluding SMNdel) and SMNdel alleles to Census-seq phenotypes for SMN protein expression and response to splicing correction by LMI070.

A) Contributions of SMN1 genes, intact SMN2 genes (SMN2', which excludes SMNdel alleles), and SMNdel alleles to the SMN protein-expression Census-seq phenotype. Bars indicate coefficients of a linear regression of the SMN protein- expression phenotype against SMN1, SMN2', and SMNdel gene copy numbers. Error bars indicate standard error. **B**) Contributions of SMN1 genes, intact SMN2 genes (SMN2', which excludes SMNdel alleles), and SMNdel alleles to the LMI070-response Census-seq phenotype. Bars indicate coefficients of a linear regression of the SMN protein-expression phenotype against SMN1, SMN2', and SMNdel gene copy numbers. Error bars indicate standard error.

3.7 - Discussion

Genetic variation shapes almost all human phenotypes, creating a profound opportunity for biological discovery and translational biology if science can begin to reveal how human alleles – individually, and in concert – shape the life of cells. To help realize this scientific possibility, we developed "village-in-a-dish" experimental systems, in which cells from scores of donors are grown, perturbed and phenotyped in a single reaction chamber; these systems enable population-genetic approaches to cell-biological questions. The analysis of these systems is enabled by three computational methods – Census-seq (**Figure 3.1**), Roll Call (**Figure 3.5**), and CSI (**Figure 3.5**) – which reveal and learn from the donor composition of cell and DNA mixtures.

The practical execution of such systems is further enabled by ways to create and maintain high-quality cell villages.

Human genomes teem with functional variation. Here, even at a single locus, Census-seq analyses revealed effects of at least three kinds of genetic variation. These included (i) *SMN1* gene copy number, which affected SMN protein levels at baseline but not responsiveness to LMI070 therapy; (ii) *SMN2* gene copy number, which affected SMN protein levels at baseline and also response to LMI070; and (iii) a cryptic *SMN2* allele (*SMNdel*), common but not routinely screened for in clinical diagnostics, which compromised SMN protein levels and abrogated LMI070 response (**Figure 3.15**).

Though cell villages and populations enable primary genetic discoveries in Census-seq, it is useful in many contexts to analyze individual control cell lines alongside villages, and to seed villages with one or more lines already known to harbor strong effects. The choice of sorting or selection criteria in Census-seq experiments can be enhanced by comparison to one or more individual cell lines expected to have a strong phenotype. For example, flow-cytometric analysis of cells from an SMA patient informed the flow-cytometric gating strategy that we then applied to the village for population-scale genetic analyses of SMN protein expression (**Figures 3.7, 3.8, 3.9**). Positive- and negative-control lines can also be derived by methods including gene editing CRISPR-i, CRISPR-a, and pharmacological perturbation of individual lines (Gasperini et al., 2019, Larson et al., 2013, Hsu et al., 2014). The further inclusion of such lines in villages enables the effects of population-genetic variation to be compared to the effects of strong laboratory perturbations. The ability to detect expected, positive-control effects also serve as a useful gate establishing that an experiment has been executed successfully. The inclusion of such positive controls may also allow the lack of variation across scores of other cell lines to be a meaningful statement about biological constraint on a cellular phenotype.

Our finding that cells from carriers of the *SMNdel* allele responded less strongly to LMI070 raises an interesting issue that will need to be addressed for many drug candidates that target RNA metabolism. Polymorphisms that influence RNA structure and regulation may be much more abundant than polymorphisms that influence the peptide sequence of the encoded protein, the traditional drug target. Census-seq could be used to identify, prior to clinical studies, individuals and genotypes who would have optimal and/or unexpected responses to such therapeutic candidates.

Additional challenges may arise in other experimental contexts. Mitotic cells present challenges in population dynamics (**Figure 3.3**) that were managed by the techniques we describe here – including genetic screening (for acquired mutations) and minimizing mechanical perturbations such as passaging. Similarly, variability in pluripotency after reprogramming or exposure to various environmental conditions across cells' lifetime may influence their differentiation potential (Cahan and Daley, 2013, Nishizawa et al., 2016). The refraction of cells down a particular pathway may even be shaped by genetic variation, a potential area for Census-seq analyses. Finally, a central challenge to the maintenance and expansion of cellular resource banks that are required to facilitate population-scale experiments has been detecting cell line contamination and validating of donor identity (Liang and Zhang, 2013, Rouhani et al., 2014). These issues have historically been of major concern and have real implications on experimental reproducibility (Neimark, 2015, Nelson-Rees et al., 1981). The methods described here (including Census-seq, Roll Call and CSI) create a greatly expanded and inexpensive tool kit for validating the identity and purity of cellular resources.

Village-in-a-dish systems make a tradeoff between detection of cell-autonomous and cell-nonautonomous genetic effects. By analyzing cells from all donors together, Census-seq normalizes most cell-non-autonomous effects across donor genotypes. Most of these cell-nonautonomous effects are common sources of experimental noise – for example, effects of cell density and its downstream effects on metabolite concentrations, cellular waste-product concentrations, cell-cell contacts, and (for neurons) synaptic stimulation and activity. The normalization of such effects in Census-seq makes cell-autonomous effects more visible above experimental noise; this was evidenced by the strong explanatory power of common alleles in almost all of our experiments (**Figures 3.9-3.11**) and the ability of these experiments to establish genotype-phenotype relationships at genome-wide statistical significance. In many contexts, though, cell-nonautonomous effects may be of real interest, so it is important to think about how to design Census-seq experiments to ascertain them. Census-seq could be readily used to map cells' responses to a non-cell-autonomous stimulus such as a ligand. However, different kinds of analysis would be needed to map genetic effects on the magnitude of the stimulus itself; this might be possible in Census-seq if the ligand's synthesis can be made the subject of a Census-seq analysis.

A key question for population-scale cell biology and systems-biological modeling involves the ability to discover and quantify modest, quantitative effects. The experiments described here ascertained strong, rare, Mendelian effects (as in cells from an SMA patient) and more modest, common effects, such as effects of common variation in gene copy number on protein expression and therapeutic response. Key to the ability to discover and characterize modest, quantitative effects is the ability to equalize technical factors across donors, reducing experimental noise and thus increasing the detectability of genetic signals relative to noise. The specific parameters chosen for selections and screens are also important and are worth careful thought and optimization. For example, the flow-cytometric gates used to select for derived villages (**Figure 3.8**) can be critical for measuring genetic effects with high sensitivity. Since Census-seq molecular and computational analyses are inexpensive (sequencing expenses are about $100 per village), we encourage researchers to experiment with a variety of parameters and observe the impact of these variables on measurement of known, positive-control effects.

Complex phenotypes with multi-locus inheritance will present interesting challenges to the design of Census-seq experiments, often requiring analysis in still-larger numbers of donors. We anticipate that the optimal approach to this will involve analyzing multiple villages each of 60-120 donors, with modest overlapping membership to inform meta-analysis. In our experience, most of the returns of Census-seq in scalability and well-controlled measurement are already realized at the scale of 40-120 donors – a level at which village assembly can also be performed by a single scientist in a single session. Villages of 40-120 donors might provide a natural unit of larger meta-analyses that seek to find genetic effects that are rare or modest in magnitude.

In selecting more complex phenotypes for study, a promising next direction may be in analyzing oligogenic phenotypes that might be affected by variation at a few loci, perhaps focusing on those pathways or molecular complexes in which genome-wide association studies (by SNP arrays or exome sequencing) have already identified risk variants or haplotypes in several different genes. Such constellations of (common and rare) genetic effects may suggest a natural integration point for cellular-genetic analysis. In schizophrenia, for example, many different genetic associations involve subunits of the L-type calcium channels, whose expression, membrane localization or even physiology could be made the subject of cellular screens and selection (Lam et al., 2019, Schizophrenia Working Group of the Psychiatric Genomics, 2014).

Polygenic phenotypes will ultimately require the largest samples and present the greatest challenge – and perhaps also the greatest reward, since such polygenetic architectures have been extremely challenging to dissect by traditional biological methods and are hypothesized to involve indirect effects through complex cellular networks (Boyle et al., 2017). An exciting possibility is that, even for polygenic illnesses, it may be possible to find cellular phenotypes that are potential convergence points of genetic effects whose functional connection was not previously appreciated. Such cellular phenotypes might associate in villages with donors' polygenic risk scores, which for many complex phenotypes identify individuals with risk equivalent to that of well-known monogenic mutations (Khera et al., 2018). Villages can in principle be designed from individuals in the two "tails" of a polygenic risk-score distribution – i.e., who have extremely high or low polygenic risk scores. It is challenging to predict how many donors will be required for such analyses, as any prediction presumes an answer to a central unknown question – the extent to which different genetic effects on a disease phenotype will converge upon a few key cellular processes. We hope population-scale experimental systems present an empirical, data-driven path toward answering these and many other questions.

3.8 - Reflection

Studying the effects of common genetic variation on cellular phenotypes requires analyzing cells from very many donors, in well-controlled experiments that can detect and measure quantitative influences on cellular phenotypes. Here we described the development of "cell village" experimental systems which enable simultaneous phenotyping and analysis of 10^4-10^6 cells from tens to hundreds of donors at once, in well-controlled experiments in which cells from all donors are grown together in a single environment (Wells et al, 2023). Such systems have enabled us to map thousands of expression QTLs and to map common genetic effects on select cellular phenotypes at genome-wide levels of statistical significance (Wells et al, 2023). Importantly, technical challenges in experimental design and workflows have initially limited the scalability and execution of these approaches and their application to additional phenotypes, especially when proliferative cells remained present in the cultures. For example, mixtures of hiPSCs from different donors cultured together in a village over several days often resulted in an imbalance of donors in the village, with just a few cell lines taking over the whole village (Mitchell et al 2020). These technical challenges could be remedied by pooling together post-mitotic cells from all donors, but this restricted village-based experiments to largely post-mitotic cell states (excluding key cell types such as astrocytes) and necessitated that all cell lines be cultured and differentiated individually. Therefore, we needed a concerted effort to understand cell line variability in village-in-a-dish systems and think about alternative approaches to ameliorating this variability.

To overcome these barriers to cell-village approaches and increase the scalability and accessibility of these approaches across different labs and their utility for genetic discovery (in which power is largely driven by the number of donors), I have greatly improved cell-culture workflows, in the process reducing the per-experiment workload necessary for generating villages. I found that much of the change in donor representation in culture arose from cell-passaging steps, which we found can unintentionally select for specific donors (potentially based on their expression levels of cell-adhesion molecules). I reduced physical manipulation of cell villages in ways that we found minimize donor population dynamics throughout culture. I validated our ability to differentiate cell villages of iPSCs with these workflows into diverse neural cell lineages, including astrocytes and neurons.

Another key challenge we overcame involved the difficulties associated with culturing 10s-100s of cell lines in parallel right before village generation. I explored whether pre-constructed villages could be frozen, saved and combined in many future experiments: in this scenario, a scientist performing a future experiment would need only to retrieve a cryovial(s) from the freezer to be able to initiate a population-scale experiment. I tested this by generating several villages of 45-48 donors and cryopreserving them. I measured donor representation at the time of banking and 24hrs after recovery, showing a high degree of correlation across donors. This indicated that large villages can be constructed, QC'd and cryopreserved, streamlining and facilitating downstream experiments. To overcome the large changes in donor representations we had initially observed, I optimized our workflow toward growing cells with minimal manipulations (including passaging of cells), enabling more-uniform donor representation throughout long-term culture and endpoint assay readouts.

With these revised and improved protocols in hand, I explored whether we could use this approach to differentiate cells from many donors together in a village but starting with a village of iPSCs (instead of a village of early post-mitotic cells or slowly proliferative cells, which had been our previous workaround). The ability to create villages at the iPSC stage would greatly reduce line-to-line variation during the most critical windows in cellular differentiation, would further decrease costs, and would enable the incorporation of larger numbers of donors and cell lines. To do so, I differentiated a village of iPSCs from 43 donors into NPCs and neurons; we found that we could now maintain individual donor representation throughout differentiation.

This new workflow enabled me, starting with a single cryovial, to perform a large-scale pharmacogenomics experiment on cells from 43 donors, differentiated into 4 cell types and exposed to 15 different pharmacological perturbations, all within one month from start to finish. This work is highlighted in Chapter 5 to better understand how common medications impact cell type interactions in the central nervous system and to identify genetic influences which may mediate these responses.

Our group has utilized these cell village-based experiments across a range of contexts focusing on brain development and disorders (Wells et al, 2023). However, much of this work has emphasized the use of neurons and neuronal progenitor cells due to their associated with brain specific disorders and the ease with which these cell types can be generated from stem cells. While neurons and their progenitor cells are of critical importance to investigating brain related illnesses, there is an increasing focus on alternative cell types, such as astrocytes, in the etiology of conditions such as schizophrenia and bipolar disorder. Currently, there are limited methods to generate astrocytes from stem cells and existing protocols are often time consuming and technically challenging. To overcome these limitations in the field, I sought to develop a new differentiation method which would allow us to generate astrocytes reliably and robustly from human stem cells. In the next chapter, I will highlight this work.

3.9 – Statement on contributions

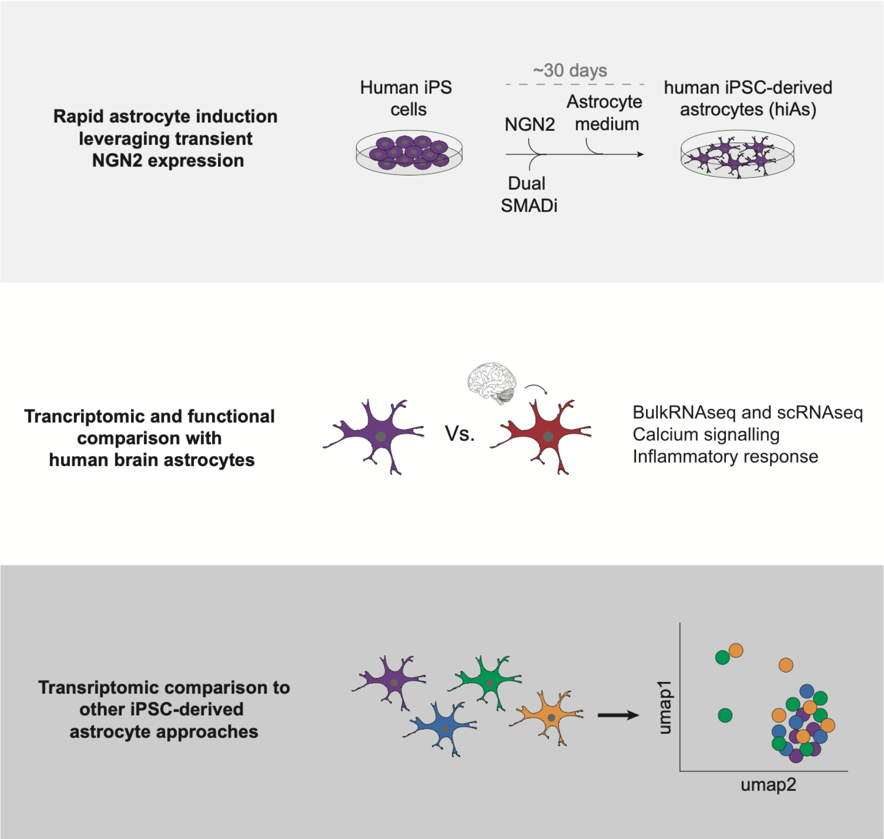
This work was co-led by Jana Mitchell and Jim Nemesh. Jana led the early development of cell culture workflows and performed experiments related to *SMN* and the pharmacogenomics in Sections **3.5** and **3.6**. Jim Nemesh developed all of the computational protocols related to Census-seq, including the underlying methodologies described above in Sections **3.2** and **3.4.** I contributed to the early development of the cell culture workflows and performed experiments related to Section **3.2**.

Chapter 4

Robust induction of functional astrocytes using NGN2 expression in human pluripotent stem cells

Robust induction of functional astrocytes using *NGN2* expression in human pluripotent stem cells

Graphical Abstract


Rapid astrocyte induction leveraging transient NGN2 expression
Human iPS cells
~30 days
NGN2
Astrocyte medium
Dual SMADi
human iPSC-derived astrocytes (hiAs)
Trancriptomic and functional comparison with human brain astrocytes
Vs.
BulkRNAseq and scRNAseq
Calcium signalling
Inflammatory response
Transriptomic comparison to other iPSC-derived astrocyte approaches
umap1
umap2

4.1 - Introduction

Astrocytes are the most abundant cell type in the human brain. They play crucial roles in regulating neuronal development, maturation, and synaptic connectivity (Allen et al, 2017; Sofroniew et al, 2010). Astrocyte dysfunction and defective astrocyte-neuron interactions have been implicated in a wide variety of disorders, including psychiatric, neurodevelopmental, and neurodegenerative disorders (Allen et al, 2017; Sofroniew et al, 2010; Pietilainen et al, 2023). Astrocytes also play an important role in regulating the cerebral microenvironment by interacting with endothelial and microglial cells that participate in the blood brain barrier (Abbott et al, 2006; Cucullo et al, 2002; Goldstein et al, 1988; Liu et al, 2020) and communicating with oligodendrocytes via direct contact and secretion of cytokines and chemokines (Amaral et al, 2013; John et al, 2012; Magnotti et al, 2011).

While brain cell types are largely thought to be conserved across species, an increasing number of studies have uncovered divergent molecular, structural, and functional features of glia. For example, transcriptional comparisons between human and rodent have revealed greater differences in glial gene expression signatures compared with neuronal-associated transcripts, suggesting that glial genes may be evolutionarily less conserved than neuronal genes (Hawrylycz et al, 2012). Moreover, while mammalian astrocytes respond to glutamate and ATP by increasing intracellular calcium concentrations, human astrocytes support different calcium wave dynamics as compared with rodent astrocytes (Han et al, 2013; Oberheim et al, 2006) which has the potential to affect subsequent release of glio-modulators. Pharmacological inhibition of the TGF**β** pathway partially prevents the synaptogenic effect of murine astrocyte-conditioned media on cortical neurons but abolishes the effect of human astrocyte-conditioned media, suggesting that human astrocytes may rely more heavily upon TGF**β** signaling than their rodent counterparts (Diniz et al, 2012). Human astrocytes also display larger cellular diameters with more elaborated and compartmented processes compared with rodent astrocytes (Oberheim et al, 2006). Indeed, while a rodent astrocyte domain can reportedly cover up to 120,000 synapses, a human astrocyte domain can cover up to 2 million synapses, suggesting greater processing complexity in the latter species (Oberheim et al, 2006). These and other species-specific features highlight the likelihood of human astrocytes to differ from rodent astrocytes in their contributions to brain function as well as brain dysfunction. Human astrocytes thus have important applications in studies of basic brain function, disease modeling and drug discovery.

With the emergence of human pluripotent stem cell (hPSC) technologies, it is now feasible to sustainably generate an array of brain cell types in vitro. Notably, numerous studies have shown that glial cells are necessary for the functional maturation of neurons (Nehme et al, 2018; Pfrieger et al, 1997; Turovsky et al, 2020). For practical considerations and ease of access, most studies supplement neuronal cultures with rodent astrocytes, and more recently, commercially available primary fetal astrocytes. However, these approaches have significant limitations. As discussed above, rodent astrocytes diverge morphologically, transcriptionally, and functionally from human astrocytes and do not allow for the investigation of the effect of human genetic variants and perturbations on biology and disease. Primary fetal astrocytes are not a sustainable resource and generally do not allow for the study of specific human genotypes of interest.

Recognizing the utility of human in vitro derived astrocytes, several protocols have been developed including those following a protracted developmental time-course in 3-dimensions (up to 20 mos.) (Sloan et al, 2017) as well as more rapid 2-dimensional protocols (Canals et al, 2018; Hedegaard et al, 2020; Santos et al, 2017; Tcw et al, 2017; Voulgaris et al, 2022). While these protocols produce cells expressing canonical astrocyte markers and are capable of recapitulating key functions such as responding to pro-inflammatory stimuli, much remains to be determined regarding: (i) the precise cell types and cell stages being generated and / or how closely they resemble fetal or adult human astrocytes from primary cultures or post-mortem preparations, (ii) the robustness of protocols across individual hPSC lines, and (iii) their utility for disease modeling applications. Furthermore, most deeply characterized approaches are either lengthy or technically complicated, involving multiple experimental steps, such as purification, replating, and culturing in different formats, rendering such protocols less amenable to implementation across multiple cell lines and manipulations.

Several astrocyte differentiation protocols have recognized the importance of robust neural progenitor cell (NPC) generation as an important foundation for efficient astrocyte differentiation (Tcw et al, 2017; Voulgaris et al, 2022). Here, we leveraged previous studies showing that NGN2 patterning with or without dual-SMAD and WNT inhibition can direct hPSCs toward diverse neural fates including forebrain neurons and peripheral neurons (Nehme et al, 2018; Limone et al, 2023; Wells et al, 2023; Lin et al, 2021) and identifying optimal media conditions to differentiate NPCs into astrocytes22. Specifically, to generate NPCs we used transient NGN2 induction combined with dual-SMAD inhibition (SB431542, LDN-193189) and WNT inhibition (XAV939), previously shown to support forebrain patterning of the NPCs (Nehme et al, 2018; Wells et al, 2023) followed by maturation in astrocyte media (ScienCell) previously screened for efficient differentiation of hPSC-derived forebrain NPCs into astrocytes (Tcw et al, 2017). This combination resulted in robust generation of astrocytes by 30 days in vitro. We then benchmarked our human induced astrocytes (hiAs) against human primary fetal astrocytes (hpAs) using a combination of immunophenotyping, RNA-sequencing and a series of functional assays assessing response to pro-inflammatory stimuli, ATP-induced calcium release and presynaptic development upon neuronal co-culture. While single-cell RNA-sequencing (scRNAseq) of in vitro derived astrocytes has been limited to 3-dimensional protocols (Sloan et al, 2017; Rapino et al, 2022), requiring dual transcription factor overexpression approaches (Leng et al, 2022) or following FACS-purification (Barbar et al, 2020), we used scRNAseq analyses of our rapid 2-dimensional protocol from eight unique parental cell lines to define their molecular signatures. This revealed a striking degree of homogeneity at the single-cell level, and high reproducibility in the differentiated product across multiple parental cell lines. Finally, we generated hiAs in a model of trisomy 21 and recapitulated key features associated with the disease.

Collectively, these analyses establish a rapid, scalable, and reproducible differentiation protocol to generate homogeneous human astrocytes with a well-defined molecular signature, with limited interventions, which can be applied to many parental cell lines and specifically for the purposes of disease modeling.

4. 2 - Robust and rapid generation of human induced astrocytes (hiAs) from NPCs.

To generate hiAs, a doxycycline inducible NGN2 expression construct was introduced into hPSCs through TALEN-mediated stable integration into the AAVS1 safe-harbor locus (**Figure 4.1A**) (Berryer et al, 2023). Following neomycin selection for construct integration, iNGN2-hPSCs were neuralized and dorsalized by switching to N2 medium with doxycycline and small molecule patterning for 48hrs to induce an NPC-like fate (**Figure 4.1B**) (Nehme et al, 2018; Wells et al, 2023). hPSCs neuralized through ectopic expression of NGN2 alongside small molecule patterning have been well-characterized, express canonical markers for NPCs such as NESTIN, PAX6, FOXG1 and SOX1 as well as dorsal rather than posterior or ventral markers and can be captured in this state for subsequent analyses as highlighted previously (Wells et al, 2023; Nehme et al, 2022). After 24hrs of Zeocin selection for NGN2 induction, and for an additional 28+ days, cells were passaged and maintained in a commercially available astrocyte medium (ScienCell) previously shown to induce astrocyte morphology, the expression of astrocyte canonical markers as well as ensure replicative competency (Tcw et al, 2017). As shown by previous work using a similar approach for hPSC-based astrocyte generation, the commercially available medium supported astrocyte maturation better than all other tested in-house recipes (Tcw et al, 2017). Indeed, by day 4 post induction, the differentiating iNGN2-NPCs acquired an astrocyte-like morphology, with flat, wide cell bodies beginning to form star-like projections (**Figure 4.1B**). By day 30, a vast majority of the differentiated cells, referred to as hiAs, expressed canonical astrocyte markers including Aquaporin 4 (AQP4), CD44, Solute Carrier family 1 member 3 (SLC1A3), S100 calcium binding protein B (S100B) and Vimentin (VIM) by immunofluorescence, paralleling results obtained with human primary astrocytes (hpAs) (**Figures 4.1C** and **4.1D**). For example, 99.05% of hiAs and 95.12% of hpAs expressed AQP4, and 93.11% of hiAs and 95.62% of hpAs expressed SLC1A3 (**Figure 4.1D**).

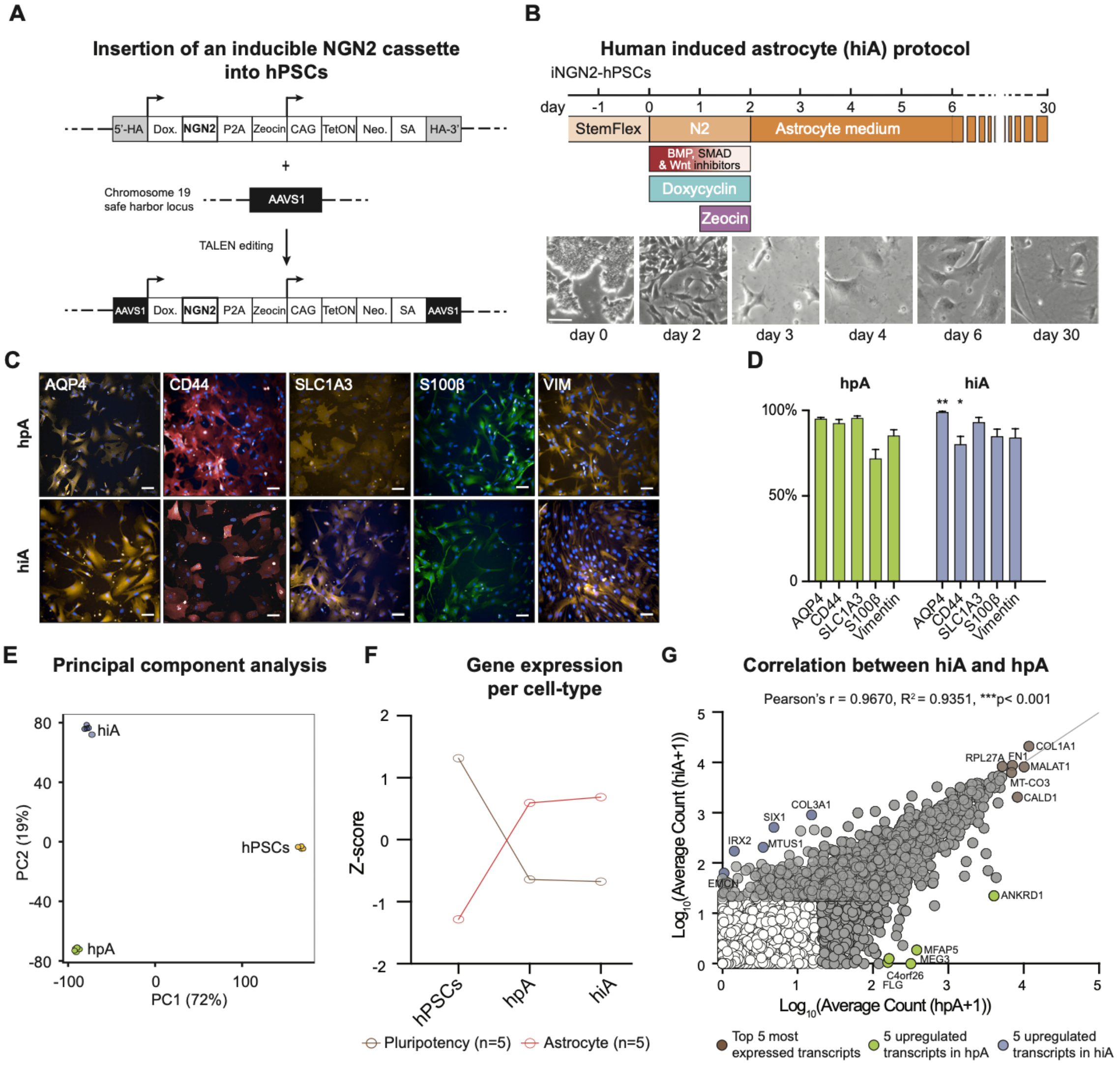


Figure 4.1. Robust and rapid generation of human induced astrocytes (hiAs) from NPCs.

(A), Integration of an inducible NGN2 cassette in the safe harbor locus of hPSCs by TALEN editing. (B), Schematic of the 30-day hiA differentiation protocol with brightfield images over the induction time-course shown below. Scale bar = 100µm. (C), Representative immunofluorescence images for AQP4, CD44, SLC1A3, S100b and VIM from human primary astrocytes (top) and human induced astrocytes (bottom). Scale bar = 25µm. (D), Quantification of each marker from hiAs and hpAs shown as % of DAPI+ cells. Data are represented as mean +/- SEM, *p<0.05, **p<0.01; unpaired non-parametric Kolmogorov Smirnov test hiA compared to hpA; n>3 biological replicates, n=3 technical replicates. (E), Principal component analysis (PCA) of bulk RNA-seq data comparing hPSCs, hpAs and hiAs. Note that a majority of variance is explained by the comparison between hPSCs and astrocytes. (F), Gene expression per cell type for hPSCs, hpAs and hiAs using canonical pluripotent and astrocyte related genes. (G), Scatterplot showing a high positive correlation between hpA and hiA expressed transcripts (TPM >20, in grey). Brown circles highlight the top 5 most expressed transcripts shared between hiAs and hpAs, while green circles highlight 5 transcripts upregulated in hpAs alone and blue circles highlight 5 transcripts upregulated in hiAs alone. Pearson's r = 0.9670, R2 = 0.9351, ***p< 0.001.

To further explore the commitment of hPSCs to an astrocyte identity, we compared the bulk transcriptomic profiles of hiAs with those of hPSCs and hpAs (**Table S4.1**). Principal component analysis of all three cell types revealed that the vast majority of transcriptomic variance could be explained by the comparison between hPSCs and astrocytes (PC1: 72%) rather than between different astrocyte populations (PC2: 19%) (**Figure 4.1E**). In addition, gene expression per cell type revealed a similar trend of decreased pluripotency genes and increased astrocyte-related genes in both hpAs and hiAs (**Figure 4.1F**). We also observed a high positive correlation between hpA and hiA expressed transcripts (Pearson's r = 0.9670; **Figure 4.1G**) suggesting a similar landscape in global transcriptome, in contrast with the correlations between hiAs and hPSCs, or hpAs and hPSCs (**Figure 4.2**). Finally, PANTHER analysis of the top enriched pathways shared by hpAs and hiAs, in contrast to hPSCs, revealed canonical astrocyte signaling pathways such as VEGF, PDGF, Angiogenesis and Integrin signaling (**Figure 4.2**). Interestingly, the top enriched pathways unique to hiAs over hpAs or unique to hpAs over hiAs were largely not astrocyte specific (**Figure 4.2**). Collectively, these analyses indicate that our hiAs harbor key molecular hallmarks of hpAs.

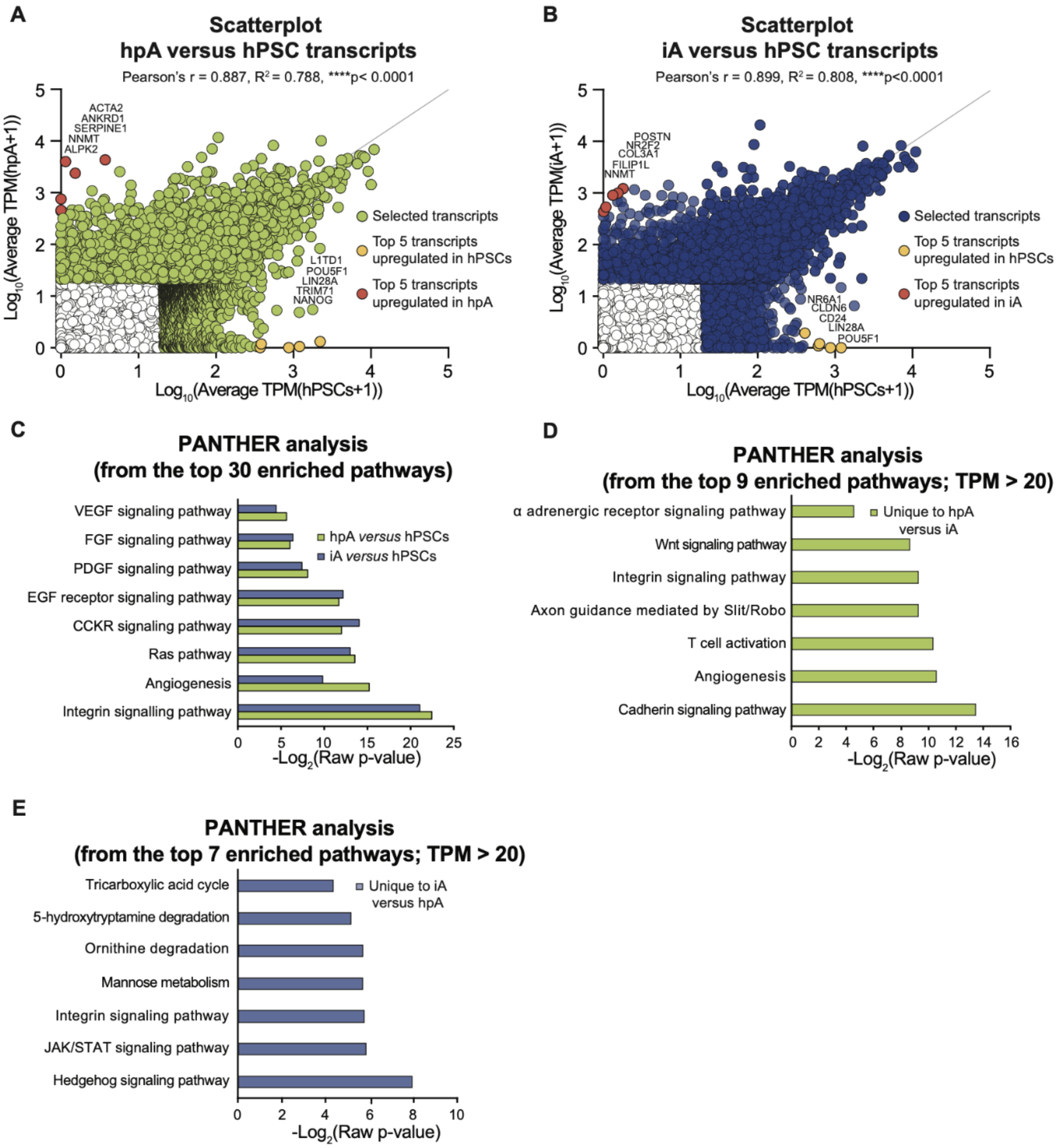


Figure 4.2. Transcriptomic comparison between hpAs, hiAs and hPSCs. Related to Figure 1 and Table S1.

(A) Scatterplot of the correlation between hPSC and hpA expressed transcripts. Pearson's r = 0.887, R2 = 0.788, ****p< 0.0001. (B) Scatterplot of the correlation between hPSC and hiA expressed transcripts. Pearson's r = 0.899, R2 = 0.808, ****p<0.0001. (C) Biological pathways enriched in differentially expressed genes from hpAs (green) and hiAs (blue) compared to hPSCs using PANTHER analysis. (D) Biological pathways enriched in hpAs compared to hiAs using PANTHER analysis. (E) Biological pathways enriched in hiAs compared to hpAs using PANTHER analysis.

4.3 - hiAs recapitulate key functional features of hpAs

We next sought to establish both cell autonomous and non-cell-autonomous functionality of hiAs as compared with hpAs, including the ability to secrete cytokines in response to pro-inflammatory stimuli, calcium oscillation dynamics in response to ATP and the capacity to promote presynaptic development of human in vitro derived neurons (hNs). The astrocyte response to pro-inflammatory stimuli is crucial for normal function, with dysfunction of the inflammatory response strongly implicated in disease (Sofroniew et al, 2010; Kam et al, 2020; Leal et al, 2013; Liddelow et al, 2017). We therefore challenged our hiAs as well as hpAs with TNFα cytokine and subsequently quantified the secretion of the pleiotropic cytokine interleukin 6 (IL-6). Specifically, hpAs and hiAs were treated with either 100ng/mL TNFα or 0.1% BSA control and the harvested supernatant was used to measure secreted IL-6 by ELISA. As expected, both hiAs and hpAs showed a robust and significant response to TNFα stimulation compared to BSA control, with hiAs secreting IL-6 at levels slightly above hpAs (**Figure 4.3A**).

Astrocytes also display spikes of cytoplasmic calcium concentration as a response to mechanical, ATP or glutamate stimulation (Turovsky et al, 2020; Allen et al, 2022; Fujii et al, 2017). To assess the excitability of our hiAs and hpAs, we evaluated their response to ATP stimulation (**Figure 4.3B-E**; Supplemental movies). Specifically, we used Fura-4 AM dye to image calcium concentration in the cytoplasm and recorded eight key features including: fluorescence level before and after stimulation as well as peak height, number, duration, rising time, falling time, peak interval, and the area under the curve. Less than a minute following ATP administration both hiAs and hpAs sharply increased their cytoplasmic calcium concentration (**Figure 4.3B**) and displayed the same clusters of typical behavior with similar distributions overall (**Figures 4.3C** and **4.4**). Data from both cell types were then pooled together and k-mean clustered (k=5) (**Figure 4.3C-E**). After separating cells based on their origin (hiA versus hpA), we also recovered similar signatures for each cluster (**Figure 4.3E**). Notably, individual traces of cells from any given cluster produced the same shape and did not depend on the sample of origin (**Figure 4.4**). These data support highly similar calcium dynamics between hiAs and hpAs.

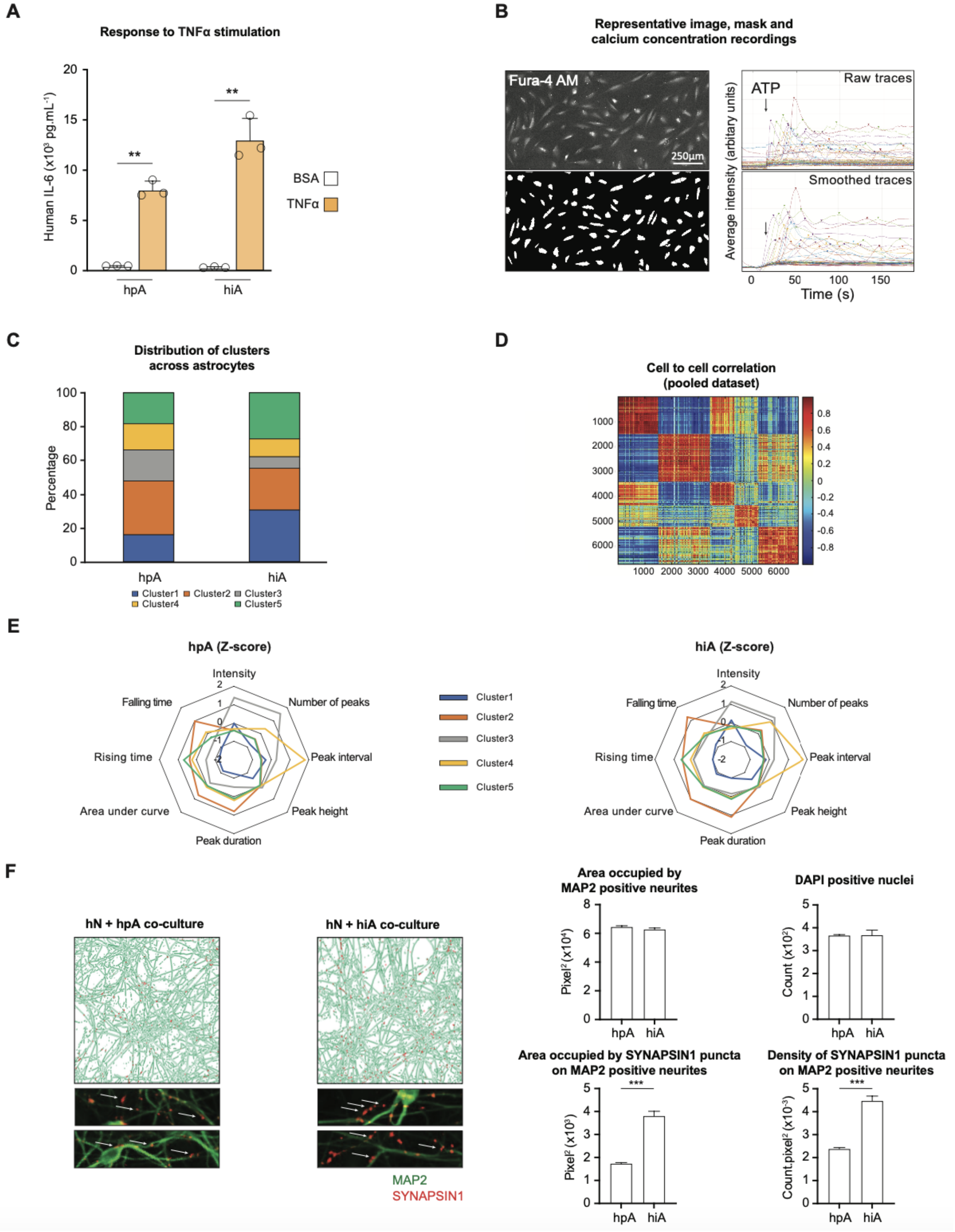
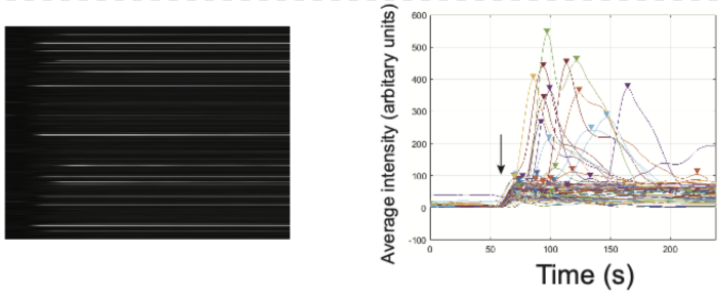


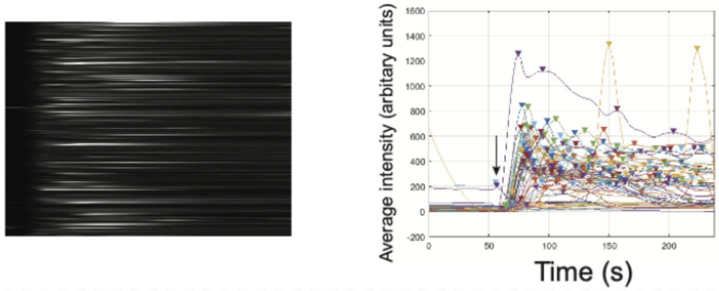
Figure 4.3. hiAs recapitulate key functional features of hpAs.

(A), Bar graph showing human IL-6 detected from hiAs and hpAs by ELISA in response to 100ng/mL human TNFa (orange) versus 0.1% BSA control (white) treatment. (Data are represented as mean +/- SEM, **p<0.01; One-way ANOVA with Tukey's multiple comparisons test; n=3 biological replicates, n=3 technical replicates). (B), ATP-induced calcium release. Representative Fura-4 image of hiAs (top left), and associated mask (bottom left), raw traces (top right) and smoothed traces (bottom right). Intensity (arbitrary unit) is measured as an average on the surface of each cell. Arrow indicates timing of ATP stimulation. Scale bar = 250μm. (C), Distribution of each cell population among the identified 5 clusters. Tables of Z-scored features were pulled together, clustered, then split into hpA and hiA. (D), Correlation heatmap for 6,745 recorded cells. Correlation is computed on the table of Z-scored values for the eight identified features. Cells are sorted by cluster, then within each cluster, are sorted based on their sample of origin, showing that the correlation is strongly dependent on calcium signature but not sample. (E), Radar plot of the mean Z-scored value taken by the 8 features used for clustering the data. Note that the shape of the typical signature of each cluster is the same for hpAs (left) and hiAs (right). (F), Left, Representative CellProfiler output images of the two conditions (hN + hpA and hN + hiA) and representative pictures of SYNAPSIN1 puncta co-localized with MAP2 positive neurites (arrows point to presynaptic puncta). Right, Quantification of the area occupied by MAP2 positive neurites, the area occupied by SYNAPSIN1 puncta colocalized on MAP2 positive neurites, the number of DAPI positive nuclei and the density of SYNAPSIN1 puncta co- localized on MAP2 positive neurites in hN + hpA and hN + hiA co-cultures. Data are represented as mean +/- SEM, n=1 technical replicate, n = 60 biological replicates (wells) per condition, ***p<0.001, unpaired t-test with Welch's correction.

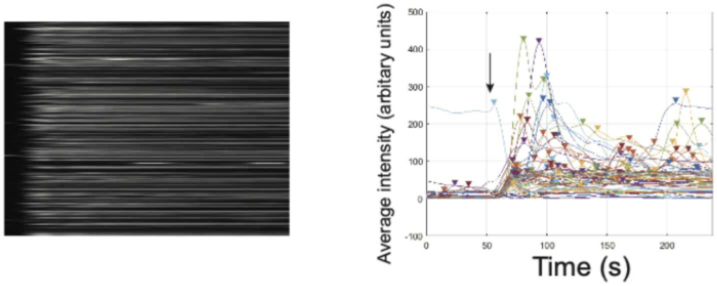
Finally, astrocytes play critical roles in the establishment of synaptic networks, a key non-cell-autonomous function of astrocytes in the human brain. Previous studies have shown that both rodent and human astrocytes can improve the maturation of hNs (Nehme et al, 2018; Canals et al, 2018; Hedegaard et al, 2020; Berryer et al, 2023). Using an established glutamatergic hN differentiation protocol15 combined with an automated synaptic quantification platform30, we analyzed presynaptic development in hN + hiA co-cultures derived from the same parental cell line, as well as hN + hpA co-cultures. Specifically, we quantified the presynaptic SYNAPSIN1 aggregates localized on and along the well described neuronal somato-dendritic marker MAP2 (microtubule associated protein 2). These analyses revealed significant increases in the density and the area of presynaptic puncta opposed to MAP2-expressing neurites in hiA + hN as compared to hpA + hN co-culture, with no difference in the area occupied by MAP2 positive neurites or the number of DAPI positive nuclei detected. This is consistent with the ability of both hiAs and hpAs to support human presynaptic network development (**Figure 4.3F**). Collectively, these results demonstrate that, compared to hpAs, our hiAs harbor similar immunocompetence, calcium transients in response to ATP and contributions to the development of neuronal networks.



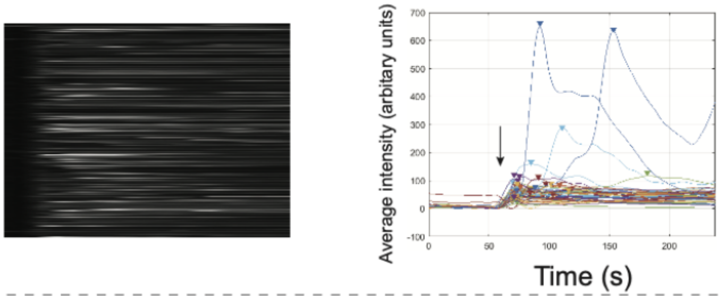
Cluster 1



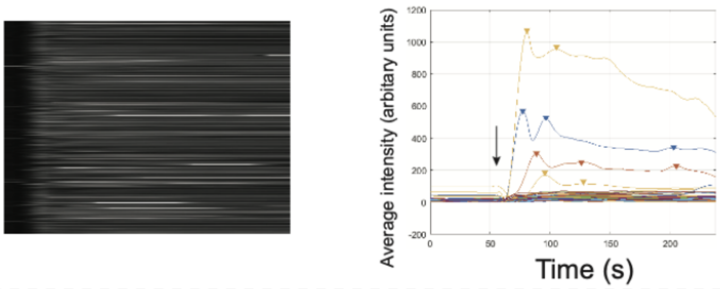
Cluster 2



Cluster 3



Cluster 4



Cluster 5

Figure 4.4. Examples of raw traces and calcium oscillations following ATP addition for typical behavior (cluster). Related to Figure 2 and Movies. Arrows indicate ATP addition. Plots show the traces of 60 randomly chosen cells from all 12 samples of origin.

4.4 - Single-cell transcriptional profiling of hiAs from multiple donors

Some current hPSC based methods for astrocyte generation require purification steps such as FACS to isolate a specific population of interest given the heterogeneity of cellular differentiation, and variability in the capacity of the differentiation to work effectively across a range of unique cell lines. To better understand the degree of homogeneity and reproducibility in our model, we performed scRNAseq on 8 unique parental cell lines (**Table S4.2**). To reduce technical variation, hiAs from each of the 8 donors were differentiated together in a "cell village" as previously described (Wells et al, 2023). In brief, hiAs from each of the 8 cell lines were induced and initially differentiated separately, then mixed in equal numbers 25 days post-induction to form a "village" and sequenced 5 days later (at day 30 post-induction). Given the proliferative nature of the cells during the early stages of differentiation, we mixed cells from each donor later in their differentiation process relative to initial village methods in an effort to mitigate variability in growth rates impacting donor representation following sequencing. Each individual cell was then assigned to its donor-of-origin using transcribed single nucleotide polymorphisms (SNPs). Uniform Manifold Approximation and Projection (UMAP) showed high homogeneity across individual cell lines, with no cell line clustering distinctly from others using leiden unsupervised clustering (**Figure 4.5A**). Our scRNAseq analyses further confirmed that hiAs expressed canonical immature astrocytes markers including VIM, GJA1, APOE, CD44, SLC1A3, and ID3 (**Figure 4.5B**). Given the propensity of in vitro differentiation protocols to have high heterogeneity and limited cell-type specificity, we explored whether using our NGN2 induction approach would produce oligodendrocyte or neuronal populations among our cells. We found that hiAs displayed minimal expression of common genes for NPCs, neurons and oligodendrocytes and continued to express limited amounts of stem cell related genes (**Figure 4.5C**). To explore the variability across cell lines, we next compared the expression of canonical immature and mature astrocyte markers across donors, as well as global gene expression. Across all 8 cell lines, we observed a high degree of correlation of global expression across donors (Pearson’s r = 0.96-0.98), highlighting the reproducibility of our method across a range of cell lines (**Figure 4.5D**).

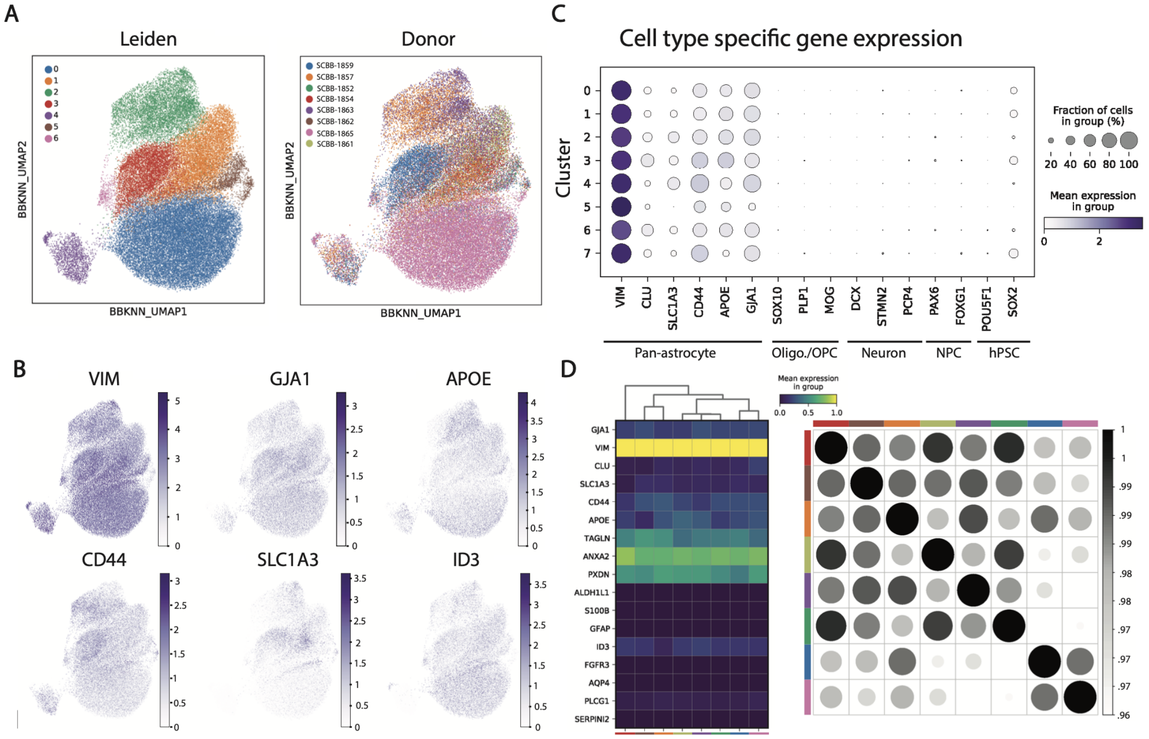


Figure 4.5. Single-cell transcriptional profiling of hiAs from multiple donors.

(A), UMAP projection of scRNAseq data from 8 unique parental lines labeled by Left, cluster using Leiden unsupervised clustering and right, cell line ID. (B), Feature plot illustrating the distribution of canonical astrocyte markers VIM, GJA1, APOE, CD44, SLC1A3, and ID3 in UMAP space. (C), Dot plot for markers of astrocytes (ID3, CLU, SLC1A3, CD44, APOE, GJA1), oligodendrocytes (SOX10, PLP1, MOG), excitatory neurons (DCX, STMN2, PCP4), neuronal progenitor cells (PAX6, FOXG1), and pluripotent stem cells (POU5F1, SOX2). (D), Left, Matrix plot displaying average expression of canonical astrocyte markers by cell line. Right, correlation matrix of average expression across all genes between cell lines.

4.5 - Comparison of hiAs with in vitro and ex vivo astrocyte datasets

An important aim of *in vitro* models is to develop cellular substrates which resemble those found in the living human brain. To understand the similarities and differences among current iPSC-based astrocyte protocols, we compared our hiAs with existing scRNAseq datasets (see

Methods; Rapino et al, 2022, Leng et al, 2022, Barbar et al, 2020). We found that UMAP displayed a high degree of similarity across each *in vitro* dataset with the majority of cells clustering together (**Figures 4.6A** and **4.6B**). hiAs showed a high degree of homogeneity, with fewer sub-populations which clustered separately from the majority compared to other datasets (**Figure 4.6C**). We next compared the overall correlation of gene expression across iPSC-astrocyte models and found a moderate to high level of correlation (Pearson's r = 0.65-0.88) (**Figure 4.6D**). Though global expression across these datasets was similar, we wondered what genes were driving differences between them. We thus performed a differential expression analysis to compare across individual models. Genes contributing to differences across models were not specific for astrocyte identity or function and included many mitochondrial and ribosomal genes (**Figure 4.6E** and **Table S4.3**). This illustrated that on average, each model resulted in comparable levels of expression of canonical astrocyte markers. While this was true when exploring average expression, there were differences in canonical astrocyte gene expression patterns across clusters, where certain models populated greater proportions (**Table S4.4**). For example, cluster 9 showed elevated expression of immature astrocyte markers such as TOP2A, CENPF and NUSAP1 relative to other clusters (**Figure 4.6F**). As shown in Figure 4C, cluster 9 predominantly contained cells from Leng et al (2022), although only representing ~ 0.2% of the total population. However, issues such as technical variation and limitations of sequencing depth may impact the detection and direct comparison of specific transcripts across datasets. We therefore examined expression across a large set of genes associated with astrocyte identity and behavior instead of focusing on single genes, creating a metagene score for each cell based on its contribution to expression of a set of genes associated with either astrocyte precursor cells or mature astrocytes as described by Zhang et al (2016). When we assessed the contribution of each model to these gene sets, we found that all models contributed similar amounts of RNA to both astrocyte precursor cell markers as well as mature astrocyte markers (**Figure 4.6G**). Consistent with these findings, hiAs showed lower standard deviations than their counterparts, further illustrating the homogeneity of our cells in their contribution to the expression of these specific markers compared to the other iPSC astrocyte models.

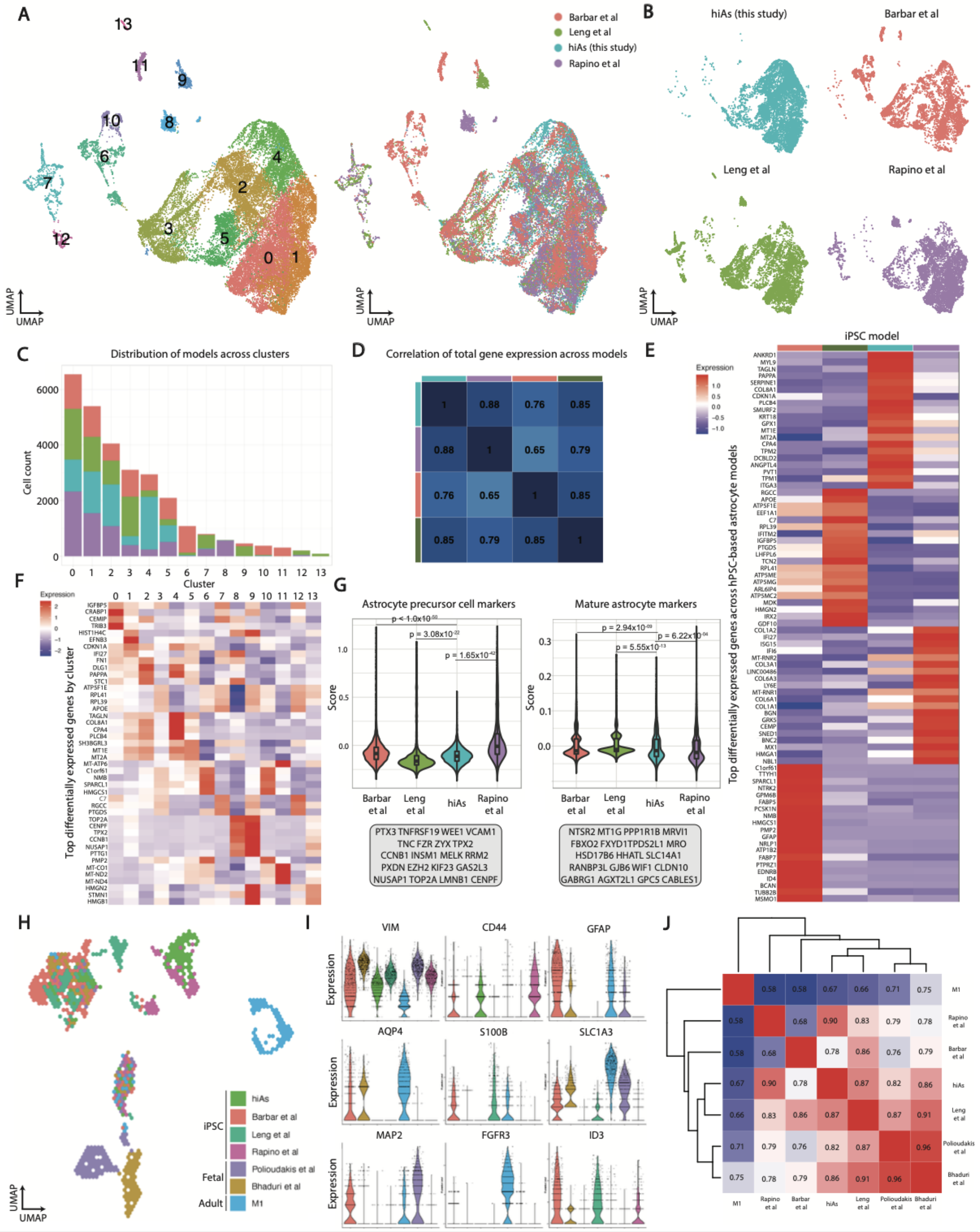


Figure 4.6. Comparison of hiAs with in vitro and ex vivo astrocyte datasets.

(A), UMAP projection of scRNAseq data from four iPSC-astrocyte datasets (Barbar et al29, Leng et al28, Rapino et al27 and hiAs generated in this study) labeled by Left, cluster using Louvain unsupervised clustering and right, by study. (B), Left, top, UMAP projection only labeling hiA data. Left, bottom, UMAP projection only labeling Leng et al28 data. Right, top, UMAP projection labeling only Barbar et al29 data. Right, bottom, UMAP projection labeling only Rapino et al27 data. (C), Distribution of cells from each iPSC-astrocyte dataset by cluster. (D), Correlation of average gene expression across all iPSC-astrocyte models. (E), Top 20 differentially expressed genes in each iPSC-astrocyte dataset. (F), Top 4 differentially expressed genes in each cluster (G), left, Metagene enrichment score for astrocyte precursor markers across iPSC-astrocyte datasets. p-value = 0.006, ANOVA. right, Metagene enrichment score for mature astrocyte markers across iPSC-astrocyte datasets. p-value < 0.005, ANOVA. Data are represented as mean +/- SEM. (H), UMAP projection of iPSC-astrocyte datasets integrated with fetal and post-mortem human brain data labeled by dataset. (I), Expression of canonical astrocyte markers (VIM, CD44, GFAP, AQP4, S100B, SLC1A3, MAP2, FGFR3, and ID3) across iPSC-astrocyte and human brain datasets. (J), Canonical correlation analysis of iPSC-astrocytes and human brain datasets for average global gene expression.

To understand the *in vivo* fidelity of our system as well as other iPSC-based astrocyte models, we integrated our scRNAseq data and published datasets with existing data from the fetal human brain prefrontal cortex (Bhaduri et al, 2021; Polioudakis et al, 2019) and M1 motor cortex from post-mortem adult brain (BICCN, 2021). UMAP analysis revealed similar clustering between in vitro datasets, with some clusters including cells from the two fetal datasets and the M1 atlas used in the analysis (**Figure 4.6H**). While some cells generated from each method overlapped with a subset of cells from the human brain tissues, most cells from the fetal and post-mortem datasets clustered separately (**Figure 4.6H**). We also observed the same pattern when including hpAs (derived from fetal tissue) in the analyses (**Figure 4.7**). Similar to hiAs and other iPSC-astrocytes, hpAs also failed to cluster with data from the fetal cortex in post-mortem data (**Figure 4.7A**), consistent with the high correlation we had noted between the gene expression profiles of hiAs and hpAs (**Figure 4.1**). The expression of some canonical astrocyte genes varied across the brain and stem cell-based datasets (**Figure 4.6I**). For instance, AQP4 and FGFR3 were more strongly expressed in human adult astrocytes compared to human primary fetal or stem cell derived astrocytes, while VIM was robustly expressed across all datasets (**Figure 4.6I**). Interestingly, we noticed discordance between the protein and mRNA abundance for several canonical astrocyte markers such as AQP4, SLC1A3, and S100B in our hiAs (**Figures 4.1C and 4.6I**). Indeed, despite low transcript levels for these genes, we could reliably detect the presence of the protein (**Figure 4.1C**). This phenotype was also observed in the human primary astrocytes (**Figure 4.7B**). We further compared the average expression across all genes between in vitro datasets and the human brain. When looking at the correlation of global gene expression across datasets, hiAs exhibit relatively high correlation with the fetal brain data, and a moderate correlation with the M1 atlas data (Pearson's r = 0.82, 0.86, 0.67, respectively) (**Figure 4.7J**). Importantly, while other iPSC-based astrocyte models exhibited similar correlations to these specific human brain datasets, hiAs showed the strongest correlation to the M1 atlas data relative to other in vitro models (Pearson's r = 0.67) (**Figure 4.7J**). Based on these analyses, we hypothesized that longer term culture of hiAs might further mature them. Transcriptomic analysis from hiAs cultured until D60 showed modest changes in gene expression space, and a small increase in correlation to the human brain datasets, suggesting hiAs exhibit a small increase in maturity with additional time in culture (**Figure 4.8A-D**). Additionally, we leveraged a published dataset of the same NGN2-driven NPC-like cells at day two (Wells et al, 2023) to examine the molecular trajectory of iPSCs, NPCs and hiAs as compared with hpAs (**Figure 4.7E**). Here, we observed an increase in early astrocyte fate regulators including NFIA, NFIB and SOX9 in hiAs and hpAs as compared with NPCs and iPSCs (**Figure 4.8E**). When assessing transcription factors implicated in astrocyte maturation (Lattke et al, 2021), we detect modest induction of RORB and LHX2, and no induction of DBX2 and FEZF2 in either hiAs or hpAs (**Figure 4.8F**). Overall, our data suggested that hiAs closely resembled other iPSC-derived astrocytes and exhibited reasonable correlation to relevant human brain datasets from the prefrontal and motor cortex, presenting a robust model for investigating astrocyte biology as early as 30 days after induction.

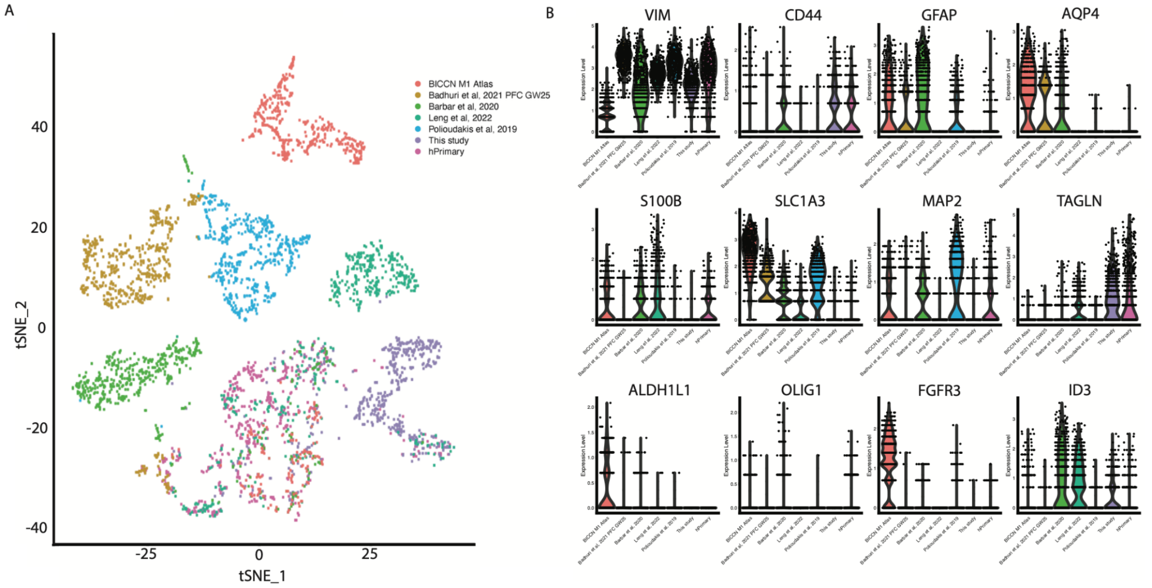


Figure 4.7. Integration of human primary astrocytes with in vitro and post-mortem brain data. Related to Figure 4.6 and Tables S4.3, S4.4 and S4.5.

(A) tSNE projection of scRNAseq data from in vitro and post-mortem astrocytes color-coded by dataset. (B) Violin plots for expression levels of canonical glia markers grouped by dataset.

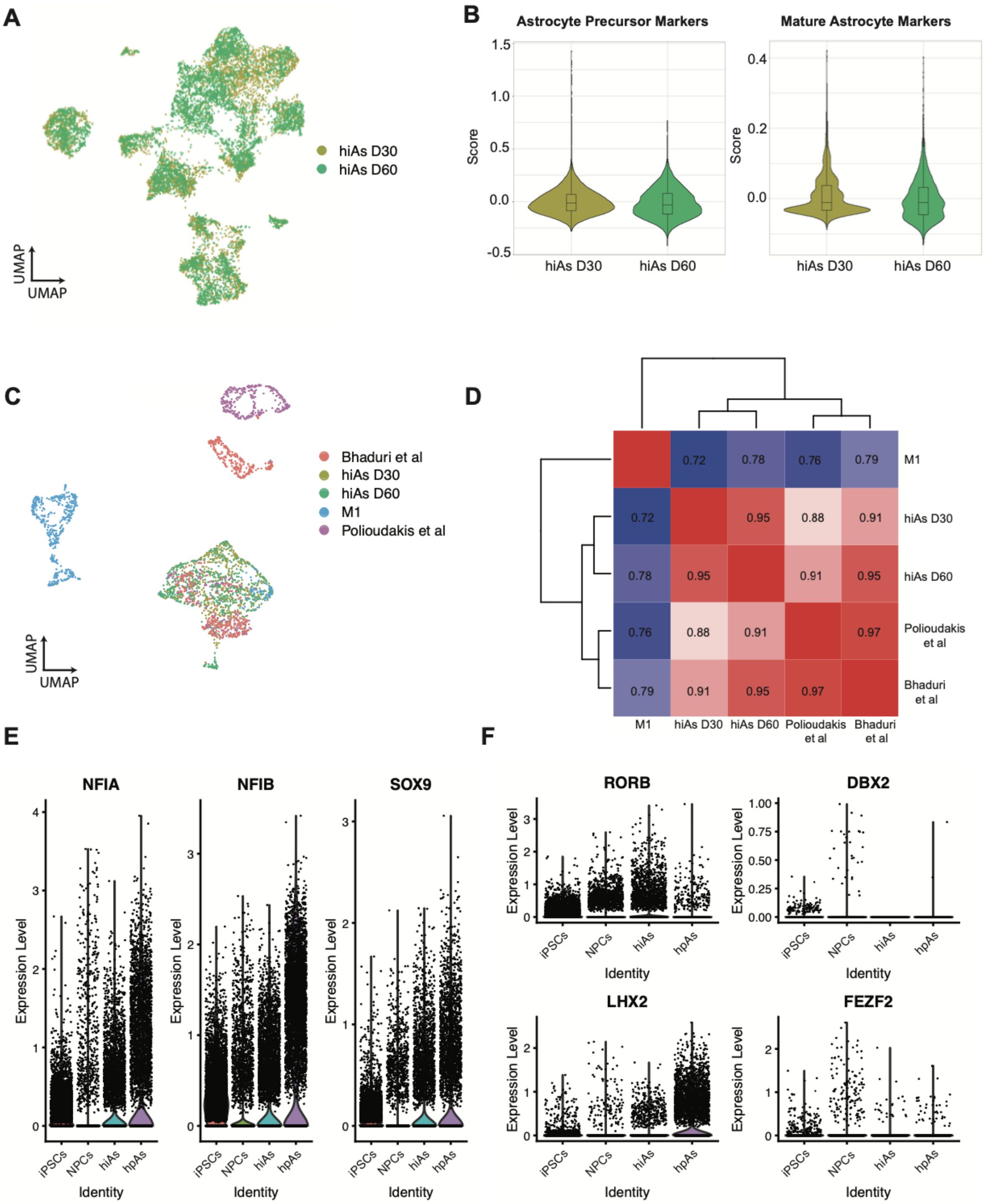


Figure 4.8. Comparison of D30 and D60 hiAs. Related to Figures 4.5 and 4.6.

(A) UMAP projection of D30 and D60 hiAs labeled by dataset. (B) Metagene enrichment score for astrocyte precursor markers and mature astrocyte markers between D30 and D60 hiAs. Data are presented as mean +/- SEM (C) UMAP projection of D30 and D60 hiAs integrated with human brain reference datasets. (D) Canonical correlation analysis between D30 and D60 hiAs integrated with human brain reference datasets. (E) Expression of early astrocyte fate transcription factors NFIA, NFIB, and SOX9 across iPSCs, NPCs1, hiAs and hpAs. (F) Expression of mature astrocyte regulators across each dataset.

4.6 - hiAs capture disease phenotypes in model of trisomy 21

We next sought to establish the utility of our hiA protocol for disease modeling, given the high relevance of this application. Specifically, numerous studies have shown deficits in astrocytes due to triplication of chromosome 21, which drives Down syndrome (Araujo et al, 2018; Chen et al, 2014; Mizuno et al, 2018; Ponroy Bally et al, 2020; Ponroy Bally et al, 2021). For example, using iPSC-derived astrocytes, Chen et al (2014) found that trisomy 21 drove global transcriptional perturbations as well as higher levels of reactive oxygen species and reduced expression of pro-synaptic factors compared to euploid astrocytes. Also using iPSC-derived astrocytes, Bally et al (2020) identified global transcriptional perturbations due to trisomy 21 coupled with chromatin accessibility analyses, revealing alterations to axon development, extracellular matrix organization and cell adhesion. We therefore generated hiAs from an isogenic pair of commercially available iPSC lines with and without trisomy 2147, referred to as DS1 (trisomy 21) and DS2U (euploid) hiAs (**Table S4.5**). As expected, hiAs derived from both euploid and trisomy 21 iPSCs expressed canonical astrocyte markers (**Figure 4.9A-B**), in accordance with our previous analyses of a control cell line (**Figure 4.1C-D**). Given the high degree of homogeneity observed in our scRNAseq datasets (**Figures 4.5** and **4.6**), we then extracted RNA for bulk transcriptional analyses (**Figure 4.9C**). As expected, a majority of expressed genes such as COL18A1, COL6A, BACE2 and ADARB1, encoded on chromosome 21 (HSA21) were upregulated in trisomy 21 hiAs compared with euploid control hiAs, with a median fold change of 1.4931 as compared to 0.9623 obtained with the same analysis performed on chromosome 3 (**Figures 4.9D** and **4.10**). Also, using a padj threshold of 0.05 and a log2FC threshold of +/-2, we observed global transcriptional perturbations, detecting 1691 significantly upregulated genes and 414 significantly downregulated genes due to trisomy 21 (**Figure 4.9E**). GO term analyses of the differentially expressed genes revealed biological processes such as cell-cell adhesion, synaptic signaling and neuron projection morphogenesis (**Figure 4.9F**), consistent with previous studies of cell-autonomous astrocyte dysfunction as well as deleterious impacts on neuronal and synaptic development in Down syndrome; of note, our analyses captured astrocyte phenotypes also detected in a 160+ day astrocyte differentiation protocol. Examples of individual genes involved in neuron projection morphogenesis and cell-cell adhesion are highlighted in Figure 4.5G and H. These data support the expected transcriptional dysregulation of hiAs when used in a model of trisomy 21.

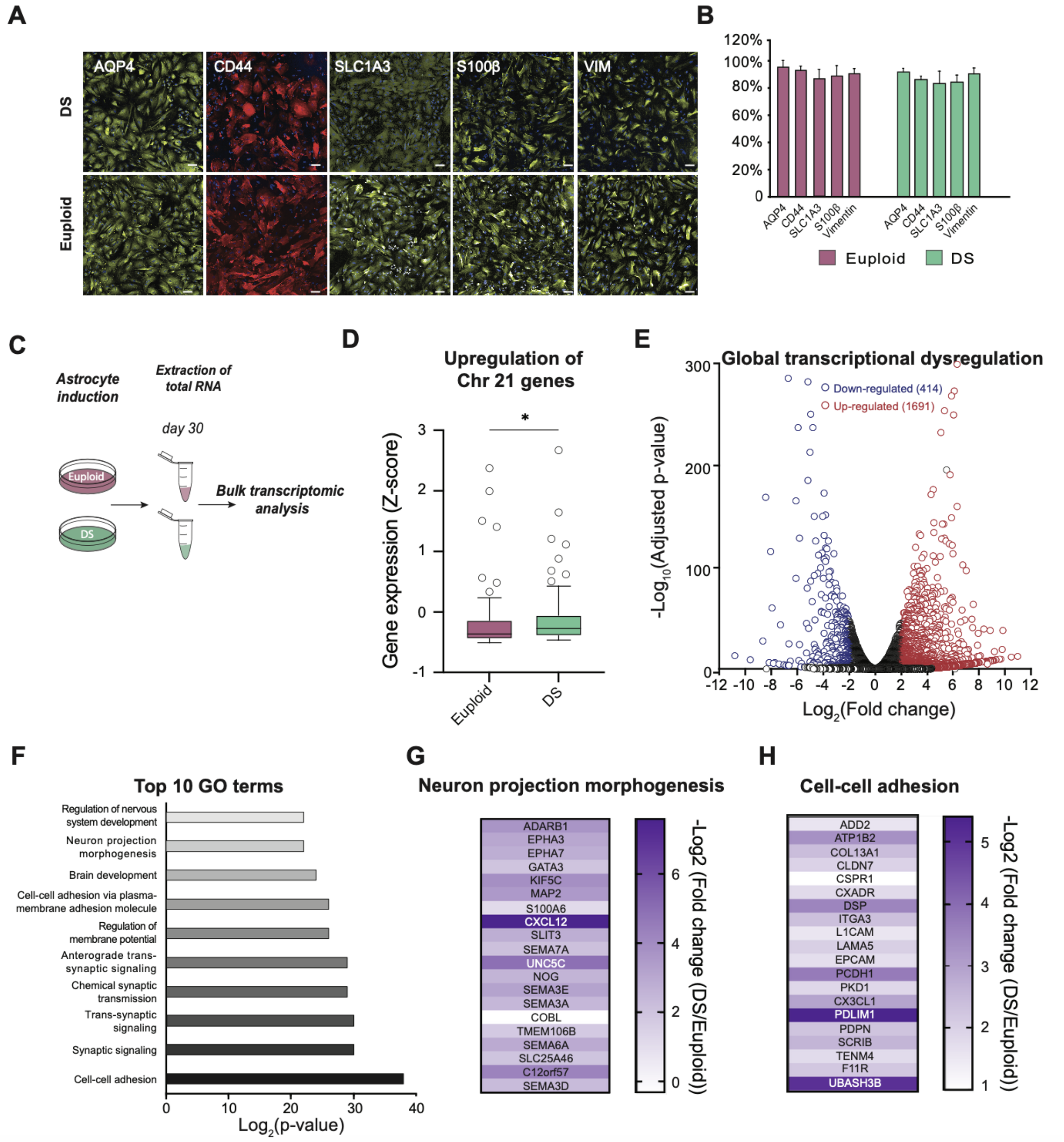


Figure 4.9. hiAs capture disease phenotypes in model of trisomy 21.

(A), Representative immunofluorescence images for AQP4, CD44, SLC1A3, S100b and VIM from astrocytes derived from DS patient cells (DS1) and astrocytes derived from euploid control cells (DS2U). Scale bar = 100 μm. (B), Quantification of each marker from DS and euploid astrocytes shown as percentage of DAPI positive cells. Data are represented as mean +/- SEM. n>3 biological replicates, n=3 technical replicates (C), Schematic of bulk transcriptional analyses for DS and euploid astrocytes. (D), Gene expression Z-scores for 77 genes encoded on chromosome 21 from DS versus euploid control cells. Data are represented mean +/- SEM, *p=0.0142, Mann Whitney two-tailed t test. (E), Volcano plot showing wide dysregulation of the transcriptome in DS astrocytes compared to isogenic euploid controls, with a bias towards upregulation. Significantly down-regulated genes are shown in blue and significantly up- regulated genes are shown in red. Log2FC is shown on the x-axis and the -log10 of the adjusted p-value is shown on the y-axis. (F), The top ten biological processes identified as enriched in differentially expressed genes, as calculated by Gene

Ontology (GO) analysis performed via Metascape. The Log2 of the p-value is shown on the x-axis. (G-H), Heat maps showing examples of genes from the 'neuron projection morphogenesis' (left) and 'cell-cell adhesion' (right) GO terms identified in (F). Scale shows the -log2FC of DS versus euploid control.

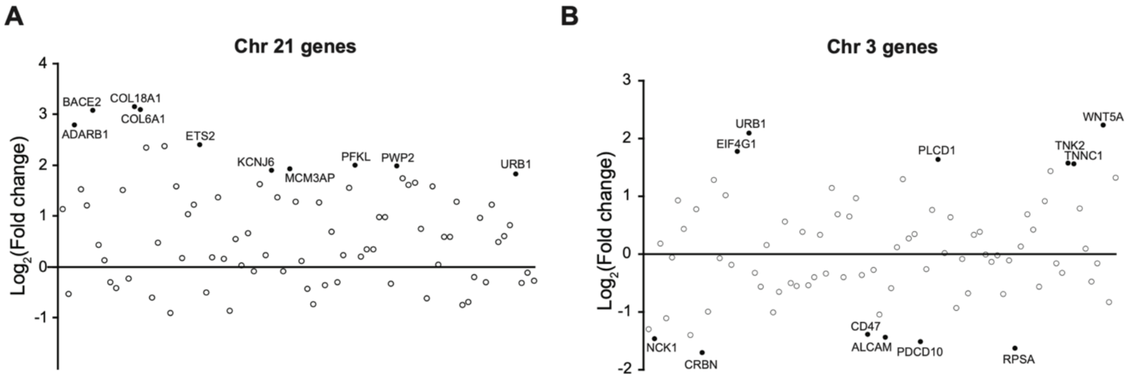


Figure 4.10. Differentially expressed genes per chromosome from DS2U versus DS1 hiAs. Related to Figure 5.

Plots showing Log2FC for genes encoded on chromosome 21 (A) and chromosome 3 (B) comparing DS1 versus DS2U hiAs. Note the expected upregulation of genes on chromosome 21 as compared with chromosome 3.

4.7 – Discussion

Despite their essentiality for modeling normal brain function as well as dysfunction in disease, glial cell types have lagged somewhat behind neuronal cell types when it comes to hPSC-based differentiation technologies. Protracted astrocyte differentiation protocols as well as those requiring 3-dimensional culture or additional purification steps complicate their utility for multi-cell line, adequately powered functional studies. Here, we present a simplified, rapid, and robust protocol to generate homogeneous astrocytes with detailed molecular and functional benchmarking compatible with many parental cell lines and for disease modeling applications. These astrocytes, generated through a simple protocol driven by transient NGN2 expression, produce inflammatory responses, elicit calcium signaling, and have pro-maturational effects on iPSC derived neurons. Further, they are transcriptionally concordant with human primary astrocytes, a common in vitro model used to study human astrocyte biology, show a great degree of overlap with existing in vitro astrocyte approaches which often require longer culture times or more elaborate interventions to isolate specific cell populations, and display strong correlations with existing human brain datasets. Finally, the hiA cell population is functionally and transcriptionally homogeneous rendering this approach amenable to genetic or pharmacological perturbation screens and circumventing the need for costly single cell sequencing approaches. These features will facilitate the study of human astrocytes for disease modeling as well as drug screening applications.

In the future, we see two key areas to build off from our current hiA differentiation protocol. First, given the high degree of reproducibility across parental cell lines as well as the ability to capture disease-relevant phenotypes and contribute to the maturation of neuronal networks, we see precision co-culture of hiAs with additional brain cell types as a logical next step. Indeed, the advantage of generating each relevant brain cell type separately is the ability to precisely control cell ratios, to generate highly reproducible preparations and to manipulate genotypes or employ genetically diverse parental cell lines to explore cell-type specific effects on network function. We show here that our hiAs can be co-cultured with human neurons and similarly contribute to their development as hpAs, and future studies to incorporate additional brain cell types will facilitate study of a host of biological questions. Second, while hiAs possess key molecular and functional features of hpAs, similar to other iPSC-derived astrocytes, they more closely resemble fetal rather than adult human astrocytes. Thus, more needs to be done to enhance their maturation in order to study the role of human astrocytes in processes beyond early development. In this regard, it is interesting to note that additional time in culture only modestly improved hiA maturity, suggesting that additional extrinsic factors may be required for further maturation. Indeed, one possibility is that a more complex co-culture system including neurons or other glial cell types will be required to recapitulate the in vivo environment more accurately and further mature the hiAs. In this regard, we recently examined transcriptional changes induced by co-culturing human neurons with murine glia, revealing enhancement of synaptic gene expression programs in neurons and increased cell adhesion molecules in glia3 consistent with the pro-maturational effects of glia on human neurons; it remains to be determined whether human neurons also have a pro-maturational effect on human astrocytes in vitro. It is also possible that additional gene networks may need to be activated in order to achieve a more mature astrocyte state, including the induction of key astrocyte fate regulators.

4.8 - Reflection

In this manuscript, we describe a strategy to generate human astrocytes quickly and efficiently from pluripotent stem cells using a combination of NGN2 expression and astrocyte differentiation media. The NGN2 overexpression approach for neuronal differentiation has been widely adopted in the field for studying neuronal biology and disease. Here we report the novel finding that neuronal progenitor cells generated through transient NGN2 induction can be differentiated into functional astrocytes, closely resembling other iPSC-derived astrocytes and with high correlation to human primary astrocytes and fetal brain datasets, making this protocol highly accessible without the need for building or adapting additional tools. This project developed an approach that filled a gap in the neuroscience field by enabling a rapid and scalable method to produce astrocytes from stem cells. One notable challenge in the field was that existing approaches for hPSC-astrocyte generation are typically lengthy, insufficiently characterized or require intermediate purification steps, limiting their utility for multi-cell line, adequately powered functional studies. We provide a novel tool which overcomes many of these limitations, increasing the field's repertoire to extend in vitro studies to a new range of diverse cell types.

Specifically, we show that our human induced astrocytes (hiAs): display remarkable homogeneity within the population and across 11 parental cell lines in the absence of additional purification steps; show high transcriptional concordance with primary human fetal astrocytes; respond to pro-inflammatory stimuli; exhibit ATP-induced calcium transients; support neuronal maturation in vitro; single-cell transcriptomic analyses reveal the generation of highly reproducible cell populations across individual donors; which are highly similar to hPSC-derived astrocytes generated using lengthier and more involved approaches; and capture key molecular hallmarks in a trisomy 21 disease model. Thus, hiAs provide a valuable and practical resource which can now be leveraged by the community for study of basic human astrocyte function and dysfunction in disease.

Now with a robust and reliable method for generating astrocytes from human stem cells, I was primed to study astrocytes in a living system and investigate how these cell types respond to perturbations of their biology. Additionally, I sought to develop an experimental and intellectual framework for understanding interactions between astrocytes and neurons *in vitro*. In the next chapter, I highlight my work investigating how astrocytes and neurons respond to various pharmacological perturbations using cell villages from many donors. Here, I unearth new insights into how antipsychotic medications impact the function of astrocytes and neurons at the cellular level and find coordinated transcriptional responses which may mirror observations from the human brain.

4.9 – Limitations

In this Chapter, I present a new method for generating astrocyte-like cells from iPSCs. There are still limitations which compromise the fidelity of the model. Absent in our data, both transcriptomics and immunostaining, is the expression or abundance of glial fibrillary acidic protein (GFAP). This has been a historic marker for assessing astrocyte identity and many protocols for generating astrocytes emphasize the activity of this gene as an indicator of success in developing a new differentiation protocol. This can become problematic for evaluating novel methods. GFAP is well established as a marker for reactive astrocytes and is known to be quite variable depending on the physiological conditions of the model system (TCW, 2017). For example, under physiological conditions, expression of GFAP in the mouse brain can be non-existent (Haim et al, 2015).

The absence of GFAP is our system could have many implications. It could suggest that we lack a population of reactive astrocytes, which may begin to upregulate GFAP activity when exposed to a given stimulus. It will be important to test this hypothesis in future experiments to better understand and physiological nature of the astrocytes we describe here. More importantly, I think the cell biology community should think more deeply about marker genes for assigning cell types. Many recent data resources for constructing an atlas of the human brain do not use GFAP as a marker for astrocyte identity (Bhaduri et al, 2021, BICCN, 2021). It will be important to generate consensus around how we identify cell types in our in vitro systems so that we can meaningfully make statements about cell-type biology.

4.10 – Statement on contributions

This work was co-led by Martin Berryer. Martin designed the Ngn2 construct and performed experiments related to Section **4.2**. I introduced the Ngn2 construct into several iPSC lines and performed all experiments and analyses related to Sections **4.4** and **4.6.** Experiments for Section **4.3** were performed by other members of the lab.

Chapter 5

Astrocyte-neuron interactions and their dynamic responses to antipsychotic medications

5.1 – Introduction

Schizophrenia is a severe brain disorder characterized by delusions and hallucinations, impairments in executive and other cognitive function, and flattened motivation, emotion, and interest (van Os et al, 2009). It affects 1% of people globally, but the biological mechanisms underlying the disorder are unknown (Owen et al, 2016). Inheritance is a major risk factor, and recent genetic discoveries have highlighted the quantitative enrichment of genes that are highly expressed by neurons and encode proteins that function at synapses (Finucane et al, 2015; Singh et al, 2022; Trubetskoy et al, 2022). These fundamental aspects of neuronal biology are dependent on interactions with glial cells, including astrocytes (Eroglu et al, 2010; Verkhratsky et al, 2018), and raise the question whether cell-nonautonomous effects of glial cells on neurons are relevant to brain disorders such as schizophrenia. Astrocytes provide neurons with homeostatic support and regulate neuronal development and maturation (Verkhratsky et al, 2018). They participate in the formation and shaping of the neuronal network by regulating synapse generation and elimination, transmission, and plasticity (Eroglu et al, 2010; Clarke et al, 2013; Pfrieger et al, 2009). Astrocytes surround neuronal cell bodies and synapses and interact with neurons through a range of contact-dependent and secreted signals that contribute to neuronal maturation (Clarke et al, 2013; Allen et al, 2014; Allen et al, 2017). However, although glial cells are necessary for the functional maturation of neurons (Nehme et al, 2018; Pfrieger et al, 1997; Ullian et al, 2001; Vierbuchen et al, 2010), many gaps remain in our understanding of the specific cellular and molecular programs that mediate these processes. Previous findings in our group revealed that upon co-culture with mouse astrocytes, stem cell derived neurons upregulate the expression of synaptic genes which are enriched for schizophrenia heritability (Pietilainen et al, 2023). Additionally, this increase in synaptic expression in the neurons was associated with an increase in synaptic cell adhesion and pro-synaptic genes, including NRXN1, in the mouse astrocytes. We observed that physical contact between neurons and astrocytes was required to induce many of the pro-synaptic effects. These data suggest that the cellular processes in glia that are associated with neuronal maturation involving synaptic programs *in vitro* are relevant to schizophrenia and provide insight into the potential role of astrocytes in psychiatric disorders.

Further, we observed that stem cell derived neurons upregulate synaptic gene expression in a coordinated manner with an increase in the expression of cholesterol-synthesis genes by mouse glia (Vartiainen & Tegtmeyer et al, *in preparation;* Nehme and Pietilainen labs*)*. It is well established in the field that the arrival of glial cells during brain development corresponds to a downregulation of cholesterol biosynthesis in neurons as astrocytes become the primary supplier of cholesterol within the central nervous system (CNS) (Li et al, 2022). In addition to these results *in vitro*, analyses performed in Steven McCarroll's lab using RNA sequencing data from human brain tissue from 191 donors uncovered tightly correlated transcriptional programs by which astrocytes and neurons couple their gene-expression investments in cholesterol biosynthesis genes (astrocytes), synaptic-adhesion molecules (astrocytes) and synaptic components (neurons); the astrocyte and neuronal gene-expression programs were both greatly enriched for schizophrenia-associated genes, and their expression was reduced in schizophrenia patients (Ling et al, *under revision*).

Systems of living, interacting cells provide ways to study dynamic responses to acute perturbations of these and other pathways, and to learn what genes and alleles regulate these responses. Here, I sought to investigate how astrocytes and neurons, derived from many individuals, respond to common drug treatments to better understand how different CNS cell types respond to various perturbations. To do so, I subjected stem cell derived neurons and astrocytes to many types of pharmacological perturbations, including perturbations of glutamate receptors and L-type calcium channels, antipsychotic drugs (clozapine and haloperidol), oxidative stress, cytokine treatment, and modulation of cholesterol biosynthesis (Simvastatin, Atorvastain, and Efavirenz). I performed scRNAseq as an initial readout of these drugs' effects. Some of the most intriguing results (on which we focus below) involve modulations of the same cholesterol-biosynthesis pathway that is under-expressed in astrocytes in schizophrenia patients.

5.2 - Pharmacogenomics in cell villages of diverse neural cell types

In order to better understand how astrocytes and neurons respond to perturbations of biological pathways which are implicated in astrocyte-neuron interactions and psychiatric disorders, I differentiated a "cell village" of iPSCs from 44 unique individuals into excitatory neurons and astrocytes and exposed them to many types of pharmacological perturbations including perturbations of glutamate receptors and L-type calcium channels, antipsychotic drugs (clozapine and haloperidol), oxidative stress, cytokine treatment, and modulation of cholesterol biosynthesis (Simvastatin, Atorvastain, and Efavirenz). To measure transcriptional responses to these perturbations, I generated single-cell RNA sequencing (**Figure 5.1A**).

I first sought to characterize the cells generated using methods previously published in our groups for differentiated iPSCs into excitatory neurons and astrocytes (Nehme et al, 2018; Berryer and Tegtmeyer et al, 2023). Unsupervised clustering and uniform manifold approximation and projection (UMAP) reduction showed distinct clusters which were separated by cell type of origin (**Figure 5.1B**). Additionally, to confirm identity of each cell type I examined the expression of canonical markers for each of the four cell types included in this experiment (iPSCs: POU5F1, NPCs:DLL3, neurons: CNTNAP2, and astrocytes: VIM).

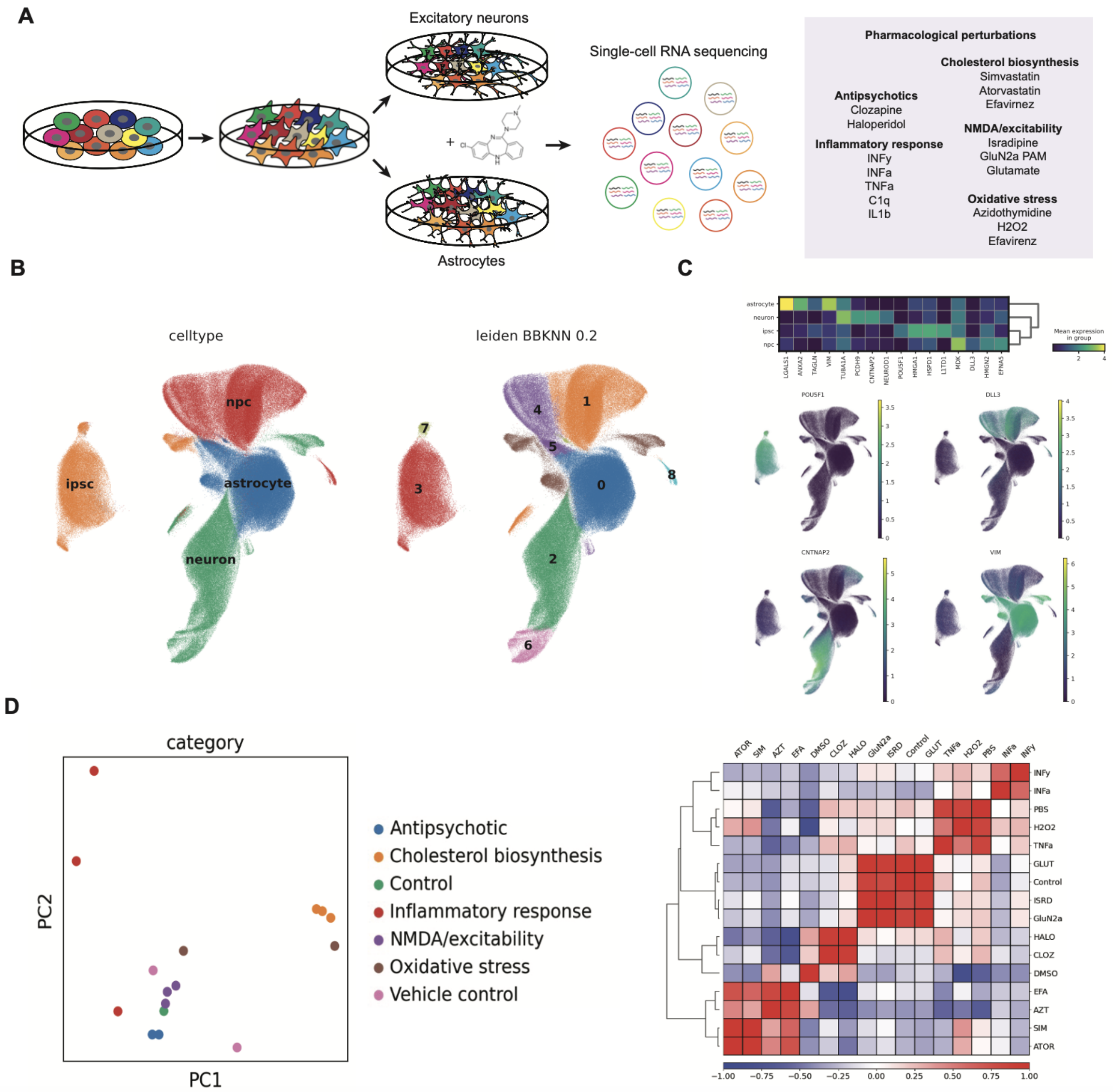


Figure 5.1. Pharmacogenomics in cell villages of diverse neural cell types.

(A) Schematic of experimental design. (B) UMAP reduction of single-cell RNA seq showing distribution of diverse cell types. (C) Expression of canonical markers for each cell type. (D) MDS plot and correlation matrix from pharmacological treatments in excitatory neurons.

Once I had validated the distinct cell types in the data, I wanted to better understand the degree to which any of the pharmacological perturbations supplied to the cells impacted their gene expression. To do so, we implemented a multi-dimensional scaling and hierarchical clustering approach to measure the global gene expression changes driven by individual perturbations and by perturbation categories. These results showed that treatment categories (for example, perturbation of the cholesterol biosynthesis pathway using the 3 drugs) elicited similar overall changes in gene expression when compared across different treatment categories (**Figure 5.1D**). We observed high correlations between the impact of a given perturbation at the gene level with other perturbations in their same category, while also identifying similarities between treatments from different categories (**Figure 5.1E**).

5.3 - Effects of antipsychotics on astrocytes

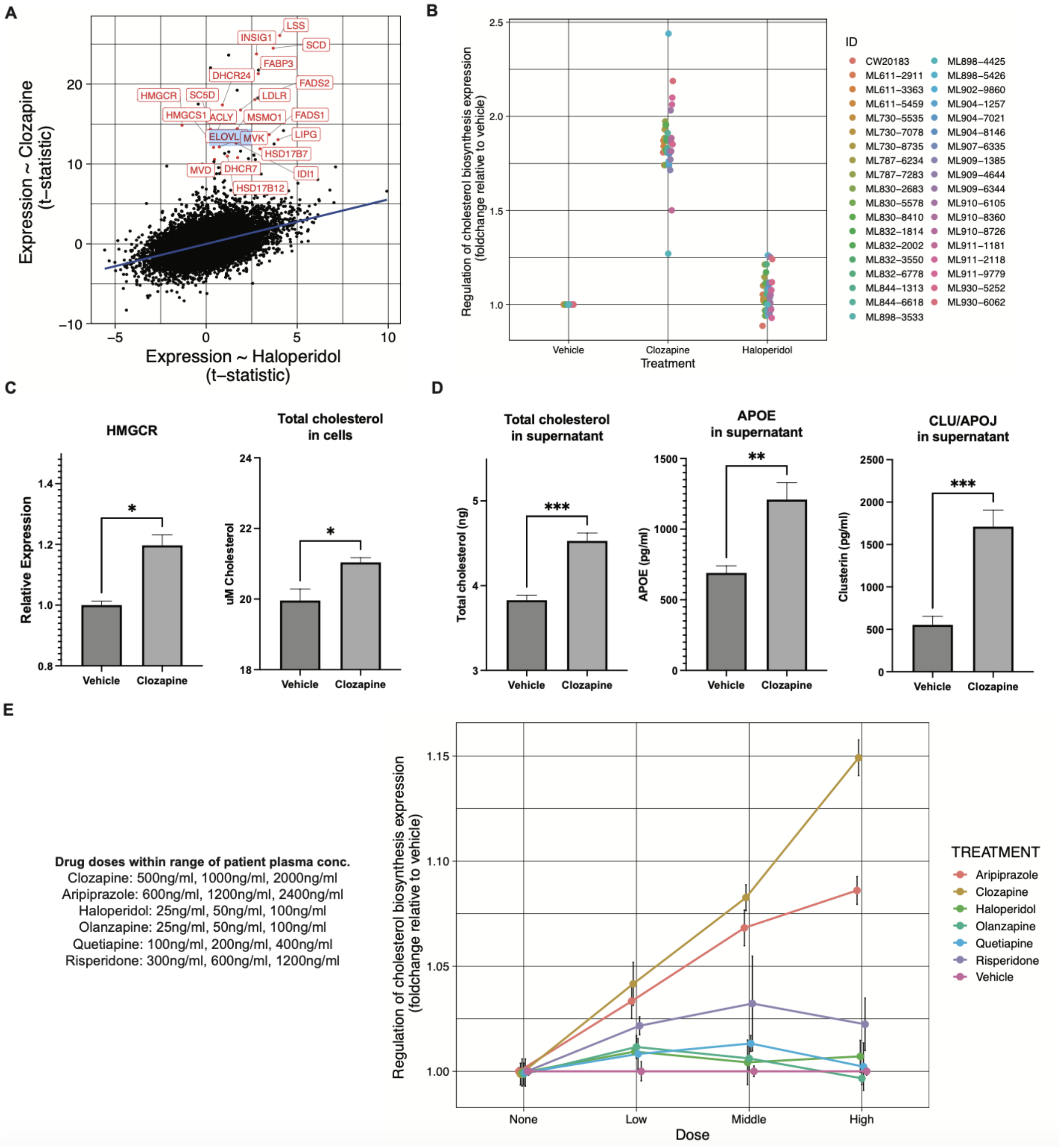
The antipsychotics haloperidol and clozapine showed high degrees of correlation in how they impacted transcription at the gene level. Relative to haloperidol, and all other antipsychotics, clozapine is particularly effective and it's often therapeutic in cases of treatment refractory or resistant schizophrenia (Kane et al, 1988). Clozapine's superior therapeutic efficacy has never been explained by affinity to dopaminergic and serotonergic receptors (which in fact is lower for clozapine) (Richtland et al, 2007) (*in fact, the cell types used in this experiment do not express dopamine receptors, indicating that any effects of clozapine and haloperidol in our system involve D2-independent mechanisms*). While the two compounds showed strong correlations, I was interested in finding differences in their cellular responses in hopes of learning more about the underlying mechanisms of clozapine efficacy. Overall changes in gene expression were similar between clozapine and haloperidol in both neurons and astrocytes, but clozapine strongly regulated genes involved in cholesterol biosynthesis when compared to haloperidol (**Figure 5.2A**).

Leveraging the "cell village" approach, I was able to ascertain that each cell line responded to the treatment in a similar way (**Figure 5.2B**). One of the powerful uses of this approach is to measure quantitative differences in phenotypes across donors. I observed that most donors responded similarly to clozapine treatment (by measuring the median logFC of genes involved in cholesterol biosynthesis) but there were several outlier donors which responded much more or much less. Being able to quantify donors in this way opens the door to mapping genetic influences on these perturbation responses.

Our initial experiments testing clozapine and haloperidol incorporated concentrations chosen based on existing literature (refs). However, these concentrations were quite extreme relative to the therapeutic doses observed in patient plasma (30uM and 10uM, respectively), and I wanted to test whether concentrations that fall within the therapeutic range might elicit similar responses. I thus performed additional experiments to validate this initial response at more therapeutically relevant levels (400ng/ml-1600ng/ml). My results showed that the impact on gene expression was also detected at these lower doses, even as early as 90 minutes after exposure (RT-qPCR for HMGCR) (**Figure 5.2C**). This increase in gene expression of HMGCR coincided with an increase in the abundance of intracellular total cholesterol (**Figure 5.2C**).

As the influx of cholesterol molecules by neurons requires them to be packaged in apolipoproteins, I measured whether treatment with clozapine would increase the amount of cholesterol exported by astrocytes and if these molecules were packaged in their respective protein transporters. In addition to an increase in cholesterol within the cell, I observed that clozapine induced export of cholesterol molecules which are packaged by apolipoprotein APOE and APOJ/CLU (**Figure 5.2D**). Interestingly, common variants in CLU are implicated in schizophrenia genetic studies.

To understand if these observations were specific to clozapine, I tested several other commonly prescribed antipsychotics (aripiprazole, quetiapine, risperidone, and olanzapine along with clozapine and haloperidol). In this experiment, I tested all six compounds across 3 different doses and 3 different durations of exposure (24hr, 48hr, and 72hrs). We replicated our initial findings of the effect of clozapine on the expression of cholesterol biosynthesis genes in a dose- and time-dependent manner and found that most other antipsychotic drugs elicited much-smaller or no responses (**Figure 5.2E**). Among the other antipsychotics tested, only aripiprazole induced cholesterol-biosynthesis gene expression to even half the extent that clozapine did (**Figure 5.2E**). Like clozapine, aripiprazole is used in some cases that have been refractory to first-line antipsychotic drugs (Kane et al, 2007).



Drug doses within range of patient plasma conc.
Clozapine: 500ng/ml, 1000ng/ml, 2000ng/ml
Aripiprazole: 600ng/ml, 1200ng/ml, 2400ng/ml
Haloperidol: 25ng/ml, 50ng/ml, 100ng/ml
Olanzapine: 25ng/ml, 50ng/ml, 100ng/ml
Quetiapine: 100ng/ml, 200ng/ml, 400ng/ml
Risperidone: 300ng/ml, 600ng/ml, 1200ng/ml

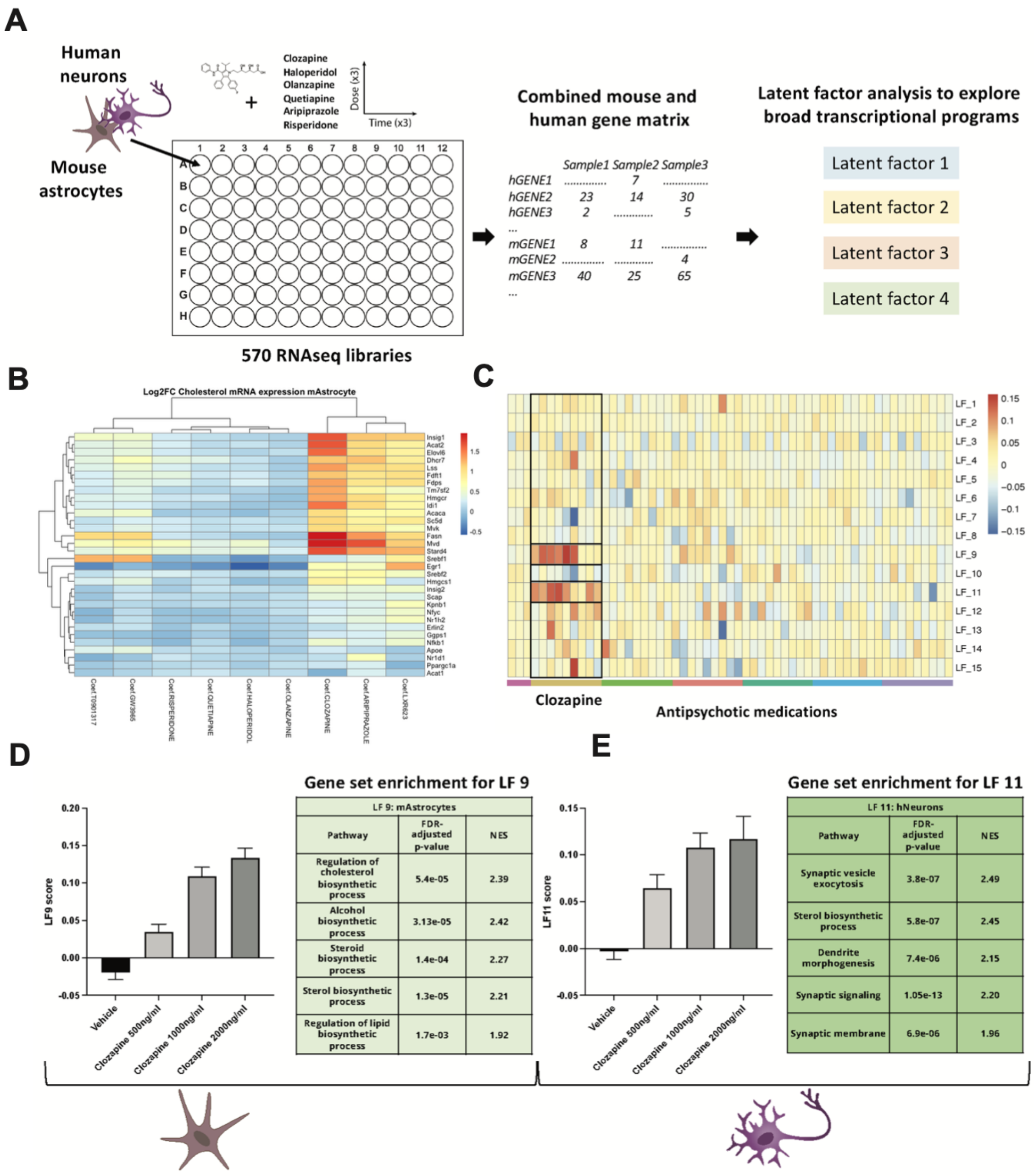
Figure 5.2. Effects of antipsychotics on astrocytes.

(A) Scatterplot of differential expression statistic of clozapine and haloperidol treatment iPSC-derived astrocytes. Genes more strongly induced by clozapine are enriched for genes regulating cholesterol biosynthesis (highlighted in red). (B) Quantitative distribution of donor specific responses to clozapine treatment. Phenotype determined by measuring median logFC of genes contained in Cholesterol Biosynthesis Gene Ontology. (C) (left) RT-qPCR of HMGCR expression following 90 min of exposure to 800ng/ml clozapine. (right) luminescence assay showing increased production of intracellular cholesterol molecules following 90 min exposure to 800ng/ml clozapine. (D) (left) ELISA for total cholesterol present in supernatant of iPSC derived astrocytes following 24hr exposure to 800ng/ml clozapine. (center) ELISA for APOE present in supernatant of iPSC derived astrocytes following 24hr exposure to 800ng/ml clozapine. (right) ELISA for CLU/APOJ present in supernatant of iPSC derived astrocytes following 24hr exposure to 800ng/ml clozapine. (E) Dose response comparison across multiple antipsychotic medications in iPSC derived astrocytes on cholesterol biosynthesis RNA expression.

5.4 - Effects of antipsychotics on astrocyte-neuron interactions

The observation that clozapine exposure regulated the production of cholesterol products in astrocytes invoked the question of whether an increase in the abundance of cholesterol as a result of treatment would impact the relationship between astrocytes and neurons as we have shown a tightly linked program between cholesterol in astrocytes and synaptic components in neurons. To test whether this axis is perturbed as a result of the treatment, I generated co-cultures with human iPSC-derived neurons and mouse astrocytes similar to the approach in Pietilainen et al (2023).

Co-cultures of stem cell derived neurons and mouse astrocytes were subjected to 6 different antipsychotics across 3 doses and 3 timepoints (**Figure 5.3A**). I generated bulk-RNA sequencing data on all conditions. With a multi-species culture, we are able to assign reads based on their species of origin. In doing so, this would allow us to map coordinated effects in the astrocytes (mouse) and neurons (human). The first step was to ensure we could observe a change in cholesterol mRNA in the mouse astrocytes across the various conditions. Consistent with our data in the iPSC-derived astrocytes, we observed that clozapine induced strong increases in the expression of cholesterol biosynthesis genes relative to other antipsychotics medications (**Figure 5.3B**).



	Sample1	Sample2	Sample3
hGENE1	....	7	....
hGENE2	23	14	30
hGENE3	2	....	5
...			
mGENE1	8	11	....
mGENE2	....	....	4
mGENE3	40	25	65
...			

Gene set enrichment for LF 9

LF 9: mAstrocytes		
Pathway	FDR-adjusted p-value	NES
Regulation of cholesterol biosynthetic process	5.4e-05	2.39
Alcohol biosynthetic process	3.13e-05	2.42
Steroid biosynthetic process	1.4e-04	2.27
Sterol biosynthetic process	1.3e-05	2.21
Regulation of lipid biosynthetic process	1.7e-03	1.92

Gene set enrichment for LF 11

LF 11: hNeurons		
Pathway	FDR-adjusted p-value	NES
Synaptic vesicle exocytosis	3.8e-07	2.49
Sterol biosynthetic process	5.8e-07	2.45
Dendrite morphogenesis	7.4e-06	2.15
Synaptic signaling	1.05e-13	2.20
Synaptic membrane	6.9e-06	1.96

Figure 5.3. Effects of antipsychotics on astrocyte-neuron interactions.

(A) Schematic of experimental design. (B) Expression of cholesterol biosynthesis genes in mouse astrocytes from co-culture experiments. (C) Latent factor analysis highlights two latent factors enriched in clozapine treated samples. (D) Latent-factor enriched in mouse astrocytes treated with clozapine are enriched for genes involved in cholesterol biosynthesis. (E) Latent factor enriched human neurons treated with clozapine are enriched for genes involved in synaptic function.

The next step was to investigate whether changes in the mouse astrocytes corresponded to changes in the human neurons. To do so, we performed a latent factor analysis using probabilistic estimation of expression residuals (PEER) from a combined mouse and human gene matrix (Stegle et al, 2012). In our latent factor analysis, we observed that two latent factors showed enrichment in samples which were exposed to clozapine relative to all other treatments (**Figure 5.3C**).

When we looked more deeply into these two latent factors (LF 9, which was enriched in mouse astrocytes and LF11, which was enriched in human neurons) we saw a striking pattern of coordinated responses. In the mouse astrocytes, consistent with our previous findings, we see an enrichment of genes which regulate cholesterol biosynthesis (**Figure 5.3D**). Intriguingly, in neurons from those same samples, we observed an enrichment in genes which encode for synaptic components (**Figure 5.3E**). These latent factors were not enriched in other samples included in our dataset, suggesting that unique among antipsychotics used in this study, clozapine regulates the astrocyte-neuron cholesterol and synaptic expression axis. Ongoing work is exploring whether these transcriptional phenotypes manifest as functional phenotypes using biochemical and physiological approaches.

5.5 - Discussion

In this chapter I explored how common antipsychotics affect astrocyte-neuron co-cultures and showed that clozapine increased the expression of cholesterol biosynthesis mRNA in astrocyte and some synaptic components in neurons, independent of D2 receptor antagonism. I am now moving onto using genomic tools to perturb genes involved in the cholesterol axis between astrocytes and neurons to better understand their roles in synaptic function and disease. Additionally, these findings suggested that antipsychotic medications oppose changes observed within the post-mortem brains in patients with schizophrenia. This work portends that perhaps the therapeutic superiority of clozapine could be due to off-target effects of regulating cholesterol biosynthesis. I hope with this work to better understand how the genes and alleles that are implicated in schizophrenia affect astrocyte-neuron interactions, and in particular the astrocyte-to-neuron cholesterol shuttle activity at synapses.

Work discussed in the previous chapters focused on cell village-based phenotypes such as gene and protein expression. However, gene expression assays (particularly single-cell methods) are still very expensive, which limits their broader adoption. It is critical that new approaches to functional genomics be leveraged to make studies more accessible and to leverage all streams of information which comes from living systems. One such approach would be to explore cell morphology through the use of advances in microscopy. Innovations in image-based profiling provides a promising potential tool by which scientists could unbiasedly measure quantitative phenotypes at a cost significantly lower than that of gene expression (Caicedo et al, 2017, Cimini et al, 2023, Bray et al, 2016). It is not yet clear that these image-based approaches would be applicable for use in stem cell research to link genetic variation to cell morphology.

In the following chapter, I set out to explore whether image-based methods would be amenable to functional genomics across 100s of genetically unique stem cell lines. Should there be links between genetic variants and cell morphology, the stem cell-based community would be able to add a new tool to their arsenal of quantitative phenotyping approaches with a cost often 1/1000 of single cell-based gene expression assays. I show that image-based methods provide a promising approach to linking human genetics to cell morphology across 100s of genetically unique cell lines.

5.6 – Statement on contributions

I performed all experiments in this Chapter. Data analyses was assisted by Jim Nemesh and Noah Pettinari.

Chapter 6

Morphological profiling for functional genomics in health and disease

High-dimensional phenotyping to define the genetic basis of cell morphology
Copyright for thesis incorporating publication
Biorxiv's "Authors retain copyright of their articles and choose from a number of distribution/reuse options."

6.1 – Introduction

Cellular morphology is an important and informative cellular trait in a variety of biological contexts, especially the study of disease. A classic example is sickle cell anemia, which is named for the sickle-like morphology of blood cells observed in patients afflicted with this condition (Gabriel et al, 2010). Like other traits such as gene expression, cellular morphology is mediated by genetic variation. Genetic studies have implicated various loci associated with red blood cell phenotypes such as mean volume and hemoglobin content (Andrews et al, 2009, Astle et al, 2016). However, there is still limited understanding of how human genetic diversity shapes cell morphology. Profiling cell morphology in different cell types and across genetically diverse populations could facilitate the identification of morphology-associated genetic variants.

Induced pluripotent stem cells (iPSCs) provide a powerful tool for capturing genetic diversity in living biological systems and large publicly or commercially available collections provide access to cell lines from donors of diverse ancestry and genetic backgrounds (Yamasaki et al, 2017, Lin et al, 2020, Streeter et al, 2017, Tegtmeyer et al, 2022, Ghosh et al, 2022, Panopoulos et al, 2017). These collections have enabled the study of how human common and rare genetic variation impacts cellular function and behavior, with a focus on gene expression and chromatin accessibility phenotypes (Baxi et al, 2022, Warren et al, 2017, Pashos et al, 2017, Carcamo-Orive et al, 2017, DeBoever et al, 2017, Kilpinen et al, 2017). Studies exploring genetic factors that drive cell morphology have shown promise but are limited by sample size and the resolution by which morphological traits are quantified (Vigilante et al, 2019). Additional efforts with increased sample sizes and greater resolution of cell morphology measurements are critical to expanding discovery power for genetic studies of cellular phenotypes.

Innovations in microscopy and image analysis have enabled the measurement of thousands of morphological traits from a single cell, constructing morphology based 'profiles'. Cell Painting, for example, leverages multiplexed dyes to enable the measurement of traits across many cellular compartments and organelles (Cimini et al, 2023, Bray et al, 2016). Cell Painting can ascertain gene function by linking expression to cellular traits and has been used to enable the prediction of functional impacts from lung cancer variants (Rohban et al, 2017, Caicedo et al, 2022). Cell morphology profiling provides a great asset for functional genomics studies compared to methods such as gene expression, being much more affordable and easily scalable at the bulk and single cell level. We hypothesized this approach could be leveraged in combination with iPSC technology to elucidate relationships more broadly between cell morphology and genetic variants.

Here, we identified the morphological impacts of genomic variants, or cell morphological quantitative trait loci (cmQTLs), by generating high-throughput morphological profiling and whole genome sequencing data on iPSCs from 297 unique donors. Leveraging Cell Painting data on >5 million iPSCs derived from these donors, we quantified 3,418 cell morphological traits and assessed their associations with rare and common genetic variants genome-wide. We identified trait-associations with rare-variant burden in several genes including WASF2, PRLR, and TSPAN15 which we then functionally validate using CRISPR interference. Additionally, we nominated one common variant convincingly associated with morphology and found suggestive evidence for over 300 loci. Finally, we leveraged these results to make predictions about sample size requirements for increasing discovery power for both common and rare variants in future cellular genetic studies. These findings show that similar to gene expression, the morphology of cells is mediated by genetic variation and highlights the utility of image-based methods for functional genomics.

6.2 - Morphological profiling and whole-genome sequencing on iPSCs from 297 unique donors

To study associations between genetic variants and morphological traits, we assembled a cohort of iPSC lines from 297 unique donors for which we generated image-based profiling and whole-genome sequencing data (**Figure 6.1**). We obtained pre-derived cell lines from the CIRM iPSC repository (Lin et al, 2020). Age, sex, medical history, ethnicity, and relatedness to other samples were recorded using questionnaires at time of enrolment and sample collection. Each iPSC line was subjected to a pluripotency test as well as genotyping to identify any abnormal karyotypes. Upon receipt, we expanded and cryo-banked each iPSC line, and performed genotyping (using the Global Screening Array (GSA)) and 30X whole-genome sequencing (WGS) on all lines. Any cell lines displaying abnormal karyotypes or genomic rearrangements > 1Mb were excluded from our study. The final cohort used in this work included 297 distinct donors of which 153 were male and 144 were female, with an average age of 21+10 (sd) years. Of the 297 donors, 207 had self-reported ancestry of European and 90 individuals reported non-European (**Table 6.1**). IPSC lines were generated from B-cells or fibroblasts using a non-integrating episomal vector system previously described (Lin et al, 2020) (**Table 6.1**). All donors included in this study have been properly consented for iPSC derivation, the experiments performed in this work, and genomic data sharing. We performed a principal component analysis (PCA) to observe the genetic diversity of cells utilized in our collection (**Figure 6.2A**). A summary breakdown of our cohort is included in Table 1 and individual cell line level metadata is included **Table S1**.

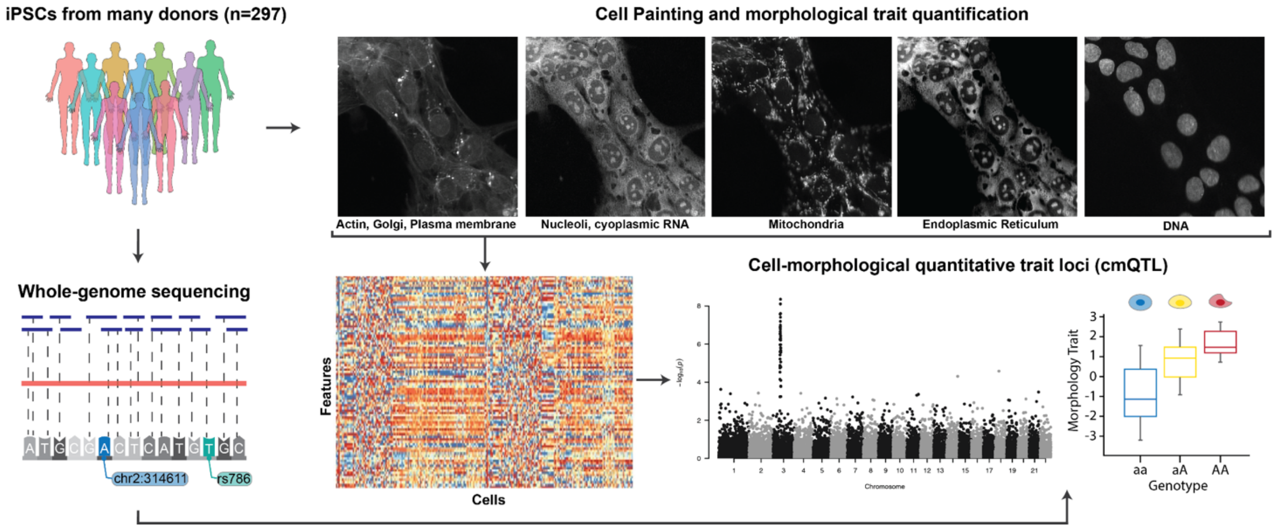


Figure 6.1. Study Overview.

iPSC lines from 297 donors were expanded, quality-control checked and then subject to both high-throughput imaging with Cell Painting and 30X whole-genome sequencing (WGS). Overall, we imaged 5.1x10^6 cells across all donors and quantified 3,418 morphological traits per cell using CellProfiler software. We inferred genetic variants from the WGS data and investigated whether individual morphological traits associated with both rare and common variation.

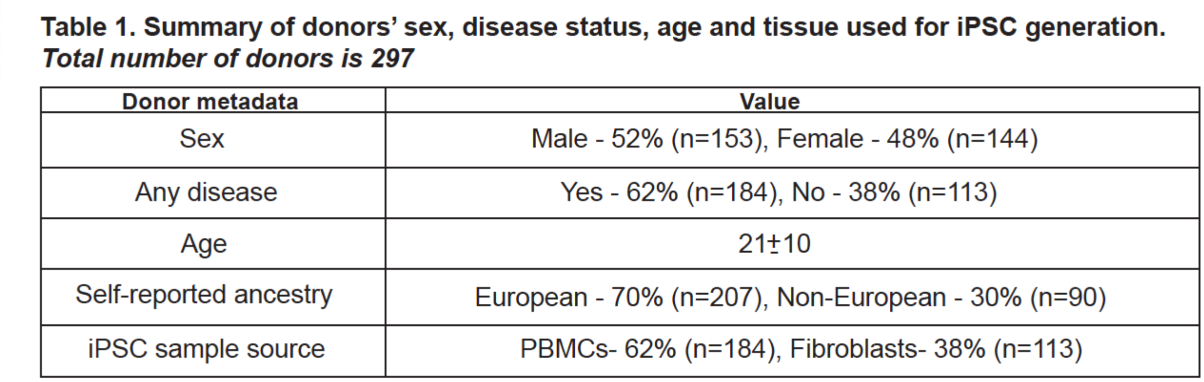


Table 1. Summary of donors' sex, disease status, age and tissue used for iPSC generation.
Total number of donors is 297

Donor metadata	Value
Sex	Male - 52% (n=153), Female - 48% (n=144)
Any disease	Yes - 62% (n=184), No - 38% (n=113)
Age	21±10
Self-reported ancestry	European - 70% (n=207), Non-European - 30% (n=90)
iPSC sample source	PBMCs- 62% (n=184), Fibroblasts- 38% (n=113)

To quantify cellular traits, we adopted the Cell Painting assay (Cimini et al, 2023, Bray et al, 2016). This multiplexing dye assay uses six stains to capture morphological characteristics for eight cellular compartments: Hoechst 33342 (DNA), wheat germ agglutinin (WGA) (golgi and plasma membrane), concanavalin A (endoplasmic reticulum), MitoTracker (mitochondria), SYTO 14 (nucleoli and cytoplasmic RNA), and phalloidin (actin). Images are processed using the open-source CellProfiler software to extract thousands of features of each cell's morphology such as shape, intensity, and texture statistics, thus forming a high-dimensional profile for each single cell (McQuin et al, 2018).

We generated Cell Painting data from all 297 donors leveraging a systematic workflow to ensure cells were treated in identical fashion across all rounds of imaging. Cell lines were thawed in batches of 48 and passaged 3 days later into a 96-well deep well plate before being transferred into a 384-well screening plate using an automated liquid handling device (**Figure 6.2B, Methods**). Cells were plated at a density of 10k cells/per well and fixed 6 hrs post-plating, so as to allow for cell attachment while minimizing differences in cell growth rates, which we observed during cell line expansion (**Figure 6.2C**). We determined these conditions through a pilot screen that contained 6 cell lines plated across various densities and fixation timepoints, which showed we could maximize differences between cell lines under these parameters (**Figure 6.2D**). Each screening plate was stained with the standard Cell Painting dyes and imaged on a Perkin Elmer Phenix automated microscope within 48hrs. We implemented this same workflow across all rounds of imaging.

Images were processed using CellProfiler to measure morphological traits and construct single-cell image-based profiles (**Methods**, McQuin et al, 2018). In total, we measured 3418 morphology traits for 5.1 million iPSCs from 297 donors after stringent QC (**Methods, Figure 6.3A, Table S2**). We classified all morphological traits based on the cellular characteristics they represented, yielding five categories: Area and shape, Granularity, Intensity, Radial distribution, and Texture (**Figure 6.4A**). Prior studies have shown that cells often displayed varied morphology in response to environmental cues and context (Vigilante et al, 2019, Schrenk-Siemens et al, 2020). To explore whether the contribution of genetic variation to cell morphology is context dependent, we segregated all cells into two groups based on whether they had any cells in contact (called colony cells, 97.48% of all cells) or not (called isolate cells, 2.52% of all cells) (**Figure 6.3B**). We note that for the purposes of our study, "colony" refers to the number of neighbors a given cell has and is distinct from the colony terminology which is often used in basic stem cell culture practices.

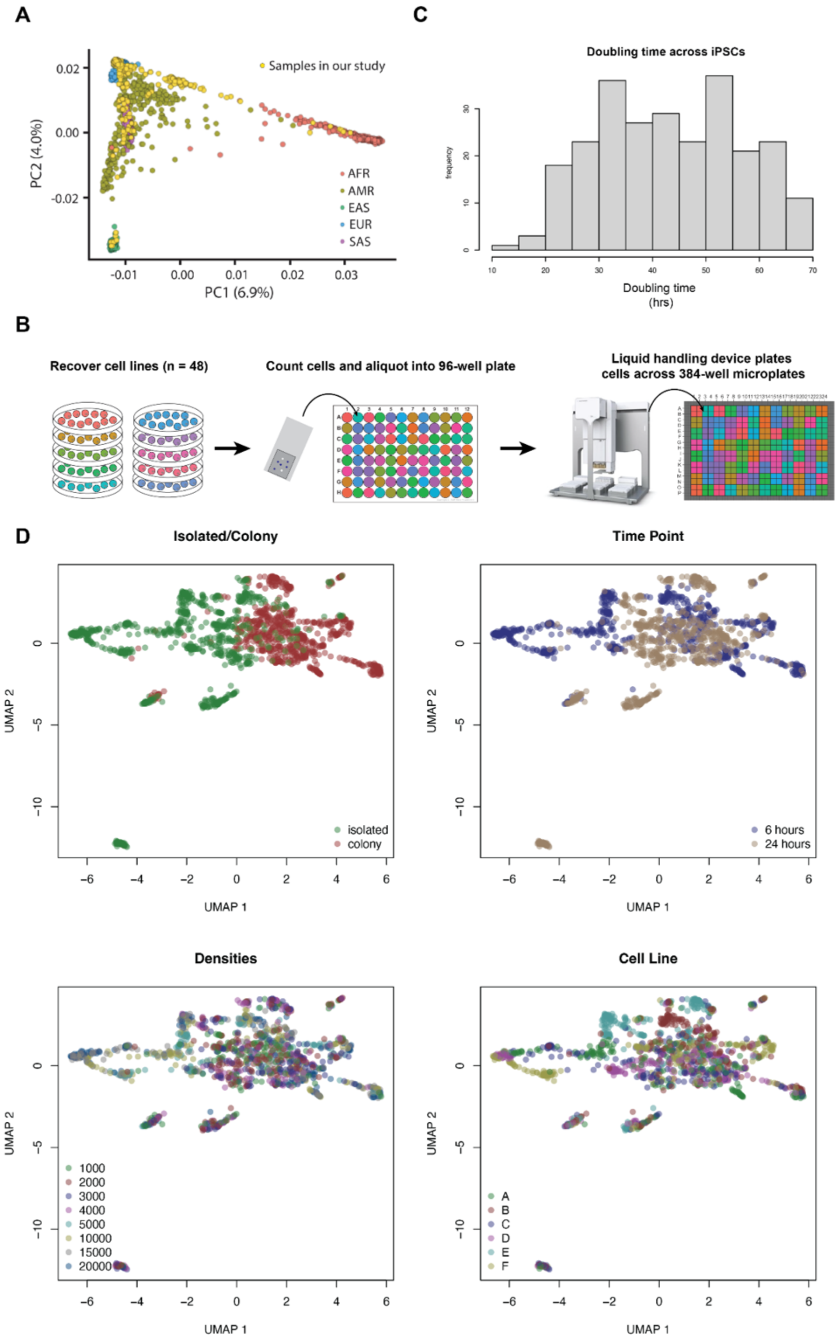
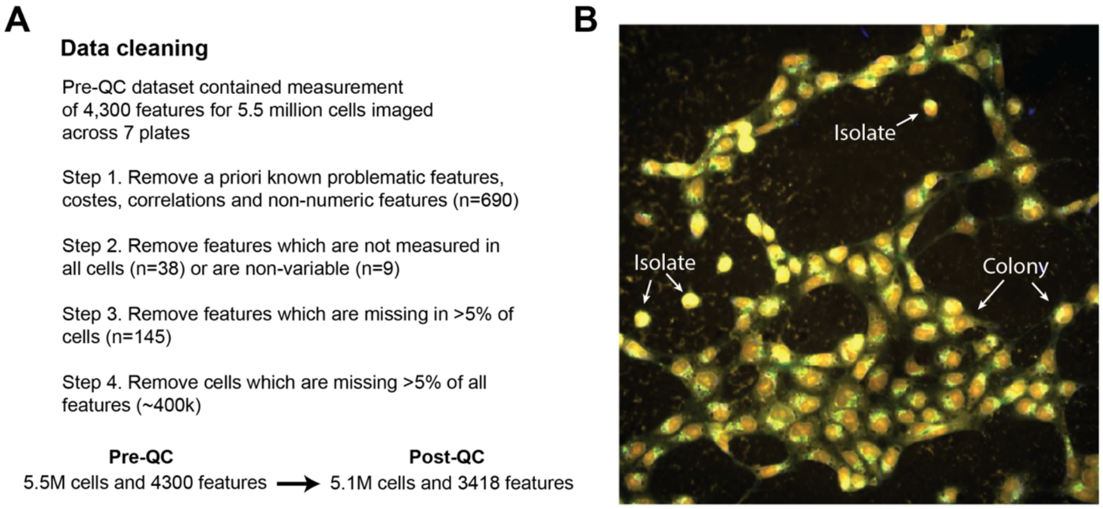


Figure 6.2. Cell line collection and pilot study.

(A) Distribution of 297 donors (yellow dots) laid over individuals from 1K genomes on PC1 and PC2 calculated from common variants (maf > 5%). Of 297 donors, 207 self-reported their ancestry as European. (B) Cells are thawed in batches of 48 and grown until they reach ~70% confluency. Then cells are dissociated and counted before being aliquoted into deep 96-well plates. A liquid handling device is then used to transfer the cell suspensions from the deep well plates into 384 well high-content screening plates. (C) Doubling time (in hrs) for all 297 cell lines used in this study. Cell growth was calculated during their standard expansion process in house after acquiring them from CIRM. (D) We tested several pilot conditions to identify the ideal parameters for our discovery cohort. We profiled cell lines derived from 6 donors and Cell Painted them under various densities (1000, 2000, 3000, 4000, 5000, 10000, 15000, and 20000 cells/well) and time points for fixation post-plating (6hr and 24hr). We leveraged this data to determine which conditions maximize our ability to measure cell line separation as well as reliably identify cells in both colony and isolate.



Data cleaning

Pre-QC dataset contained measurement of 4,300 features for 5.5 million cells imaged across 7 plates

Step 1. Remove a priori known problematic features, costes, correlations and non-numeric features (n=690)

Step 2. Remove features which are not measured in all cells (n=38) or are non-variable (n=9)

Step 3. Remove features which are missing in >5% of cells (n=145)

Step 4. Remove cells which are missing >5% of all features (~400k)

Pre-QC 5.5M cells and 4300 features → **Post-QC** 5.1M cells and 3418 features

Figure 6.3. Data QC.

(**A**) A total of 4318 cell morphology traits were quantified across all 5.5 million iPSCs cells from 297 donors. Morphology traits a priori known to be problematic, not measured across all cells or non-variable across cells were removed. Also, cells missing measurement for >5% of traits were removed, yielding 3418 traits across 5.1 million cells. (**B**) This image is to highlight the variability in cell contexts which we categorize in our study. Cells which do not come in contact with other cells are classified as isolate cells. Those which touch 1 or more cells are classified as colony.

We next performed 30X whole-genome sequencing (WGS) on all iPSC lines. Following quality control (QC, see Methods), we retained 7,020,633 common (minor allele frequency (MAF) > 5%) and 122,256 rare (MAF < 1%) variants for downstream analyses.

6.3 - Cell line characteristics and technical factors drive variability in morphological traits

Previous studies have shown that technical factors, including plate and well position can alter morphology-based readouts (Schiff et al, 2022). To explore the presence of cmQTLs in our data, we sought to identify technical factors which may confound our morphological phenotypes and remove these sources of variance from our downstream association tests. We performed a variance component analysis using well-level data to quantify the observed variance that can be attributed to each morphological trait by technical factors and cell line characteristics (**Methods**). We assessed the significance for each variance component, correcting for the number of tests, which was the product of the traits (n=3418) and factors (n=9) which include technical features such as imaging plate, well position, the number of cell neighbors, and whether the well was positioned on the edge of the plate (onEdge) in addition to demographic characteristics for the cell lines including genetic sex, reprogramming sample source, age of donor at time of sample collection, and the clinical diagnosis for our tissue donors. We observed strong batch effects across imaging plates, which contributed the greatest degree of variance to our morphology traits (61.8+17%, **Figure 6.4B, Figure 6.5A**). Several other confounders contributed varying levels of effect on different morphological traits (**Figure 6.4C**). After correcting for these covariates, the remaining difference among cell line donors was significantly associated with all traits, explaining 16.7+11% of the variance. (**Figure 6.4B**). This indicated the potential for a genetic basis to the variability in morphology traits. Residual is the remaining (technical) variance in morphological traits which is unexplained by the factors discussed above. Interestingly, the difference among donors explained a greater degree of variance in the trait category of AreaShape relative to the other trait categories (Wilcoxon rank sum test P = 1.1x10-55, **Figure 6.5B**). We note that some of the shared variance may be explained by non-genetic factors, such as stable epigenetic modifications.

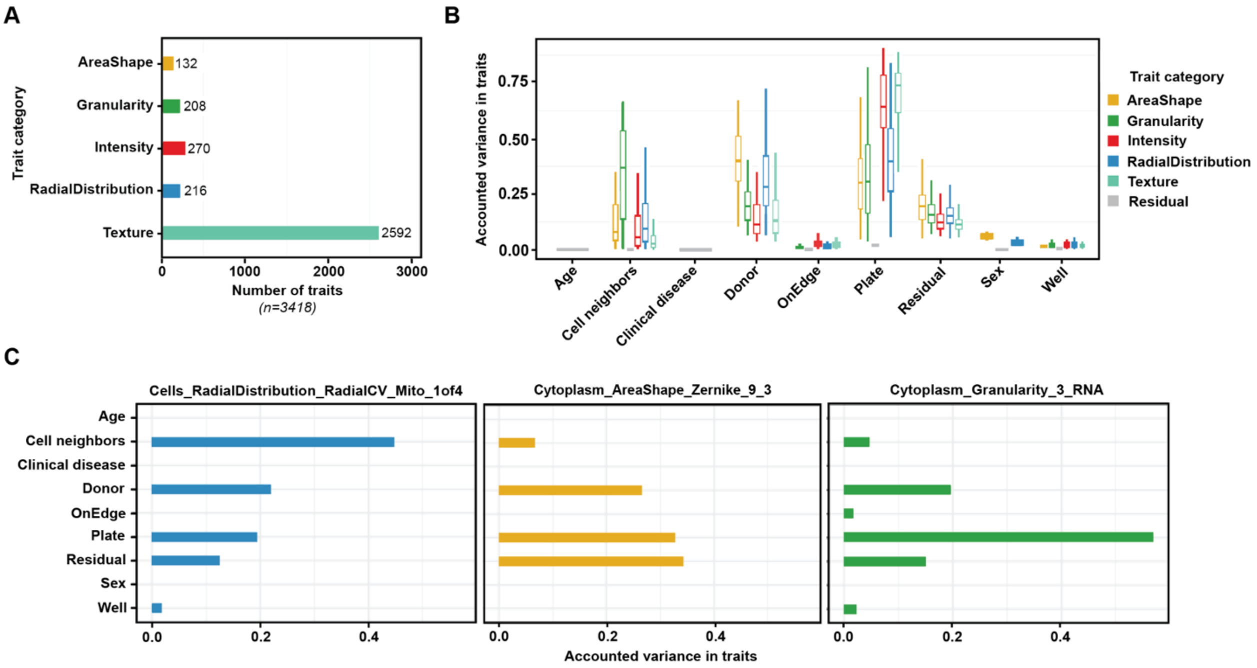


Figure 6.4. Summary of morphological traits and variant component analysis.

(**A**) Summary of five categories of morphological traits captured in our data (n=3418). (**B**) Explained variance across all morphological traits (n=3418). (**C**) Exploring explained Variation in individual traits, namely distribution of mitochondria around nucleus, cytoplasmic Zernike shape metric 9_3, and cytoplasmic granularity in the RNA channel at scale 3, showed differences in sources of variance, including technical effects such as plate and well of imaging, whether the well was situated on the row or column on the edge of plate (onEdge), biological sources such as donor. Donor ID represents the difference among donors after accounting for their age, sex, disease-status, and above-mentioned imaging-related technical factors. Residual is the remaining unaccounted variation in traits.

We observed that many traits had very high pairwise correlation (Pearson r > 0.9) with one or more traits (**Figure 6.5C**). To reduce redundancy in our downstream analyses, we selected a common set of 246 traits having r < 0.9 with each other by iteratively selecting a single representative trait for the set of correlated traits (r > 0.9) (**Methods**). We refer to this common set of 246 traits as "composite traits", which were used for our rare and common variant association tests (**Table S3**). We next summarized well-level morphology data into donor-level values (i.e., pseudo-bulk) by mean-averaging individual morphology traits across all wells for a given donor, resulting in one measurement per trait per donor (N = 246 traits and 297 donors) for both isolate and colony cells. These donor-level trait values were used for our quantitative association tests.

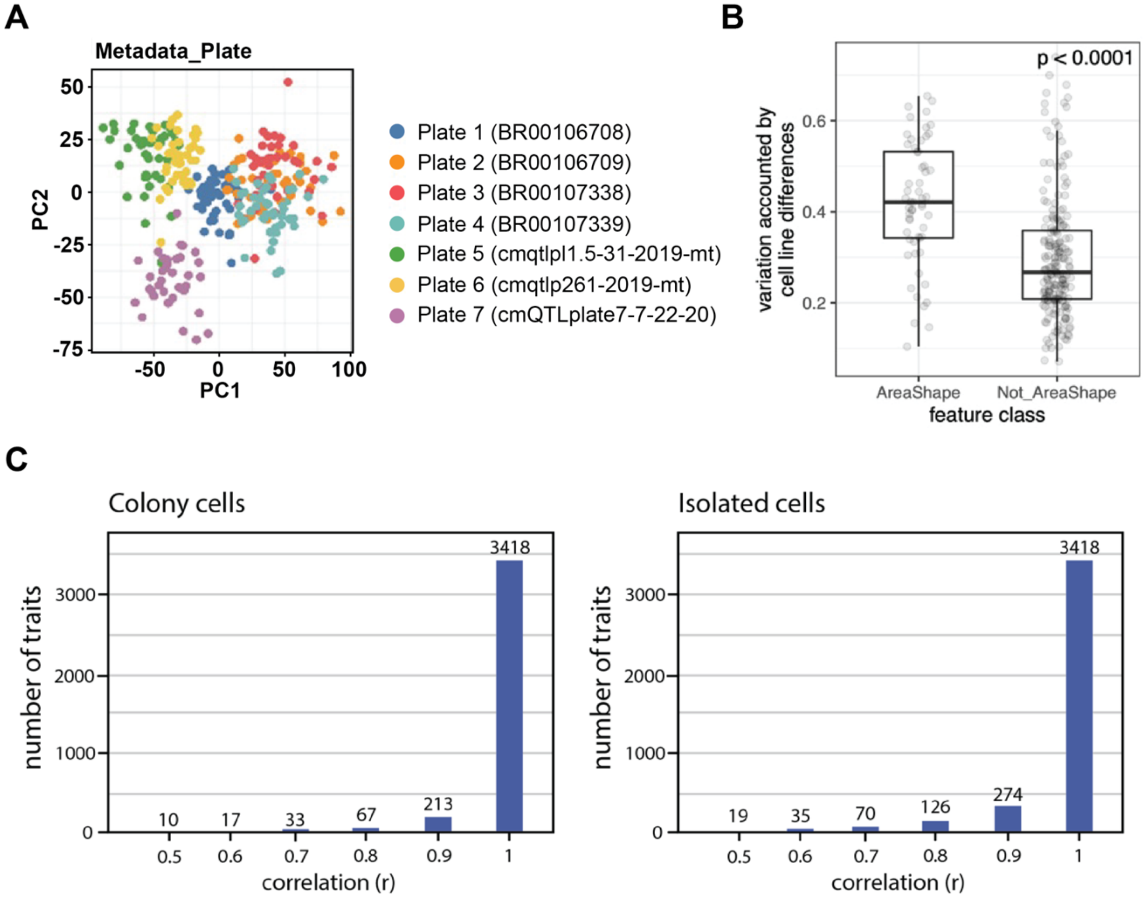


Figure 6.5. Variance component analysis.

(**A**) Distribution of 297 donors on PC1 and PC2 calculated from morphology traits (n=3418) colored by 7 plates on which iPSCs from donors were imaged, showing the batch (plate) effect in the measurement of morphology traits. (**B**) The comparison of variation explained by genetic difference among donors in traits belonging to Area and Shape category and other categories. P-value from Wilcoxon rank sum test is shown. (**C**) The number of traits having correlation (Pearson r) of up to 0.5, 0.6, 0.7, 0.8, 0.9 and 1 (on x-axis) with at-least one other trait is shown for cells in colonies and cells which are isolated.

6.4 - Rare variant burden for cell morphological traits

Sequencing studies have identified hundreds of genes containing rare coding variants with association to disease burden (Sun et al, 2022, Wang et al, 2021, Backman et al, 2021, Karczewski et al, 2022). These variants often have large effect sizes but explain a modest degree of total disease heritability (Weiner, Nadig et al, 2023). To explore the effect of rare genetic variation on cellular morphology, we analyzed the association of composite traits (n = 246) with gene-level burden of protein-altering rare variants (MAF < 0.01). To ensure well-powered investigation, we examined 9105 genes that had rare variants in at least 2% of donors (n >= 6). We modeled individual morphology traits as a function of rare protein-altering variant burden in a gene, controlling for plate, well, and donor sex using linear regression (**Methods**). We performed our analysis separately for both colony and isolated cells. We identified 4 genome-wide significant associations between morphological traits and rare variant burden (P < 2.2x10-8, Bonferroni correction for 246 traits and 9105 genes) (**Figure 6.6A**). These associations included one trait in colony cells and three traits in isolate cells. We did not observe any inflation in association statistics for these traits (Lambda (**λ**) = 1.01 for the association in colony cells and **λ** = 1.01, 0.96, 0.98 for the associations in isolate cells) (**Figure 6.7A**). While the top feature associations (using a stringent Bonferroni correction) are quite different between isolate and colony cells, there is a modest overall correlation between the associations of morphology traits and genetic variants (**Figure 6.7B**).

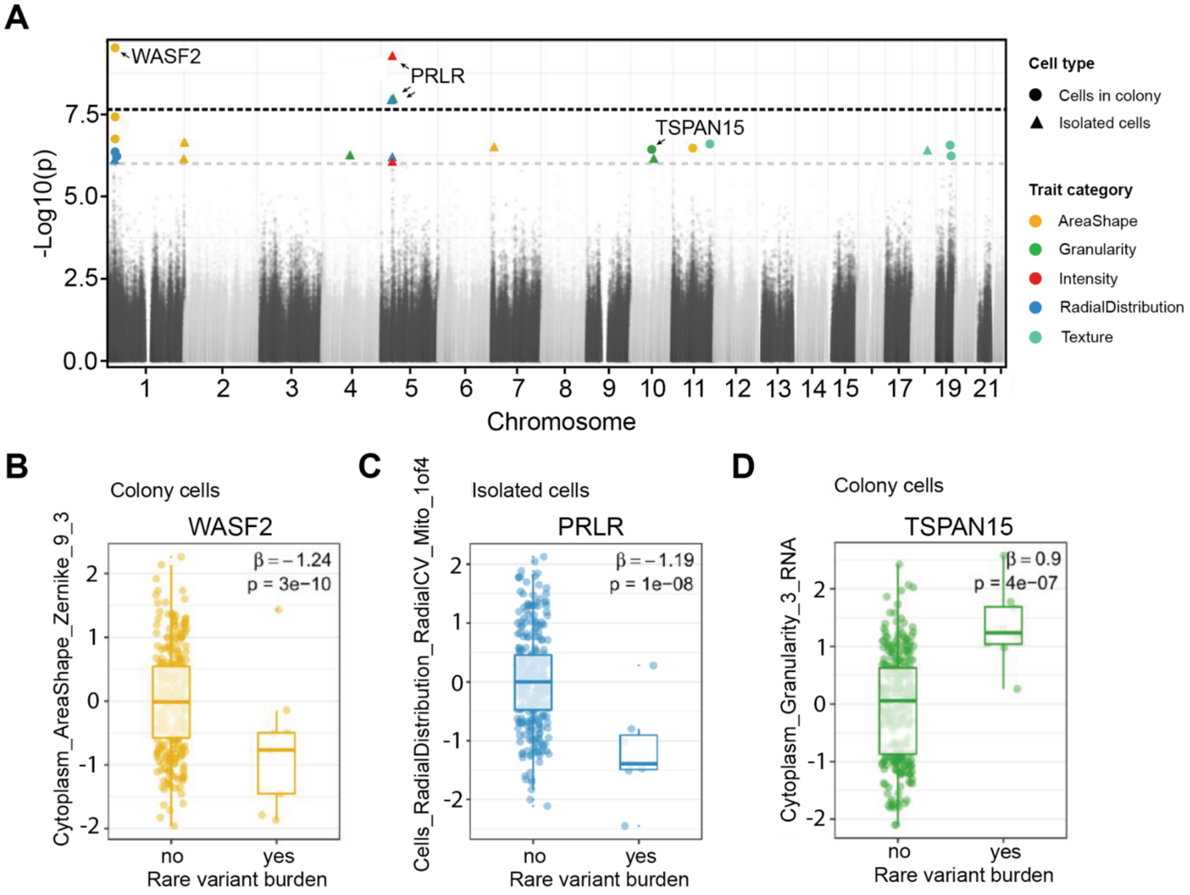


Figure 6.6. Association between morphology and rare variant burden.

(**A**) Manhattan plot showing association between morphological traits (n=246) and rare variant burden in candidate genes (n=9105). Black dotted line represents the p-value threshold after Bonferroni correction for the number of tested traits and genes (P = 0.05/246x9105, i.e., 2.2x10^-8). Grey dotted line represents the p-value threshold for suggestive evidence of association (P = 10^-6). (**B-D**) Box plots displaying the association between the Zernike_9_3 cytoplasm shape metric and rare variant burden in WASF2 gene (B), distribution of mitochondria around the nucleus and rare variant burden in PRLR gene (C) and cytoplasmic granularity measure in the RNA channel and rare variant burden in TSPAN15 gene. We provide the effect size (**β** estimate) and raw p-value of the association for each trait.

Rare variant burden in WASF2 was negatively associated with a Zernike shape measure of the cytoplasm (Cytoplasm_AreaShape_Zernike_9_3) in colony cells (n = 3 missense and 1 in-frame deletion rare variants, **β** or effect size (se) = -1.24 (0.18), P = 3.1x10-10; **Figure 6.6B**). Zernike features represent polynomial reconstructions of an organelle or object of cells. WASF2 is named for its association to Wiskott-Aldrich syndrome, a rare genetic disorder which greatly increases the risk of various cancers (Ding et al, 2023, Yang et al, 2022, Rana et al, 2021, and

Rana et al, 2023). WASF2 protein binds profilin, a G-actin-binding protein, promoting the exchange of ADP/ATP on actin and the formation of actin filament clusters (Insall et al, 2009, Takenawa et al, 2001). The disruption of WASF2 impairs actin formation and organization that could lead to their polarized distribution and spindle-shaped cells (Kiger et al, 2003). In representative images of cells with rare variants in WASF2 it is difficult to identify this polarized and spindle-like shape by eye when compared to reference lines (**Figure 6.7C**). In addition to our genome-wide association, rare variant burden in WASF2 had nominal association ($P < 0.05$) with 90 other traits including 27 traits of area and shape category, suggesting WASF2 may contribute to a range of cell morphological characteristics (**Table S4**).

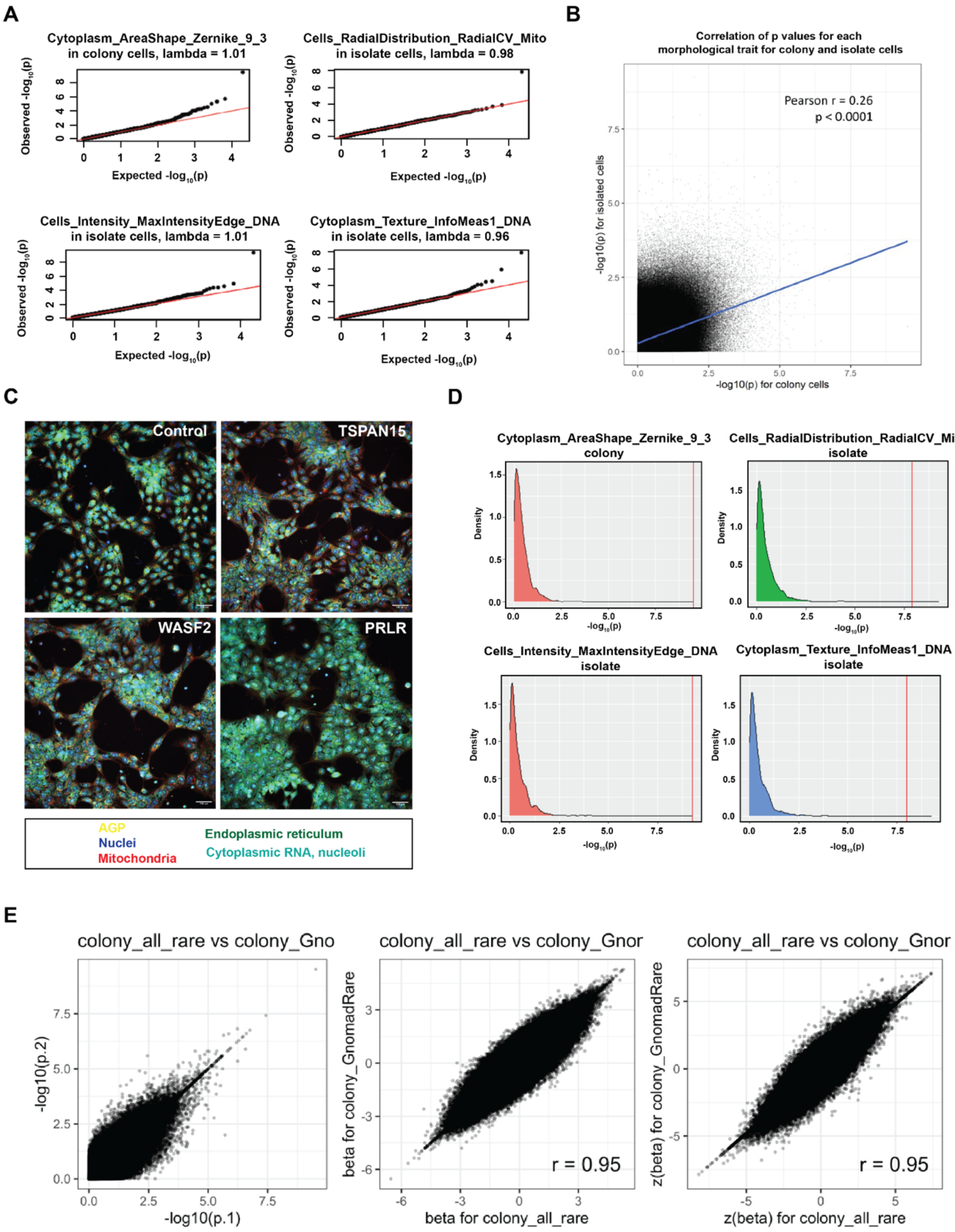


Figure 6.7. Analyses of rare variant associations.

(**A**) QQ plots show the distribution of expected and observed p-value of association with all tested genes for 4 morphology traits. Each dot is a tested gene. Lambda statistic (**λ**), a measure of inflation in observed p-values, is shown. (**B**) This figure shows the overall correlation between the p vals for each morphological trait and its association with a gene between cells in isolate and cells in colony. This shows there is modest overall correlation between both cell contexts. (**C**) Randomly selected representative images from wells containing cell lines harboring rare variants in WASF2, PRLR, and TSPAN15 compared to reference cell lines with no detected variants. (**D**) The distribution of p values following permutations of our rare variant associations. The distribution of p values observed suggests that our associations are unlikely to have occurred by chance. (**E**) Comparison of p-value and z-score of effect size (beta) of associations between individual morphology traits and rare variant burden in individual genes using all rare variants in our dataset and those rare variants (out of all) which are also present in gnomAD dataset is shown for colony cells (A) and isolated cells (B). Pearson r is shown.

Three traits were significantly associated with rare variant burden in the PRLR gene (n = 6 missense rare variants, **β** (se) = -1.17 (0.2), P = 1.2x10-8; **Figure 6.6C**). The most interesting among these included asymmetries in the distribution of mitochondria in the perinuclear space (Cells_RadialDistribution_RadialCV_Mito_1of4). PRLR function has been linked to several forms of cancers, including breast cancer and lymphoma (Kavarthapu et al, 2022, López Fontana et al, 2021, Gharbaran et al, 2021). PRLR encodes membrane-anchored receptors for a prolactin ligand and is a part of the class-I cytokine receptor superfamily and regulator of JAK-STAT5 pathway activity, regulating autocrine/paracrine loops present in stem cells, which mediate their quiescence and proliferation (Sackmann-Sala et al, 2015, Bole-Feysot et al, 1998). Previous findings in adipocytes showed PRLR knockout alters mitochondrial packing and distribution throughout the cell (Viengchareun et al, 2008). Moreover, rare variant burden in PRLR had nominal association (P < 0.05) with 118 other traits, providing more support to PRLR as a genetic determinant of cellular morphology (**Table S5**).

We also inspected the associations with suggestive evidence, i.e., P < 10-6. There was a total of 12 and 13 associations in colony and isolated cells, respectively, which passed this threshold (**Table S6**). One of the strongest associations in our suggestive results was between the distribution in size of RNA particles in the cytoplasm (Cytoplasm_Granularity_3_RNA) and rare variant burden in TSPAN15 gene (n=2 missense and 1 splice region rare variants in the gene, **β** (se) = 0.9 (0.17), P = 3.7x10-7; **Figure 6.6D**). TSPAN15 is expressed in all human tissues and encodes for a cell surface protein (GTEx, 2017). A member of the tetraspanin family of transmembrane segments, TSPAN15 has been implicated in tumor related conditions (Huang et al, 2022). TSPAN15 plays a role in cell activation and self-renewal through negative regulation of

Notch-signaling (Jouannet et al, 2016). Disruption of TSPAN15 could lead to increased cell proliferation and transcriptional activation which may be represented in our data by an increase in the measurable RNA content in the cytoplasm.

To ensure our observed associations were not an artifact of our nonparametric regression model, we permuted the data by randomly assigning trait values across donors. These results suggested that our observed significance of association between rare variant burden and cell morphological traits was unlikely to have occurred by chance (**Figure 6.7D**). To confirm that our observed associations were not driven by somatic variation introduced during iPSC reprogramming or those which arise in cell culture (despite the short culture time in our study), we repeated our association test while restricting to only those variants that were previously observed in the gnomAD database (106,590 of 122,256 variants) (Karczewski et al, 2020). All of our observed associations were recapitulated (significant after Bonferroni correction for multiple testing and with suggestive evidence) with concordant effect size and statistical significance (p-value) (**Figure 6.7E**). Taken together, our findings suggest we could reliably identify associations between morphological traits and rare protein coding variants.

6.5 - Functional validation of rare variant associations

CRISPR-based gene editing has been shown to be a viable mechanism for validating gene expression phenotypes resulting from rare variation (Li et al, 2017). To corroborate our rare-variant burden associations, we examined whether knockdown of these genes impacted the same morphological traits for which we identified a rare-variant burden association. We transfected iPSCs from a single cell line expressing constitutive dCas9-KRAB CRISPRi machinery with sgRNAs targeting the transcriptional start site (TSS) for WASF2, PRLR, and TSPAN15 (**Figure 6.8A**). Each gene was targeted by 2 different sgRNA sequences, which were validated for knockdown of their target gene (25-95% efficiency) (**Figure 6.9, Table S7**). Cells transfected with non-targeting sgRNAs were included as controls. We generated per-well (population-averaged) morphological profiling using the same methods for our discovery cohort.

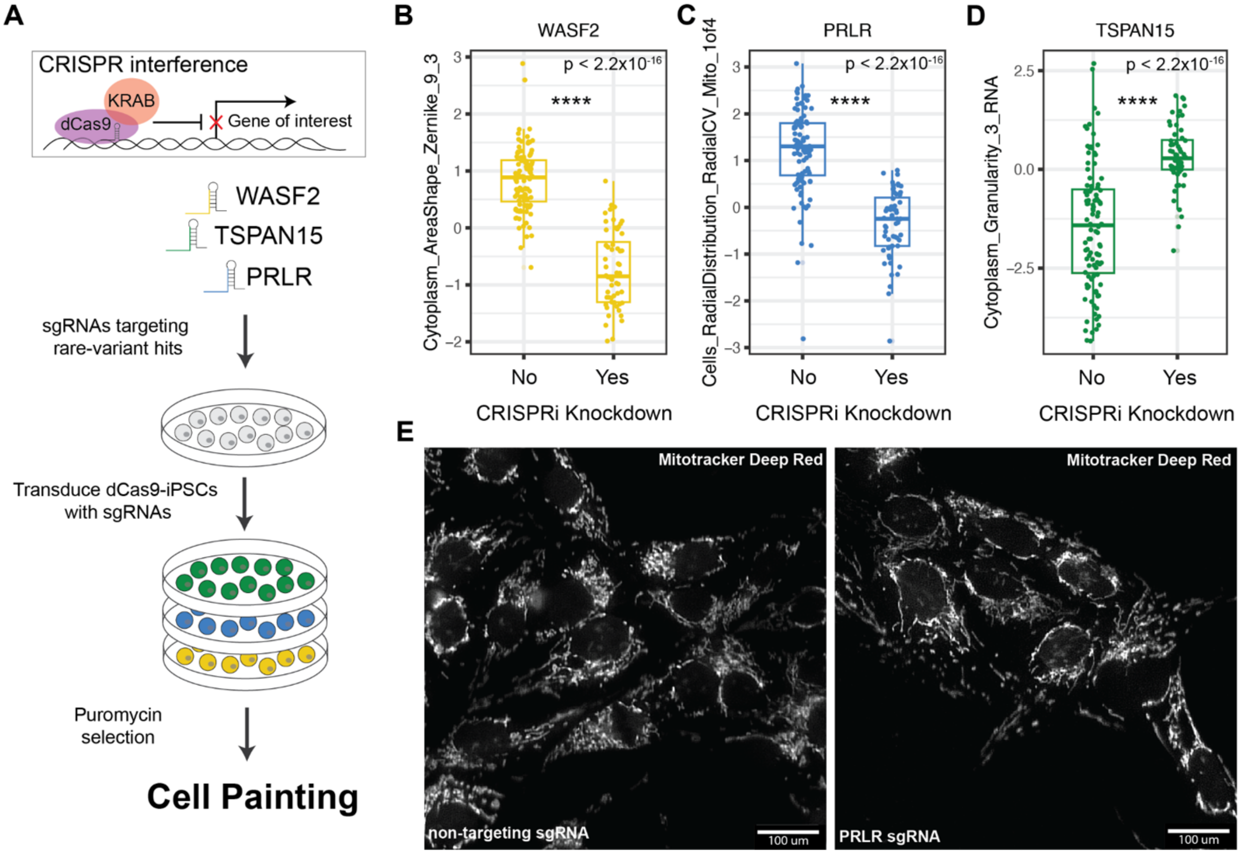


Figure 6.8. Functional validation of rare-variant burden associations.

(**A**) Workflow for knockdown of rare-variant genes using CRISPR interference in iPSCs expressing constitutive dCas9-KRAB. (**B-D**) Box plots displaying quantification of traits between control non-targeting sgRNAs and sgRNAs targeting WASF2, TSPAN15, and PRLR on a per-well level (n = 52 wells per non-targeting sgRNAs, n = 56 wells per targeting sgRNAs, $P < 2.2x10^{-16}$, Welch's Two-Sample T-Test). Effect on the trait score is consistent with what we observed in our rare-variant burden association. Data is presented in a Tukey-style boxplot with the median (Q2) and the first and the second quartiles (Q2, Q3) and error bars defined by the last data point within +/− 1.5-times the interquartile range. (E) Representative image of an observable gene-trait association for PRLR. Cells_RadialDistribution_RadialCV_Mito_1of4 relates to the asymmetric distribution of mitochondria in the ring right around the nucleus. In the non-targeting controls (left) we observed clustering of mitochondria on a particular side of the nucleus, whereas in the PRLR knockdown sgRNA (right) we observed a more distributed presence of mitochondria around the nucleus.

We compared the morphological trait values for our rare-variant associations between non-targeting sgRNAs and those targeting our genes of interest. For each gene tested, we observed the predicted changes, and in the same direction, for each individual trait relative to controls (n = 28 wells per targeting sgRNA and 52 wells per non-targeting sgRNA, Welch's Two Sample T-Test, P < 2.2x10-16) (**Figure 6.8B-D**). Specifically, knockdown of WASF2 resulted in a decreased normalized score for the trait Cytoplasm_AreaShape_Zernicke_9_3 (**Figure 6.8B**). We observed that a reduction in the expression of TSPAN15 coincided with an increase in trait score for Cytoplasm_Granularity_3_RNA (**Figure 6.8D**). Lastly, knockdown of PRLR decreased Cells_RadialDistribution_RadialCV_Mito_1of4, which defines the relationship between the radial distribution of mitochondria around the nucleus (**Figure 6.8C**). This effect is highlighted in representative images, whereby cells transfected with a PRLR targeting sgRNA displayed more uniform distribution of mitochondria around the nucleus when compared to non-targeting sgRNA cells where mitochondria tend to colocalize to one side of the nucleus (**Figure 6.8E**).

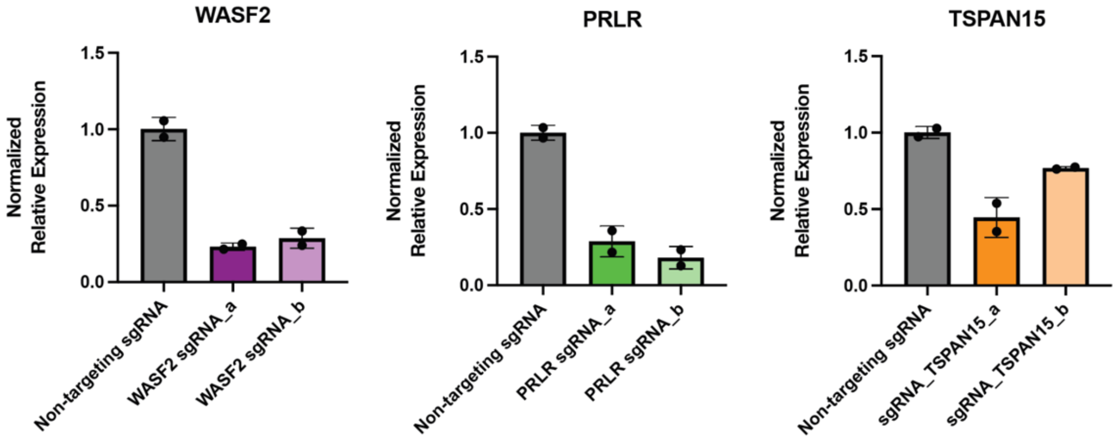


Figure 6.9. qPCR knockdown of rare-variant associations using CRISPR interference.

Relative expression of sgRNA target genes compared to GAPDH and RPL10 between iPSCs transfected with gene targeting sgRNAs and non-targeting control sgRNAs.

6.6 - Common variant associations for cell morphological traits

Genome-wide association studies (GWAS) have identified thousands of common variants that are associated with common diseases and traits. These variants have small effect sizes at the individual level but combine to explain a large degree of common disease heritability (Trubetskoy et al, 2022, Yang et al, 2010, O'Connor et al, 2021, Weiner, Nadig, et al, 2023). To identify common variants that are implicated in cell morphology, we performed 246 genome-wide association analyses, one for each composite trait. Each association was tested in colony and isolated cells separately (**Figure 6.10A, Figure 6.11C**). With our set of 297 donors, only one variant, rs315506, overlapping the chr17q11.2 locus, passed the genome-wide significance threshold (Bonferroni correction for 246 morphology traits, 5x10-8/246 = 2x10-10). rs315506 is an intergenic variant and was associated with spatial distribution of endoplasmic reticulum (ER) in the cytoplasm (Cytoplasm_RadialDistribution_RadialCV_ER_3of4) in colonies (MAF = 0.08, **β** (se) = -0.52 (0.08), P = 1.4x10-10, **Figure 6.11A**). This variant also showed suggestive evidence of association (P < 10-5/246 = 4.1x10-8) with spatial distribution of ER near the periphery of cells (Cells_RadialDistribution_MeanFrac_ER_4of4). rs315506 lies in the center of a 400kb window containing the genes NF1, CORPS, UTP6 and SUZ12. Chromosomal alterations on chr17q11.2 cause NF1 microdeletion syndrome, which has been shown to impair protein localization to the ER (Serra et al, 2019, Shih et al, 2020). To corroborate this observation, we analyzed the publicly available JUMP-Cell Painting data from U2OS cells that have perturbed NF1 and SUZ12 using CRISPR interference (Chandrasekaran et al., 2023). In this data, we see a significant change in our associated trait when NF1 and SUZ12 expression is decreased (**Figure 6.10B**).

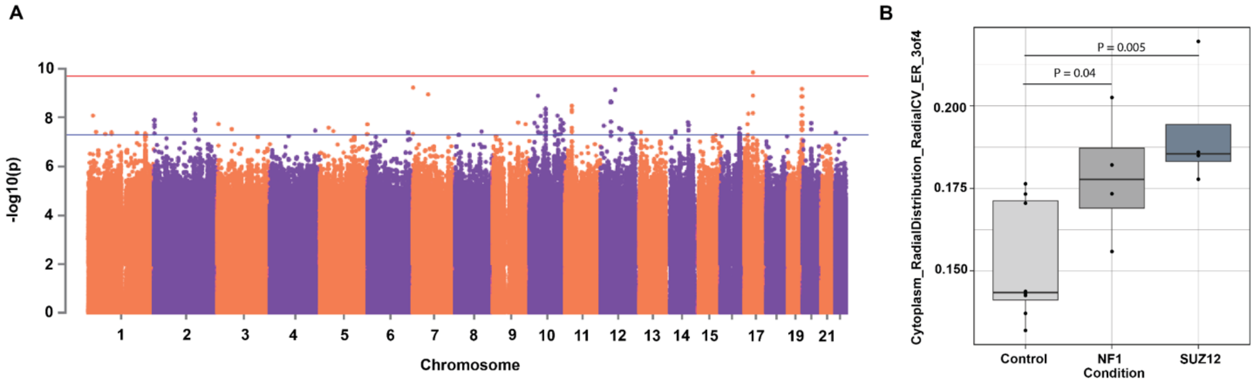


Figure 6.10. Common variant analysis.

(**A**) Manhattan plot for trait association test in colony cells. Red line represents the p-value threshold after Bonferroni correction for the number of tested traits and genes ($P < 2x10^{-10}$) Blue line represents the p-value threshold for hits with suggestive evidence ($P < 4.1x10^{-8}$). (**B**) Impact of knockdown of NF1 and SUZ12 in U2OS cells on Cytoplasm_RadialDistrubtion_RadialCV_ER_3of4 (P = 0.04 NF1, P=0.005 SUZ12, Welch's Two-Sample T-Test). Data is presented in a Tukey-style boxplot with the median (Q2) and the first and the second quartiles (Q2, Q3) and error bars defined by the last data point within +/− 1.5-times the interquartile range.

In colony cells, the second strongest association was on chromosome 7 (between Nuclei_Granularity_9_AGP and rs36036340, MAF = 0.08, **β** (SE) = 0.38 (0.06), P = 6x10-10). rs36036340 lies within the gene PRKAR1B. Variants in PRKAR1B have been linked to neurodevelopmental disorders and activity of PRKAR1B has been shown to regulate tumorigenesis (Elsayed et al, 2021, Feng et al, 2021, Marbach et al, 2021). PRKAR1B mediates PI3K/AKT/mTOR pathway signaling through direct interactions between PRKAR1B and PI3K-110alpha (Elsayed et al, 2021). We were unable to link perturbations in PRKAR1B to morphological changes for this feature using publicly available data (**Figure 6.11B**).

The most significant association in isolated cells was found on chromosome 13 (between Nuclei_RadialDistribution_RadialCV_Brightfield_2of4 and rs9301897, MAF = 0.13, **β** (se) = -0.31 (0.05), P = 4.5x10-10) (**Figure 6.11C**). rs9301897 lies within the gene GPC6, which is known to play a role in cell growth and division through the activation of cell surface receptors (Filmus et al, 2008, Veugelers et al, 1999).

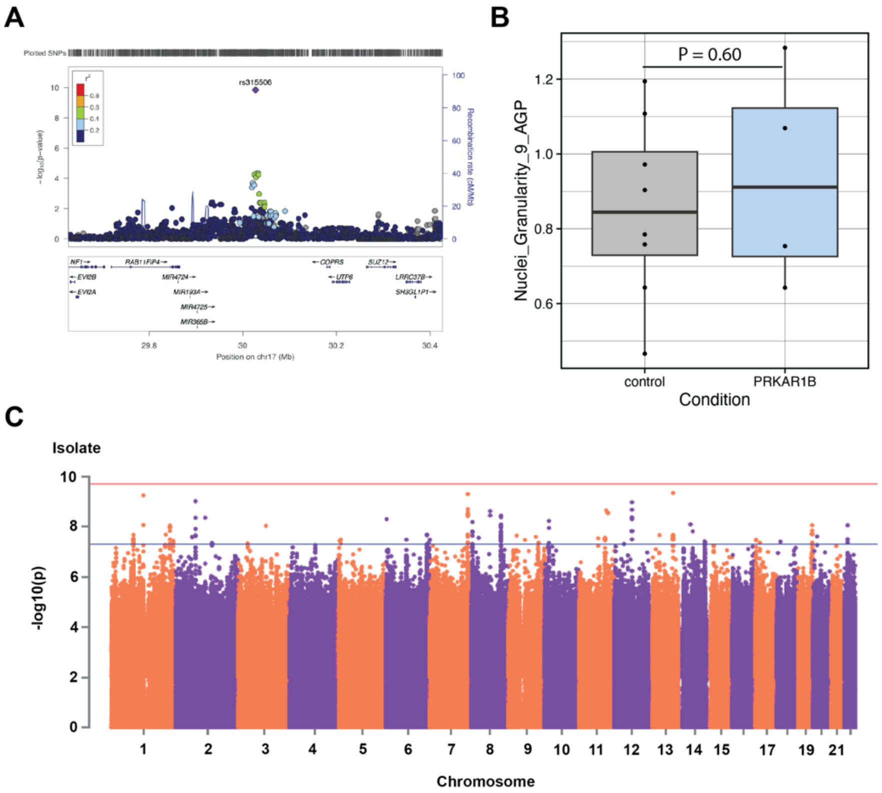


Figure 6.11. Analyses of common variant associations.

(A) LocusZoom plot for the association signal at chr17q11.2. rs315506, was significantly associated with spatial distribution of cytoplasm (Cytoplasm_RadialDistribution_RadialCV_ER_3of4) in colonies (MAF = 0.08, effect size (se) = -0.52 (0.08), P = 1.4x10^{-10}). (**B**) Impact of knockdown of PRKAR1B in U2OS cells on Nuclei_Granularity_9_AGP (P=0.60, Welch's Two-Sample T-Test). (**C**) Manhattan plot for trait association test in isolate cells. Red line represents the p-value threshold after Bonferroni correction for the number of tested traits and genes (P < 2x10^{-10}) Blue line represents the p-value threshold for hits with suggestive evidence (P < 4.1x10^{-8}).

Over 300 loci reached the suggestive genome-wide significance threshold (P < 4.1x10-8, **Table S8**) indicating that a larger sample size and improved statistical power would be able to identify additional common variants associated with cell morphology. To confirm our observed associations were not attributable to noise, we permuted the data by shuffling genotype labels and repeating the association tests. These results suggested that our observed significance of association between common variants and cell morphological traits was unlikely to have occurred by chance (**Figure 6.12**). Moreover, several loci (**Table S8**) showed suggestive association with more than one trait suggesting shared genetic etiology among different morphological traits.

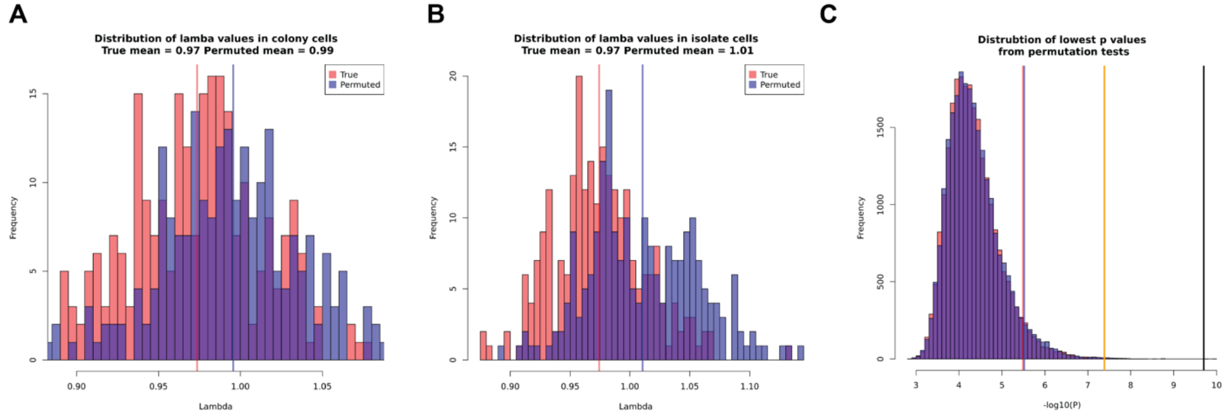


Figure 6.12. **Permutation of common variant analyses.**

(**A**) Comparison of lambda value distributions between true and permuted genotypes in colony cells (mean lambda; true = 0.97, permuted = 0.99). (**B**) Comparison of lambda value distributions between true and permuted genotypes in isolate cells (mean lambda; true = 0.97, permuted = 1.01). (**C**) Distribution of the lowest p values and 5% FDR threshold based on our perturbation analysis (red = colony cells, blue = isolate cells). Genome-wide (black) and suggestive (orange) threshold from our discovery associations.

6.7 - Sample size requirements and predictions for future cellular genetic studies

There has been an emergence of cellular genetics studies that aim to uncover the biological function of genetic variation (Wolter et al, 2023, Jerber et al, 2021, Wells et al, 2023, Miller, 2016, Mitchell et al, 2020). When compared to genetics studies of quantitative traits, cellular GWAS are limited in sample sizes, which provide a barrier to the discovery of significant associations. We sought to understand the power of our study with a sample size of nearly 300 individuals, and a distribution of effect sizes spanning common and rare variants. Our findings suggest that genetic discovery for cell morphological phenotypes was achievable at our current sample size. However, the small number of significant discoveries in our analysis begs the question of how many discoveries can be made at larger sample sizes. If discovery of many hits required a few thousand samples, such experiments would be feasible and worthwhile; if such discovery required hundreds of thousands to millions of samples, it may be out of reach for the foreseeable future. To answer this question, we estimated the distribution of common and rare variant effect sizes using Fourier Mixture Regression noLD (FMR-noLD) (**Methods**; O'Connor, 2021). Briefly, FMR-noLD fits a flexible mixture model to the distribution of effect sizes. This mixture model can be used to simulate effect sizes at various sample sizes, predicting how many significant discoveries will be made.

For common variants, we analyzed summary statistics from a pruned set of approximately 350,000 variants (**Methods**). We found that our dataset was underpowered for this analysis: FMR-noLD inferred that essentially all common variant effect sizes are 0, which is implausible and the expected behavior in the low power regime. For rare variants, we analyzed summary statistics from the main burden association analysis described earlier. In contrast to the common variant analysis, we predict that many discoveries will be made at feasible larger sample sizes, with more than 250 significant discoveries at N=1000 and more than 2000 discoveries at N=2000 (**Figure 6.13A, B**). Overall, our study was underpowered to detect a significant number of associations between common haplotypes and cell morphological traits, but our rare variant analyses provide a promising path for future studies exploring the impact of rare genetic variation on cell morphology.

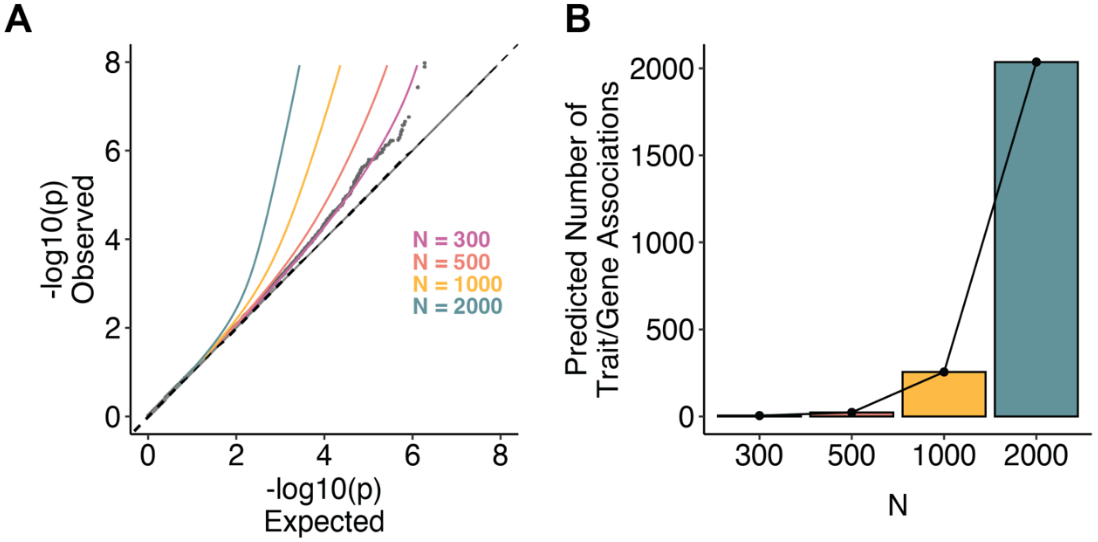


Figure 6.13. Distribution of effect sizes and predictions for future discovery.

A) QQ plot with current observed tests, model fit, and three predicted lines corresponding to N = 500, 1000, 2000.
B) Boxplots showing the number of discoveries per trait at current N, then projected N = 500, 1000, 2000.

6.8 - Discussion

Previous studies linking genetic variants to cellular function have largely focused on human genes and alleles which mediate molecular phenotypes, such as gene or protein expression and chromatin accessibility (Liang et al, 2021, Panopoulos et al, 2017, Wells et al, 2023, Jerber et al, 2021). To expand on these studies, we combined high-throughput cell culture techniques with cost-effective and high-dimensional image-based cell profiling (i.e., Cell Painting) to link genetic variants to their morphological function in 297 donors.

Our work provides the largest to date exploration of genetic influences on cell morphology (what we term cmQTLs). Where previous studies have been limited by both sample size and the scale of morphological measurements, we combined whole genome sequence analysis with Cell Painting to define relationships between genetic variants and morphological traits extracted from >5M iPSCs. We identified confounding factors that drive variation in cellular phenotypes which are important to address when performing similar studies. In particular, attenuating batch effects across plates and well position is critical in imaging-based assays. To address this challenge, we incorporated automated liquid handling devices to maximize plate distribution of cell lines across 384 well microplates.

We measured associations between rare variant burden and morphological traits, identifying novel associations between WASF2, PRLR, TSPAN15 and morphological phenotypes related to cytoplasmic area and shape, nucleic granularity, and the distribution of mitochondria around the nucleus. These associations were validated by CRISPR-mediated knockdown and supported by mechanistic information about these genes from the literature. Even though our effective knockdown had a range of efficiency (25-95%) we were able to measure meaningful changes in morphological features even at the lower range. This is consistent with previous work showing that gene expression is often stochastic and subtle changes in expression may lead to large changes in functional protein (Svenningsen et al, 2019). Each of the genes nominated in our rare variant analysis have been implicated in various cancers. We find this result interesting and somewhat unsurprising, given that pluripotent stem cells exhibit self-renewal properties which closely resemble cancer cells. These results suggest that genetic studies in iPSCs may shed meaningful insights into cancer-linked genes. It will be important for future studies to measure whether these associations are cell-type specific, or if they would be retained using differentiated cells. We extended our analysis to look for associations between morphological traits and common haplotypes. We found one significant result and 300 potential associations. We corroborated our significant common variant association with publicly available data showing that perturbations of nearby genes impact the associated morphological trait.

Interestingly, we observed no overlap in traits and associated variants between colony and isolated cells, suggesting a differential effect of genetic variation based on the environmental context of the cells. This is consistent with previous studies that have shown that intrinsic properties of cells may only come to light in the context of altering the cellular environment (Vigilante et al, 2019, Schrenk-Siemens et al, 2015, Nathan et al, 2022). In this study, we pseudo-bulked single cell profiling data to generate per-donor trait scores. It will be interesting for future work to examine morphology at the single cell level, similar to single-cell RNA sequencing approaches to better understand genetic influences on cellular heterogeneity.

The small number of significant discoveries in our work highlights that in vitro genetic studies still require substantial increases in sample sizes to saturate discovery potential. Our common variant analysis suggests that we are vastly underpowered to measure genetic associations to cell morphology and our estimated effect size distributions infer that cell morphology may behave similarly to quantitative traits. As discovery potential for quantitative traits often scales linearly with the number of samples included in the study, our data suggests that even with 3000 genetically unique cell lines, we may still only yield 10 genome-wide significant common variant cmQTLs. This is a sobering result, as it suggests that tens of thousands of cell lines would be needed to begin mapping SNP-trait associations for cell morphology. In contrast, our analysis of rare variant effect sizes suggests that with modest increases in sample sizes, we are well-positioned to detect many rare variant cmQTLs. Future studies which can leverage 1000-2000 unique cell lines may yield many 1000s of genome-wide significant gene-trait associations. While scaling in vitro studies to 2000 cell lines will still be a large hurdle, it is one which can be feasibly overcome with current iPSC collections.

Our work has several limitations that highlight directions for future research. This study focused on how germline genetic variation influences stem cell morphology. While we applied a rigorous quality control to identify and remove cells with abnormal karyotypes or large genomic rearrangements, it is possible that new somatic mutations may arise over time in culture. It will be important for additional studies to explore how recurring somatic mutations mediate cell morphology. Furthermore, the cell types utilized in this study are in a basal, undifferentiated state. It will be valuable for future studies to explore these associations in more physiologically relevant contexts, where disease-associated variants are enriched (Finucane et al, 2018). These findings suggest this framework could be applied to relevant cells and tissues such as iPSC-derived differentiated cells, post-mortem brain samples or excisable somatic cells. Moreover, we did not find any cell morphological traits associated with clinical disease categories from the cell line donors (data not shown). Similar to exploring common variant associations, we are likely underpowered in any single disease category to identify significant associations. There have been many studies elucidating morphological features associated with various diseases, but they often contained larger sample sizes and incorporated more specialized cell types (Gharaba et al, 2023, Antony et al, 2020, Schiff, Migliori, Chen, Carter et al, 2022). Extending our current study to diverse cell types and increasing the number of samples for clinical disease categories will be a critical next step in efforts to link cell morphology to human illnesses.

This approach holds significant promise for future studies leveraging human-derived, disease-relevant cell types for modeling the impact of genetic variation on cellular function. The use of imaging to capture phenotypes is particularly attractive in experimental designs for several reasons, such as the low cost per cell for imaging, and the ease of processing data and preparation of the cells or tissues as compared to the generation of other molecular data such as RNA-sequencing or epigenetic assays (Caicedo et al, 2017). Moreover, large imaging datasets provide tools for developing robust statistical models for combined analysis of morphological profiling data with additional modalities such as gene expression to comprehensively interrogate genetic variants and their function (Haghighi et al, 2022). Taken together, we demonstrate cellular morphology can be a cost-effective readout for modeling the biological function of human genetic variation.

6.9 – Reflection

A fundamental challenge in human biology is to understand the molecular and cellular function of human variation. Tremendous progress has been made linking molecular signatures to both rare and common genetic variation, but molecular phenotypes only account for one aspect of cellular function. It is therefore critical to develop and optimize alternative data modalities to gain a more comprehensive understanding of the role genetic variation plays in modulating human health and disease.

In this chapter, I describe a coordinated strategy to scale *in vitro* cellular studies in combination with state-of-the-art imaging and analysis techniques to nominate genetic variants which mediate cellular morphology, which we term cell morphological quantitative trait loci (cmQTLs). We find several variants with strong associations to morphological traits, which we validate using functional knockdown experiments.

This project bridges a gap between two fields, human genetics, and microscopy. These findings provide a new avenue for downstream exploration of human genetics. As imaging-based approaches are considerably cheaper than modern next generation sequencing techniques, our method is scalable to larger cohorts and more broadly applicable across labs and institutions, increasing accessibility to the study of functional genomics.

This approach may be beneficial across several fields, from human genetics to stem cell biology, as this framework can be leveraged to explore a range of diverse cell types and physiologically relevant tissues as a cost-effective way for cellular genetic studies. Similar to eQTL discovery, cmQTLs can be used in fine-mapping efforts to narrow down causal variants in genetic association studies, facilitating mechanistic insights and therapeutic discovery.

Now that I had added a new tool in my repertoire, I aimed to combine the methods described in the aforementioned chapters to gain insights into convergence between rare and common genetic variants which confer risk for psychiatric phenotypes. In the following two chapters, I will describe work incorporating gene expression, chromatin architecture, and morphological profiling to understand how common genetic risk and the 22q11.2 deletion converge to influence molecular and cellular phenotypes implicated in stem cell derived neural cell types.

6.10 - Statement of contributions

This work was co-led by Jatin Arora and Samira Asgari. Jatin performed the variance component analysis and rare variant burden test in Section **6.3** and **6.4**. Samira Asgari performed the common variant test in Section **6.6** with some assistant from myself. Ajay Nadig and Luke O'Connor performed the effect size distribution and sample predictions in Section **6.7**. I performed all data generation for this project and all of the CRISPRi validation analyses highlighted in Section **6.2** and **6.5**.

Chapter 7

Convergence of genetic risk factors for neuropsychiatric conditions at chromosome 22q

7.1 - Introduction

Thousands of genetic variants are now confidently associated with trait and disease variation in humans (Abdellaoui et al, 2023), and the functional annotation of those variants is a primary bottleneck in human genomics (Lappalainen & Macarthur, 2021). Most variant-to-function paradigms focus on connecting variants to nearby genes ("cis-genes"), such as genes close to disease-associated single nucleotide polymorphisms (SNPs; Nasser et al., 2021) or genes within the boundaries of disease-associated genomic copy number variants (CNVs; Collins et al., 2022). At the same time, a separate body of work has characterized the state and 3D conformation of chromatin (Kundaje et al., 2015; Lieberman-Aiden et al., 2009), identifying supra-megabase genomic clusters that interact physically and have similar chromatin states (Yaffe & Tanay, 2011; Rao et al., 2014). This regulatory architecture, which connects genetic variants to faraway genes in large genomic compartments, raises the possibility that the cis-gene approach may sometimes fail to identify the more distal functional impacts of disease-associated genetic variation. We aim here to integrate low-dimensional, regionally-patterned properties of the 3D genome into variant-to-function mapping in human genetics, and to improve variant-to-function characterization in genomic regions characterized by particularly high degrees of 3D contact.

Within the 3D genome literature, an enrichment of intrachromosomal contact and transcriptional activity has been consistently observed in the q arm of chromosome 22 (chr22q).

Yaffe and Tanay (2011), for example, clustered the genome by intrachromosomal contact and epigenetic modifications, and assigned the entirety of chr22q to a "high activity cluster". Rao et al. (2014) refined these results with deeper HiC data and found that chr22q has the highest proportion of the A1 sub-compartment (marked by high transcriptional activity and an enrichment for activating histone modifications) of any chromosome. These observations, in combination with our previous suggestive findings regarding autism-associated genetic variation at chr16p (Weiner et al, 2022; Huguet et al, 2022), led us to hypothesize that the richly interacting, densely connected regulatory architecture of chr22q may shape the effects of both rare and common disease-associated genetic variation therein. Chr22q is home to the recurrent 22q11.2 microdeletion (22q11.2del), a heavily studied CNV which causes a multi-system disorder – DiGeorge Syndrome – characterized by a diverse series of neuropsychiatric, cardiac, and immune phenotypes (McDonald-McGinn et al. 2015). 22q11.2del is the genetic risk factor most statistically associated with schizophrenia at a population level (Marshall et al., 2017). Notably, while the cardiac effects of 22q11.2del have been mapped to a putative driver gene within the canonical deletion interval, *TBX1* (Yagi et al., 2003), the deletion's neuropsychiatric effects remain unexplained (Motahari et al., 2019; but see also Khan et al., 2020, and Nehme et al., 2022), perhaps implying a non-canonical mechanism of risk. Like all other chromosomes, chr22q also contains thousands of common SNPs which act additively to influence human complex trait variation (e.g., Loh et al., 2015), including the traits associated with DiGeorge Syndrome. It is unclear how region-level epigenomic and 3D conformational features shape the functional consequences of common polygenic variation and the transcriptional relationship between a trait's common and rare variant influences.

Here, we integrate several functional genomic and genetic association datasets to investigate the impact of disease-associated genetic variation within chr22q in human neural cells, and how that impact is shaped by this region's exceptional regulatory architecture. Additionally, we expanded these analyses to morphological traits to understand how the 22q11.2 deletion impacts cellular phenotypes. We focused our analyses on neural cells, given 22q11.2del's cryptic association with neuropsychiatric disorders (e.g., schizophrenia, intellectual disability, autism, attention deficit hyperactivity disorder; Olsen et al, 2018). We find that

22q11.2del decreases the average expression of genes across the 35 megabase (MB) chr22q, far beyond the ~3MB contained by canonical deletion boundaries. This decrease is replicable in multiple brain cell types and correlated with the patterning of distal intrachromosomal contacts within chr22q. We show that common, polygenic risk for neuropsychiatric phenotypes on 22q similarly decreases average gene expression across chr22q, in a manner convergent with the deletion effects at a per gene level. This broad downregulation nominates novel mechanistic explanations for 22q11.2del-associated neuropsychiatric phenotypes, at the levels of individual genes (notably *SMARCB1*), disease-associated genomic regions (22q13.3), and dysregulation of the full human transcriptome. More broadly, our findings suggest that region-level (e.g., spanning tens of megabases) genomic features shape the functional impact of genetic variation, with implications for fundamental genome biology.

7.2 - Unusual regulatory architecture of 22q and 22q11.2

We observed that across similarly sized (33 MB) genomic partitions, chr22q was enriched for intrachromosomal physical contacts across both lymphoblastoid cell lines (Lieberman-Aiden et al., 2009) and fetal neocortex (Won et al., 2016) (Mean Z = 2.07; **Figure 7.1A**). Based on this elevated degree of intrachromosomal physical contact, we hypothesized that chr22q would be enriched for distal regulatory elements and computed the fraction of each partition annotated as being in the "enhancer" state by ChromHMM in dorsolateral prefrontal cortex (DLPFC) samples from the Roadmap Epigenome Project (Kundaje et al., 2015). We found that chr22q had the highest density of active enhancer elements of any similarly sized genomic partition (Z = 2.24; **Figure 7.2**). These findings replicate and build on previous reports identifying extraordinary transcriptional activity and intrachromosomal physical interaction at chr22q (Rao et al., 201; Yaffe & Tanay, 2011).

To further explore the regulatory architecture of chr22q, we generated HiC data from previously described wildtype human pluripotent stem cell lines (hPSCs) (Nehme, Pietlanen et al., 2022), meta-analyzing across three *in vitro* cell types including hPSCs, neuronal progenitor cells (NPCs), and excitatory neurons. We found evidence of widespread, distal intrachromosomal contact on chr22q (**Figure 7.1B**). Additionally, we noted that the 22q11.2del deletion interval (**Figure 7.1B, orange box**) seemed to be particularly enriched for distal intrachromosomal contact. Indeed, when we examined the median degree of contact between 1Mb bins on chr22q and the rest of chr22q, we found that bins overlapping the 22q11.2del region comprised one of three peaks across the chromosome (**Figure 7.1C**). Taken together, these findings demonstrate that chr22q has a densely connected regulatory architecture, and that within chr22q, 22q11.2 is a key hub of distal intrachromosomal contact. These observations led us to hypothesize that genetic variation on chr22q, especially deletion of a contact hub such as 22q11.2, may impact gene expression across the entire chromosome arm.

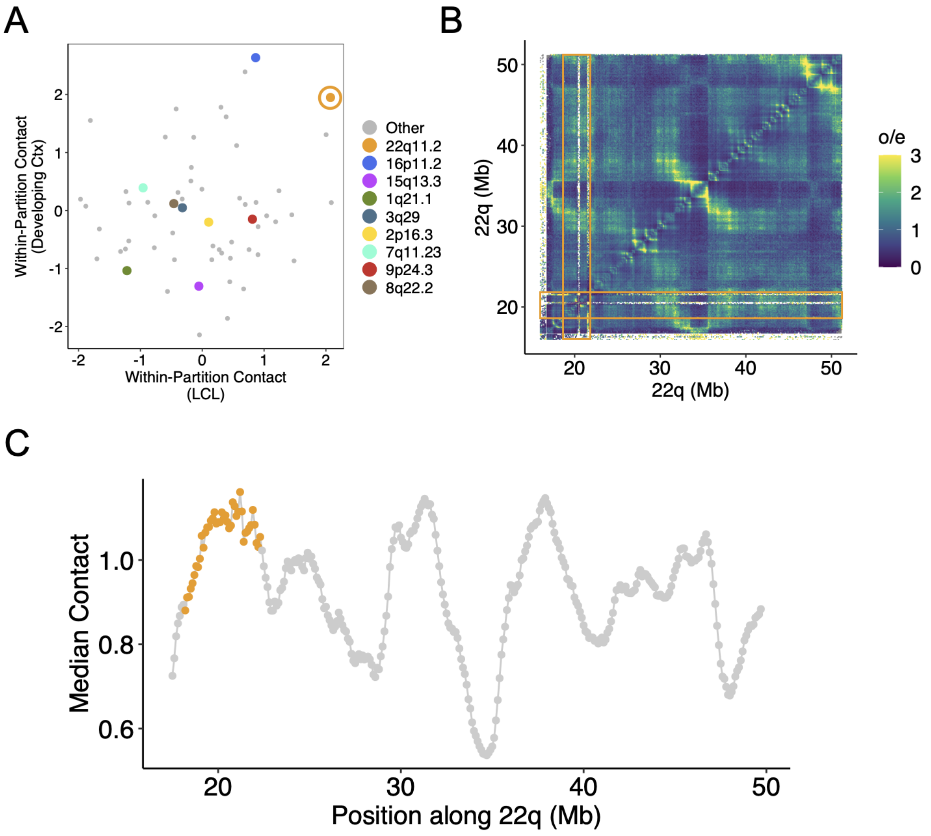


Figure 7.1. Unusual regulatory architecture of 22q and 22q11.2.

(A) Degree of chromosomal contact within 33-Mb partitions tiling the human genome. Partitions containing CNVs associated with schizophrenia and other neuropsychiatric diseases (Marshall et al, 2017, Nat Gen) are shown in color.

(B) Intrachromosomal contact map for chr22q, generated by pooling data between 3 cell types; the canonical 22q11.2 deletion interval is highlighted in orange. (C) Median contact between positions on chr22q and the rest of chr22q, computed with a sliding window approach. Positions within the 22q11.2 deletion boundaries are highlighted in orange.

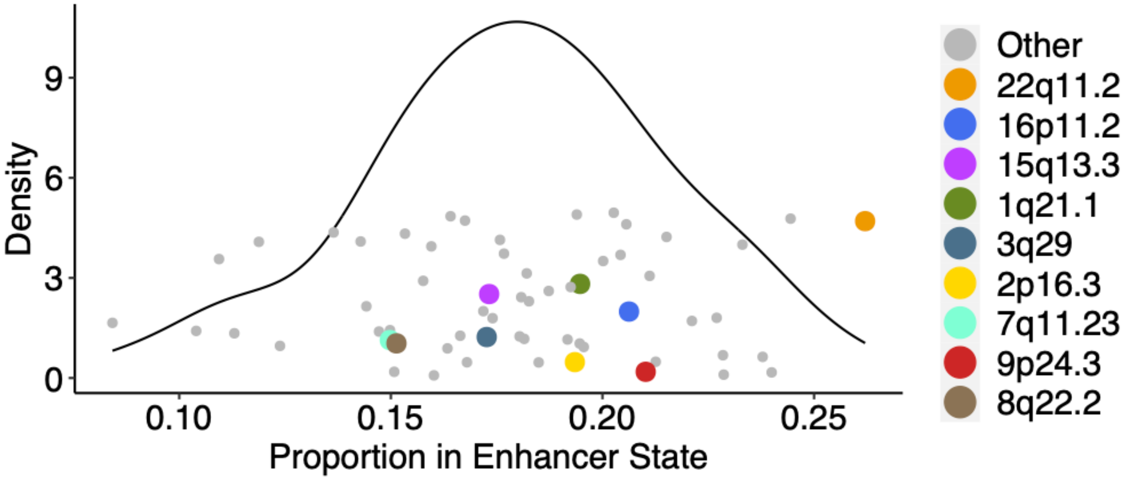


Figure 7.2. Density of enhancer elements in large genomic partitions.

For each ~33 Mb genomic partition, we used ChromHMM predictions from the Roadmap Epigenome Project dorsolateral prefrontal cortex samples (Kundaje et al, 2015) for active enhancers (EnhG1, EnhG2, EnhA1, EnhA2) to compute the proportion of each partition in the enhancer state. Colored points represent partitions containing a CNV significantly associated with schizophrenia (Marshall et al, 2017). Y-axis position of points is randomly jittered to prevent overplotting.

7.3 - 22q11.2del decreases expression of genes across chromosome 22q

To test this, we revisited bulk RNA sequencing data from isogenic 22q11.2del and control hPSC lines that we previously generated (Nehme, Pietilainen et al., 2022). Including all three cell-types (hPSCs; neuronal progenitor cells, NPCs; excitatory neuron, ExN) in a single differential expression model, we found that 22q11.2del was associated with subtle downregulation of gene expression across chr22q (mean t-statistic = -0.88, mean fold change = 0.93, -log$_{10}$(p) = 19, two-sided t-test, **Figure 7.3A**). We unpacked and assessed the robustness of chromosome-wide expression downregulation with several follow-up analyses. Differential expression effects across chr22q were more negative than in the rest of the genome, suggesting that expression downregulation is region specific (**Figure 7.4A**). We found a similar downward shift of differential expression statistics when analyzing raw versus normalized read counts, confirming that chromosome-wide expression downregulation is not an artifact of expression normalization (**Figure 7.4B**). Additionally, we found the directionally consistent, but less statistically robust, evidence for expression downregulation in a case-control cohort comparing 22q11.2del carriers to control individuals (p = 0.06; Figure **7.4C**). The loss of statistical power was expected given the genetic background variability that is present in case-control samples but absent in the isogenic setting. When we analyzed each of the four cell types individually, we found evidence for expression downregulation within each cell-type, suggesting that this phenomenon is not strongly cell-type dependent within this lineage (**Figure 7.5**). Lastly, per-gene differential expression effects were correlated between hPSCs, NPCs, and ExN (**Figure 7.6**).

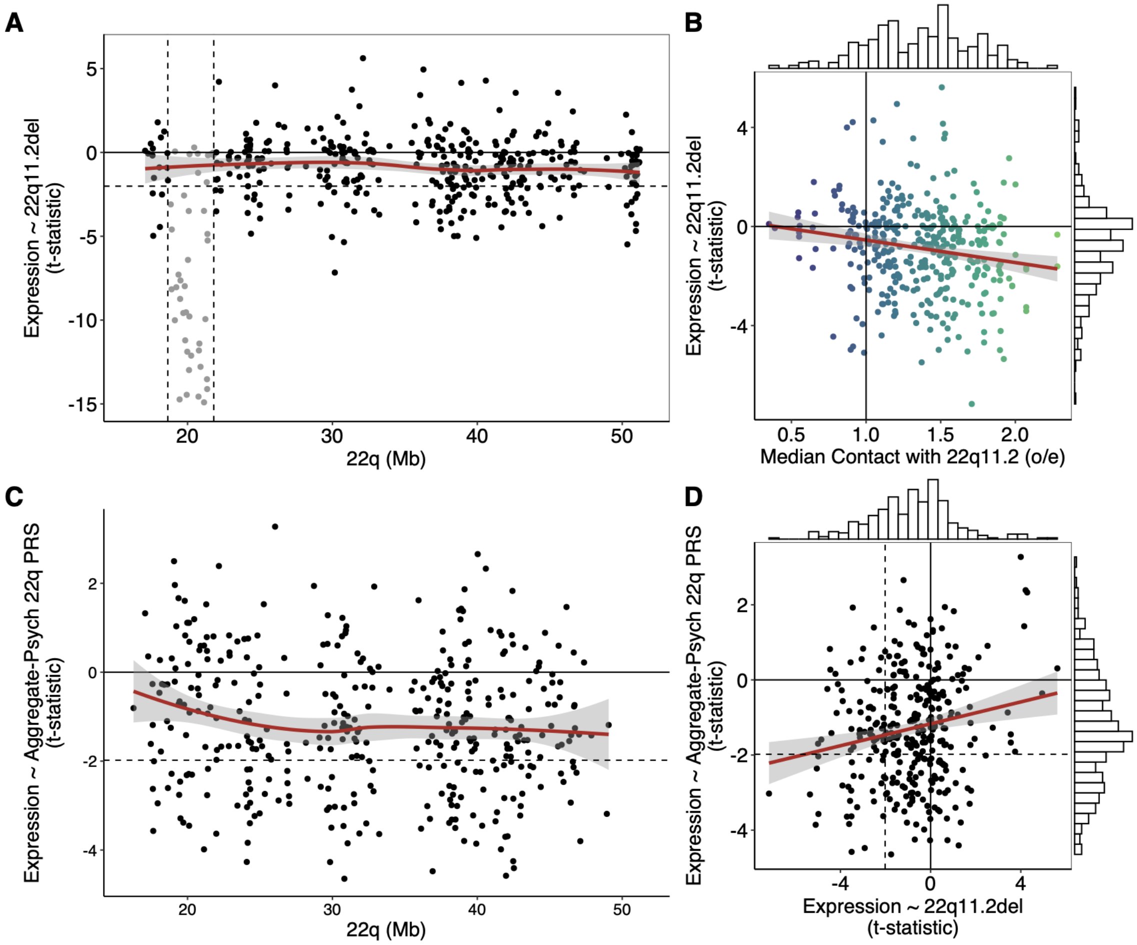


Figure 7.3. Schizophrenia-associated rare deletions and common polygenic risk convergently decrease gene expression at chromosome 22q.

(**A**) 22q11.2del-induced differential expression across chr22q, integrating data from three cell-types. (**B**) Association between differential expression statistics and wild-type degree of chromatin contact with the canonical 22q11.2

deletion region (**C**) Association between PGS-SCZ$_{22q}$ and expression of chr22q genes in post-mortem glutamatergic neurons. (**D**) Association between gene-wise differential expression statistics from (A) and (C).

The expression downregulation in isogenic cell lines is consistent with our previous finding of extended expression effects of the autism-associated 16p11.2del across Chr16p (Weiner et al, 2022). The relatively limited scope of 16p11.2del data resources however, hampered our ability to characterize per-gene expression effects, and the relationship between those effects and the 3D genome. In contrast, the better-powered 22q11.2del data resource was able to identify significant extended expression effects at individual genes across chr22q (**Figure 7.3A, points below horizontal dashed line**), and included HiC data in the relevant *in vitro* cell types, allowing us to ask whether distal expression effects were explained by 3D genome conformation. For each gene on 22q, we computed the degree of physical contact between the 100kB window containing the gene and the 22q11.2del deletion interval in control cell lines, aggregated across hPSCs, NPCs, and ExNs. Firstly, we found that genes across chr22q are in physical contact with 22q11.2 to a much greater degree than expected by chance (mean observed/expected = 1.35, -log$_{10}$(p) = 61, two-sided t-test; **Figure 7.3B**). Secondly, we found that physical contact was predictive of differential expression at the per gene level, such that genes with greater contact with 22q11.2 in control cell lines had greater expression downregulation in 22q11.2del (Pearson's r = -0.24, p = 4.3 * 10^{-6}; **Figure 7.3B**). When we examined cell types individually, we found the same effect in three of four cell types (**Figure 7.5**). Our findings suggest that, in wild-type cells, 22q11.2 interacts physically with large swaths of chr22q in a manner that positively regulates gene expression. In 22q11.2del, expression of chr22q genes is diminished proportionally to the wild-type physical contact.

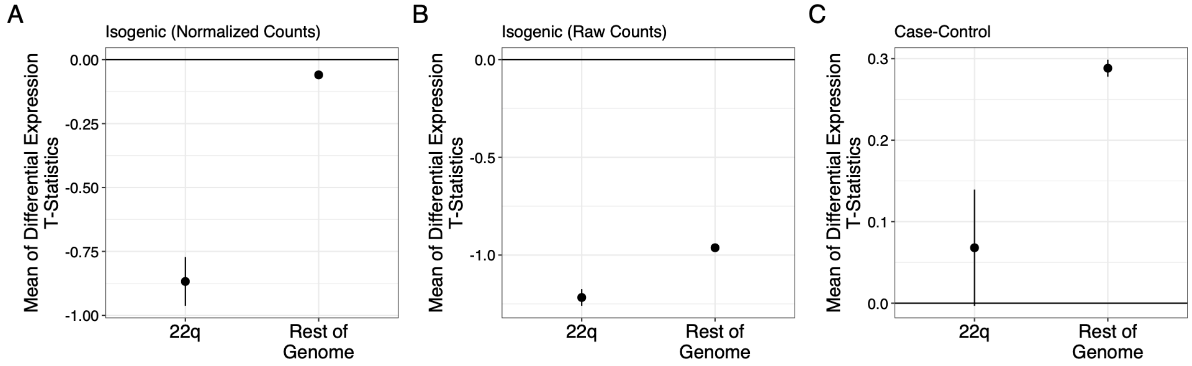


Figure 7.4. Comparison of 22q11.2 Differential Expression Effects on 22q versus in rest of genome.

Mean differential expression statistic and standard error of the mean are shown for the genes on chr22q, and all other genes. (A) Normalized count using the standard size-factor based approach implemented in DESeq2 (e.g., same isogenic analysis shown in Figure 2A) (B) Raw counts, e.g., isogenic analysis from Figure 1 with size-factors set to 1 (C) Case-Control analysis, from a previously described collection of 28 22q11.2del cases and 20 controls.

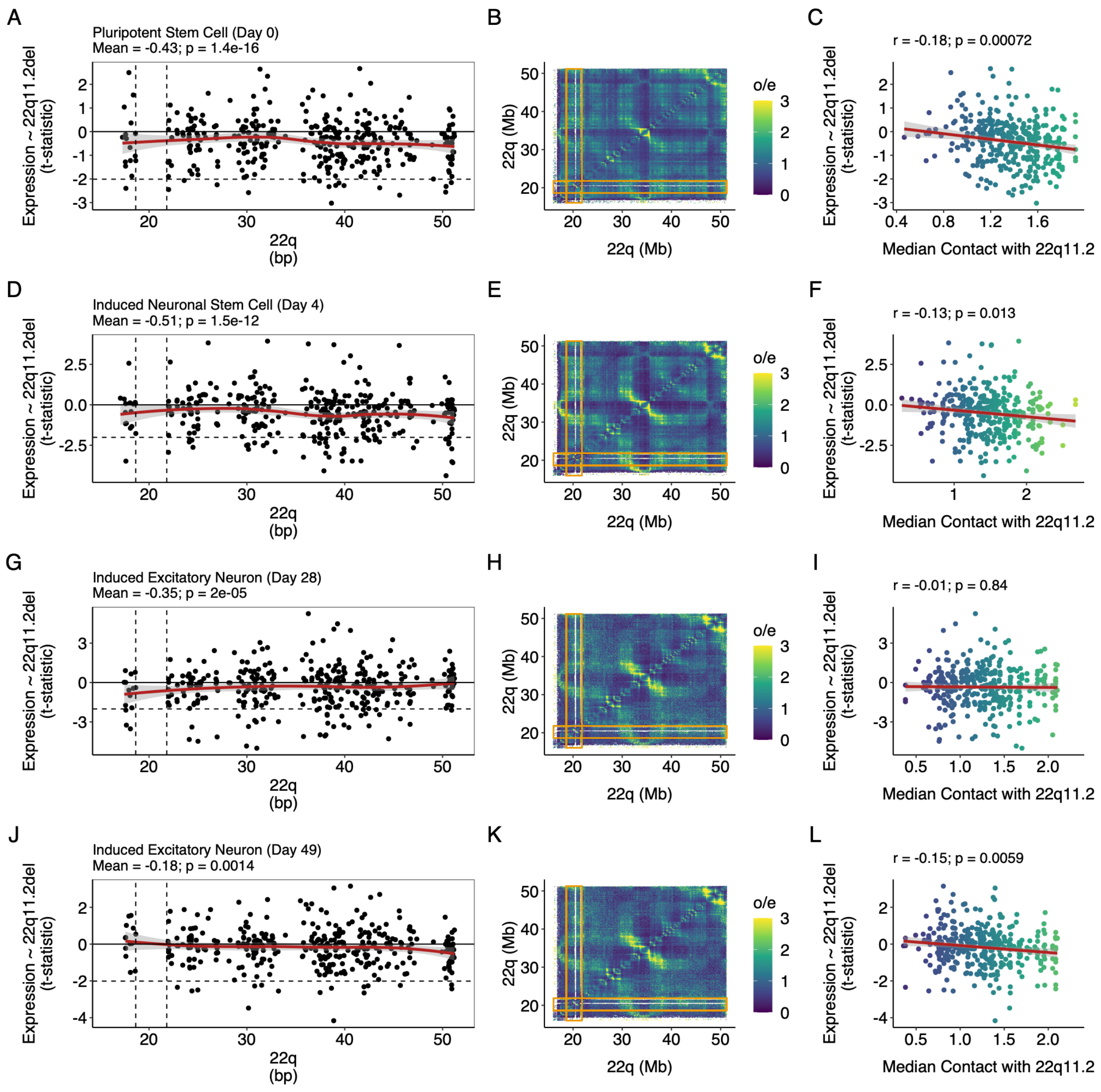


Figure 7.5. Effects of 22q11.2del by cell type.

Analyses stratified by cell type. Shown for (A-C) human pluripotent stem cells, (D - F) neuronal progenitor cells, (G - I) excitatory neurons sampled at day 28 of differentiation protocol, (J - L) excitatory neurons sampled at day 49 of differentiation protocol.

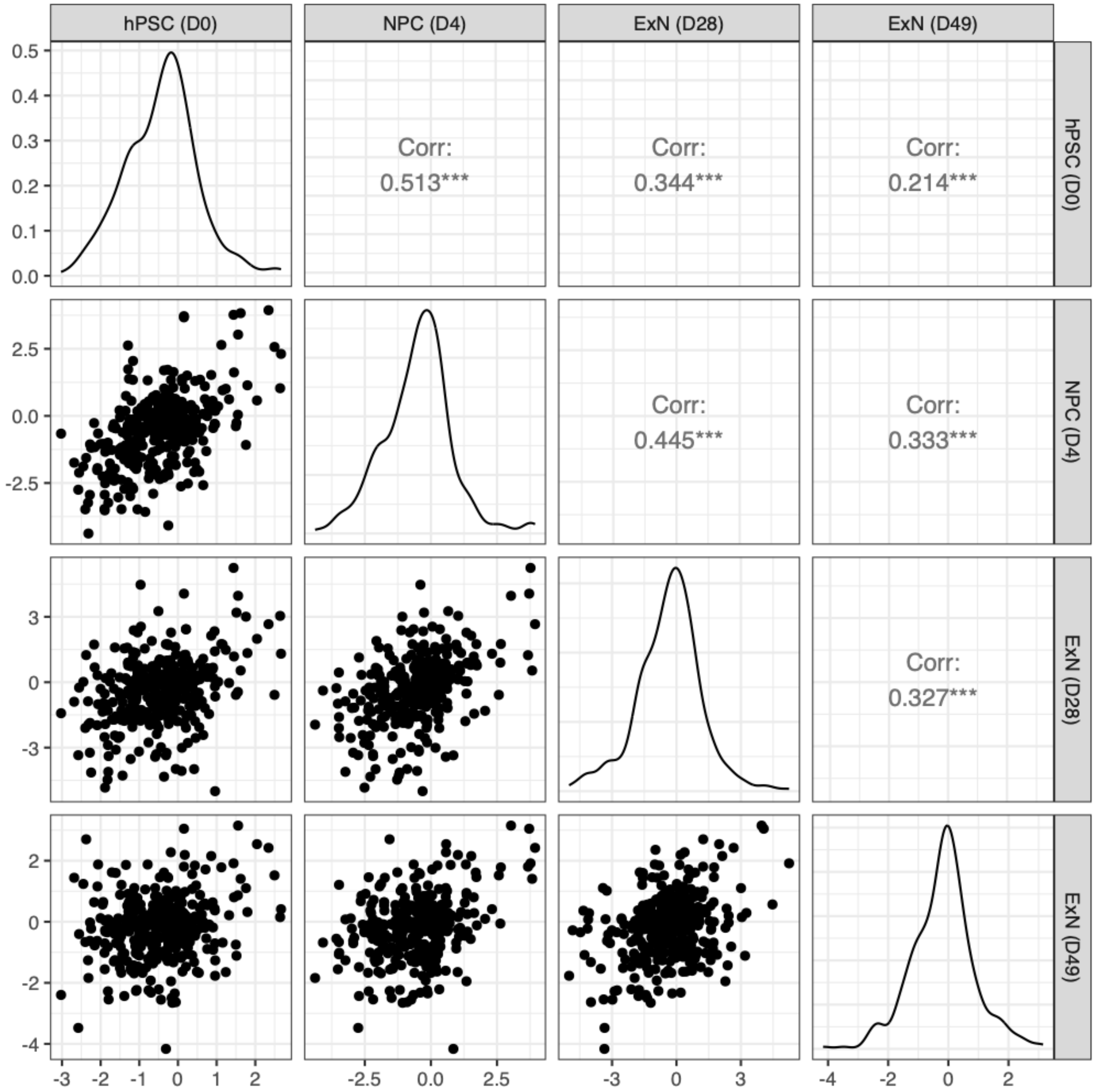


Figure 7.6. Correspondence between 22q11.2 differential expression effects between cell types.

We ran the combined analysis shown in Figure 2A for each of the four cell types in our dataset. Lower triangle shows the correlation in gene-wise differential expression statistics for chr22q genes across cell types. Diagonal shows the distribution of these effects for each cell type. Upper triangle shows Pearson's correlations related to data in lower triangle.

7.4 - Local polygenic burden for neuropsychiatric traits decreases expression of genes across chromosome 22q

In addition to rare deletion events such as 22q11.2del, common polygenic variation makes a substantial contribution to variation in complex traits, including neuropsychiatric disorders (Sullivan et al., 2017). After observing that the effects of 22q11.2del are shaped by the 3D conformation of chr22q, we wondered whether similar principles may apply to common polygenic variation, which explains a much larger proportion of trait heritability than rare genetic variants (Gaugler et al., 2014; Weiner, Nadig et al., 2023). We selected 8 traits with well-powered GWAS: four neuropsychiatric traits (schizophrenia, autism, attention-deficit hyperactivity disorder, and intelligence) and four non-neuropsychiatric traits (height, body-mass index, red blood cell count, serum low-density lipoprotein). For each trait, we aggregated individually small common variant effects across chr22q into a local polygenic score (PGS_{22q}) and assessed the association between these regional polygenic scores and gene expression across chr22q. We hypothesized that common polygenic risk at chr22q would decrease gene expression, in a manner convergent with 22q11.2del. To test this hypothesis, we leveraged a resource pairing genotyping and single-nucleus RNA sequencing in human post-mortem brain tissue from 122 individuals. We computed PGS_{22q} for each trait using publicly available summary statistics and associated these scores with expression of genes across chr22q in glutamatergic neurons. We found that elevated regional polygenic risk for each of the neuropsychiatric traits was associated with decreased expression of genes across 22q in glutamatergic neurons, with weaker associations for the non-neuropsychiatric traits (**Figure 7.7**). A notable exception to this trend was body-mass index, a biometric trait whose regional polygenic risk score was associated with decreased expression of genes across chr22q. This observation may reflect the previously described strong enrichment of body mass index heritability in brain expressed genes (Finucane et al., 2015).

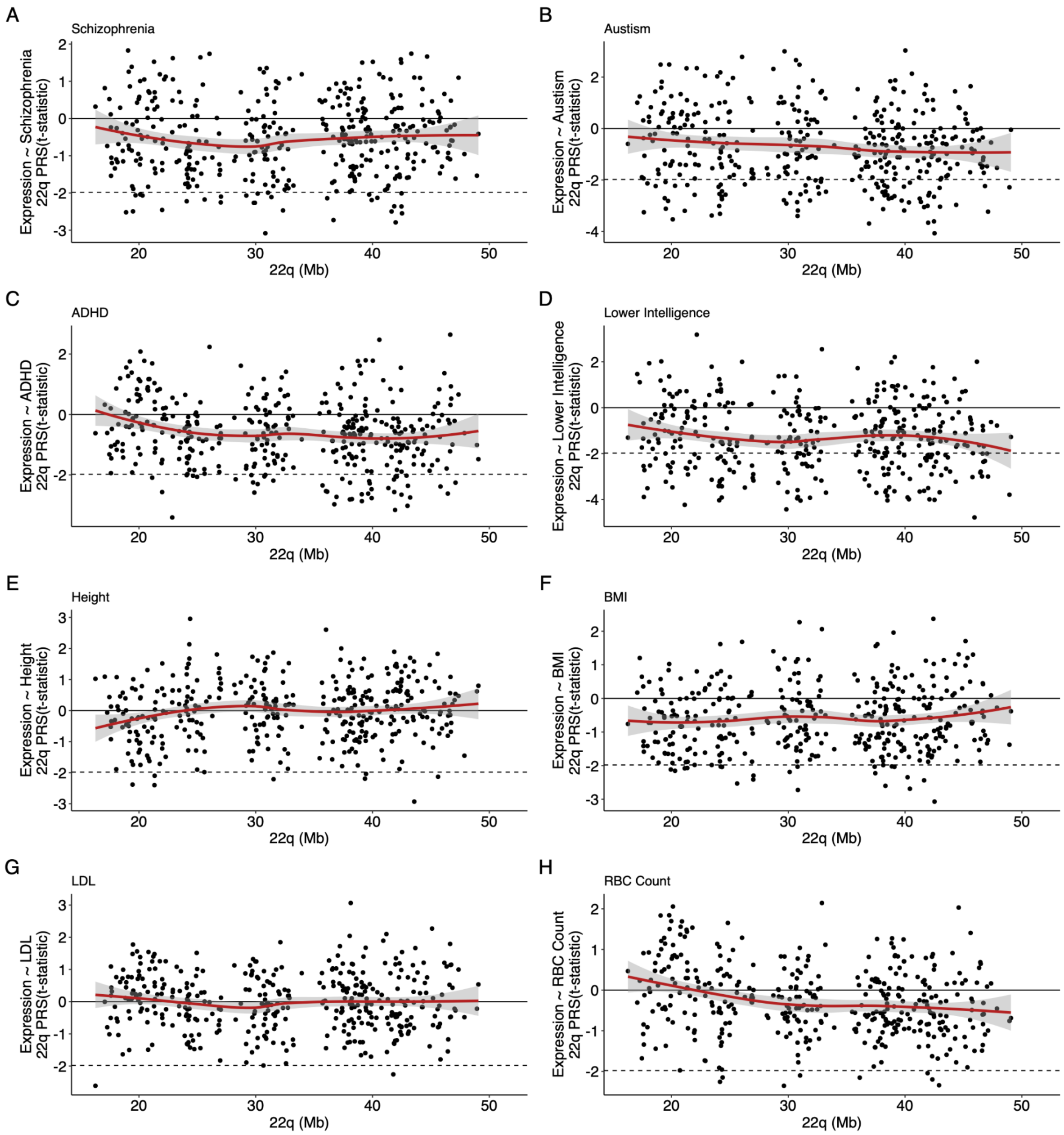


Figure 7.7: Association between various local PGS and gene expression in glutamatergic neurons.

(A) Schizophrenia (Trubetskoy et al, 2022) (B) Autism (unpublished, **Methods**) (C) Attention-Deficit Hyperactivity Disorder (Demontis et al, 2019). (D) Lower Intelligence (Savage-Jansen et al, 2018), (E - H) UK Biobank traits (Neale Lab, **URLs**).

Motivated by these findings, we collapsed the four neuropsychiatric PGS_{22q} into an aggregate score, which gave us greater power to detect this expression downregulation in glutamatergic neurons (mean t-statistic = -1.18; -log(p) = 42, two-sided t-test; **Figure 7.3C**). We note that this approach and result are concordant with a recent report demonstrating that combinations of trait PGS yield greater prediction accuracy for neuropsychiatric disorders (Albiñana et al, 2022). The aggregate score was also associated with decreased chr22q gene expression in the other three major cell types of the cerebral cortex (**Figure 7.8**), again suggesting that this phenomenon is not strongly cell-type dependent. We replicated our findings in European-ancestry bulk RNA sequencing samples from the CommonMind consortium (**Figure 7.8**). We did not replicate these results in the African-ancestry samples from the same collection, likely reflecting bias towards European ancestry individuals in the individual GWAS (Martin et al., 2019).

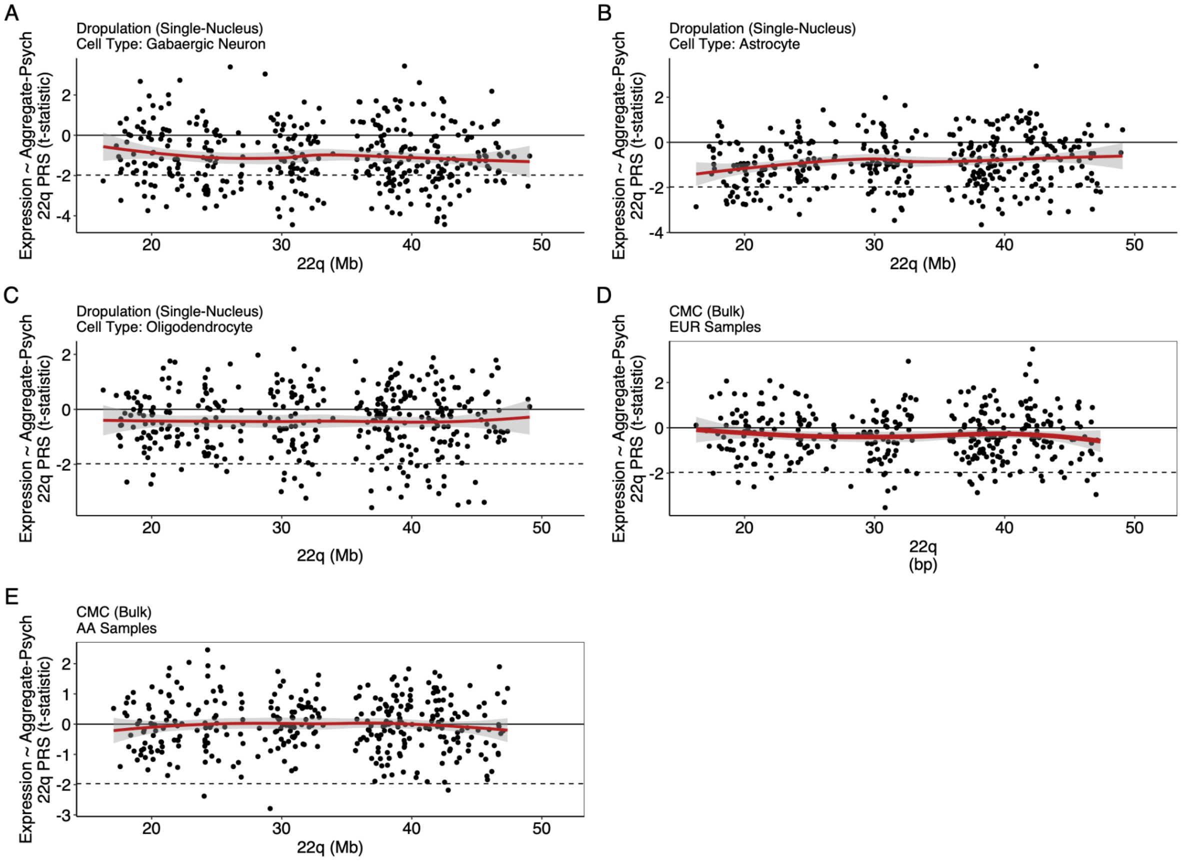


Figure 7.8. Association between PGS-SCZ$_{22q}$ in other cell types, and in bulk sequencing data.

(A-D) Relationship between PGS$_{22q}$ and chr22q gene expression in gabaergic neurons, astrocytes, and oligodendrocytes in post-mortem DLPFC single-nucleus RNAseq data. (D-E) Relationship between PGS$_{22q}$ and chr22q gene expression in bulk RNAseq data from the DLPFC, in European ancestry (EUR) and African-American ancestry (AA) samples.

Integrating our results for rare and common genetic variation, we found that per-gene differential expression effects of 22q11.2del and aggregated neuropsychiatric PGS_{22q} were correlated (Pearson's r = 0.16, p = 0.003; **Figure 7.3D**). The effect size correlation between these dramatically different study designs (experimentally induced deletion in *in vitro* isogenic setting versus naturally occurring common alleles in post-mortem samples) strongly implies that rare and common genetic risk factors for neuropsychiatric disease at chr22q mechanistically converge at the level of gene expression. This finding suggests that the richly interacting regulatory architecture of chr22q shapes the impact of both small- and large-effect genetic variants. Additionally, these convergent findings suggest that lower average expression of genes across chr22q is a risk factor for a diverse set of neuropsychiatric conditions. However, these effects also raise a challenging question for functional genomics: how do we investigate the consequences of subtle downregulation of hundreds of genes? We now explore this question through three analyses at different and complementary genomic scales: single genes, groups of genes, and the entire transcriptome.

7.5 - 22q11.2del and local polygenic risk decrease the expression of SMARCB1

Among the 27 chr22q genes that are significantly and convergently downregulated by 22q11.2del and regional polygenic influences on neuropsychiatric traits, *SMARCB1* (chr22: 24129150-24176703) is particularly notable. *SMARCB1* encodes a component of the mammalian SWI/SNF (mSWI/SNF) complex, which is essential for the development and maintenance of the active enhancer landscape of the genome (Nakayama et al., 2017, Wang et al., 2017). The phenotypic landscape associated with genetic variation in *SMARCB1* is complex. Missense variants in SMARCB1 that cause hypomorphic nucleosome remodeling activity are associated with Coffin-Siris Syndrome, a severe neurodevelopmental disorder (Tsurusaki et al., 2012; Mashtalir et al., 2020; Valencia et al., 2019). Germline loss-of-function variants in *SMARCB1* are associated with infancy-onset rhabdoid tumors with a two-year survival of 7.6% (Bordeaut et al., 2011); the association of such loss-of-function variants with neurodevelopmental outcomes in later childhood is consequently unknown. It is abundantly clear that *SMARCB1* is extremely dosage sensitive, with population genetic analyses revealing profound depletion of both missense (Z = 3.6) and loss-of-function (pLI = 1) variation in the genome aggregation database (gnomAD; Karczewski et al, 2020), as well as strong evidence for haploinsufficiency from copy number variant association data (pHaplo = 0.99; Collins et al, 2022).

We found that decreased *SMARCB1* expression was associated with both 22q11.2del (p = 0.002) and aggregated neuropsychiatric PGS_{22q} (p = 0.0002) (**Figure 7.9A, B**). We confirmed the effect of 22q11.2del on *SMARCB1* relative expression via qPCR and found that 22q11.2del is associated with a 28% decrease in expression (p = 0.001; **Figure 7.10A**). We further confirmed that this decrease in expression is associated with a decrease in protein level, finding via western blot that 22q11.2del is associated with a ~37% decrease in SMARCB1 relative protein level (p = 0.001; **Figure 7.11**). In light of the severe *SMARCB1*-depletion-associated phenotypes described above, this fractional decrease in expression could potentially exert substantial effects on risk for neurodevelopmental disorders as well as malignancy, which has an ambiguous association with 22q11.2del (Lambert et al., 2017). However, whether such fractional decreases phenocopy the larger dosage decreases described above is difficult to ascertain from existing phenotypic data resources. Instead, leveraging a detailed body of work on the regulatory role of *SMARCB1* (Nakayama et al., 2017), we aimed to understand the cellular consequences of partial *SMARCB1* depletion.

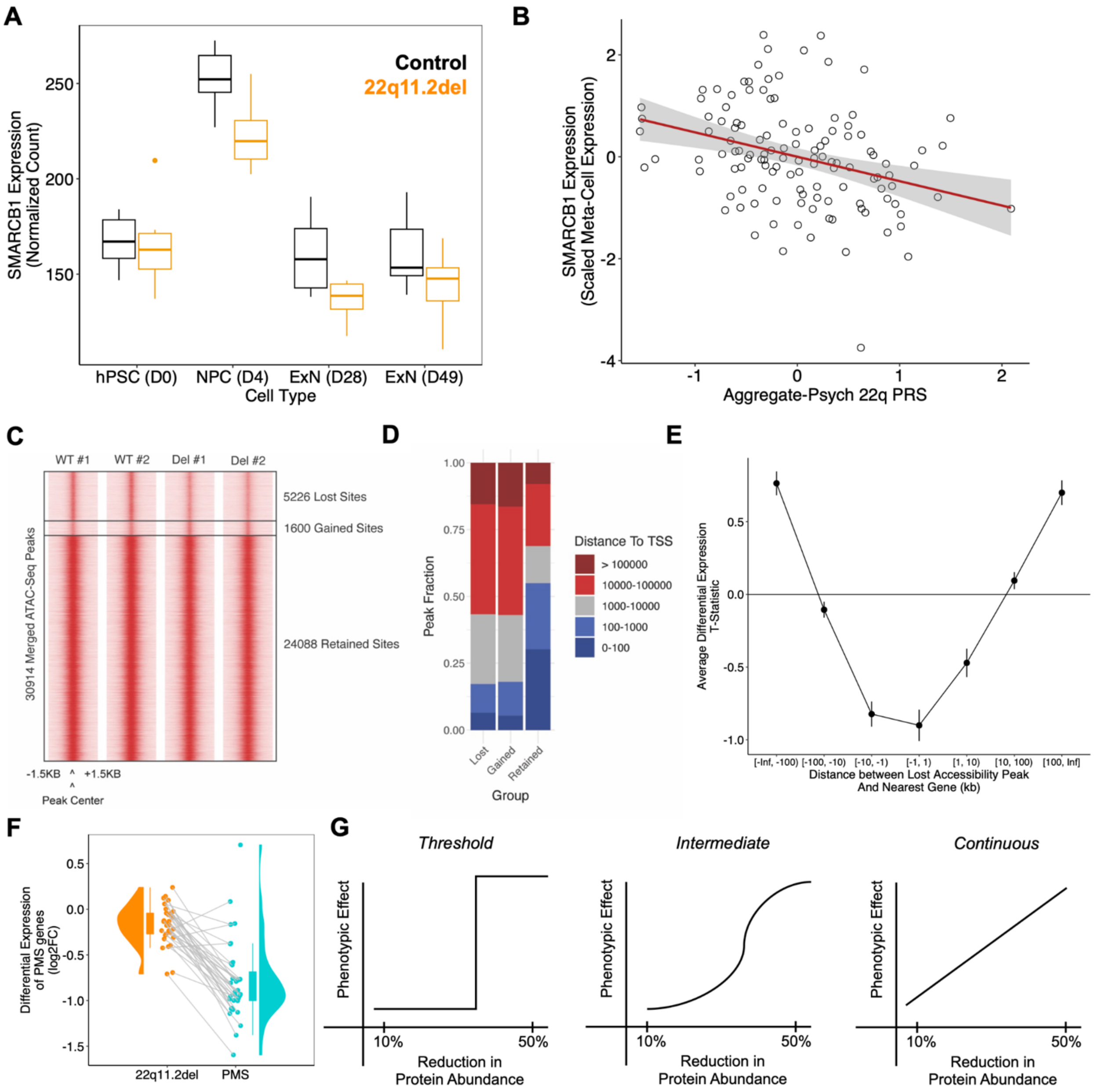


Figure 7.9. Effect of 22q11.2del on SMARCB1 and Phelan-McDermid Syndrome Genes.

(A) Differential expression of SMARCB1 due to 22q11.2del in isogenic cell lines, across three cell types. Within ExN, we sampled two different time points, Day 28 (D28) and Day 49 (D49) of the differentiation protocol. (B) Association between glutamatergic neuron meta-cell SMARCB1 expression and Aggregate-Psych 22q PRS in snRNA-seq data from pooled, multi-donor sample of post-mortem human brains. (C) Heatmaps of chromatin accessibility in 22q11.2del cell lines and isogenic controls, for sites with significantly lost or gained accessibility in 22q11.2del and retained sites. (D) Distance to nearest transcription start site (TSS) for lost, gained, and retained accessibility peaks. (E) Differential expression due to 22q11.2del of genes near lost chromatin accessibility peaks, stratified by distance to nearest gene. (F) Differential expression of 22q13.3 region genes in 22q11.2del (orange) and Phelan-McDermid Syndrome (blue). (G) Models for relationship between gene dosage and phenotypic effect.

We hypothesized that sub-hemizygous downregulation of *SMARCB1* may impact genome-wide regulatory architecture in 22q11.2del. To assess this hypothesis, we performed sequencing of transposase-accessible chromatin (ATAC-Seq) in 22q11.2del and isogenic control induced NPCs. ATAC-Seq results revealed a genome-wide loss of 5226 accessible sites versus gain of 1600 accessible sites, i.e., a net loss of ~17% of accessible sites genome-wide (**Figure 7.9C**). Close examination of retained sites revealed a subtle decrease in accessibility, possibly reflecting broader accessibility downregulation that we are underpowered to detect (**Figure 7.10**). Lost sites were further from transcription start sites than retained sites, consistent with a selective loss of enhancers (**Figure 7.9D**). When we examined the genes closest to lost accessible sites, we found that genes near lost sites had preferentially down-regulated expression in 22q11.2del, consistent with these lost accessible elements having regulatory activity (**Figure 7.9E**). Taken together, these results point to genome-wide enhancer dysregulation in 22q11.2del, which is potentially explained by downregulation of *SMARCB1*.

7.6 - 22q11.2del and local polygenic risk decrease the expression of genes in the Phelan McDermid Syndrome region

In addition to 22q11.2del, chr22q is home to the microdeletion at 22q13.3, a causal deletion for Phelan-McDermid Syndrome (PMS). PMS is characterized by neonatal hypotonia, delayed or absent speech, developmental delay, and several dysmorphic features (Phelan & McDermid, 2011). In our 22q11.2del isogenic cell lines, we assessed whether 22q11.2del was associated with decreased average expression of genes in the PMS region, which are separated from the deletion by approximately 30 Mb. We found that 22q11.2del was associated with a significant, 10.5% decrease in PMS region genes (p = 0.001; **Figure 7.9F**); this change is similar to the mean downregulation across chr22q. To further contextualize this finding, we repeated our analysis in a similar resource of cell lines carrying the PMS-associated 22q13.3 deletion (Breen et al, 2020). We found that 22q13.3 deletion caused a 58% expression reduction in PMS region genes (i.e., close to the expected 50% decrease due to hemizygous deletion); this implies that 22q11.2del exerts 24.8% of the PMS-associated change in expression of these genes. This finding raises the possibility that genes at 22q13.3 may in part explain the phenotypic effects of 22q11.2del.

PMS is unlike 22q11.2del in that a putative causal driver gene, *SHANK3*, has been identified. Individuals with protein-truncating variants in *SHANK3* recapitulate symptoms of PMS (Bonaglia et al, 2001; Durand et al, 2007). 22q11.2del was not associated with significantly decreased *SHANK3* expression in our dataset, a finding that we confirmed with qPCR (**Figure 7.10B**). However, individuals with interstitial 22q13.3 deletions excluding *SHANK3* can also manifest the classic PMS phenotype (Wilson et al, 2008), and individuals with large deletions at 22q13.3 have been found to have more severe neurodevelopmental delay than individuals with single deletion of *SHANK3* (Levy et al) These observations suggest that genes in this region other than *SHANK3* may contribute to the 22q13.3del phenotype, and could contribute quantitatively smaller phenotypic impacts in 22q11.2del.

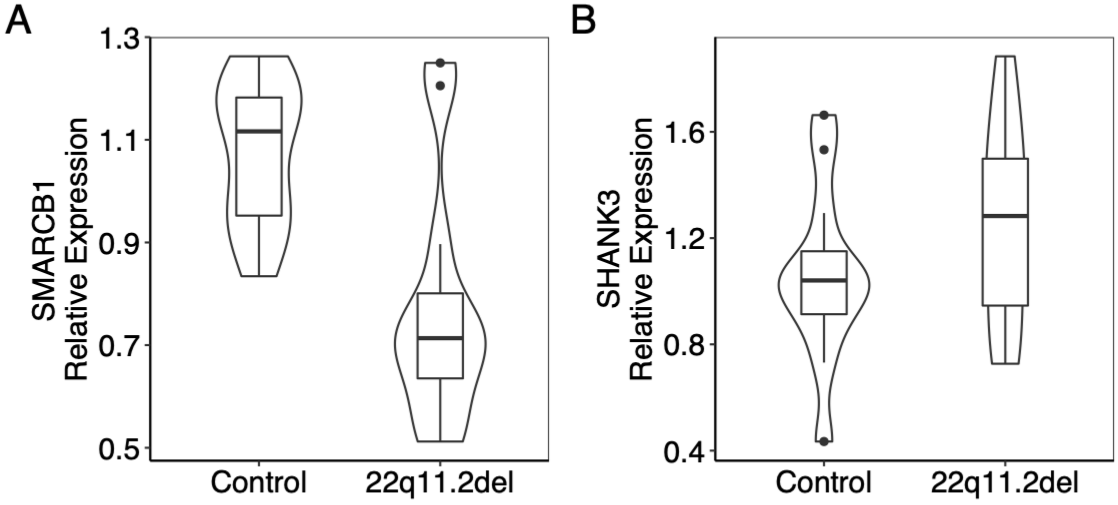


Figure 7.10: qPCR validation for SMARCB1 and SHANK3.

qPCR was performed in induced neural progenitor cells differing only in 22q11.2del induced with CRISPR-Cas9. (A) Replication of significant finding for SMARCB1 (B) replication of null finding for SHANK3.

These findings raise a critical, yet unanswered question in human genetics: given that a gene/region is associated with a phenotype via rare, damaging hemizygous mutations, what is the expected effect of a small decrease in expression of that gene/region? In a threshold model, such small effects induce phenotypes in a binary manner depending on the effect size and threshold, implying buffering of gene regulatory networks against smaller changes in gene dosage (Naqvi et al, 2023). In a continuous model, small effects induce quantitatively smaller phenotypic changes, implying that there is little to no such buffering (**Figure 7.9G**). Although which model is correct probably varies between genes, the generally observed convergence between phenotypic associations of rare loss-of-function mutations and common, mostly non-coding SNPs suggests that a continuous model is accurate for many genes (Backman et al, 2011). If the continuous model applies to *SMARCB1*, genes in the PMS region, or other yet to be identified causal genes on chr22q, the subtle expression changes we observe likely contribute at least partially to the 22q11.2del phenotype.

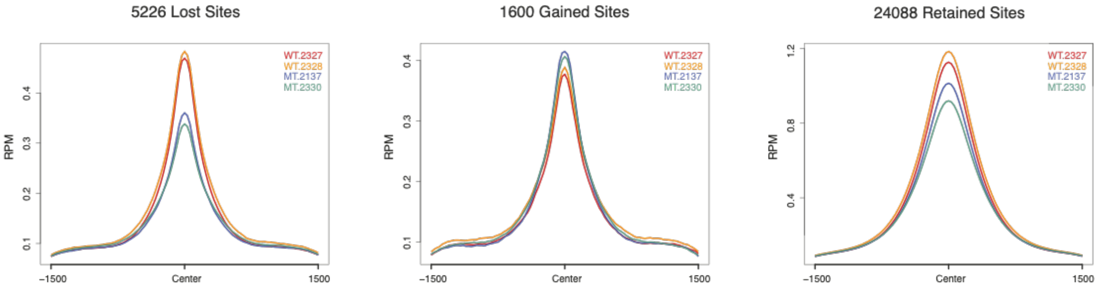


Figure S7.10: Lost, Gained, and Retained Accessibility Peaks in 22q11.2del.

7.7 - 22q11.2del induces dispersed effects on gene expression across the transcriptome

Genes are connected by dense cellular networks that describe regulatory, physical, and functional relationships (Barabási et al, 2010). We hypothesized that by inducing gene expression changes across chr22q, 22q11.2del may strongly perturb these networks, and cause dispersed expression up- and down-regulation of many or even the majority of genes across the entire transcriptome. For example, by decreasing expression of *SMARCB1* and genes in the PMS region, 22q11.2del likely indirectly induces changes in the expression of many genes across the transcriptome (**Figure 7.9E**). Decreased expression of other genes on chr22q may also lead to such distributed effects. Such changes, in aggregate, could potentially cause major changes in cellular physiology, analogously to how small-effect SNPs across the genome can collectively cause severe diseases (Boyle et al, 2017). We operationalize this "transcriptome-wide impact" as transcriptome-wide overdispersion of differential expression statistics and estimate this overdispersion across multiple cell-type contexts and neuropsychiatric-associated CNVs (**Figure 7.11A**)

We applied this method to gene expression data from isogenic induced neural stem cell and excitatory neuron cell lines with the following disease-associated deletions induced via CRISPR-Cas9: 22q11.2 deletion, 16p11.2 deletion, 15q13.3 deletion. While these deletions all raise risk for neurodevelopmental disorders, they vary in features such as effect size, deletion size, and relative phenotypic preference among schizophrenia, autism, and intellectual disability. For these analyses, we excluded genes within deleted regions for each copy number variant.

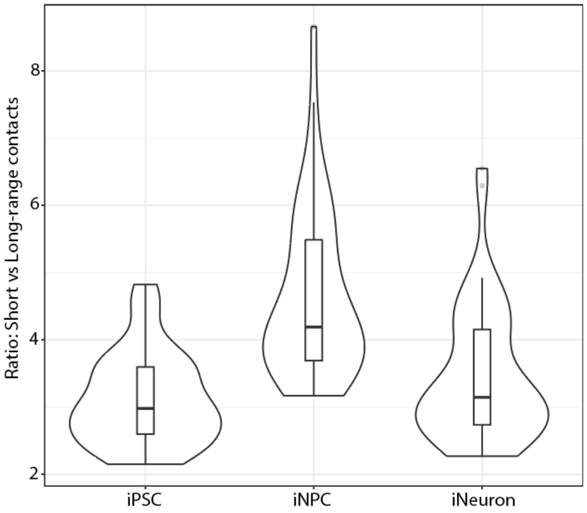


Figure S7.11: Short vs long-range HiC contacts across cell types

We found that 16p11.2del, 15q13.3del, and 22q11.2del induced significantly over-dispersed genome-wide effects in neural progenitor cells and induced neural stem cells (**Figure 7.12B**). Intriguingly, we additionally found that 22q11.2del was the only deletion examined that induced significant genome-wide overdispersion of differential expression effects in excitatory neurons (**Figure 7.12B**). These results suggest that, at a transcriptome-wide level, neural progenitor cells are more sensitive to these perturbations than excitatory neurons. This, in part, may be explained by the 3D architecture of the genome at different cell stages, where the ratio of short to long range contacts are increased during the neuronal progenitor stage when compared to terminally differentiated neurons (**Figure 7.11**). Furthermore, among the three deletions, 22q11.2del appears to have uniquely strong effects in mature excitatory neurons (**Figure 7.12B**). This differential expression overdispersion in 22q11.2del persisted when removing all chr22q genes and all significant differentially expressed genes (**Figure 7.13**), suggesting that the overdispersion is pervasive across the transcriptome. Additionally, we observed this overdispersion when limiting to genes that are intolerant to loss-of-function variation in the general population (**Figure 7.13**; Karczewski et al, 2020), which are generally sensitive to both increases and decreases in dosage (Collins et al, 2022). These dispersed effects could cause changes in cellular state, or hamper developmental transitions between cell types, in a manner that may explain phenotypes associated with 22q11.2del. Additionally the extent to which these indirect effects are similar across individuals with 22q11.2del, modulated by genetic background, or impacted by chance, is unclear. Variability of dispersed indirect effects, in which genes are affected or the degree of effect, could potentially explain the remarkable phenotypic heterogeneity that is observed in 22q11.2del, from fetal lethality (Maisenbacher et al, 2017) to minimal morbidity and inclusion in the UK Biobank (Crawford et al, 2019). Future experiments should use large samples of 22q11.2del carriers varying in phenotype to assess these questions.

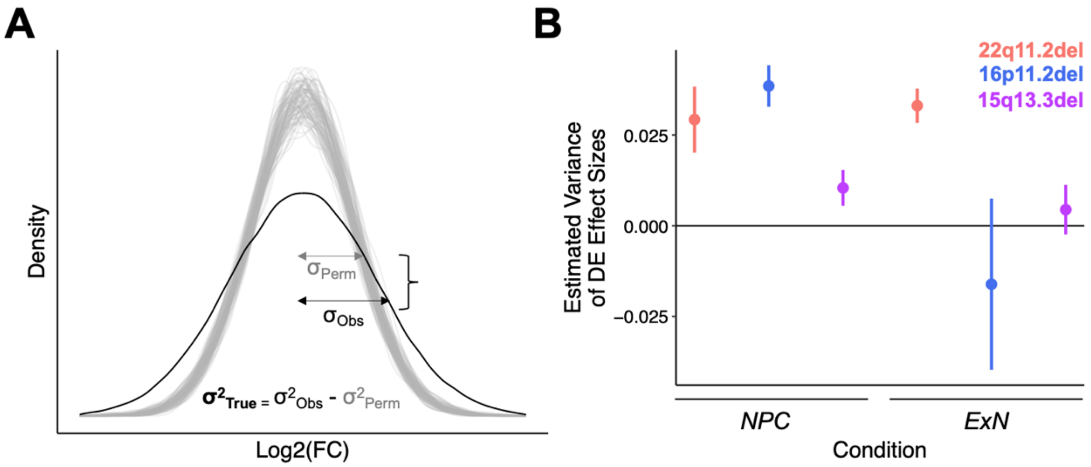


Figure 7.12. Overdispersion of Transcriptome-wide Differential Expression Effects in 22q11.2del and other Genomic Disorders.

(A) Approach to quantifying transcriptome-wide overdispersion of differential expression effects. Observed effects are shown in black, and effects from many permutations of condition labels are shown in grey. The variance of the true effects is the difference between the variances of the observed and permutation distribution. (B) DE effect variance (e.g., overdispersion) estimates for three genomic deletions, in two cell type contexts.

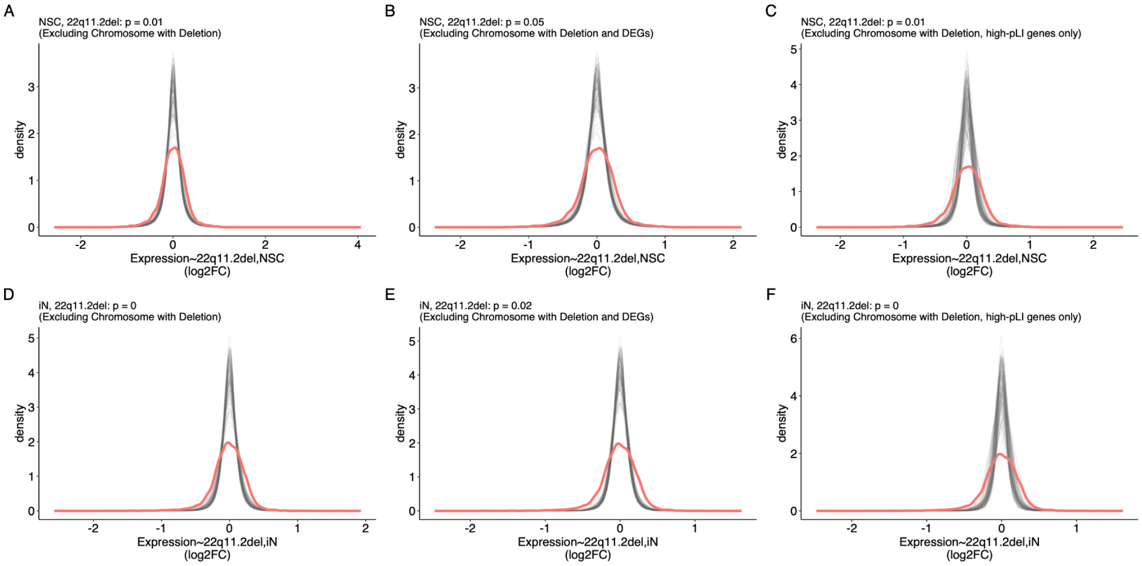


Figure 7.13: Transcriptome-wide impact with gene subsets.

Analysis depicted in Figure 7.4, performed in both NPC and ExN, with different subsets of genes.

7.8 – Discussion

We have demonstrated that genetic variation at 22q induces unusually dispersed effects on gene expression, in a manner influenced by the unusually active, physically interacting architecture of this genomic region. These findings have several implications for the mechanistic study of disease-associated copy number variants, convergence between common and rare genetic risk factors, and genome biology more broadly.

Firstly, our findings offer substantial new traction in the study of 22q11.2del. Although this variant has been subject to decades of intensive study, the search for a causal gene for its psychiatric manifestations has been equivocal, leading many to hypothesize there is not a single driver gene at this locus at all (Motahari et al, 2019). Rather than focusing on individual genes, we integrated 3D genome conformation data to reveal that 22q11.2del broadly decreases gene expression on chr22q. This previously unappreciated phenomenon opens the door to qualitatively novel mechanistic hypotheses regarding 22q11.2del. For example, we find that the protein level of *SMARCB1*, a causal gene for Coffin-Siris syndrome and key regulator of enhancer architecture, is diminished in 22q11.2del, with measurable impact on genome-wide chromatin accessibility. We additionally identify a subtle decrease in expression of genes in the Phelan-McDermid Syndrome region (22q13.3; but notably without a decrease in expression of *SHANK3*), suggesting a potential shared basis of these two disorders, whose deleted regions are separated by tens of megabases. These particularly interesting novel candidates, as well as the larger group of perturbed genes, could potentially yield considerable insight into the pathogenic mechanism of 22q11.2del. Our findings expand on existing literature characterizing individual candidate genes at 22q11.2 (e.g., *DGCR8* in Khan et al, 2020, Forsyth et al, 2021 and *TBX1* in Nehme et al, 2022, among many others), and propose that phenotypic effects of 22q11.2del likely arise from a combination of deleted gene effects and the presently described distal expression dysregulation. We note that while we have focused on brain-related resources in this study, similar investigations in other tissues affected by 22q11.2del (developing heart, thymus, craniofacial structures, among others) may be similarly informative.

Secondly, we find that the effects of common variants associated with neuropsychiatric traits have similarly dispersed associations with gene expression. This effect size dispersion suggests that in areas of high distal intrachromosomal contact, proximal cis-genes do not fully capture the effects of common risk variants. Additionally, the observed correlation between expression associations of 22q11.2del and polygenic risk suggests that lower average expression of all genes across chr22q is a risk factor for diverse neuropsychiatric conditions. We previously identified a similar phenomenon at chromosome chr16p (Weiner et al, 2022), with the autism-associated 16p11.2del and local polygenic risk for autism exerting dispersed, correlated effects on expression across 16p. Identification of a similar phenomenon at chr22q raises the question of how pervasive such regional effects are, and their broader relevance to neuropsychiatric diagnosis. Indeed, chr16p and chr22q are not the only genomic hotspots of physical interaction and transcriptional activity (Rao et al, 2014); our focus on these regions largely owes to the fact that they both contain high sequence identity segmental duplications (Vollger et al, 2022), leading to recurrent, large genomic deletions (e.g., 16p11.2del and 22q11.2del) with well-developed data resources. Future efforts should query the influence of genomic compartment architecture on genetic variant effects genome-wide and use genome editing technology to introduce and study large deletions across the genome, including areas with no recurrent copy number variation. We speculate that such studies may reveal a genome-wide connection between highly active, physically interacting compartments and neuropsychiatric traits, based on our observations in the schizophrenia, autism, and ID associated 16p11.2del and 22q11.2del.

More broadly, our results draw a new connection in genome biology, between the low-dimensional architecture of the 3D genome and the effects of phenotype-associated genetic variation. While the low-dimensional compartment architecture of the genome has been observed at least since the invention of HiC (Lieberman-Aiden et al, 2009,), usage of 3D genome conformation in variant interpretation, and variant-to-function analysis, has been largely limited to proximal interactions between genes and nearby regulatory elements (Nasser et al, 2021). Here, we demonstrate that broader patterns of 3D genome conformation shape the influences of genetic variation. The finding that variants may regulate genes that are distal in sequence space, but relatively more proximal in 3D space, has consequences for the many methods in statistical genetics that assume variant effect decays with genomic distance, including estimators of expression quantitative trait loci (eQTLs; GTEx Consortium 2016) and gene-set tests for GWAS (de Leeuw et al, 2015). Notably, this model, where compartment architecture facilitates variant effects across large genomic distances, is supported by the recent observation that even the 11th-through-20th furthest genes from GWAS associations mediate significant fractions of variant heritability (Weiner et al, 2022).

Lastly, our work demonstrates that the dispersed effects of genetic variation in richly interacting regions can induce transcriptome-wide dysregulation of gene expression. The biological consequences of such transcriptome-wide effects are difficult to predict, but we suspect that this dysregulation may compromise cell-state and cell-type-specific function, as well as normative developmental transitions between cell-types. Methods that focus on individual genes will likely be insufficient to disentangle these effects. We hypothesize that as in GWAS (Maier et al, 2017), methods that analyze effects genome-wide will be necessary to meet the challenge of variant effect interpretation.

7.9 – Limitations & Non-replication

Several experiments related to Chapter 7 were repeated to validate our initial findings, and to explore more specific biological findings such as the dysregulation of SMARCB1. To validate our initial observations, I generated several new datasets. First, I generated new bulkRNA-sequencing data on NPCs derived from our isogenic 22q11.2 deletion lines. With this data, we were unable to replicate two key findings. We did not observe the cis-extended effect (whereby deletion of 22q11.2 impacts genes along Chr22q) nor could we detect reduced expression of SMARCB1. Next, I generated scRNA-seq data on iPSC-derived neurons from the discovery cohort, including 30 control samples and 21 samples with 22q11.2 deletions. With this data, we also were unable to replicate the cis-extended effect.

I am currently performing more validation experiments to explore whether we can reliably recapitulate our initial observations. The lack of replication could be due to several factors. These could be technical, and the RNA-seq assays which I have used in the more recent data generations are different from those utilized in the data generation for our 2022 paper. The method in which we generate our cells is also different. In our previous paper, we use lentivirus to introduce our inducible Ngn2. Currently we integrate Ngn2 into the genome using electroporation. While the cells we produce with the integrated construct are as, if not more, mature than the lentiviral counterpart, we cannot rule out that this may have subtle effects. This may especially be the case in our cis-extended analysis where the per-gene effects are very low. There is, of course, the possibility that our initial results are false. For this reason, I have performed several additional validation experiments to confirm whether this is the case. If it is the case, we will have to think about other mechanistic explanations for our genome-wide observations and restructure our findings.

7.10 – Statement on Contributions

This work was co-led by Ajay Nadig. Ajay performed all computational analyses within the study. I generated all of the data within this study.

Chapter 8

Convergence of morphological features in 22q11.2 deletion and severe psychiatric disorders

8.1 – Introduction

Though severe psychiatric disorders are highly heterogeneous, common physiological differences within the brain have been identified Individuals with schizophrenia have altered connectivity in the brain and reduced dendritic arborization when compared to healthy individuals (Nejad et al., 2012; Glausier and Lewis, 2013). Brain volume abnormalities have also been implicated in schizophrenia; large-scale brain MRI scans from patients with schizophrenia, compared with healthy controls, have found reduced hippocampus, amygdala, and thalamus volumes, as well as several other regions of the brain (van Erp et al., 2016). Individuals at high risk for psychosis show a steeper decline in cortical thickness as psychosis develops. This decline has been observed prior to treatment with antipsychotics, demonstrating that early tissue loss is not caused by antipsychotics (Cannon et al., 2015). Bipolar disorder, with or without psychosis, is also associated with thinner cortical gray matter in the front, temporal, and parietal regions of both hemispheres of the brain (Hibar et al., 2017). Pharmacological medications, including lithium, antiepileptics, and antipsychotics, may cause changes in brain morphology, but this area of research remains highly debated (Huhtaniska et al., 2017).

Although numerous physiological differences have been observed in brain structures, there still remains questions as to whether morphological differences and phenotypes may be seen at the cellular level. If so, it would provide a tool whereby we could study morphological differences in living systems which we can perturb to attempt to rescue those differences.

There are many diseases in which cellular morphology serves as an indicator of disease. Most notably, in sickle cell disease, red blood cells have a sickle-like shape, rather than a circular shape. Changes to cell morphology can also correlate with disease stage. In patients with Nasal NK/T-cell lymphoma, p53 mutations were associated with larger transformed cells and presented with more advanced stage disease (Quintanilla-Martinez et al., 2001). Changes to cell morphology are also not limited to physical disorders; reactive astrogliosis is a common feature observed in many neurological disorders, such as Alzheimer's disease and Huntington's disease. Recent evidence suggests, however, that subtle changes to astrocyte morphology precede reactive astrogliosis and may drive disease progression (Zhou et al., 2019). Identifying microscopic differences in morphology may then provide insight into the underlying mechanisms of disease and pathology.

In psychotic disorders, some morphological differences have been observed at a cellular level. Reduced synaptic density and dendritic spine alterations have long been implicated in disease pathology (Feinberg, 1982; Glausier and Lewis, 2013). Subcellular features, such as organelle arrangement, have not been systematically investigated in an unbiased approach. If morphological phenotypes of disease can be defined at a cellular and subcellular scale, then morphological profiles could serve as a tool for testing drug treatments which may rescue phenotypes of disease.

Following on our functional investigation on Chr22q which highlights shared risk factors for psychosis between rare and common variant risk in this locus, I reasoned that we may be able to identify common cellular phenotypes between rare and common variants cases of psychotic disorders such as schizophrenia and bipolar disorder. Here, I developed Neural Cell Painting and applied this to a cohort of 65 cell lines of which 24 are affected with neuropsychiatric conditions and carry the 22q11.2 deletion, and 12 are affected with neuropsychiatric conditions with no known genetic risk variants. We identify morphological features which are altered between cases and control in both rare and common variant cell lines. Further, we find several morphological features which are altered in both conditions, suggesting common cellular pathology occurring from both rare and common genetic risk for psychiatric disorders.

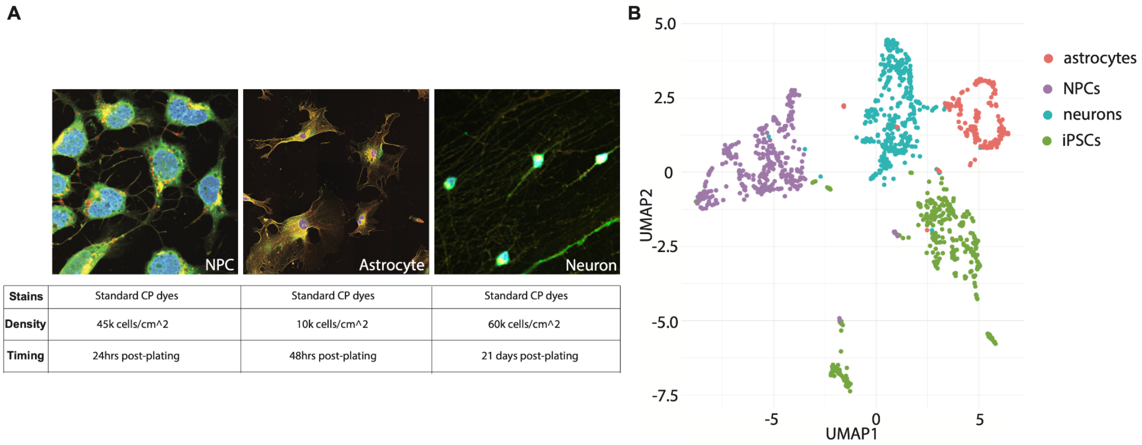
8.2 - Neural Cell Painting

To analyze cellular morphology, I established the Cell Painting assay for three different iPSC-derived neural cell types (NPCs, neurons, and astrocytes) and optimized for seeding density and fixation at the desired developmental timepoints. NPCs were seeded into 384-well imaging plates coated with Poly-d-lysine and laminin. It was necessary to use Poly-d-lysine and laminin coating for imaging, as Geltrex is a basement membrane-like product and absorbs the stains used in the Cell Painting assay, which causes a high degree of background fluorescence. Cells were stained on the final day post-induction; iPSCs on D0, NPCs were fixed at D4, excitatory neurons co-cultured with mouse glia were fixed at D28 upon terminal differentiation, and astrocytes were fixed at D30 upon terminal differentiation (**Figure 8.1A**).

Cells were stained using the Cell Painting assay developed by Bray et al. (2016) using an optimized version by Cimini et al. (2022). To stain for the cellular and subcellular features, the following stains were used: Hoechst 33342 (DNA), WGA (golgi and plasma membrane), concanavalin A (endoplasmic reticulum), MitoTracker (mitochondria), SYTO 14 (nucleoli and cytoplasmic RNA), and phalloidin (actin). Representative images of iPSC-derived NPCs stained using the Cell Painting assay were taken using the Zeiss Airyscan 980 LSM (**Figure 8.1A**).

For feature extraction and data analysis, cells were imaged using the Opera Phenix high-content screening microscope to capture both wide-field and confocal images. 5 channels were imaged, and 9 fields of view (FOV) were captured per well, producing 90 total images per cell line per cell type. All neuronal cell types were successfully stained and imaged using this assay (**Figure 8.1A**). Collectively, these results have demonstrated the efficacy of employing Cell Painting in iPSC derived neural cell types from multiple donors to produce high-throughput and high-content image data for subsequent analysis.

Commonly, the use of dimensional reduction approaches such as PCA or UMAP are utilized to assess global structure in gene expression or other data methodologies. I was curious if exploring morphological data in a similar fashion would elucidate unique morphological features in various cell types. While this is easily ascertained by the human eye upon examining cells through brightfield images, a quantitative approach to understanding how cellular organelles are distributed in different cellular contexts, and whether this was altered in different genotypes, was needed. I found that analogous to gene expression data, morphological profiling data sufficiently delineated between the various cell types explored in our study, highlighting the ability to detect subtle differences in organelle behavior across differentiation lineages (**Figure 8.1B**).



Stains	Standard CP dyes	Standard CP dyes	Standard CP dyes
Density	45k cells/cm^2	10k cells/cm^2	60k cells/cm^2
Timing	24hrs post-plating	48hrs post-plating	21 days post-plating

Figure 8.1. Neural Cell Painting for investigating morphological signatures in psychiatric disorders.

A) We identified optimal conditions for cells at every stage. Right: Representative Cell Painting images from NPCs, astrocytes, and neurons, used to extract morphological features. **B)** UMAP embedding of morphological features showing that cell types can be successfully distinguished from one another using morphological features, analogous to separation based on RNA-sequencing.

8.3 - Morphological profiling in cell lines with 22q11.2 deletion syndrome

To understand how genetic variation associated with psychiatric disorders impacts cellular morphology, I performed Neural Cell Painting on a cohort of cell lines harboring a 22q11.2 deletion in addition to age and ancestry matched controls. These 48 cell lines were differentiated into NPCs, neurons, and astrocytes using published protocols, including those described in Chapter 4 (Nehme et al, 2018; Berryer and Tegtmeyer et al, 2023). Differentiated cells were plated into 384-well plates and matured until the desired time point before being stained with the canonical Cell Painting dyes (Cimini et al, 2023; Bray et al, 2016). After fixation and staining, cells were imaged on a Perkin Elmer Phenix high-content imaging system (**Figure 8.1A**). Once images were acquired, I used computational pipelines to extract morphological features from each cell and identify morphological features which distinguish affected cells from unaffected cells to investigate how the 22q11.2 deletion alters cell morphology across cell types **(Figure 8.1B).**

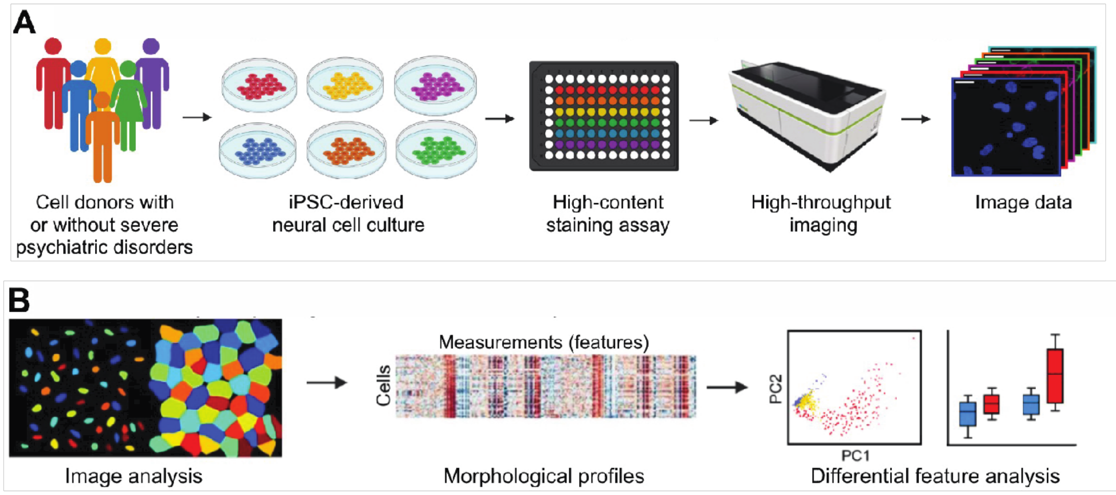

A
Cell donors with or without severe psychiatric disorders
iPSC-derived neural cell culture
High-content staining assay
High-throughput imaging
Image data
B
Image analysis
Measurements (features)
Cells
Morphological profiles
PC2
PC1
Differential feature analysis

Figure 8.2. Schematic study workflow.

(A) I used the Cell Painting assay to generate high-throughput and high-content image data from neural cell types derived from patients with or without idiopathic severe psychiatric disorders and those with 22q11.2 deletion syndrome **(B)** I analyzed images with CellProfiler, generate morphological profiles, then identify differential features between case and control cell lines to see if there are overlaps in which morphological traits are affected by the 22q11.2 deletion. Figure adapted from Bray et al. (2016).

8.4 - Morphological feature quantification

After cell lines were imaged, CellProfiler (4.2.4) was used to analyze and quantify the morphological features of the neural cell types. A single pipeline was created to standardize measurements across cell lines and cell types. Images were pre-processed to improve accuracy of morphological measurements. Microscopy optics can result in uneven illumination, so all images were processed using an illumination function to correct varied illumination and shading in images. Image intensities for each image were also calculated. Image quality was assessed by measuring saturation, intensity, and blur, which allowed us to identify any abnormalities.

After images were standardized and assessed for quality, the primary objects (nuclei) were segmented. The channel stained for DNA was enhanced to improve nuclei segmentation. A global thresholding method and minimum cross-entropy thresholding strategy was used to identify and segment the nuclei. Clumped objects were distinguished by shape. To segment the nuclei of D28 neurons co-cultured with glia, the size and intensity of the nuclei stain was used to differentiate the iPSC-derived neurons from mouse glia.

To identify the secondary objects (cell bodies), we used the channel stained for actin, Golgi body, and plasma membrane. These images were enhanced and intensified to improve cell body segmentation. Secondary objects were segmented using a global thresholding method and Otsu three-class thresholding strategy. Secondary objects were only segmented if a primary object was within, connecting each cell body with a nucleus (**Figure 8.3A**). Following segmentation of the nucleus and cell body, the tertiary object (cytoplasm) could be additionally segmented by subtracting the nucleus from the cell body.

The three objects–nucleus, cell body, and cytoplasm–were the foundation for measuring the individual morphological features between and within each cell. Each object was measured for size, shape, and total area. The segmented cell bodies were used to measure neighboring cells, including directly adjacent cells and cells within a given proximity. Measuring object neighbors served as an additional readout of cell density and clustering. To quantify neurites and cell branching, a morphological skeleton was created and seeded from the nuclei (**Figure 8.3B**).

The images containing the different cellular and subcellular features–the endoplasmic reticulum, cytoplasmic RNA, nucleoli, mitochondria, actin cytoskeleton, Golgi body, and plasma membrane–underwent a series of measurements to develop a highly granular morphological profile for each cell. The colocalization and correlation between intensities in different images were measured within objects. For example, the colocalization and correlation between the mitochondria and the nucleus of every cell could be explicitly measured and inform mitochondria location in the cell. Granularity, intensity, and texture of each channel was also measured in respect to each object. The process of measuring colocalization, granularity, and texture was applied to each channel and each object, generating an immense amount of data.

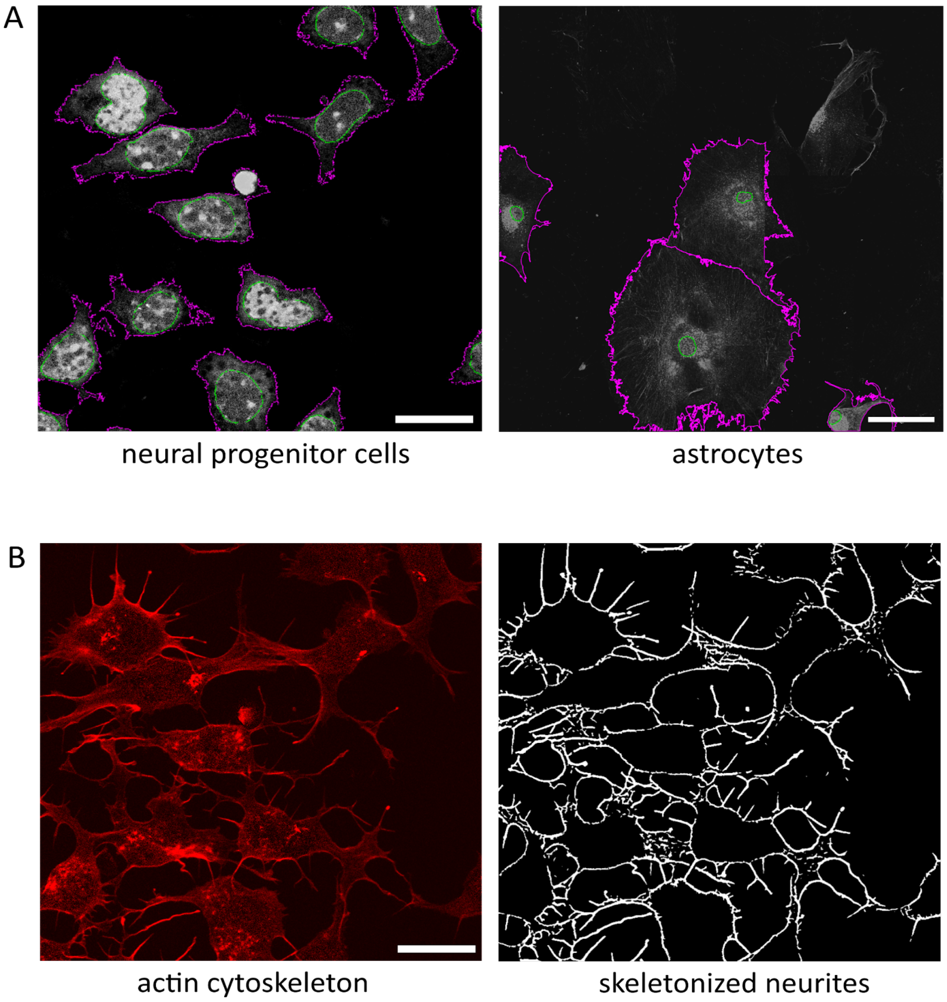


Figure 8.3. Automatic segmentation using CellProfiler.

(A) Outlines of the automatically segmented nuclei, shown in green, and cell body, shown in magenta, overlaid the endoplasmic reticulum channel in neural progenitor cells (left) and astrocytes (right). Left scale bar: 20 μm. Right scale bar: 100 μm. **(B)** Neural progenitor cells with skeletonized neurites identified using the actin cytoskeleton channel. Scale bar: 20 μm.

In total, 4347 features were measured for each cell, with an equal number of measurements made for each channel (**Figure 8.4A**). The complete list of feature types are as follows: area and shape, children, correlation, granularity, intensity, location, neighbors, number, parent, radial distribution, texture, and object skeleton. The majority of feature types measured were texture (64.6%), followed by correlation (10.4%), granularity (6.6%), and intensity (6.2%) (**Figure 8.4B**).

Overall, our approach enabled us to establish the Cell Painting assay in different neural cell types, and in cells from multiple donors, for the first time. This high-content and unbiased approach to image analysis was crucial in extracting information at a scale that would be impossible to quantify manually. Furthermore, these results demonstrate the viability of using CellProfiler to extract morphological features effectively and accurately from neural cell types stained with the Cell Painting assay. Once this data was obtained, we sought to identify morphological features that may differentiate between cases with 22q11.2 deletion syndrome and controls.

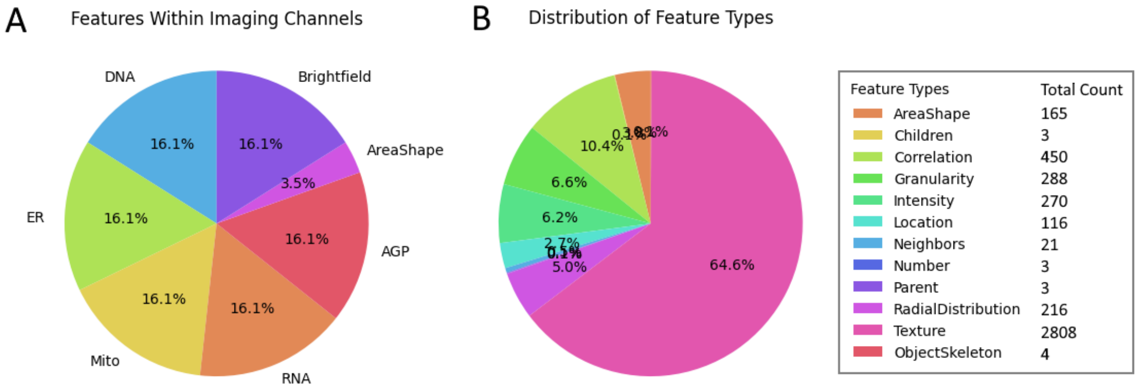


Figure 8.4. Morphological features extracted with CellProfiler.

(A) Percentage of features identified from each imaging channel. **(B)** Distribution of different feature types in the data. The total number of each feature type is displayed in the legend.

8.5 - Differential feature analysis in 22q11.2 deletion cell lines

To identify differential features across cell types and to find case/control differences, I calculated the per-well averages for each feature across cell lines and cell types. In order to perform differential feature analyses, it was critical to first confirm that profiles generated from individual replicate wells from the same cell line were highly correlated. To confirm the morphological profiles of each well within a cell line were consistent, I calculated the Spearman's correlation for each cell line. Results showed the correlation coefficient was very high across all cell lines for each cell type, suggesting within cell line profiles are highly similar. The average correlation coefficient for morphological profiles was 0.9940.

I observed that deletion and control lines clustered separately based on morphological features across the various cell types, indicating that we could reliably identify signatures which distinguish between wild-type and affected genotypes (**Figure 8.5A**). I next applied a two-sample t-test on a per feature basis with a p-value threshold of <0.001 to identify features which were significantly different between 22q11.2 samples when compared to controls. The total number of differentiating features identified for stem cells, NPCs, neurons, and astrocytes were 79, 31, 8, and 69, respectively. For example, in neurons derived from individuals with 22q11.2 deletion, there was a reduction in the granularity of mitochondria (*Cytoplam_Granularity_2_Mito*) when compared to healthy controls (**Figure 8.5B**). In astrocytes, we observed a change in the radial distribution of the Golgi apparatus (*Cells_RadialDistribution_RadialCV_AGP_3of4*) in cells from 22q11.2 deletion cell lines, suggesting the mutation resulted in changes in intracellular localization of the Golgi (**Figure 8.5B**). Interestingly, we observed no overlap between the significant differential features across all four cell types. However, many features impacted by the 22q11.2 deletion alter various morphological characteristics associated with the mitochondria across the diverse range of cell types (**Figure 8.5C**).

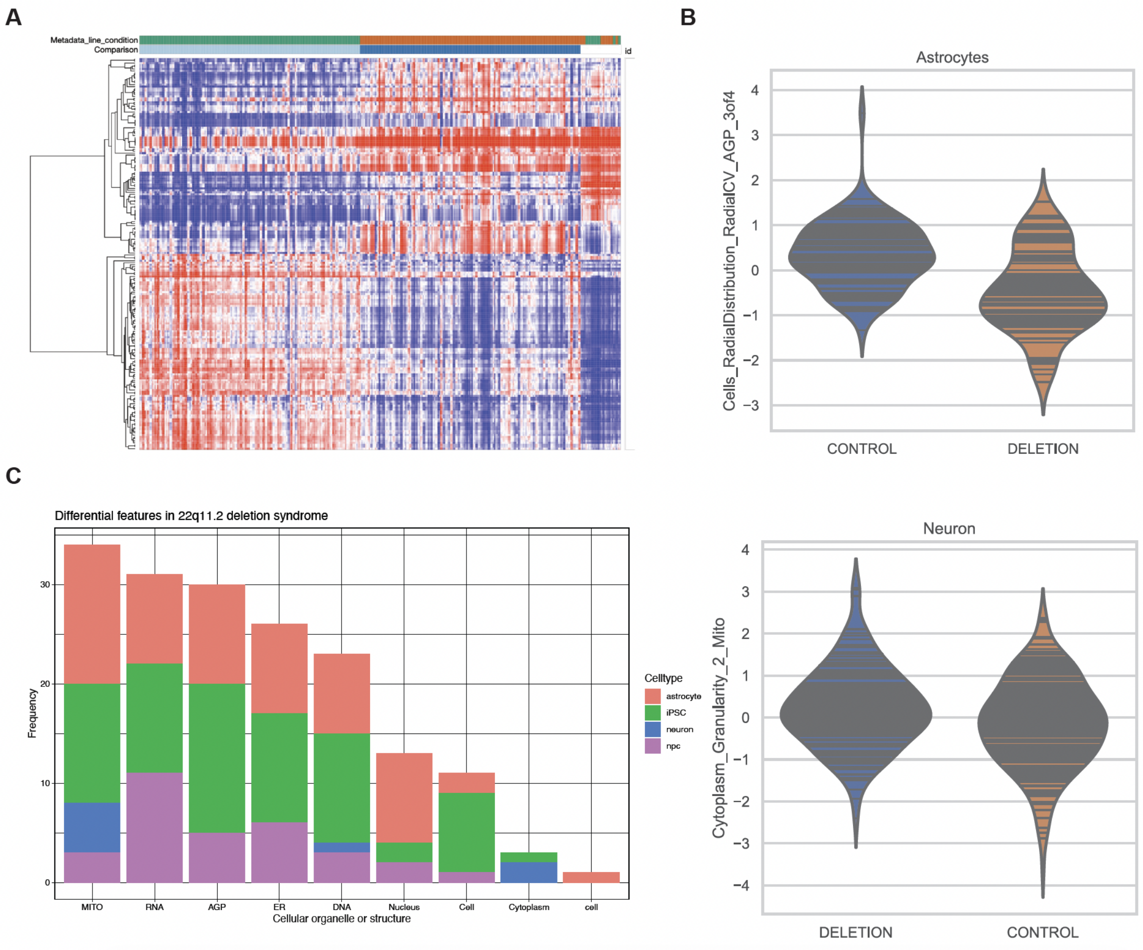


Figure 8.5. Differential feature analysis in 22q11.2 deletion cell lines.

A) We identified differential morphological signatures in samples with 22q11.2 deletion when compared to controls. **B)** Examples of differential features in neurons and astrocytes with 22q11.2del. *Each horizontal line represents an experimental well.* **C)** Organelles and structures enriched for differential features across cell types.

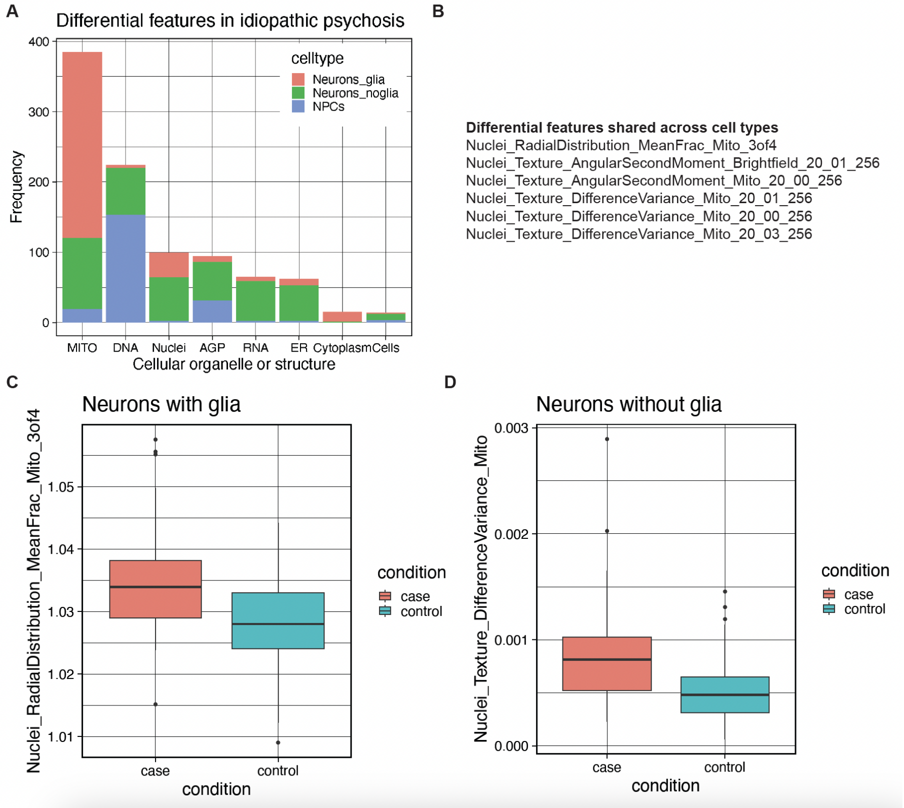
8.6 - Morphological profiling in cell lines from individuals with idiopathic psychosis

After identifying differential features in the 22q11.2 dataset, I wondered whether we could observe similar morphological deficits in cell lines from patients diagnosed with psychiatric phenotypes but where no highly penetrant genetic variant has been identified.

To do so, I repeated Neural Cell Painting on a new cohort of cells from individuals who have been diagnosed with schizophrenia or bipolar disorder along with age and ancestry match controls. Morphological profiles were generated across NPCs, neurons without mouse glia, neurons with mouse glia, and astrocytes. The feature analysis in the astrocyte data is ongoing and is not included in this thesis.

8.7 - Differential feature analysis in idiopathic psychosis cell lines

I identified differential features between cases and controls for each cell type as I had done in the 22q11.2 data set. Using a two-sample t-test on a per feature basis with a p-value threshold of <0.001, the total number of differentiating features identified for NPCs, neurons without glia, and neurons with glia were 212, 403, and 343 respectively.



Differential features shared across cell types
Nuclei_RadialDistribution_MeanFrac_Mito_3of4
Nuclei_Texture_AngularSecondMoment_Brightfield_20_01_256
Nuclei_Texture_AngularSecondMoment_Mito_20_00_256
Nuclei_Texture_DifferenceVariance_Mito_20_01_256
Nuclei_Texture_DifferenceVariance_Mito_20_00_256
Nuclei_Texture_DifferenceVariance_Mito_20_03_256

Figure 8.6. **Differential feature analysis in idiopathic psychosis cell lines.**

A) Organelles and structures enriched for differential features across cell types. B) Features which are significantly different across all cell types. C) Boxplot for Nuclei_RadialDistribution_MeanFrac_Mito_3of4 in neurons cultured with mouse glia. D) Boxplot for Nuclei_Texture_DifferenceVariance_Mito in neurons cultured without mouse glia.

Similarl to what I had observed in the 22q11.2 deletion samples, many features altered in cell lines derived from individuals with psychiatric phenotypes were associated with the mitochondria across the diverse range of cell types (**Figure 8.6A**). However, unlike the samples with 22q11.2 deletion, we did find several features which were significantly different across each of the cell types (**Figure 8.6B**). These features were:

Nuclei_RadialDistribution_MeanFrac_Mito_3of4,
Nuclei_Texture_AngularSecondMoment_Brightfield_20_01_256,
Nuclei_Texture_AngularSecondMoment_Mito_20_00_256,
Nuclei_Texture_DifferenceVariance_Mito_20_01_256,
Nuclei_Texture_DifferenceVariance_Mito_20_00_256,
Nuclei_Texture_DifferenceVariance_Mito_20_03_256

The differential features of mitochondrial texture, homogeneity, and radial distribution relative to the nucleus across cell types indicate broad changes in the distribution of mitochondria in the neural cell types of individuals with severe psychiatric disorders (representative boxplots shown in **Figure 8.6C, D**). Intriguingly, associations between mitochondrial abnormalities and psychiatric disorders have been described in the literature. These abnormalities have included oxidative stress, altered Ca^{2+} homeostasis, and energy deficits (Srivastava et al., 2019).

These features are consistent with those implicated in our 22q11.2 case and control data highlighted above, suggesting that rare and common variant risk for psychiatric phenotypes may converge on mitochondria pathology, which is a long-standing observation in schizophrenia research (Cataldo et al., 2010; Roberts, 2017). My findings provide new evidence for morphological abnormalities of the mitochondria in individuals with psychiatric disorders.

8.8 – Discussion

To understand how complex psychiatric disorders impact cellular function, I generated image-based profiling data on a cohort of cell lines derived from patients with a range of severe psychiatric phenotypes (both rare and common variant cases) alongside age and ancestry matched controls. These results show that as early in development as neuronal progenitor cells, we can detect meaningful differences in cell morphology, and these differences are maintained throughout neuronal differentiation into mature excitatory neurons and astrocytes.

The discovery of shared differential features in mitochondrial morphology across iPSC-derived neural cell types is highly interesting in the context of psychiatric disease. Neurons are highly dependent on mitochondrial functions, which include ATP-production, intracellular Ca^{2+} signaling, reactive oxygen species homeostasis, and regulation of synaptic activity (Zimmermann, 1994; Babcock and Hille, 1998; Massaad and Klann, 2011; Li et al., 2004). As such, mitochondrial dysfunction has long been associated with a wide range of psychiatric disorders (Kung and Roberts, 1999; Daniels et al., 2020). These mitochondrial deficits are not exclusive to idiopathic cases of severe psychiatric disorders. We observed similar mitochondrial traits altered in the presence of the 22q11.2 deletion. The 22q11.2 region also encodes for 9 proteins that affect mitochondrial function (Maynard et al., 2008; Napoli et al., 2015). iPSC-derived excitatory neurons from patients with both the 22q11.2 deletion and a diagnosis of schizophrenia exhibited mitochondrial deficits (Li et al., 2019). The similarities in disease phenotype between rare genetic variants and idiopathic cases of severe psychiatric disease indicate there may exist shared mechanisms of disease.

Severe psychiatric disorders have been linked with mitochondrial deficits in patients and postmortem samples (Cataldo et al., 2010; Roberts, 2017). Among schizophrenic patients, no singular mitochondrial phenotype of disease in the brain has been observed, as the anatomical differences appear to depend on the region and cell type (Roberts, 2017). In oligodendrocytes, changes to mitochondrial size, shape, and number have been observed in patients with schizophrenia (Uranova et al., 2007; Vikhreva et al., 2016). In the basal ganglia, significantly fewer mitochondria per axon terminal have also been observed in patients with schizophrenia. (Kung and Roberts, 1999).

Mitochondrial deficits have also been observed in multiple *in vitro* models of severe psychiatric disorders. For example, forebrain-like NPCs derived from individuals with schizophrenia display mitochondrial damage and oxidative stress (Brennand et al., 2015). Decreased mitochondrial membrane potential has been observed in iPSC-derived glutamatergic neurons from patients with schizophrenia. Furthermore, iPSC-derived dopaminergic neural precursor cells from the same patients display abnormal intracellular mitochondria distribution (Robicsek et al., 2013). Hippocampal dentate gyrus-like neurons derived from iPSCs of patients with bipolar disorder also exhibit mitochondrial dysfunction and changes to mitochondrial morphology (Mertens et al., 2015). The multiple studies cited here provide compelling evidence for mitochondrial deficits in psychiatric disorders but are limited by the small number of patient cell lines used. As such, these findings may not be considered conclusive. My observations on changes to mitochondrial morphology provide further support for mitochondrial deficits across different neural cell types at multiple stages of development in people with severe psychiatric disorders.

8.9 – Statement on Contributions

I performed all experiments and analyses within this Chapter with some assistance from Kathryn Boit and Dhara Liyanage. Image processing was performed by the Imaging Platform at the Broad Institute.

Chapter 9

Discussion and future directions

9.1 - Discussion

9.2 – Conclusions

Since their discovery more than 15 years ago, human induced pluripotent stem cells have provided an exciting new tool for being able to model human genetics in a living system. To take full advantage of this genetic variation, it has become evident that sample sizes for in vitro studies must climb to match pace with rapid advances in human genetic studies (Nehme et al, 2022, Schrode et al, 2019, Brennand et al, 2011). Further, it is necessary to continue to adopt and develop new methods and approaches which enable scientists to take full advantage of these tools. I have set out to develop, adapt, and utilize state-of-the-art methods to highlight the utility of human induced pluripotent stem cells for modeling human genetics to gain new insights in the biological consequences of what makes each of us different. I sought to do this by addressing several critical elements: (1) to optimize experimental approaches (cell-villages) to demonstrate the feasibility of performing population scale experiments including 10s-100s of genetically unique cell lines, (2) implement new imaging techniques for their use in human cellular models, specifically within the context of neurological and neuropsychiatric conditions, and (3) use the advances to investigate how rare and common genetic variations mediate cellular phenotypes across a range of contexts.

9.2.1 - Optimizing cell-village experiments

The cell-village approaches to scaling iPSC-based studies could constitute a key innovation in translational genomics. It presented the opportunity to overcome nearly all of the technical limitations to cellular studies including the costs and complexity of performing experiments with sufficient sample sizes, and the technical variation that often plagues large experiments which often leads to either under-discoveries or false positives. However, in the first several years following the development of these methods, we realized there were still tremendous hurdles which may intimidate many labs and altogether inhibit their adoption.

I have been able to overcome many of these hurdles and show that these types of approaches can be used effectively and simply, allowing individual scientists to culture 100s of cell lines at once in order to map genetic influences on cellular phenotypes. The first innovation was to make minor adjustments in how we treated the cells when in culture. Fewer mechanical manipulations prevented drastic differences in donor growth dynamics within an experiment. This enabled a fairly uniform distribution across many individuals cultured together so that downstream experiments weren't overcome by single cell lines taking over the environment. Next, I reasoned that cell-villages could be made in advance and cryo-preserved to be thawed and placed into experiments later on. This improvement has made it quite simple for other scientists to perform population-scale experiments without having sophisticated skills in stem cell and *in vitro* biology. This immediately reduces the challenge of having to thaw dozens to hundreds of cell lines every time one wishes to perform a large experiment. Lastly, with a focus on neuropsychiatric conditions, I showed how useful this approach was for differentiating large numbers of cell lines simultaneously. Moreover, it ensures that every single cell line is exposed to the same environment during the critical stages of differentiation, which provides a stronger basis for comparisons across donors.

As large iPSC collections continue to grow and become more widely available to researchers, I believe there will be broad adoption of cell-village type approaches in order to be intentional about large-scale experiments and phenotyping. It will be imperative to encourage their adoption and continued innovation to reduce the gap between genetic associations and our understanding of the biological consequences of such associations.

9.2.2 - Astrocyte-neuron interactions

Astrocytes make critical contributions to the formation, stabilization, and maturation of synapses throughout life. I am using experimental systems of iPSC-derived neurons and astrocytes to better understand how genes accomplish, and genetic variations affect, neuron-astrocyte interactions. My focus on neuron-astrocyte interactions at synapses is shaped by two sets of results. First, our groups previous analyses of human brain tissue from 191 donors uncovered tightly correlated transcriptional programs by which astrocytes and neurons couple their gene-expression investments in cholesterol biosynthesis genes (astrocytes), synaptic-adhesion molecules (astrocytes) and synaptic components (neurons); the astrocyte and neuronal gene-expression programs were both greatly enriched for schizophrenia-associated genes, and their expression was reduced in schizophrenia patients. Second, additional experiments on neuron-astrocyte co-cultures reveal many of these same relationships, and also provide a way to study their dynamics and genetics *in vitro*: upon co-culture with mouse glia, iPSC-derived neurons upregulate synaptic gene expression in a coordinated manner with increased expression of cholesterol-synthesis genes by mouse glia (Pietilainen et al, 2023).

Systems of living, interacting cells provide ways to study dynamic responses to acute perturbations of these and other pathways, and to learn what genes and alleles regulate these responses. I subjected iPSC-derived neurons and astrocytes to many types of pharmacological perturbations, including perturbations of glutamate receptors and L-type calcium channels, antipsychotic drugs (clozapine and haloperidol), oxidative stress, cytokine treatment, and modulation of cholesterol biosynthesis (Simvastatin, Atorvastain, and Efavirenz). I performed scRNAseq as an initial readout of these drugs' effects. Some of the most intriguing results so far (on which we focus below) involve modulations of the same cholesterol-biosynthesis pathway that is under-expressed in astrocytes in schizophrenia patients.

These experiments helped answer a question about the brain-transcriptomics results last year, which was whether the observed reductions in cholesterol-synthesis gene expression by astrocytes in schizophrenia patients were potentially just an effect of antipsychotic drugs or statins. We found that antipsychotics in fact oppose these changes: antipsychotics increased expression of cholesterol-synthesis genes in both neurons and astrocytes. This effect was common to all 43 donors in the village (i.e., it did not represent the eccentric response of an individual cell line).

Intriguingly, clozapine induced the expression of cholesterol-synthesis genes more robustly than haloperidol did. Clozapine's superior therapeutic efficacy has never been explained by affinity to dopaminergic and serotonergic receptors (which in fact is lower for clozapine) (Richtland et al, 2007). I am now seeking to understand whether an impact on CNS cholesterol could contribute to its efficacy.

To replicate and elaborate on these findings, I tested several other commonly prescribed antipsychotics (aripiprazole, quetiapine, risperidone, and olanzapine along with clozapine and haloperidol) across various treatment durations and doses. I replicated our initial findings of the effect of clozapine on the expression of cholesterol biosynthesis genes in a dose- and time-dependent manner and found that most other antipsychotic drugs elicited much-smaller or no responses. Among the other antipsychotics tested, only aripiprazole induced cholesterol-biosynthesis gene expression to even half the extent that clozapine did. (Like clozapine, aripiprazole is used in some cases that have been refractory to first-line antipsychotic drugs.)

In current experiments in astrocyte-neuron co-cultures and in mice, I am investigating whether clozapine also induces the changes in synaptic gene expression that characterize the Synaptic Neuron-Astrocyte Program in brain tissue, and whether it does so in a DRD2-independent manner. I hope with this work to better understand how the genes and alleles that are implicated in schizophrenia affect astrocyte-neuron interactions, and in particular the astrocyte-to-neuron cholesterol shuttle activity at synapses.

9.2.3 - Cell painting for modeling functional genomics

The morphology of cells is dynamic and mediated by genetic and environmental factors. Characterizing how genetic variation impacts cell morphology can provide an important link between disease association and cellular function. To understand if we could uncover genetic influences on cell morphology, we combined genomic and high-content imaging approaches on iPSCs from 297 unique donors to map what we term cell morphological quantitative trait loci (cmQTLs) (Tegtmeyer et al, 2023). Conventional methods for measuring cell morphology are often biased and follow discrete biological hypotheses, which restrict discovery power in large-scale unbiased studies. To overcome this, we utilized new innovations in imaging-based methods such as Cell Painting, which is a multiplexing dye assay (Bray et al, 2016, Cimini et al, 2023). This approach enables the simultaneous phenotyping of 8 cellular organelles and their relationships to one another to create a comprehensive map of cellular morphology. Leveraging Cell Painting on nearly 300 iPSCs, we identified novel associations between rare protein altering variants in *WASF2*, *TSPAN15*, and *PRLR* with several morphological traits related to cell shape, nucleic granularity, and mitochondrial distribution. Knockdown of these genes by CRISPRi confirmed their role in cell morphology. Analysis of common variants yielded one significant association and nominated over 300 variants with suggestive evidence ($P<10^{-6}$) of association with one or more morphology traits. Our results showed that, similar to other molecular phenotypes, morphological profiling can yield insight about the function of genes and variants and could be leveraged to explore the effect of risk variants implicated in psychiatric disorders on cellular morphological traits.

9.2.4 - Common and rare variant convergence in neuropsychiatric disorders

Strong impact variants offer a compelling opportunity to understand the biology of schizophrenia risk. 22q11.2del was the first such variant to be discovered, and the ~3 megabase deletion remains one of the strongest genetic risk factors for schizophrenia. While pleiotropy and incomplete penetrance are more rule than exception in psychiatric genetics, 22q11.2del is extraordinary in its heterogeneity of human impact. Individuals carrying the 22q11.2 deletion are at increased risk for several additional neuropsychiatric phenotypes, including intellectual disability (ID), autism, and attention deficit hyperactivity disorder (ADHD). They are also at risk for cardiac defects, immune dysfunction, cleft palate, endocrine dysfunction, and many other medical diagnoses. Few individuals with 22q11.2del will meet criteria for all these diagnoses. Some individuals with 22q11.2del do not have any apparent medical conditions and are unaware that they carry the variant.

The mechanisms through which 22q11.2del confers neuropsychiatric disease risk have long been cryptic. Dozens of labs have examined the genic content of the ~3 megabase (MB) 22q11.2del, but efforts to conclusively map the psychiatric effects to specific driver genes in the region have been equivocal (but see also Khan et al., 2020, and Nehme et al., 2022 for recent work nominating driver genes). The lack of such clear evidence for psychiatric driver genes stands in contrast to the cardiac phenotypes associated with 22q11.2del, which have been mapped to the gene TBX1, which sits within the canonical 22q11.2 deletion region. This perhaps implies a non-canonical mechanism links 22q11.2del to risk for schizophrenia and other neuropsychiatric disorders.

Over the last two years, we discovered that genomic neighborhoods characterized by uncommonly dense chromatin contact are highly relevant to neuropsychiatric disease. The q arm of chromosome 22 (chr22q) is the neighborhood maximally dense in both intra-chromosomal contact and active enhancers. In a previous effort, our colleagues discovered highly dispersed effects of the autism-associated microdeletion at 16p11.2 on regional gene expression, that were mirrored by effects of local polygenic autism risk, and potentially explained by the unusual density of physical intrachromosomal contact on chromosome 16p (Weiner et al., 2022). The only other genomic region with a higher density of intrachromosomal contact was chr22q (Figure 1), leading us to hypothesize that a similar non-canonical mechanism may explain the effects of this 22q11.2del.

We integrated several functional genomic and genetic association datasets to investigate risk factors for neuropsychiatric disease at chr22q. We found that the rare 22q11.2 deletion decreases expression of genes across chr22q, far beyond canonical deletion boundaries. This decrease is replicable in multiple brain cell types and was explained by the patterning of distal intrachromosomal contacts with the 22q11.2 deletion region. We additionally found that common, polygenic risk for schizophrenia, autism, ADHD, and lower IQ at 22q similarly decreased gene expression, in a manner convergent with the rare deletion effects. The convergent rare and common variant signals nominated specific genes (notably *SMARCB1*), gene-sets (22q13), and cell-types (excitatory neurons) in the biology of 22q11.2del syndrome and neuropsychiatric disease. They show a previously undocumented, highly consistent neurodevelopmental and psychiatric preference for collective upregulation of genes distributed across a vast region of the genome, the q arm of chromosome 22.

Among the non-deleted genes on 22q most affected by 22q11.2del, *SMARCB1* stands out as particularly interesting (22q11.2del-induced *SMARCB1* fold change = 0.72, via qPCR). *SMARCB1* is a key component of the BAF complex, which is essential for the function of active enhancers, genome-wide (Nakayama *et al.*, 2017). To explore whether 22q11.2del is associated with enhancer dysfunction, we performed ATAC-Seq in isogenic cell lines with or without a CRISPR engineered 22q11.2 deletion (Nehme *et al.*, 2022), and found a genome-wide net loss of accessibility peaks. These lost peaks were biased towards sites distal to transcription start sites, and genes near lost peaks had more negative differential expression, consistent with these lost peaks being regulatory elements (Figure 2). These findings suggest that 22q11.2del is associated with genome-wide dysregulation of enhancer architecture.

Motivated by these findings, we hypothesized that 22q11.2del would be associated with transcriptome- wide expression dysregulation. To test this hypothesis, we devised a simple permutation testing procedure to assess whether differential expression statistics are over-dispersed relative to null expectation. We found that 22q11.2del was associated with transcriptome-wide overdispersion of differential expression effects ($p < 0.01$), suggesting expression dysregulation that is not limited to a small set of genes. The observations are likely to create strong risk for developmental and psychiatric disorders, and to result in highly variable patient outcomes. We are now developing a series of collaborative projects to expand upon these qualitatively novel observations, understand this newly-discovered genomic model, and characterize the mechanisms through which 22q11.2 deletion syndrome creates risk for schizophrenia and other neuropsychiatric disorders.

I next applied Neural Cell Painting to a large cohort of stem cells derived from individuals carrying the 22q11.2 deletion (Nehme et al, 2022) as well as individuals with idiopathic psychosis. We generated high-dimensional morphological profiles from 60 cell lines from this cohort across four cell types; iPSCs, NPCs, neurons and astrocytes across controls, idiopathic cases, and cell lines harboring the 22q11.2 deletion (36 control, 10 idiopathic case, 24 22q11.2 deletion). We established computational pipelines to extract cell type-specific features, which are relevant for studying the unique biology in various cell states, and to delineate cell types from one another in a co-culture system. We found that cell types can be successfully distinguished from one another using morphological features, analogous to separation based on RNA-sequencing data. Excitingly, we identified differential morphological signatures in samples from patients with psychiatric disorders when compared to controls which show similarity to the features which distinguish samples with 22q11.2 deletion. For example, we found that mitochondria in neurons from idiopathic cases and 22q11.2 cell lines displayed higher granularity, suggesting oxidative stress phenotypes. Collectively, our findings demonstrate the ability of morphological profiling to characterize diverse cell types by their cellular features and identify differential traits in cells of different genotypes.

9.3 – Strengths

This work offers a range of technical, conceptual, and intellectual advances in the field of stem cell biology and genomics. Our study to investigate the genetic influences of cell morphology is, to the best of my knowledge, the largest iPSC-based study to date. This work also provides a framework for morphology-based studies of common genetic variation and will complement future work incorporating increased sample sizes and more disease-relevant cell types. We developed a new protocol to generate astrocytes from iPSCs. Our approach is simple, fast, and highly reproducible relative to other methods currently used for generating such cell types. While some other methods may produce overall more mature astrocytes, in order to make meaningful progress understanding human genetic variation, it's critical to have methods which are scalable across hundreds of genetically unique cell lines and that can be in a cost-efficient manner. Our work on chromosome 22q provides a valuable conceptual advance in how we think about and understand relationships between rare and common variants in neuropsychiatric conditions. Moreover, we provide a new gene-non-centric model for how copy number variants may be pathogenic. These findings suggest that the penetrance of rare variants may be dictated somewhat by the rest of an individual's genetic background, which may explain phenotypic heterogeneity observed in individuals with 22q11.2 deletion syndrome.

9.4 – Limitations

The scope of my work focused on developing, implementing, and utilizing various approaches for leveraging human derived stem-cells for modeling functional genomics. The main goals are to understand to what extent the genetic variation which makes each of us unique impacts brain function at the cellular level. There are many throat clearings one must make when drawing conclusions from stem-cell studies.

Stem-cells, while they retain the genetic signal of their donor, are amenable to various mutations which arise throughout their lifespan in culture which may act upon certain phenotypes (Kilpinen et al, 2017, Vigilante et al, 2019, Vickers et al, 2021). All cell lines utilized here in this study underwent diligent genotyping to ensure their genomic integrity, but it is not impossible for mutations to have arisen during a given experiment.

Our efforts to map morphology traits to genetic variation were only performed on undifferentiated stem-cells. It's becoming increasingly evident that disease associated variants are enrichment for cell types implicated in their respective disorders (Finucane et al, 2018). Therefore, it will be important for future work similar to ours to focus on alternative cell types in order to understand how human genetics influences cellular morphology more meaningfully. This work also requires stronger approaches to validating trait associations through either the inclusion of more cell lines not contained within the discovery cohort, or to use more advanced methods of genome editing such as base-editing.

9.5 - Future directions

9.5.1 - Increasing samples sizes for investigating how common variants mediate biological function

Much of the work described above incorporates many dozens to hundreds of cell lines. While these projects may seem well powered from an in vitro standpoint, they were sufficiently under-powered to make meaningful progress to elucidate the biological function to common variants implicated in human illnesses.

The next iteration of these works, in terms of both transcriptomic and morphological profiling, will need to include 500 or more genetically unique cell lines in order to capture the true effects of subtle variation. This can be readily accomplished as stem cell collections across the world continue to expand and provide substantial resources to scientists.

Expanding to so many cell lines will be met with its own technical and cost challenges. Particularly, as we observed in our cell morphology study, many technical factors confounded our morphological traits. Similar to our approaches for transcriptional profiling, cell-village methods are also amenable to image-based profiling. Current work has shown that "pooled-optical profiling" can be an effective tool for exploring how genetic perturbations impact cellular phenotypes (Feldman et al, 2019). This same method can be implemented for using many genetically unique donors through either fluorescent tag introduced into each unique cell line or more advanced in situ sequencing technologies which simultaneously capture imaging and sequencing data. This technique is currently cost prohibitive and technically challenging but will become increasingly less expensive and easy to implement in the near future and will be an important innovation to help scale many of these approaches. Additionally, expanding these approaches to many more cell types will be critical to understand how genetic variants act in cell type specific manners, which will be important for mapping their interpretations back to human health and disease.

9.5.2 - Progress towards fully humanized experimental systems for studying neuroscience

Utilized and presented in this work are methods to generate a range of neural cell types from iPSCs. However, much of the function of our brains occurs through cell type-cell type dynamic interactions. Therefore, it will be critical to optimize approaches across the various cell types I highlight in this work so that downstream assays can capture how genetic variants mediate these cell type-cell type interactions, which are increasingly implicated in a range of brain disorders.

9.5.3 - Genetic perturbations of astrocyte-neuron interactions

Having identified these cellular responses and ways to elicit them *in vitro*, we are now using genetic-perturbation experiments (pooled CRISPRi screens) and cell-village experiments to uncover how astrocyte-neuron interactions are shaped by genes and genetic variation implicated in schizophrenia.

The common haplotypes implicated by schizophrenia genome-wide association studies involve SNPs at or near the genes that regulate cholesterol synthesis by astrocytes (e.g. *SREBF1*, *SREBF2*), the astrocyte-to-neuron cholesterol shuttle (*CLU*, which encodes Apolipoprotein J), or uptake of these cholesterol-apolipoprotein particles by neurons (*LRP1*, *LDLR*). We are currently designing cell-village experiments with which to detect and measure quantitative effects of these common haplotypes on cholesterol synthesis, cholesterol uptake, and clozapine response. Enabling these experiments are innovations in our cell-village workflow that we developed over the past year (see previous section).

I am also using genetic perturbation (pooled CRISPRi screening on 100 target genes) to uncover 1) the genes/proteins required for clozapine to regulate cholesterol biosynthesis in astrocytes; 2) the effect of disrupting cholesterol transfer between astrocytes and neurons through manipulation of genes with crucial roles in this interaction (13 astrocyte specific, 8 neuron specific); and 3) the functional consequences of knocking down the expression of genes of interest, including genes with rare variants identified through SCHEMA, and SNAP genes described above (Figure 1D). To assess the impact of these genetic perturbations across diverse cell types and development, I have generated single-cell RNA-expression profiles of the effects of these perturbations in neuronal progenitor cells, glutamatergic neurons, and astrocytes. Additionally, I am coupling the genetic perturbations with pharmacological perturbations to identify the genes/proteins that are required for clozapine to regulate cholesterol-synthesis gene expression. To do so, cells with knockdown of genes involved in the mammalian mevalonate pathway and synaptic-astrocyte program will be exposed to clozapine, aripiprazole, and haloperidol. The primary goal will be to identify specific genetic perturbations which inhibit clozapine and aripiprazole's upregulation of cholesterol mRNA.

9.5.4 - Incorporation of clinical outcomes with genomic phenotypes to gain insights into heterogeneous penetrance of risk variants

22q11.2 deletion (22q11.2del) is associated with an exceptionally variable series of developmental, behavioral, and medical phenotypes (e.g., intellectual disability, autism, and cardiac defects). It is also the genetic event most associated with schizophrenia (SCZ) at a population level. 22q11.2del's neuropsychiatric associations remain poorly understood, and do not resolve to a single gene in the deletion region. As introduced in Chapter 6 above, our team recently discovered that 22q11.2del acts through a previously undescribed genomic model, influencing genes and cellular phenotypes relevant to neuropsychiatric disorders through 3D genome conformation. Current human cellular resources are too small to advance this discovery and leverage it to understand psychiatric disease. In the future, we will expand on these efforts to establish a cohort of a few hundred individuals with 22q11.2 deletions to link variability in the 22q11.2dels genomic and neurodevelopmental impacts. Further, having a sufficient sample size with relevant clinical information may help to identify processes mediating 22q11.2del's observed genomic impact along with complementary mechanistic experiments.

Reference list

- "Clinical Trials Need More Diversity" in Scientific American 319, 3, 10 (September 2018) doi: 10. 1038/scientificamerican0918-10.
- "Sickle-Cell Anemia: Haplotype." n.d. Accessed November 29, 2022. https://www.nature.com/scitable/topicpage/sickle-cell-anemia-a-look-at-global-8756219/.
- Abbott, N. J., Ronnback, L. & Hansson, E. Astrocyte-endothelial interactions at the blood-brain barrier. Nat Rev Neurosci 7, 41-53 (2006). https://doi.org:10.1038/nrn1824
- Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: Realizing the promise. Am J Hum Genet. 2023 Feb 2;110(2):179-194. doi: 10.1016/j.ajhg.2022.12.011. Epub 2023 Jan 11. PMID: 36634672; PMCID: PMC9943775.
- Abud, E.M. et al, (2017). iPSC-Derived Human Microglia- like Cells to Study Neurological Diseases. Neuron, 94 (2), 278–293 e279.
- Adelmann, C. H., Wang, T., Sabatini, D. M. & Lander, E. S. 2019. Genome-Wide CRISPR/Cas9 Screening for Identification of Cancer Genes in Cell Lines. Methods Mol Biol, 1907, 125-136.
- Alarcón, M., Abrahams, B. S., Stone, J. L., Duvall, J. A., Perederiy, J. V., Bomar, J. M., Sebat, J., Wigler, M., Martin, C. L., Ledbetter, D. H., Nelson, S. F., Cantor, R. M., & Geschwind, D. H. (2008). Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. American journal of human genetics, 82(1), 150–159. https://doi.org/10.1016/j.ajhg.2007.09.005
- Allen, M. et al. Astrocytes derived from ASD individuals alter behavior and destabilize neuronal activity through aberrant Ca(2+) signaling. Mol Psychiatry 27, 2470-2484 (2022). https://doi.org:10.1038/s41380-022-01486-x
- Allen, N. J. & Eroglu, C. Cell Biology of Astrocyte-Synapse Interactions. Neuron 96, 697-708 (2017). https://doi.org:10.1016/j.neuron.2017.09.056
- Amaral, A. I., Meisingset, T. W., Kotter, M. R. & Sonnewald, U. Metabolic aspects of neuron-oligodendrocyte-astrocyte interactions. Front Endocrinol (Lausanne) 4, 54 (2013). https://doi.org:10.3389/fendo.2013.00054
- American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). https://doi.org/10.1176/appi.books.9780890425596
- Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics (Oxford, England) 31, 166-169 (2015). https://doi.org:10.1093/bioinformatics/btu638
- Andrews, Nancy C. 2009. "Genes Determining Blood Cell Traits." Nature Genetics 41 (11): 1161–62.

- Araujo, B. H. S. et al. Down Syndrome iPSC-Derived Astrocytes Impair Neuronal Synaptogenesis and the mTOR Pathway In Vitro. Mol Neurobiol 55, 5962-5975 (2018). https://doi.org:10.1007/s12035-017-0818-6
- Astle, William J., Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L. Mann, Daniel Mead, et al. 2016. "The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease." Cell 167 (5): 1415–29.e19.
- Babcock, D. F., & Hille, B. (1998). Mitochondrial oversight of cellular Ca2+ signaling. Current opinion in neurobiology, 8(3), 398–404. https://doi.org/10.1016/s0959-4388(98)80067-6
- Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, Benner C, Liu D, Locke AE, Balasubramanian S, Yadav A, Banerjee N, Gillies CE, Damask A, Liu S, Bai X, Hawes A, Maxwell E, Gurski L, Watanabe K, Kosmicki JA, Rajagopal V, Mighty J; Regeneron Genetics Center; DiscovEHR; Jones M, Mitnaul L, Stahl E, Coppola G, Jorgenson E, Habegger L, Salerno WJ, Shuldiner AR, Lotta LA, Overton JD, Cantor MN, Reid JG, Yancopoulos G, Kang HM, Marchini J, Baras A, Abecasis GR, Ferreira MAR. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature. 2021 Nov;599(7886):628-634. doi: 10.1038/s41586-021-04103-z. Epub 2021 Oct 18. PMID: 34662886; PMCID: PMC8596853.
- Bain, G., Kitchens, D., Yao, M., Huettner, J., Gottlieb, D., (1995). Embryonic Stem Cells Express Neuronal Properties in Vitro. Dev. Biol.,. https://doi.org/10.1006/dbio.1995.1085.
- Balding, D. J. 2003. Likelihood-based inference for genetic correlation coefficients. Theor Popul Biol, 63, 221-30.
- Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011 Jan;12(1):56-68. doi: 10.1038/nrg2918. PMID: 21164525; PMCID: PMC3140052.
- Barbar, L., Rusielewicz, T., Zimmer, M., Kalpana, K. & Fossati, V. Isolation of Human CD49f(+) Astrocytes and In Vitro iPSC-Based Neurotoxicity Assays. STAR Protoc 1, 100172 (2020). https://doi.org:10.1016/j.xpro.2020.100172
- Baxi, Emily G., Terri Thompson, Jonathan Li, Julia A. Kaye, Ryan G. Lim, Jie Wu, Divya Ramamoorthy, et al. 2022. "Answer ALS, a Large-Scale Resource for Sporadic and Familial ALS Combining Clinical and Multi-Omics Data from Induced Pluripotent Cell Lines." Nature Neuroscience 25 (2): 226–37.
- Becker, D. et al, (2006). Sex differences in platelet reactivity and response in low-dose aspirin therapy. J. Am. Med. Assoc., 295 (12), 1420–1427.
- Benito-Kwiecinski, S. et al, (2021). An early cell shape transition drives evolutionary expansion of the human forebrain. Cell, 184, 2084–2102.e19.
- Benito-Kwiecinski, S., Giandomenico, S. L., Sutcliffe, M., Riis, E. S., Freire-Pritchett, P., Kelava, I., ... & Lancaster, M. A. (2021). An early cell shape transition drives evolutionary

expansion of the human forebrain. Cell, 184(8), 2084-2102. https://doi.org/10.1016/j.cell.2021.02.050

- Berryer, M. H. et al. High-content synaptic phenotyping in human cellular models reveals a role for BET proteins in synapse assembly. Elife 12 (2023). https://doi.org:10.7554/eLife.80168
- Berryer, M. H., Tegtmeyer, M., Binan, L., Valakh, V., Nathanson, A., Trendafilova, D., Crouse, E., Klein, J., Meyer, D., Pietiläinen, O., Rapino, F., Farhi, S. L., Rubin, L. L., McCarroll, S. A., Nehme, R., & Barrett, L. E. (2022). Robust induction of functional astrocytes using NGN2 expression in human pluripotent stem cells. bioRxiv. https://doi.org/10.1101/2022.09.07.507028
- Bhaduri, A. et al, (2020). Cell stress in cortical organoids impairs molecular subtype specification. Nature, 578 (7793), 142–148.
- Bhaduri, A. et al. An atlas of cortical arealization identifies dynamic molecular signatures. Nature 598, 200-204 (2021). https://doi.org:10.1038/s41586-021-03910-8
- Bijur, P. et al, (2008). Response to Morphine in Male and Female Patients: Analgesia and Adverse Events. T. Clin. J. Pain, 24 (3), 192–198. https://doi.org/10.1097/AJP.0b013e31815d3619.
- Birger, A. et al, (2019). Human iPSC-derived astrocytes from ALS patients with mutated C9ORF72 show increased oxidative stress and neurotoxicity. EBioMedicine, 50, 274–289.
- Bole-Feysot, C., V. Goffin, M. Edery, N. Binart, and P. A. Kelly. 1998. "Prolactin (PRL) and Its Receptor: Actions, Signal Transduction Pathways and Phenotypes Observed in PRL Receptor Knockout Mice." Endocrine Reviews 19 (3): 225–68.
- Bonaglia MC, Giorda R, Borgatti R, Felisari G, Gagliardi C, Selicorni A, Zuffardi O. Disruption of the ProSAP2 gene in a t(12;22)(q24.1;q13.3) is associated with the 22q13.3 deletion syndrome. Am J Hum Genet. 2001 Aug;69(2):261-8. doi: 10.1086/321293. Epub 2001 Jun 18. PMID: 11431708; PMCID: PMC1235301.
- Bonifaz-Pena, V. et al, (2014). Exploring the distribution of genetic markers of pharmacogenomics relevance in Brazilian and Mexican populations. PLoS ONE, 9, (11) e112640
- Bourdeaut F, Lequin D, Brugières L, Reynaud S, Dufour C, Doz F, André N, Stephan JL, Pérel Y, Oberlin O, Orbach D, Bergeron C, Rialland X, Fréneaux P, Ranchere D, Figarella-Branger D, Audry G, Puget S, Evans DG, Pinas JC, Capra V, Mosseri V, Coupier I, Gauthier-Villars M, Pierron G, Delattre O. Frequent hSNF5/INI1 germline mutations in patients with rhabdoid tumor. Clin Cancer Res. 2011 Jan 1;17(1):31-8. doi: 10.1158/1078-0432.CCR-10-1795. PMID: 21208904.
- Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017 Jun 15;169(7):1177-1186. doi: 10.1016/j.cell.2017.05.038. PMID: 28622505; PMCID: PMC5536862.

- Boyle, E. A., Li, Y. I. & Pritchard, J. K. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell, 169, 1177-1186.
- Bray, M. A., Singh, S., Han, H., Davis, C. T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S. M., Gibson, C. C., & Carpenter, A. E. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nature protocols, 11(9), 1757–1774. https://doi.org/10.1038/nprot.2016.105
- Bray, Mark-Anthony, Shantanu Singh, Han Han, Chadwick T. Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M. Gustafsdottir, Christopher C. Gibson, and Anne E. Carpenter. 2016. "Cell Painting, a High-Content Image-Based Assay for Morphological Profiling Using Multiplexed Fluorescent Dyes." Nature Protocols 11 (9): 1757–74.
- Breen MS, Browne A, Hoffman GE, Stathopoulos S, Brennand K, Buxbaum JD, Drapeau E. Transcriptional signatures of participant-derived neural progenitor cells and neurons implicate altered Wnt signaling in Phelan-McDermid syndrome and autism. Mol Autism. 2020 Jun 19;11(1):53. doi: 10.1186/s13229-020-00355-0. PMID: 32560742; PMCID: PMC7304190.
- Brennand, K., Savas, J. N., Kim, Y., Tran, N., Simone, A., Hashimoto-Torii, K., Beaumont, K. G., Kim, H. J., Topol, A., Ladran, I., Abdelrahim, M., Matikainen-Ankney, B., Chao, S. H., Mrksich, M., Rakic, P., Fang, G., Zhang, B., Yates, J. R., 3rd, & Gage, F. H. (2015). Phenotypic differences in hiPSC NPCs derived from patients with schizophrenia. Molecular psychiatry, 20(3), 361–368. https://doi.org/10.1038/mp.2014.22
- Brennand, K.J. et al, (2011). Modelling schizophrenia using human induced pluripotent stem cells. Nature, 473 (7346), 221–225.
- Brown, C.A. et al, (2021). Estimated Prevalence and Incidence of Amyotrophic Lateral Sclerosis and SOD1 and C9orf72 Genetic Variants. Neuroepidemiology,, 1– 12.
- Burchard, E.G. et al, (2003). The importance of race and ethnic background in biomedical research and clinical practice. N. Engl. J. Med., 348 (12), 1170–1175.
- Burnett, B. G., Munoz, E., Tandon, A., Kwon, D. Y., Sumner, C. J. & Fischbeck, K. H. 2009. Regulation of SMN protein stability. Mol Cell Biol, 29, 1107-15.
- Cahan, P. & Daley, G. Q. 2013. Origins and implications of pluripotent stem cell variability and heterogeneity. Nat Rev Mol Cell Biol, 14, 357-68.
- Cai, Z., Li, S., Matuskey, D., Nabulsi, N., & Huang, Y. (2019). PET imaging of synaptic density: a new tool for investigation of neuropsychiatric diseases. Neuroscience letters, 691, 44-50. https://doi.org/10.1016/j.neulet.2018.07.038
- Caicedo, Juan C., John Arevalo, Federica Piccioni, Mark-Anthony Bray, Cathy L. Hartland, Xiaoyun Wu, Angela N. Brooks, et al. 2022. "Cell Painting Predicts Impact of Lung Cancer Variants." Molecular Biology of the Cell 33 (6): ar49.

- Caicedo, Juan C., Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S. Vasilevich, et al. 2017. "Data-Analysis Strategies for Image-Based Cell Profiling." Nature Methods 14 (9): 849–63.
- Canals, I. et al, (2018). Rapid and efficient induction of functional astrocytes from human pluripotent stem cells. Nature Methods, 15 (9), 693–696.
- Canals, I. et al. Rapid and efficient induction of functional astrocytes from human pluripotent stem cells. Nat Methods (2018). https://doi.org:10.1038/s41592-018-0103-2
- Cannon, T. D., Chung, Y., He, G., Sun, D., Jacobson, A., van Erp, T. G., McEwen, S., Addington, J., Bearden, C. E., Cadenhead, K., Cornblatt, B., Mathalon, D. H., McGlashan, T., Perkins, D., Jeffries, C., Seidman, L. J., Tsuang, M., Walker, E., Woods, S. W., Heinssen, R., North American Prodrome Longitudinal Study Consortium (2015). Progressive reduction in cortical thickness as psychosis develops: a multisite longitudinal neuroimaging study of youth at elevated clinical risk. Biological psychiatry, 77(2), 147–157. https://doi.org/10.1016/j.biopsych.2014.05.023
- 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. 2015. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature, 527, 192-7.
- Carcamo-Orive, Ivan, Ngan F. Huang, Thomas Quertermous, and Joshua W. Knowles. 2017. "Induced Pluripotent Stem Cell-Derived Endothelial Cells in Insulin Resistance and Metabolic Syndrome." Arteriosclerosis, Thrombosis, and Vascular Biology 37 (11): 2038–42.
- 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. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome biology, 7(10), R100. https://doi.org/10.1186/gb-2006-7-10-r100
- Cataldo, A. M., McPhie, D. L., Lange, N. T., Punzell, S., Elmiligy, S., Ye, N. Z., Froimowitz, M. P., Hassinger, L. C., Menesale, E. B., Sargent, L. W., Logan, D. J., Carpenter, A. E., & Cohen, B. M. (2010). Abnormalities in mitochondrial structure in cells from patients with bipolar disorder. The American journal of pathology, 177(2), 575–585. https://doi.org/10.2353/ajpath.2010.081068
- Cataldo, Anne M., Donna L. McPhie, Nicholas T. Lange, Steven Punzell, Sarah Elmiligy, Nancy Z. Ye, Michael P. Froimowitz, et al. 2010. "Abnormalities in Mitochondrial Structure in Cells from Patients with Bipolar Disorder." The American Journal of Pathology 177 (2): 575–85.
- Centeno, E.G.Z. et al, (2018). 2D versus 3D human induced pluripotent stem cell-derived cultures for neurodegenerative disease modelling. Mol. Neurodegener., 13 (1), 27.
- Chan, Y. et al, (2018). Enabling multiplexed testing of pooled donor cells through whole-genome sequencing. Genome Med., 10 (1), 31.

- Chang, D. et al, (2017). A meta-analysis of genome-wide association studies identifies 17 new Parkinson’s disease risk loci. Nature Genet., 49 (10), 1511–1516.
- Chang, E.A. et al, (2015). Derivation of Ethnically Diverse Human Induced Pluripotent Stem Cell Lines. Sci. Rep., 5, 15234.
- Chen, C. et al. Role of astroglia in Down's syndrome revealed by patient-derived human-induced pluripotent stem cells. Nat Commun 5, 4430 (2014). https://doi.org:10.1038/ncomms5430
- Chen, T. H. 2020. New and Developing Therapies in Spinal Muscular Atrophy: From Genotype to Phenotype to Treatment and Where Do We Stand? Int J Mol Sci, 21.
- Cheung, A. K., Hurley, B., Kerrigan, R., Shu, L., Chin, D. N., Shen, Y., O'brien, G., Sung, M. J., Hou, Y., Axford, J., Cody, E., Sun, R., Fazal, A., Fridrich, C., Sanchez, C. C., Tomlinson, R. C., Jain, M., Deng, L., Hoffmaster, K., Song, C., Van Hoosear, M., Shin, Y., Servais, R., Towler, C., Hild, M., Curtis, D., Dietrich, W. F., Hamann, L. G., Briner, K., Chen, K. S., Kobayashi, D., Sivasankaran, R. & Dales, N. A. 2018. Discovery of Small Molecule Splicing Modulators of Survival Motor Neuron-2 (SMN2) for the Treatment of Spinal Muscular Atrophy (SMA). J Med Chem, 61, 11021-11036.
- Cheung, V. G., Spielman, R. S., Ewens, K. G., Weber, T. M., Morley, M. & Burdick, J. T. 2005. Mapping determinants of human gene expression by regional and genome- wide association. Nature, 437, 1365-9.
- Chiaradia, I., Lancaster, M.A., (2020). Brain organoids for the study of human neurobiology at the interface of in vitro and in vivo. Nature Neurosci., 23, 1496–1508.
- Cimini, B. A., Chandrasekaran, S. N., Kost-Alimova, M., Miller, L., Goodale, A., Fritchman, B., ... & Carpenter, A. E. (2022). Optimizing the Cell Painting assay for image-based profiling. bioRxiv. https://doi.org/10.1101/2022.07.13.499171
- Cimini, Beth A., Srinivas Niranj Chandrasekaran, Maria Kost-Alimova, Lisa Miller, Amy Goodale, Briana Fritchman, Patrick Byrne, et al. 2022. “Optimizing the Cell Painting Assay for Image-Based Profiling.” bioRxiv. https://doi.org/10.1101/2022.07.13.499171.
- Clayton, T. M., Whitaker, J. P., Sparkes, R. & Gill, P. 1998. Analysis and interpretation of mixed forensic stains using DNA STR profiling. Forensic Sci Int, 91, 55-70.
- Clinicaltrials.gov, Keyword search: “ipsc”.
- Collins RL, Glessner JT, Porcu E, Lepamets M, Brandon R, Lauricella C, Han L, Morley T, Niestroj LM, Ulirsch J, Everett S, Howrigan DP, Boone PM, Fu J, Karczewski KJ, Kellaris G, Lowther C, Lucente D, Mohajeri K, Nõukas M, Nuttle X, Samocha KE, Trinh M, Ullah F, Võsa U; Epi25 Consortium; Estonian Biobank Research Team; Hurles ME, Aradhya S, Davis EE, Finucane H, Gusella JF, Janze A, Katsanis N, Matyakhina L, Neale BM, Sanders D, Warren S, Hodge JC, Lal D, Ruderfer DM, Meck J, Mägi R, Esko T, Reymond A, Kutalik Z, Hakonarson H, Sunyaev S, Brand H, Talkowski ME. A cross-disorder dosage sensitivity map of the human genome. Cell. 2022 Aug 4;185(16):3041-3055.e25. doi: 10.1016/j.cell.2022.06.036. Epub 2022 Aug 1. PMID: 35917817; PMCID: PMC9742861.

- Cornwell, M. et al. VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. BMC Bioinformatics 19, 135 (2018). https://doi.org:10.1186/s12859-018-2139-9
- Crawford K, Bracher-Smith M, Owen D, Kendall KM, Rees E, Pardiñas AF, Einon M, Escott-Price V, Walters JTR, O'Donovan MC, Owen MJ, Kirov G. Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. J Med Genet. 2019 Mar;56(3):131-138. doi: 10.1136/jmedgenet-2018-105477. Epub 2018 Oct 20. PMID: 30343275.
- Crawford, M.B., DeLisi, L.E., (2016). Issues related to sex differences in antipsychotic treatment. Curr. Opin. Psychiatry, 29 (3), 211–217. https://doi.org/10.1097/ YCO.0000000000000243.
- Cucullo, L. et al. A new dynamic in vitro model for the multidimensional study of astrocyte-endothelial cell interactions at the blood-brain barrier. Brain Res 951, 243-254 (2002). https://doi.org:10.1016/s0006-8993(02)03167-0
- Daniels, T. E., Olsen, E. M., & Tyrka, A. R. (2020). Stress and Psychiatric Disorders: The Role of Mitochondria. Annual review of clinical psychology, 16, 165–186. https://doi.org/10.1146/annurev-clinpsy-082719-104030
- Davis, R., (2020). Mechanism of Actionand Target Identification: A Matter of Timing in Drug Discovery. iScience, 23, 101487
- de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015 Apr 17;11(4):e1004219. doi: 10.1371/journal.pcbi.1004219. PMID: 25885710; PMCID: PMC4401657.
- DeBoever, Christopher, He Li, David Jakubosky, Paola Benaglio, Joaquin Reyna, Katrina M. Olson, Hui Huang, et al. 2017. “Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells.” Cell Stem Cell 20 (4): 533–46.e7.
- Di Giorgio, F. P., Boulting, G. L., Bobrowicz, S. & Eggan, K. C. 2008. Human embryonic stem cell-derived motor neurons are sensitive to the toxic effect of glial cells carrying an ALS-causing mutation. Cell Stem Cell, 3, 637-48.
- Dimos, J.T. et al, (2008). Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. Science, 321 (5893), 1218–1221.
- Diniz, L. P. et al. Astrocyte-induced synaptogenesis is mediated by transforming growth factor beta signaling through modulation of D-serine levels in cerebral cortex neurons. J Biol Chem 287, 41432-41445 (2012). https://doi.org:10.1074/jbc.M112.380824
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) 29, 15-21 (2013). https://doi.org:10.1093/bioinformatics/bts635
- Du, Z.W., Chen, H., Liu, H., et al., (2015). Generation and expansion of highly pure motor neuron progenitors from human pluripotent stem cells. Nature Commun., 6, 6626.

- Durand CM, Betancur C, Boeckers TM, Bockmann J, Chaste P, Fauchereau F, Nygren G, Rastam M, Gillberg IC, Anckarsäter H, Sponheim E, Goubran-Botros H, Delorme R, Chabane N, Mouren-Simeoni MC, de Mas P, Bieth E, Rogé B, Héron D, Burglen L, Gillberg C, Leboyer M, Bourgeron T. Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. Nat Genet. 2007 Jan;39(1):25-7. doi: 10.1038/ng1933. Epub 2006 Dec 17. PMID: 17173049; PMCID: PMC2082049.
- E.M. Wigdor et al., The female protective effect against autism spectrum disorder, n.d. doi: 10.1101/2021.03.29. 21253866.
- E.P. Terlizzi, B. Zablotsky, Mental Health Treatment Among Adults: United States, 2019, NCHS Data Brief 380 (2020) 1–8.
- Egeland, T., Dalen, I. & Mostad, P. F. 2003. Estimating the number of contributors to a DNA profile. Int J Legal Med, 117, 271-5.
- Ehrlich, M. et al, (2017). Rapid and efficient generation of oligodendrocytes from human induced pluripotent stem cells using transcription factors. Proc. Natl. Acad. Sci. U. S. A., 114 (11), E2243–E2252.
- Elsayed, Abdelrahman M., Emine Bayraktar, Paola Amero, Salama A. Salama, Abdelaziz H. Abdelaziz, Raed S. Ismail, Xinna Zhang, et al. 2021. "PRKAR1B-AS2 Long Noncoding RNA Promotes Tumorigenesis, Survival, and Chemoresistance via the PI3K/AKT/mTOR Pathway." International Journal of Molecular Sciences 22 (4). https://doi.org/10.3390/ijms22041882.
- Evans, M.J., Kaufman, M.H., (2007). Establishment in culture of pluripotential cells from mouse embryos. Nature,. https://doi.org/10.1038/292154a0.
- Fakunle, E.S., Loring, J.F., (2012). Ethnically diverse pluripotent stem cells for drug development. Trends Mol. Med.,. https://doi.org/10.1016/j.molmed.2012.10.007.
- Fan, L.Z. et al, (2018). All-optical synaptic electrophysiology probes mechanism of ketamine- induced disinhibition. Nature Methods, 15, 823–831.
- Feinberg, I. (1982). Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence?. Journal of psychiatric research, 17(4), 319-334. https://doi.org/10.1016/0022-3956(82)90038-3
- Feldman D, Singh A, Schmid-Burgk JL, Carlson RJ, Mezger A, Garrity AJ, Zhang F, Blainey PC. Optical Pooled Screens in Human Cells. Cell. 2019 Oct 17;179(3):787-799.e17. doi: 10.1016/j.cell.2019.09.016. PMID: 31626775; PMCID: PMC6886477.
- Feng, Zhen-Hua, Lin Zheng, Teng Yao, Si-Yue Tao, Xiao-An Wei, Ze-Yu Zheng, Bing-Jie Zheng, et al. 2021. "EIF4A3-Induced Circular RNA PRKAR1B Promotes Osteosarcoma Progression by miR-361-3p-Mediated Induction of FZD4 Expression." Cell Death & Disease 12 (11): 1025.
- Filmus, Jorge, and Mariana Capurro. 2008. "The Role of Glypican-3 in the Regulation of Body Size and Cancer." Cell Cycle 7 (18): 2787–90.

- Finkel, R. S., Chiriboga, C. A., Vajsar, J., Day, J. W., Montes, J., De Vivo, D. C., Yamashita, M., Rigo, F., Hung, G., Schneider, E., Norris, D. A., Xia, S., Bennett, C. F. & Bishop, K. M. 2016. Treatment of infantile-onset spinal muscular atrophy with nusinersen: a phase 2, open-label, dose-escalation study. Lancet, 388, 3017-3026.
- Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, Ripke S, Day FR; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; RACI Consortium; Purcell S, Stahl E, Lindstrom S, Perry JR, Okada Y, Raychaudhuri S, Daly MJ, Patterson N, Neale BM, Price AL. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015 Nov;47(11):1228-35. doi: 10.1038/ng.3404. Epub 2015 Sep 28. PMID: 26414678; PMCID: PMC4626285.
- Finucane, H. et al, (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nature Genet., 50, 621–629.
- Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P. R., Lareau, C., Shoresh, N., Genovese, G., Saunders, A., Macosko, E., Pollack, S., Brainstorm Consortium, Perry, J. R. B., Buenrostro, J. D., Bernstein, B. E., Raychaudhuri, S., McCarroll, S., Neale, B. M., Price, A. L. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nature genetics, 50(4), 621–629. https://doi.org/10.1038/s41588-018-0081-4
- Finucane, Hilary K., Yakir A. Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, et al. 2018. "Heritability Enrichment of Specifically Expressed Genes Identifies Disease-Relevant Tissues and Cell Types." Nature Genetics 50 (4): 621–29.
- Forsyth JK, Mennigen E, Lin A, Sun D, Vajdi A, Kushan-Wells L, Ching CRK, Villalon-Reina JE, Thompson PM; 22q11.2 ENIGMA Consortium; Bearden CE. Prioritizing Genetic Contributors to Cortical Alterations in 22q11.2 Deletion Syndrome Using Imaging Transcriptomics. Cereb Cortex. 2021 Jun 10;31(7):3285-3298. doi: 10.1093/cercor/bhab008. PMID: 33638978; PMCID: PMC8196250.
- Frick, L. R., Williams, K., & Pittenger, C. (2013). Microglial dysregulation in psychiatric disease. Clinical & developmental immunology, 2013, 608654. https://doi.org/10.1155/2013/608654
- Fujii, Y., Maekawa, S. & Morita, M. Astrocyte calcium waves propagate proximally by gap junction and distally by extracellular diffusion of ATP released from volume-regulated anion channels. Sci Rep 7, 13115 (2017). https://doi.org:10.1038/s41598-017-13243-0
- Gasperini, M., Hill, A. J., Mcfaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N. & Shendure, J. 2019. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell, 176, 377-390 e19.

- Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, Mahajan M, Manaa D, Pawitan Y, Reichert J, Ripke S, Sandin S, Sklar P, Svantesson O, Reichenberg A, Hultman CM, Devlin B, Roeder K, Buxbaum JD. Most genetic risk for autism resides with common variation. Nat Genet. 2014 Aug;46(8):881-5. doi: 10.1038/ng.3039. Epub 2014 Jul 20. PMID: 25038753; PMCID: PMC4137411.
- Gava-Junior, G. et al, (2020). A Cell Culture Model for Studying the Role of Neuron-Glia Interactions in Ischemia. J. Vis. Exp., 165
- Gecz, J. et al, (2009). The genetic landscape of intellectual disability arising from chromosome X. Trends Genet., 25 (7), 308–316.
- Genolet, O. et al, (2021). Identification of X-chromosomal genes that drive sex differences in embryonic stem cells through a hierarchical CRISPR screening approach. Genome Biol., 22 (1), 110.
- Giandomenico, S. L., Sutcliffe, M., & Lancaster, M. A. (2021). Generation and long-term culture of advanced cerebral organoids for studying later stages of neural development. Nature protocols, 16(2), 579-602. https://doi.org/10.1038/s41596-020-00433-w
- Giandomenico, S.L., Sutcliffe, M., Lancaster, M.A., (2021). Generation and long-term culture of advanced cerebral organoids for studying later stages of neural development. Nature Protoc., 16, 579–602.
- Glausier, J. R., & Lewis, D. A. (2013). Dendritic spine pathology in schizophrenia. Neuroscience, 251, 90–107. https://doi.org/10.1016/j.neuroscience.2012.04.044
- Glessner, J. T., Li, J., Wang, D., March, M., Lima, L., Desai, A., Hadley, D., Kao, C., Gur, R. E., Cohen, N., Sleiman, P. M. A., Li, Q., Hakonarson, H., & Janssen-CHOP Neuropsychiatric Genomics Working Group (2017). Copy number variation meta-analysis reveals a novel duplication at 9p24 associated with multiple neurodevelopmental disorders. Genome medicine, 9(1), 106. https://doi.org/10.1186/s13073-017-0494-1
- Goldstein, G. W. Endothelial cell-astrocyte interactions. A cellular model of the blood-brain barrier. Ann N Y Acad Sci 529, 31-39 (1988). https://doi.org:10.1111/j.1749-6632.1988.tb51417.x
- Goncalves, E. et al, (2020). Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. Mol. Syst. Biol., 16, e9405
- Grimm, Amandine, and Anne Eckert. 2017. "Brain Aging and Neurodegeneration: From a Mitochondrial Point of View." Journal of Neurochemistry 143 (4): 418–31.
- Groen, E. J. N., Talbot, K. & Gillingwater, T. H. 2018. Advances in therapy for spinal muscular atrophy: promises and challenges. Nat Rev Neurol, 14, 214-224.
- Grove, J. et al, (2019). Identification of common genetic risk variants for autism spectrum disorder. Nature Genet., 51 (3), 431–444.
- GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx

(eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, et al. 2017. "Genetic Effects on Gene Expression across Human Tissues." Nature 550 (7675): 204–13.

- GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts:; Laboratory, Data Analysis &Coordinating Center (LDACC):; NIH program management:; Biospecimen collection:; Pathology:; eQTL manuscript working group:; Battle A, Brown CD, Engelhardt BE, Montgomery SB. Genetic effects on gene expression across human tissues. Nature. 2017 Oct 11;550(7675):204-213. doi: 10.1038/nature24277. Erratum in: Nature. 2017 Dec 20;: PMID: 29022597; PMCID: PMC5776756.
- Guo, L. et al, (2021). Sex Differences in Alzheimer's Disease: Insights From the Multiomics Landscape. Biol. Psychiatry,.
- Gusella, J.F. et al, (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature, 306 (5940), 234–238.
- Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 20, 296 (2019). https://doi.org:10.1186/s13059-019-1874-1
- Haghighi, Marzieh, Juan C. Caicedo, Beth A. Cimini, Anne E. Carpenter, and Shantanu Singh. 2022. "High-Dimensional Gene Expression and Morphology Profiles of Cells across 28,000 Genetic and Chemical Perturbations." Nature Methods, November, 1–8.
- Han, X. et al. Forebrain engraftment by human glial progenitor cells enhances synaptic plasticity and learning in adult mice. Cell Stem Cell 12, 342-353 (2013). https://doi.org:10.1016/j.stem.2012.12.015
- Handsaker, R. E., Korn, J. M., Nemesh, J. & Mccarroll, S. A. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet, 43, 269-76.
- Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M. & Mccarroll, S. A. 2015. Large multiallelic copy number variations in humans. Nat Genet, 47, 296-303.
- Hawrot, J. et al, (2020). Modeling cell-autonomous motor neuron phenotypes in ALS using iPSCs. Neurobiol. Dis., 134, 104680
- Hawrylycz, M. J. et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature 489, 391-399 (2012). https://doi.org:10.1038/nature11405

- He et al, (2020). Lineage recording reveals dynamics of cerebral organoid regionalization. Biorxiv,. https://doi.org/ 10.1101/2020.06.19.162032.
- Hedegaard, A. et al. Pro-maturational Effects of Human iPSC-Derived Cortical Astrocytes upon iPSC-Derived Cortical Neurons. Stem Cell Reports 15, 38-51 (2020). https://doi.org:10.1016/j.stemcr.2020.05.003
- Hibar, D. P., Westlye, L. T., Doan, N. T., Jahanshad, N., Cheung, J. W., Ching, C., Versace, A., Bilderbeck, A. C., Uhlmann, A., Mwangi, B., Krämer, B., Overs, B., Hartberg, C. B., Abé, C., Dima, D., Grotegerd, D., Sprooten, E., Bøen, E., Jimenez, E., Howells, F. M., … Andreassen, O. A. (2018). Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. Molecular psychiatry, 23(4), 932–942. https://doi.org/10.1038/mp.2017.73
- Hilker, R., Helenius, D., Fagerlund, B., Skytthe, A., Christensen, K., Werge, T. M., Nordentoft, M., & Glenthøj, B. (2018). Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. Biological psychiatry, 83(6), 492–498. https://doi.org/10.1016/j.biopsych.2017.08.017
- Hino, K. et al, (2018). An mTOR Signaling Modulator Suppressed Heterotopic Ossification of Fibrodysplasia Ossificans Progressiva. Stem Cell Rep., 11, 1106–1119.
- Hoffman, G.E. et al, (2017). Transcriptional signatures of schizophrenia in hiPSC-derived NPCs and neurons are concordant with post-mortem adult brains. Nature Commun., 8 (1), 2225.
- Hsu, J. Y., Fulco, C. P., Cole, M. A., Canver, M. C., Pellin, D., Sher, F., Farouni, R., Clement, K., Guo, J. A., Biasco, L., Orkin, S. H., Engreitz, J. M., Lander, E. S., Joung, J. K., Bauer, D. E. & Pinello, L. 2018. CRISPR-SURF: discovering regulatory elements by deconvolution of CRISPR tiling screen data. Nat Methods, 15, 992-993.
- Hsu, P. D., Lander, E. S. & Zhang, F. 2014. Development and applications of CRISPR-Cas9 for genome engineering. Cell, 157, 1262-78.
- Hu, B.Y., Zhang, S.C., (2009). Differentiation of spinal motor neurons from pluripotent human stem cells. Nature Protoc., 4 (9), 1295–1304.
- Hu, Y. Q. & Fung, W. K. 2003. Interpreting DNA mixtures with the presence of relatives. Int J Legal Med, 117, 39-45.
- Hua, Y., Sahashi, K., Hung, G., Rigo, F., Passini, M. A., Bennett, C. F. & Krainer, A. R. 2010. Antisense correction of SMN2 splicing in the CNS rescues necrosis in a type III SMA mouse model. Genes Dev, 24, 1634-44.
- Hua, Y., Vickers, T. A., Okunola, H. L., Bennett, C. F. & Krainer, A. R. 2008. Antisense masking of an hnRNP A1/A2 intronic splicing silencer corrects SMN2 splicing in transgenic mice. Am J Hum Genet, 82, 834-48.
- Huang, C.Y. et al, (2019). Human iPSC banking: barriers and opportunities. J. Biomed. Sci., 26 (1), 87.

- Huhtaniska, S., Jääskeläinen, E., Hirvonen, N., Remes, J., Murray, G. K., Veijola, J., Isohanni, M., & Miettunen, J. (2017). Long-term antipsychotic use and brain changes in schizophrenia - a systematic review and meta-analysis. Human psychopharmacology, 32(2), 10.1002/hup.2574. https://doi.org/10.1002/hup.2574
- Imamura, K. et al, (2021). iPSC screening for drug repurposing identifies anti-RNA virus agents modulating host cell susceptibility. FEBS Open Bio, 11 (5), 1452– 1464.
- Insall, Robert H., and Laura M. Machesky. 2009. "Actin Dynamics at the Leading Edge: From Simple Machinery to Complex Networks." Developmental Cell 17 (3): 310–22.
- Isasi, R. et al, (2014). Identifiability and privacy in pluripotent stem cell research. Cell Stem Cell, 14 (4), 427–430. https://doi.org/10.1016/j.stem.2014.03.014.
- Ito, H., Uchida, T., Makita, K., (2015). Ketamine Causes Mitochondrial Dysfunction in Human Induced Pluripotent Stem Cell-Derived Neurons. PLoS ONE, 10, e0128445
- J.V. McGivern, A.D. Ebert, Exploiting pluripotent stem cell technology for drug discovery, screening, safety, and toxicology assessments. Adv. Drug Delivery Rev. (n.d.). https://doi.org/10.1016/j.addr.2013.11.012.
- Jansen, I.E. et al, (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nature Genet., 51 (3), 404–413.
- Jerber, J. et al, (2021). Population-scale single-cell RNA- seq profiling across dopaminergic neuron differentiation. Nature Genet., 53, 304–312.
- Jerber, Julie, Daniel D. Seaton, Anna S. E. Cuomo, Natsuhiko Kumasaka, James Haldane, Juliette Steer, Minal Patel, et al. 2021. "Population-Scale Single-Cell RNA-Seq Profiling across Dopaminergic Neuron Differentiation." Nature Genetics 53 (3): 304–12.
- John, G. R. Investigation of astrocyte - oligodendrocyte interactions in human cultures. Methods Mol Biol 814, 401-414 (2012). https://doi.org:10.1007/978-1-61779-452-0_27
- Jones, C. A., Watson, D. J., & Fone, K. C. (2011). Animal models of schizophrenia. British journal of pharmacology, 164(4), 1162–1194. https://doi.org/10.1111/j.1476-5381.2011.01386.x
- Jouannet, Stéphanie, Julien Saint-Pol, Laurent Fernandez, Viet Nguyen, Stéphanie Charrin, Claude Boucheix, Christel Brou, Pierre-Emmanuel Milhiet, and Eric Rubinstein. 2016. "TspanC8 Tetraspanins Differentially Regulate the Cleavage of ADAM10 Substrates, Notch Activation and ADAM10 Membrane Compartmentalization." Cellular and Molecular Life Sciences: CMLS 73 (9): 1895–1915.
- Kam, T. I., Hinkle, J. T., Dawson, T. M. & Dawson, V. L. Microglia and astrocyte dysfunction in parkinson's disease. Neurobiol Dis 144, 105028 (2020). https://doi.org:10.1016/j.nbd.2020.105028
- Kamitaki, N. et al, (2020). Complement genes contribute sex-biased vulnerability in diverse disorders. Nature, 582, 577–581.
- Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA,

Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME; Genome Aggregation Database Consortium; Neale BM, Daly MJ, MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434-443. doi: 10.1038/s41586-020-2308-7. Epub 2020 May 27. Erratum in: Nature. 2021 Feb;590(7846):E53. Erratum in: Nature. 2021 Sep;597(7874):E3-E4. PMID: 32461654; PMCID: PMC7334197.

- Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." Nature 581 (7809): 434–43.
- Khan TA, Revah O, Gordon A, Yoon SJ, Krawisz AK, Goold C, Sun Y, Kim CH, Tian Y, Li MY, Schaepe JM, Ikeda K, Amin ND, Sakai N, Yazawa M, Kushan L, Nishino S, Porteus MH, Rapoport JL, Bernstein JA, O'Hara R, Bearden CE, Hallmayer JF, Huguenard JR, Geschwind DH, Dolmetsch RE, Paşca SP. Neuronal defects in a human cellular model of 22q11.2 deletion syndrome. Nat Med. 2020 Dec;26(12):1888-1898. doi: 10.1038/s41591-020-1043-9. Epub 2020 Sep 28. PMID: 32989314; PMCID: PMC8525897.
- Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T. & Kathiresan, S. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet, 50, 1219-1224.
- Kiger, A. A., B. Baum, S. Jones, M. R. Jones, A. Coulson, C. Echeverri, and N. Perrimon. 2003. "A Functional Genomic Analysis of Cell Morphology Using RNA Interference." Journal of Biology 2 (4): 27.
- Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., Bala, S., Bensaddek, D., Casale, F. P., Culley, O. J., Danecek, P., Faulconbridge, A., Harrison, P. W., Kathuria, A., Mccarthy, D., Mccarthy, S. A., Meleckyte, R., Memari, Y., Moens, N., Soares, F., Mann, A., Streeter, I., Agu, C. A., Alderton, A., Nelson, R., Harper, S., Patel, M., White, A., Patel, S. R., Clarke, L., Halai, R., Kirton, C. M., Kolb-Kokocinski, A., Beales, P., Birney, E., Danovi, D., Lamond, A. I., Ouwehand, W. H., Vallier, L., Watt, F. M., Durbin, R., Stegle, O. & Gaffney, D. J. 2017. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature, 546, 370-375.

- Kilpinen, Helena, Angela Goncalves, Andreas Leha, Vackar Afzal, Kaur Alasoo, Sofie Ashford, Sendu Bala, et al. 2017. "Common Genetic Variation Drives Molecular Heterogeneity in Human iPSCs." Nature 546 (7658): 370–75.
- Kiskinis, E. et al, (2018). All-Optical Electrophysiology for High-Throughput Functional Characterization of a Human iPSC-Derived Motor Neuron Model of ALS. Stem Cell Rep., 10, 1991–2004.
- Kung, L., & Roberts, R. C. (1999). Mitochondrial pathology in human schizophrenic striatum: a postmortem ultrastructural study. Synapse, 31(1), 67-75. https://doi.org/10.1002/(SICI)1098-2396(199901)31:1<67::AID-SYN9>3.0.CO;2-%23
- Kunkle, Brian W., Michael Schmidt, Hans-Ulrich Klein, Adam C. Naj, Kara L. Hamilton-Nelson, Eric B. Larson, Denis A. Evans, et al. 2021. "Novel Alzheimer Disease Risk Loci and Pathways in African American Individuals Using the African Genome Resources Panel: A Meta-Analysis." JAMA Neurology 78 (1): 102–13.
- Lam, M., Chen, C. Y., Li, Z., Martin, A. R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B. C., Liu, R., Zhou, W., Guan, L., Kamatani, Y., Kim, S. W., Kubo, M., Kusumawardhani, A., Liu, C. M., Ma, H., Periyasamy, S., Takahashi, A., Xu, Z., Yu, H., Zhu, F., Schizophrenia Working Group Of The Psychiatric Genomics, C., Indonesia Schizophrenia, C., Genetic, R. O. S. N.-C., The, N., Chen, W. J., Faraone, S., Glatt, S. J., He, L., Hyman, S. E., Hwu, H. G., Mccarroll, S. A., Neale, B. M., Sklar, P., Wildenauer, D. B., Yu, X., Zhang, D., Mowry, B. J., Lee, J., Holmans, P., Xu, S., Sullivan, P. F., Ripke, S., O'donovan, M. C., Daly, M. J., Qin, S., Sham, P., Iwata, N., Hong, K. S., Schwab, S. G., Yue, W., Tsuang, M., Liu, J., Ma, X., Kahn, R. S., Shi, Y. & Huang, H. 2019. Comparative genetic architectures of schizophrenia in East Asian and European populations. Nat Genet, 51, 1670-1678.
- Lambert MP, Arulselvan A, Schott A, Markham SJ, Crowley TB, Zackai EH, McDonald-McGinn DM. The 22q11.2 deletion syndrome: Cancer predisposition, platelet abnormalities and cytopenias. Am J Med Genet A. 2018 Oct;176(10):2121-2127. doi: 10.1002/ajmg.a.38474. Epub 2017 Sep 22. PMID: 28940864.
- Lappalainen T, MacArthur DG. From variant to function in human disease genetics. Science. 2021 Sep 24;373(6562):1464-1468. doi: 10.1126/science.abi8207. Epub 2021 Sep 23. PMID: 34554789.
- Larson, M. H., Gilbert, L. A., Wang, X., Lim, W. A., Weissman, J. S. & Qi, L. S. 2013. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nat Protoc, 8, 2180-96.
- Lattke, M. et al. Extensive transcriptional and chromatin changes underlie astrocyte maturation in vivo and in culture. Nat Commun 12, 4335 (2021). https://doi.org:10.1038/s41467-021-24624-5

- Leal, M. C., Casabona, J. C., Puntel, M. & Pitossi, F. J. Interleukin-1beta and tumor necrosis factor-alpha: reliable targets for protective therapies in Parkinson's Disease? Front Cell Neurosci 7, 53 (2013). https://doi.org:10.3389/fncel.2013.00053
- Lee, Y. et al, (2018). Excitatory and inhibitory synaptic dysfunction in mania: an emerging hypothesis from animal model studies. Exp. Mol. Med., 50 (4), 1–11.
- Lefebvre, S., Burglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., Benichou, B., Cruaud, C., Millasseau, P., Zeviani, M. & Et Al. 1995. Identification and characterization of a spinal muscular atrophy-determining gene. Cell, 80, 155-65.
- LeGates, T.A. et al, (2019). Sex differences in antidepressant efficacy. Neuropsychopharmacology, 44, 140–154.
- Lehmkuhl, Erik M., Suvithanandhini Loganathan, Eric Alsop, Alexander D. Blythe, Tina Kovalik, Nicholas P. Mortimore, Dianne Barrameda, et al. 2021. "TDP-43 Proteinopathy Alters the Ribosome Association of Multiple mRNAs Including the Glypican Dally-like Protein (Dlp)/GPC6." Acta Neuropathologica Communications 9 (1): 52.
- Leng, K. et al. CRISPRi screens in human iPSC-derived astrocytes elucidate regulators of distinct inflammatory reactive states. Nat Neurosci 25, 1528-1542 (2022). https://doi.org:10.1038/s41593-022-01180-9
- Li, J., Ryan, S. K., Deboer, E., Cook, K., Fitzgerald, S., Lachman, H. M., Wallace, D. C., Goldberg, E. M., & Anderson, S. A. (2019). Mitochondrial deficits in human iPSC-derived neurons from patients with 22q11.2 deletion syndrome and schizophrenia. Translational psychiatry, 9(1), 302. https://doi.org/10.1038/s41398-019-0643-y
- Li, Xin, Yungil Kim, Emily K. Tsang, Joe R. Davis, Farhan N. Damani, Colby Chiang, Gaelen T. Hess, et al. 2017. "The Impact of Rare Variation on Gene Expression across Tissues." Nature 550 (7675): 239–43.
- Li, Y. et al, (2011). Generation of iPSCs from mouse fibroblasts with a single gene, Oct4, and small molecules. Cell Res., 21 (1), 196–204.
- Li, Y. et al, (2017). Cell sex affects extracellular matrix protein expression and proliferation of smooth muscle progenitor cells derived from human pluripotent stem cells. Stem Cell Res. Ther., 8 (1), 156.
- Li, Z., Okamoto, K., Hayashi, Y., & Sheng, M. (2004). The importance of dendritic mitochondria in the morphogenesis and plasticity of spines and synapses. Cell, 119(6), 873–887. https://doi.org/10.1016/j.cell.2004.11.003
- Liang, Dan, Angela L. Elwell, Nil Aygün, Oleh Krupa, Justin M. Wolter, Felix A. Kyere, Michael J. Lafferty, et al. 2021. "Cell-Type-Specific Effects of Genetic Variation on Chromatin Accessibility during Human Neuronal Differentiation." Nature Neuroscience 24 (7): 941–53.
- Liang, G. & Zhang, Y. 2013. Genetic and epigenetic variations in iPSCs: potential causes and implications for application. Cell Stem Cell, 13, 149-59.

- Liddelow, S. A. & Barres, B. A. Reactive Astrocytes: Production, Function, and Therapeutic Potential. Immunity 46, 957-967 (2017). https://doi.org:10.1016/j.immuni.2017.06.006
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009 Oct 9;326(5950):289-93. doi: 10.1126/science.1181369. PMID: 19815776; PMCID: PMC2858594.
- Limone, F. et al. Efficient generation of lower induced motor neurons by coupling Ngn2 expression with developmental cues. Cell Rep 42, 111896 (2023). https://doi.org:10.1016/j.celrep.2022.111896
- Lin, H. C. et al. NGN2 induces diverse neuron types from human pluripotency. Stem Cell Reports 16, 2118-2127 (2021). https://doi.org:10.1016/j.stemcr.2021.07.006
- Lin, Hung-Yu, Chia-Wei Liou, Shang-Der Chen, Te-Yao Hsu, Jiin-Haur Chuang, Pei-Wen Wang, Sheng-Teng Huang, et al. 2015. “Mitochondrial Transfer from Wharton’s Jelly-Derived Mesenchymal Stem Cells to Mitochondria-Defective Cells Recaptures Impaired Mitochondrial Function.” Mitochondrion 22 (May): 31–44.
- Lin, S. S., Delaura, S. & Jones, E. M. 2020. The CIRM iPSC repository. Stem Cell Res, 44, 101671.
- Lin, Stephen S., Susan DeLaura, and Eugenia M. Jones. 2020. “The CIRM iPSC Repository.” Stem Cell Research 44 (April): 101671.
- Linden D. E. (2012). The challenges and promise of neuroimaging in psychiatry. Neuron, 73(1), 8–22. https://doi.org/10.1016/j.neuron.2011.12.014
- Lisman, J.E. et al, (2008). Circuit-based framework for understanding neurotransmitter and risk gene interactions in schizophrenia. Trends Neurosci., 31, 234–242.
- Liu, L. R., Liu, J. C., Bao, J. S., Bai, Q. Q. & Wang, G. Q. Interaction of Microglia and Astrocytes in the Neurovascular Unit. Front Immunol 11, 1024 (2020). https://doi.org:10.3389/fimmu.2020.01024
- Liu, Q. et al, (2012). Human neural crest stem cells derived from human ESCs and induced pluripotent stem cells: induction, maintenance, and differentiation into functional schwann cells. Stem Cells Transl. Med., 1 (4), 266–278.
- Liu, W., Deng, Y., Liu, Y., Gong, W., Deng, W., (2013). Stem cell models for drug discovery and toxicology studies. J. Biochem. Mol. Toxicol.,. https://doi.org/10.1002/jbt.21470.
- Lo Sardo, V., Ferguson, W., Erikson, G. A., Topol, E. J., Baldwin, K. K. & Torkamani, A. 2017. Influence of donor age on induced pluripotent stem cells. Nat Biotechnol, 35, 69-74.

- Loh PR, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ; Schizophrenia Working Group of Psychiatric Genomics Consortium; de Candia TR, Lee SH, Wray NR, Kendler KS, O'Donovan MC, Neale BM, Patterson N, Price AL. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet. 2015 Dec;47(12):1385-92. doi: 10.1038/ng.3431. Epub 2015 Nov 2. PMID: 26523775; PMCID: PMC4666835.
- Loh, P. R., Genovese, G., Handsaker, R. E., Finucane, H. K., Reshef, Y. A., Palamara, P. F., Birmann, B. M., Talkowski, M. E., Bakhoum, S. F., Mccarroll, S. A. & Price, A. L. 2018. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature, 559, 350-355.
- Lorson, C. L., Hahnen, E., Androphy, E. J. & Wirth, B. 1999. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. Proc Natl Acad Sci U S A, 96, 6307-11.
- Lotsch, J. et al, (2009). Cross-sectional analysis of the influence of currently known pharmacogenetic modulators on opioid therapy in outpatient pain centers. Pharmacogenet. Genom., 19 (6), 429–436.
- 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). https://doi.org:10.1186/s13059-014-0550-8
- Luo, Z. et al, (2008). Race differences in nicotine dependence in the Collaborative Genetic study of Nicotine Dependence (COGEND). Nicotine Tob. Res., 10 (7), 1223–1230. https://doi.org/10.1080/ 14622200802163266.
- Macosko, E. Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202-1214 (2015). https://doi.org:10.1016/j.cell.2015.05.002
- Magnotti, L. M., Goodenough, D. A. & Paul, D. L. Functional heterotypic interactions between astrocyte and oligodendrocyte connexins. Glia 59, 26-34 (2011). https://doi.org:10.1002/glia.21073
- Mahajani, S. et al, (2019). Homogenous generation of dopaminergic neurons from multiple hiPSC lines by transient expression of transcription factors. Cell Death Dis., 10 (12), 898.
- Mahajani, S., Raina, A., Fokken, C., Ku ⍰ gler, S., Ba ⍰ hr, M., (2019). Homogenous generation of dopaminergic neurons from multiple hiPSC lines by transient expression of transcription factors. Cell Death Dis., 10, 898.
- Maier RM, Visscher PM, Robinson MR, Wray NR. Embracing polygenicity: a review of methods and tools for psychiatric genetics research. Psychol Med. 2018 May;48(7):1055-1067. doi: 10.1017/S0033291717002318. Epub 2017 Aug 29. PMID: 28847336; PMCID: PMC6088780.

- Maisenbacher MK, Merrion K, Pettersen B, Young M, Paik K, Iyengar S, Kareht S, Sigurjonsson S, Demko ZP, Martin KA. Incidence of the 22q11.2 deletion in a large cohort of miscarriage samples. Mol Cytogenet. 2017 Mar 9;10:6. doi: 10.1186/s13039-017-0308-6. PMID: 28293297; PMCID: PMC5345148.
- Marbach, Felix, Georgi Stoyanov, Florian Erger, Constantine A. Stratakis, Nikolaos Settas, Edra London, Jill A. Rosenfeld, et al. 2021. "Variants in PRKAR1B Cause a Neurodevelopmental Disorder with Autism Spectrum Disorder, Apraxia, and Insensitivity to Pain." Genetics in Medicine: Official Journal of the American College of Medical Genetics 23 (8): 1465–73.
- Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, Antaki D, Shetty A, Holmans PA, Pinto D, Gujral M, Brandler WM, Malhotra D, Wang Z, Fajarado KVF, Maile MS, Ripke S, Agartz I, Albus M, Alexander M, Amin F, Atkins J, Bacanu SA, Belliveau RA Jr, Bergen SE, Bertalan M, Bevilacqua E, Bigdeli TB, Black DW, Bruggeman R, Buccola NG, Buckner RL, Bulik-Sullivan B, Byerley W, Cahn W, Cai G, Cairns MJ, Campion D, Cantor RM, Carr VJ, Carrera N, Catts SV, Chambert KD, Cheng W, Cloninger CR, Cohen D, Cormican P, Craddock N, Crespo-Facorro B, Crowley JJ, Curtis D, Davidson M, Davis KL, Degenhardt F, Del Favero J, DeLisi LE, Dikeos D, Dinan T, Djurovic S, Donohoe G, Drapeau E, Duan J, Dudbridge F, Eichhammer P, Eriksson J, Escott-Price V, Essioux L, Fanous AH, Farh KH, Farrell MS, Frank J, Franke L, Freedman R, Freimer NB, Friedman JI, Forstner AJ, Fromer M, Genovese G, Georgieva L, Gershon ES, Giegling I, Giusti-Rodríguez P, Godard S, Goldstein JI, Gratten J, de Haan L, Hamshere ML, Hansen M, Hansen T, Haroutunian V, Hartmann AM, Henskens FA, Herms S, Hirschhorn JN, Hoffmann P, Hofman A, Huang H, Ikeda M, Joa I, Kähler AK, Kahn RS, Kalaydjieva L, Karjalainen J, Kavanagh D, Keller MC, Kelly BJ, Kennedy JL, Kim Y, Knowles JA, Konte B, Laurent C, Lee P, Lee SH, Legge SE, Lerer B, Levy DL, Liang KY, Lieberman J, Lönnqvist J, Loughland CM, Magnusson PKE, Maher BS, Maier W, Mallet J, Mattheisen M, Mattingsdal M, McCarley RW, McDonald C, McIntosh AM, Meier S, Meijer CJ, Melle I, Mesholam-Gately RI, Metspalu A, Michie PT, Milani L, Milanova V, Mokrab Y, Morris DW, Müller-Myhsok B, Murphy KC, Murray RM, Myin-Germeys I, Nenadic I, Nertney DA, Nestadt G, Nicodemus KK, Nisenbaum L, Nordin A, O'Callaghan E, O'Dushlaine C, Oh SY, Olincy A, Olsen L, O'Neill FA, Van Os J, Pantelis C, Papadimitriou GN, Parkhomenko E, Pato MT, Paunio T; Psychosis Endophenotypes International Consortium; Perkins DO, Pers TH, Pietiläinen O, Pimm J, Pocklington AJ, Powell J, Price A, Pulver AE, Purcell SM, Quested D, Rasmussen HB, Reichenberg A, Reimers MA, Richards AL, Roffman JL, Roussos P, Ruderfer DM, Salomaa V, Sanders AR, Savitz A, Schall U, Schulze TG, Schwab SG, Scolnick EM, Scott RJ, Seidman LJ, Shi J, Silverman JM, Smoller JW, Söderman E, Spencer CCA, Stahl EA, Strengman E, Strohmaier J, Stroup TS, Suvisaari J, Svrakic DM, Szatkiewicz JP, Thirumalai S, Tooney PA, Veijola J, Visscher PM, Waddington J, Walsh D, Webb BT, Weiser M, Wildenauer

DB, Williams NM, Williams S, Witt SH, Wolen AR, Wormley BK, Wray NR, Wu JQ, Zai CC, Adolfsson R, Andreassen OA, Blackwood DHR, Bramon E, Buxbaum JD, Cichon S, Collier DA, Corvin A, Daly MJ, Darvasi A, Domenici E, Esko T, Gejman PV, Gill M, Gurling H, Hultman CM, Iwata N, Jablensky AV, Jönsson EG, Kendler KS, Kirov G, Knight J, Levinson DF, Li QS, McCarroll SA, McQuillin A, Moran JL, Mowry BJ, Nöthen MM, Ophoff RA, Owen MJ, Palotie A, Pato CN, Petryshen TL, Posthuma D, Rietschel M, Riley BP, Rujescu D, Sklar P, St Clair D, Walters JTR, Werge T, Sullivan PF, O'Donovan MC, Scherer SW, Neale BM, Sebat J; CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet. 2017 Jan;49(1):27-35. doi: 10.1038/ng.3725. Epub 2016 Nov 21. Erratum in: Nat Genet. 2017 Mar 30;49(4):651. Erratum in: Nat Genet. 2017 Sep 27;49(10):1558. PMID: 27869829; PMCID: PMC5737772.

- Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019 Apr;51(4):584-591. doi: 10.1038/s41588-019-0379-x. Epub 2019 Mar 29. Erratum in: Nat Genet. 2021 May;53(5):763. PMID: 30926966; PMCID: PMC6563838.
- Martin, A.R. et al, (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genet., 51, 584–591.
- Martin, G.R., (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc. Natl. Acad. Sci., 78, 7634–7638.
- Marx, V., (2020). Reality check for organoids in neuroscience. Nature Methods, 17 (10), 961–964.
- Masdeu J. C. (2011). Neuroimaging in psychiatric disorders. Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics, 8(1), 93–102. https://doi.org/10.1007/s13311-010-0006-0
- Mashtalir N, Suzuki H, Farrell DP, Sankar A, Luo J, Filipovski M, D'Avino AR, St Pierre R, Valencia AM, Onikubo T, Roeder RG, Han Y, He Y, Ranish JA, DiMaio F, Walz T, Kadoch C. A Structural Model of the Endogenous Human BAF Complex Informs Disease Mechanisms. Cell. 2020 Oct 29;183(3):802-817.e24. doi: 10.1016/j.cell.2020.09.051. Epub 2020 Oct 13. PMID: 33053319; PMCID: PMC7717177.
- Massaad, C. A., & Klann, E. (2011). Reactive oxygen species in the regulation of synaptic plasticity and memory. Antioxidants & redox signaling, 14(10), 2013–2054. https://doi.org/10.1089/ars.2010.3208
- Matume, N.D. et al, (2019). Characterization of APOBEC3 variation in a population of HIV-1 infected individuals in northern South Africa. BMC Med. Genet., 20 (1), 21.
- Maynard, T. M., Meechan, D. W., Dudevoir, M. L., Gopalakrishna, D., Peters, A. Z., Heindel, C. C., Sugimoto, T. J., Wu, Y., Lieberman, J. A., & Lamantia, A. S. (2008).

Mitochondrial localization and function of a subset of 22q11 deletion syndrome candidate genes. Molecular and cellular neurosciences, 39(3), 439–451. https://doi.org/10.1016/j.mcn.2008.07.027

- McDonald-McGinn DM, Sullivan KE, Marino B, Philip N, Swillen A, Vorstman JA, Zackai EH, Emanuel BS, Vermeesch JR, Morrow BE, Scambler PJ, Bassett AS. 22q11.2 deletion syndrome. Nat Rev Dis Primers. 2015 Nov 19;1:15071. doi: 10.1038/nrdp.2015.71. PMID: 27189754; PMCID: PMC4900471.
- Mcfarland, J. M., Paolella, B. R., Warren, A., Geiger-Schuller, K., Shibue, T., Rothberg, M., Kuksenko, O., Jones, A., Chambers, E., Dionne, D., Bender, S., Wolpin, B. W., Ghandi, M., Tirosh, I., Rozenblatt-Rosen, O., Roth, J. A., Tgolub, T. R., Regev, A., Aguirre, A. J., Vazquez, F. & Tsherniak, A. 2019. Multiplexed single-cell profiling of post-perturbation transcriptional responses to define cancer vulnerabilities and therapeutic mechanism of action. bioRxiv.
- McNeish, J. et al, (2010). High-throughput screening in embryonic stem cell-derived neurons identifies potentiators of alpha-amino-3-hydroxyl-5-methyl-4-isoxazolepropionate-type glutamate receptors. J. Biol. Chem., 285 (22), 17209–17217.
- McNeish, J., Gardner, J.P., Wainger, B.J., Woolf, C.J., Eggan, K., (2015). From Dish to Bedside: Lessons Learned While Translating Findings from a Stem Cell Model of Disease to a Clinical Trial. Cell Stem Cell, 17, 8– 10.
- McPhie, D.L. et al, (2018). Oligodendrocyte differentiation of induced pluripotent stem cells derived from subjects with schizophrenias implicate abnormalities in development. Transl. Psychiatry, 8 (1), 230.
- McQuin, Claire, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, Beth A. Cimini, Kyle W. Karhohs, Minh Doan, et al. 2018. “CellProfiler 3.0: Next-Generation Image Processing for Biology.” PLoS Biology 16 (7): e2005970.
- Merkle FT, Ghosh S, Genovese G, Handsaker RE, Kashin S, Meyer D, Karczewski KJ, O'Dushlaine C, Pato C, Pato M, MacArthur DG, McCarroll SA, Eggan K. Whole-genome analysis of human embryonic stem cells enables rational line selection based on genetic variation. Cell Stem Cell. 2022 Mar 3;29(3):472-486.e7. doi: 10.1016/j.stem.2022.01.011. Epub 2022 Feb 16. PMID: 35176222; PMCID: PMC8900618.
- Merkle, F. T., Ghosh, S., Kamitaki, N., Mitchell, J., Avior, Y., Mello, C., Kashin, S., Mekhoubad, S., Ilic, D., Charlton, M., Saphier, G., Handsaker, R. E., Genovese, G., Bar, S., Benvenisty, N., Mccarroll, S. A. & Eggan, K. 2017. Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations. Nature, 545, 229-233.
- Merkle, F.T. et al, (2017). Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations. Nature, 545, 229–233.
- Mertens, J. et al, (2015). Directly Reprogrammed Human Neurons Retain Aging-Associated Transcriptomic Signatures and Reveal Age-Related Nucleocytoplasmic Defects. Cell Stem Cell, 17, 705–718.

- Mertens, J., Wang, Q. W., Kim, Y., Yu, D. X., Pham, S., Yang, B., Zheng, Y., Diffenderfer, K. E., Zhang, J., Soltani, S., Eames, T., Schafer, S. T., Boyer, L., Marchetto, M. C., Nurnberger, J. I., Calabrese, J. R., Ødegaard, K. J., McCarthy, M. J., Zandi, P. P., Alda, M., … Yao, J. (2015). Differential responses to lithium in hyperexcitable neurons from patients with bipolar disorder. Nature, 527(7576), 95–99. https://doi.org/10.1038/nature15526
- Meyer, K., Marquis, J., Trub, J., Nlend Nlend, R., Verp, S., Ruepp, M. D., Imboden, H., Barde, I., Trono, D. & Schumperli, D. 2009. Rescue of a severe mouse model for spinal muscular atrophy by U7 snRNA-mediated splicing modulation. Hum Mol Genet, 18, 546-55.
- Miranda-Azpiazu, P. et al, (2018). A novel dynamic multicellular co-culture system for studying individual blood-brain barrier cell types in brain diseases and cytotoxicity testing. Sci. Rep., 8 (1), 8784.
- Mitchell, J.M. et al, (2020). Mapping genetic effects on cellular phenotypes with "cell villages.". Biorxiv,. https:// doi.org/10.1101/2020.06.29.174383. 2020.06.29.174383.
- Mizuno, G. O. et al. Aberrant Calcium Signaling in Astrocytes Inhibits Neuronal Excitability in a Human Down Syndrome Stem Cell Model. Cell Rep 24, 355-365 (2018). https://doi.org:10.1016/j.celrep.2018.06.033
- Monani, U. R., Lorson, C. L., Parsons, D. W., Prior, T. W., Androphy, E. J., Burghes, A. H. & Mcpherson, J. D. 1999. A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2. Hum Mol Genet, 8, 1177-83.
- Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S. & Cheung, V. G. 2004. Genetic analysis of genome-wide variation in human gene expression. Nature, 430, 743-7.
- Mortensen, P. B., Pedersen, M. G., & Pedersen, C. B. (2010). Psychiatric family history and schizophrenia risk in Denmark: which mental disorders are relevant?. Psychological medicine, 40(2), 201–210. https://doi.org/10.1017/S0033291709990419
- Motahari Z, Moody SA, Maynard TM, LaMantia AS. In the line-up: deleted genes associated with DiGeorge/22q11.2 deletion syndrome: are they all suspects? J Neurodev Disord. 2019 Jun 7;11(1):7. doi: 10.1186/s11689-019-9267-z. PMID: 31174463; PMCID: PMC6554986.
- Mouton, J.P. et al, (2016). Adverse Drug Reactions Causing Admission to Medical Wards. Medicine, 95, e3437.
- Nadadhur, A.G. et al, (2019). Neuron-Glia Interactions Increase Neuronal Phenotypes in Tuberous Sclerosis Complex Patient iPSC-Derived Models. Stem Cell Rep., 12 (1), 42–56.
- Nagar, S. et al, (2020). Population structure and pharmacogenomic risk stratification in the United States. BMC Biol., 18, 140. https://doi.org/10.1186/s12915-020- 00875-4.

- Nakayama RT, Pulice JL, Valencia AM, McBride MJ, McKenzie ZM, Gillespie MA, Ku WL, Teng M, Cui K, Williams RT, Cassel SH, Qing H, Widmer CJ, Demetri GD, Irizarry RA, Zhao K, Ranish JA, Kadoch C. SMARCB1 is required for widespread BAF complex-mediated activation of enhancers and bivalent promoters. Nat Genet. 2017 Nov;49(11):1613-1623. doi: 10.1038/ng.3958. Epub 2017 Sep 25. PMID: 28945250; PMCID: PMC5803080.
- Napoli, E., Tassone, F., Wong, S., Angkustsiri, K., Simon, T. J., Song, G., & Giulivi, C. (2015). Mitochondrial Citrate Transporter-dependent Metabolic Signature in the 22q11.2 Deletion Syndrome. The Journal of biological chemistry, 290(38), 23240–23253. https://doi.org/10.1074/jbc.M115.672360
- Naqvi S, Kim S, Hoskens H, Matthews HS, Spritz RA, Klein OD, Hallgrímsson B, Swigut T, Claes P, Pritchard JK, Wysocka J. Precise modulation of transcription factor levels identifies features underlying dosage sensitivity. Nat Genet. 2023 May;55(5):841-851. doi: 10.1038/s41588-023-01366-2. Epub 2023 Apr 6. PMID: 37024583; PMCID: PMC10181932.
- Naryshkin, N. et al, (2014). Motor neuron disease. SMN2 splicing modifiers improve motor function and longevity in mice with spinal muscular atrophy. Science,. https://doi.org/10.1126/science.1250127.
- Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, Jones TR, Nguyen TH, Ulirsch JC, Lekschas F, Mualim K, Natri HM, Weeks EM, Munson G, Kane M, Kang HY, Cui A, Ray JP, Eisenhaure TM, Collins RL, Dey K, Pfister H, Price AL, Epstein CB, Kundaje A, Xavier RJ, Daly MJ, Huang H, Finucane HK, Hacohen N, Lander ES, Engreitz JM. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021 May;593(7858):238-243. doi: 10.1038/s41586-021-03446-x. Epub 2021 Apr 7. PMID: 33828297; PMCID: PMC9153265.

- Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, Jones TR, Nguyen TH, Ulirsch JC, Lekschas F, Mualim K, Natri HM, Weeks EM, Munson G, Kane M, Kang HY, Cui A, Ray JP, Eisenhaure TM, Collins RL, Dey K, Pfister H, Price AL, Epstein CB, Kundaje A, Xavier RJ, Daly MJ, Huang H, Finucane HK, Hacohen N, Lander ES, Engreitz JM. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021 May;593(7858):238-243. doi: 10.1038/s41586-021-03446-x. Epub 2021 Apr 7. PMID: 33828297; PMCID: PMC9153265.

- Nehme R, Pietiläinen O, Artomov M, Tegtmeyer M, Valakh V, Lehtonen L, Bell C, Singh T, Trehan A, Sherwood J, Manning D, Peirent E, Malik R, Guss EJ, Hawes D, Beccard A, Bara AM, Hazelbaker DZ, Zuccaro E, Genovese G, Loboda AA, Neumann A, Lilliehook C, Kuismin O, Hamalainen E, Kurki M, Hultman CM, Kähler AK, Paulo JA, Ganna A, Madison J, Cohen B, McPhie D, Adolfsson R, Perlis R, Dolmetsch R, Farhi S, McCarroll S, Hyman S, Neale B, Barrett LE, Harper W, Palotie A, Daly M, Eggan K. The 22q11.2 region regulates presynaptic gene-products linked to schizophrenia. Nat Commun. 2022 Jun

27;13(1):3690. doi: 10.1038/s41467-022-31436-8. PMID: 35760976; PMCID: PMC9237031.

- Nehme, R. et al, (2018). Combining NGN2 Programming with Developmental Patterning Generates Human Excitatory Neurons with NMDAR-Mediated Synaptic Transmission. Cell Rep., 23 (8), 2509–2523.
- Nehme, R. et al, (2018). Combining NGN2 Programming with Developmental Patterning Generates Human Excitatory Neurons with NMDAR-Mediated Synaptic Transmission. Cell Rep., 23, 2509–2523.
- Nehme, R. et al. Combining NGN2 Programming with Developmental Patterning Generates Human Excitatory Neurons with NMDAR-Mediated Synaptic Transmission. Cell Reports 23, 2509-2523 (2018).
- Nehme, R. et al. The 22q11.2 region regulates presynaptic gene-products linked to schizophrenia. Nat Commun 13, 3690 (2022). https://doi.org:10.1038/s41467-022-31436-8
- Nehme, R., Barrett, L., (2020). Using human pluripotent stem cell models to study autism in the era of big data. Mol. Autism, 11, 21.
- Nehme, R., Pietiläinen, O., Artomov, M., Tegtmeyer, M., Valakh, V., Lehtonen, L., Bell, C., Singh, T., Trehan, A., Sherwood, J., Manning, D., Peirent, E., Malik, R., Guss, E. J., Hawes, D., Beccard, A., Bara, A. M., Hazelbaker, D. Z., Zuccaro, E., Genovese, G., … Eggan, K. (2022). The 22q11.2 region regulates presynaptic gene-products linked to schizophrenia. Nature communications, 13(1), 3690. https://doi.org/10.1038/s41467-022-31436-8
- Nehme, R., Zuccaro, E., Ghosh, S. D., Li, C., Sherwood, J. L., Pietilainen, O., Barrett, L. E., Limone, F., Worringer, K. A., Kommineni, S., Zang, Y., Cacchiarelli, D., Meissner, A., Adolfsson, R., Haggarty, S., Madison, J., Muller, M., Arlotta, P., Fu, Z., Feng, G. & Eggan, K. 2018. Combining NGN2 Programming with Developmental Patterning Generates Human Excitatory Neurons with NMDAR-Mediated Synaptic Transmission. Cell Rep, 23, 2509-2523.
- Nehme, R., Zuccaro, E., Ghosh, S. D., Li, C., Sherwood, J. L., Pietilainen, O., Barrett, L. E., Limone, F., Worringer, K. A., Kommineni, S., Zang, Y., Cacchiarelli, D., Meissner, A., Adolfsson, R., Haggarty, S., Madison, J., Muller, M., Arlotta, P., Fu, Z., Feng, G., Eggan, K. (2018). Combining NGN2 Programming with Developmental Patterning Generates Human Excitatory Neurons with NMDAR-Mediated Synaptic Transmission. Cell reports, 23(8), 2509–2523. https://doi.org/10.1016/j.celrep.2018.04.066
- Neimark, J. 2015. Line of attack. Science, 347, 938-40.
- Nejad, A. B., Ebdrup, B. H., Glenthøj, B. Y., & Siebner, H. R. (2012). Brain connectivity studies in schizophrenia: unravelling the effects of antipsychotics. Current neuropharmacology, 10(3), 219–230. https://doi.org/10.2174/157015912803217305

- Nelson-Rees, W. A., Daniels, D. W. & Flandermeyer, R. R. 1981. Cross- contamination of cells in culture. *Science,* 212**,** 446-52.
- Nestler, E. J., & Hyman, S. E. (2010). Animal models of neuropsychiatric disorders. Nature neuroscience, 13(10), 1161–1169. https://doi.org/10.1038/nn.2647
- Network, B. I. C. C. A multimodal cell census and atlas of the mammalian primary motor cortex. Nature 598, 86-102 (2021). https://doi.org:10.1038/s41586-021-03950-0
- NIH ASTHMA 8.4 10-28-B (1995).
- Nishizawa, M., Chonabayashi, K., Nomura, M., Tanaka, A., Nakamura, M., Inagaki, A., Nishikawa, M., Takei, I., Oishi, A., Tanabe, K., Ohnuki, M., Yokota, H., Koyanagi-Aoi, M., Okita, K., Watanabe, A., Takaori-Kondo, A., Yamanaka, S. & Yoshida, Y. 2016. Epigenetic Variation between Human Induced Pluripotent Stem Cell Lines Is an Indicator of Differentiation Capacity. *Cell Stem Cell,* 19**,** 341-54.
- Novikova, G. et al, (2021). Integration of Alzheimer’s disease genetics and myeloid genomics identifies disease risk regulatory elements and genes. Nature Commun., 12 (1), 1610.
- O. Schmetzer, A. Flo ⍰rcken, Sex Differences in the Drug Therapy for Oncologic Diseases, Springer, Berlin, Heidelberg, n.d. doi: https://doi.org/10.1007/978-3-642-30726-3_19.
- Obashi, Kazuki, and Shigeo Okabe. 2013. “Regulation of Mitochondrial Dynamics and Distribution by Synapse Position and Neuronal Activity in the Axon.” The European Journal of Neuroscience 38 (3): 2350–63.
- Oberheim, N. A., Wang, X., Goldman, S. & Nedergaard, M. Astrocytic complexity distinguishes the human brain. Trends Neurosci 29, 547-553 (2006). https://doi.org:10.1016/j.tins.2006.08.004
- Odeh, H. et al, (2015). The Biobank Economic Modeling Tool (BEMT): Online Financial Planning to Facilitate Biobank Sustainability. Biopreserv. Biobank, 13 (6), 421–429.
- Olsen L, Sparsø T, Weinsheimer SM, Dos Santos MBQ, Mazin W, Rosengren A, Sanchez XC, Hoeffding LK, Schmock H, Baekvad-Hansen M, Bybjerg-Grauholm J, Daly MJ, Neale BM, Pedersen MG, Agerbo E, Mors O, Børglum A, Nordentoft M, Hougaard DM, Mortensen PB, Geschwind DH, Pedersen C, Thompson WK, Werge T. Prevalence of rearrangements in the 22q11.2 region and population-based risk of neuropsychiatric and developmental disorders in a Danish population: a case-cohort study. Lancet Psychiatry. 2018 Jul;5(7):573-580. doi: 10.1016/S2215-0366(18)30168-8. Epub 2018 Jun 7. PMID: 29886042; PMCID: PMC6560180.
- Onwordi, E. C., Halff, E. F., Whitehurst, T., Mansur, A., Cotel, M. C., Wells, L., Creeney, H., Bonsall, D., Rogdaki, M., Shatalina, E., Reis Marques, T., Rabiner, E. A., Gunn, R. N., Natesan, S., Vernon, A. C., & Howes, O. D. (2020). Synaptic density marker SV2A is reduced in schizophrenia patients and unaffected by antipsychotics in rats. Nature communications, 11(1), 246. https://doi.org/10.1038/s41467-019-14122-0

- Pak, C. et al, (2021). Cross-platform validation of neurotransmitter release impairments in schizophrenia patient-derived NRXN1-mutant neurons. Proc. Natl. Acad. Sci. U. S. A., 118 (22)
- Palacino, J., Swalley, S. E., Song, C., Cheung, A. K., Shu, L., Zhang, X., Van Hoosear, M., Shin, Y., Chin, D. N., Keller, C. G., Beibel, M., Renaud, N. A., Smith, T. M., Salcius, M., Shi, X., Hild, M., Servais, R., Jain, M., Deng, L., Bullock, C., Mclellan, M., Schuierer, S., Murphy, L., Blommers, M. J., Blaustein, C., Berenshteyn, F., Lacoste, A., Thomas, J. R., Roma, G., Michaud, G. A., Tseng, B. S., Porter, J. A., Myer, V. E., Tallarico, J. A., Hamann, L. G., Curtis, D., Fishman, M. C., Dietrich, W. F., Dales, N. A. & Sivasankaran, R. 2015. SMN2 splice modulators enhance U1-pre-mRNA association and rescue SMA mice. *Nat Chem Biol,* 11**,** 511-7.
- Panopoulos, Athanasia D., Matteo D'Antonio, Paola Benaglio, Roy Williams, Sherin I. Hashem, Bernhard M. Schuldt, Christopher DeBoever, et al. 2017. "iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types." Stem Cell Reports 8 (4): 1086–1100.
- Parikh, T., Walkup, J.T., (2021). The Future of Ketamine in the Treatment of Teen Depression. Am. J. Psychiat., 178, 288–289.
- Pasca, S.P. et al, (2011). Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome. Nature Med., 17, 1657–1662.
- Pashos, Evanthia E., Yoson Park, Xiao Wang, Avanthi Raghavan, Wenli Yang, Deepti Abbey, Derek T. Peters, et al. 2017. "Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci." Cell Stem Cell 20 (4): 558–70.e10.
- Pei, Y. et al, (2016). Comparative neurotoxicity screening in human iPSC-derived neural stem cells, neurons and astrocytes. Brain Res., 1638, 57–73.
- Pernas, Lena, and Luca Scorrano. 2016. "Mito-Morphosis: Mitochondrial Fusion, Fission, and Cristae Remodeling as Key Mediators of Cellular Function." Annual Review of Physiology 78: 505–31.
- Pfrieger, F. W. & Barres, B. A. Synaptic efficacy enhanced by glial cells in vitro. Science 277, 1684-1687 (1997). https://doi.org:10.1126/science.277.5332.1684
- Phelan K, McDermid HE. The 22q13.3 Deletion Syndrome (Phelan-McDermid Syndrome). Mol Syndromol. 2012 Apr;2(3-5):186-201. doi: 10.1159/000334260. Epub 2011 Nov 22. PMID: 22670140; PMCID: PMC3366702.
- Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J. B., Stephens, M., Gilad, Y. & Pritchard, J. K. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature,* 464**,** 768-72.

- Pietilainen, O. et al. Astrocytic cell adhesion genes linked to schizophrenia correlate with synaptic programs in neurons. Cell Rep 42, 111988 (2023). https://doi.org:10.1016/j.celrep.2022.111988
- Pietiläinen, O., Trehan, A., Meyer, D., Mitchell, J., Tegtmeyer, M., Valakh, V., Gebre, H., Chen, T., Vartiainen, E., Farhi, S. L., Eggan, K., McCarroll, S. A., & Nehme, R. (2023). Astrocytic cell adhesion genes linked to schizophrenia correlate with synaptic programs in neurons. Cell reports, 42(1), 111988. https://doi.org/10.1016/j.celrep.2022.111988
- Pintacuda, G., Martin, J.M., Eggan, K.C., (2021). Mind the translational gap: using iPS cell models to bridge from genetic discoveries to perturbed pathways and therapeutic targets. Mol. Autism, 12, 10.
- Polanski, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics (Oxford, England) 36, 964-965 (2020). https://doi.org:10.1093/bioinformatics/btz625
- Polioudakis, D. et al. A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. Neuron 103, 785-801 e788 (2019). https://doi.org:10.1016/j.neuron.2019.06.011
- Ponroy Bally, B. & Murai, K. K. Astrocytes in Down Syndrome Across the Lifespan. Front Cell Neurosci 15, 702685 (2021). https://doi.org:10.3389/fncel.2021.702685
- Ponroy Bally, B. et al. Human iPSC-derived Down syndrome astrocytes display genome-wide perturbations in gene expression, an altered adhesion profile, and increased cellular dynamics. Hum Mol Genet 29, 785-802 (2020). https://doi.org:10.1093/hmg/ddaa003
- Powell, C. M., & Miyakawa, T. (2006). Schizophrenia-relevant behavioral testing in rodent models: a uniquely human disorder?. Biological psychiatry, 59(12), 1198–1207. https://doi.org/10.1016/j.biopsych.2006.05.008
- Quadrato, G. et al, (2017). Cell diversity and network dynamics in photosensitive human brain organoids. Nature, 545, 48–53.
- 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. (2017). Cell diversity and network dynamics in photosensitive human brain organoids. Nature, 545(7652), 48–53. https://doi.org/10.1038/nature22047
- Quintanilla-Martinez, L., Kremer, M., Keller, G., Nathrath, M., Gamboa-Dominguez, A., Meneses, A., Luna-Contreras, L., Cabras, A., Hoefler, H., Mohar, A., & Fend, F. (2001). p53 Mutations in nasal natural killer/T-cell lymphoma from Mexico: association with large cell morphology and advanced disease. The American journal of pathology, 159(6), 2095–2105. https://doi.org/10.1016/S0002-9440(10)63061-1

- Quist, E. et al, (2021). Transcription Factor Programming of Human Pluripotent Stem Cells to Functionally Mature Astrocytes for Monocultures and Cocultures with Neurons. Methods Mol. Biol., 2352, 133–148.
- Ramdas, S. & Servais, L. 2020. New treatments in spinal muscular atrophy: an overview of currently available data. *Expert Opin Pharmacother,* 21, 307-315.
- Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014 Dec 18;159(7):1665-80. doi: 10.1016/j.cell.2014.11.021. Epub 2014 Dec 11. Erratum in: Cell. 2015 Jul 30;162(3):687-8. PMID: 25497547; PMCID: PMC5635824.
- Rapino, F. et al. Small-molecule screen reveals pathways that regulate C4 secretion in stem cell-derived astrocytes. Stem Cell Reports (2022). https://doi.org:10.1016/j.stemcr.2022.11.018
- Rathorpe, S. et al, (2002). Sex-based differences in the effect of digoxin for the treatment of heart failure. N. Engl. J. Med., 347 (18)
- Raya, A. et al, (2009). Disease-corrected haematopoietic progenitors from Fanconi anaemia induced pluripotent stem cells. Nature, 460 (7251), 53–59.
- Raz, L., Miller, V.M., (2012). Considerations of sex and gender differences in preclinical and clinical trials. Handb. Exp. Pharmacol., 214, 127–147.
- Ripke et al, (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. Medrxiv,. https://doi.org/10.1101/2020.09.12.20192922.
- Roadmap Epigenomics Consortium; Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb 19;518(7539):317-30. doi: 10.1038/nature14248. PMID: 25693563; PMCID: PMC4530010.
- Roberts R. C. (2017). Postmortem studies on mitochondria in schizophrenia. Schizophrenia research, 187, 17–25. https://doi.org/10.1016/j.schres.2017.01.056

- Robicsek, O., Karry, R., Petit, I., Salman-Kesner, N., Müller, F-J., Klein, E., Aberdam, D., & Ben-Shachar, D. (2013). Abnormal neuronal differentiation and mitochondrial dysfunction in hair follicle-derived induced pluripotent stem cells of schizophrenia patients. Molecular psychiatry, 18(10), 1067-1076. https://doi.org/10.1038/mp.2013.67
- Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) 26, 139-140 (2010). https://doi.org:10.1093/bioinformatics/btp616
- Rohban, Mohammad Hossein, Shantanu Singh, Xiaoyun Wu, Julia B. Berthet, Mark-Anthony Bray, Yashaswi Shrestha, Xaralabos Varelas, Jesse S. Boehm, and Anne E. Carpenter. 2017. "Systematic Morphological Profiling of Human Gene and Allele Function via Cell Painting." eLife 6 (March): e24060.
- Ronen, D., Benvenisty, N., (2014). Sex-dependent gene expression in human pluripotent stem cells. Cell Rep., 8 (4), 923–932.
- Rouhani, F., Kumasaka, N., De Brito, M. C., Bradley, A., Vallier, L. & Gaffney, D. 2014. Genetic background drives transcriptional variation in human induced pluripotent stem cells. *PLoS Genet,* 10**,** e1004432.
- Rowe, R.G., Daley, G.Q., (2019). Induced pluripotent stem cells in disease modelling and drug discovery. Nature Rev. Genet., 20, 377–388.
- Sabolic, I. et al, (2007). Gender differences in kidney function. Pflugers Arch., 455 (3), 397–429.
- Sackmann-Sala, Lucila, Jacques-Emmanuel Guidotti, and Vincent Goffin. 2015. "Minireview: Prolactin Regulation of Adult Stem Cells." Molecular Endocrinology 29 (5): 667–81.
- Santos, R. et al. Differentiation of Inflammation-Responsive Astrocytes from Glial Progenitors Generated from Human Induced Pluripotent Stem Cells. Stem Cell Reports 8, 1757-1769 (2017). https://doi.org:10.1016/j.stemcr.2017.05.011
- Satterstrom, F.K. et al, (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell, 180, 568–584.e23.
- Schiff, Lauren, Bianca Migliori, Ye Chen, Deidre Carter, Caitlyn Bonilla, Jenna Hall, Minjie Fan, et al. 2022. "Integrating Deep Learning and Unbiased Automated High-Content Screening to Identify Complex Disease Signatures in Human Fibroblasts." Nature Communications 13 (1): 1–13.
- Schizophrenia Working Group Of The Psychiatric Genomics, C. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature,* 511**,** 421-7.
- Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. Science (New York, N.Y.), 372(6537), eabf4740. https://doi.org/10.1126/science.abf4740

- Schrenk-Siemens, Katrin, Hagen Wende, Vincenzo Prato, Kun Song, Charlotte Rostock, Alexander Loewer, Jochen Utikal, Gary R. Lewin, Stefan G. Lechner, and Jan Siemens. 2015. "PIEZO2 Is Required for Mechanotransduction in Human Stem Cell-Derived Touch Receptors." Nature Neuroscience 18 (1): 10–16.
- Serra, Gregorio, Vincenzo Antona, Giovanni Corsello, Federico Zara, Ettore Piro, and Raffaele Falsaperla. 2019. "NF1 Microdeletion Syndrome: Case Report of Two New Patients." Italian Journal of Pediatrics 45 (1): 138.
- Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelson, T., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G. & Zhang, F. 2014. Genome- scale CRISPR-Cas9 knockout screening in human cells. *Science,* 343**,** 84-87.
- Shih, Yu-Tzu, Tzyy-Nan Huang, Hsiao-Tang Hu, Tzu-Li Yen, and Yi-Ping Hsueh. 2020. "Vcp Overexpression and Leucine Supplementation Increase Protein Synthesis and Improve Fear Memory and Social Interaction of Nf1 Mutant Mice." Cell Reports 31 (13): 107835.
- Sidhaye, J., Knoblich, J., (2006). Brain organoids: an ensemble of bioassays to investigate human neurodevelopment and disease. Cell Death Differ.,. https://doi.org/10.1038/s41418-020-0566-4.
- Singh, R. N. & Singh, N. N. 2018. Mechanism of Splicing Regulation of Spinal Muscular Atrophy Genes. *Adv Neurobiol,* 20**,** 31-61.
- Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J. D., Bass, N., Bigdeli, T. B., Breen, G., Bromet, E. J., Buckley, P. F., Bunney, W. E., Bybjerg-Grauholm, J., Byerley, W. F., Chapman, S. B., Chen, W. J., Churchhouse, C., Craddock, N., Cusick, C. M., DeLisi, L., … Daly, M. J. (2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature, 604(7906), 509–516. https://doi.org/10.1038/s41586-022-04556-w
- Sirugo, G., Williams, S.M., Tishkoff, S.A., (2019). The Missing Diversity in Human Genetic Studies. Cell, 177, 26–31.
- Sloan, S. A. et al. Human Astrocyte Maturation Captured in 3D Cerebral Cortical Spheroids Derived from Pluripotent Stem Cells. Neuron 95, 779-790 e776 (2017). https://doi.org:10.1016/j.neuron.2017.07.035
- Smits, L.M., Schwamborn, J.C., (2020). Midbrain Organoids: A New Tool to Investigate Parkinson's Disease. Front. Cell Dev. Biol., 8, 359.
- Sofroniew, M. V. & Vinters, H. V. Astrocytes: biology and pathology. Acta Neuropathol 119, 7-35 (2010). https://doi.org:10.1007/s00401-009-0619-8
- Soldner, F. et al, (2009). Parkinson's disease patient- derived induced pluripotent stem cells free of viral reprogramming factors. Cell, 136 (5), 964–977.
- Srivastava, R., Faust, T., Ramos, A., Ishizuka, K., & Sawa, A. (2018). Dynamic Changes of the Mitochondria in Psychiatric Illnesses: New Mechanistic Insights From Human

Neuronal Models. Biological psychiatry, 83(9), 751–760. https://doi.org/10.1016/j.biopsych.2018.01.007

- Staerk, J. et al, (2010). Reprogramming of human peripheral blood cells to induced pluripotent stem cells. Cell Stem Cell, 7 (1), 20–24.
- Stirling, D. R., Swain-Bowden, M. J., Lucas, A. M., Carpenter, A. E., Cimini, B. A., & Goodman, A. (2021). CellProfiler 4: improvements in speed, utility and usability. BMC bioinformatics, 22, 1-11. https://doi.org/10.1186/s12859-021-04344-9
- Stirling, David R., Madison J. Swain-Bowden, Alice M. Lucas, Anne E. Carpenter, Beth A. Cimini, and Allen Goodman. 2021. "CellProfiler 4: Improvements in Speed, Utility and Usability." BMC Bioinformatics 22 (1): 433.
- Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., De Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavare, S., Deloukas, P., Hurles, M. E. & Dermitzakis, E. T. 2007a. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science,* 315**,** 848-53.
- Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavare, S., Deloukas, P. & Dermitzakis, E. T. 2007b. Population genomics of human gene expression. *Nat Genet,* 39**,** 1217-24.
- Streeter, Ian, Peter W. Harrison, Adam Faulconbridge, The HipSci Consortium, Paul Flicek, Helen Parkinson, and Laura Clarke. 2017. "The Human-Induced Pluripotent Stem Cell Initiative-Data Resources for Cellular Genetics." Nucleic Acids Research 45 (D1): D691–97.
- Suarez-Kurtz, G., Botton, M.R., (2013). Pharmacogenomics of warfarin in populations of African descent. Brit. J. Clin. Pharmaco., 75, 334–346.
- Sullivan PF, Agrawal A, Bulik CM, Andreassen OA, Børglum AD, Breen G, Cichon S, Edenberg HJ, Faraone SV, Gelernter J, Mathews CA, Nievergelt CM, Smoller JW, O'Donovan MC; Psychiatric Genomics Consortium. Psychiatric Genomics: An Update and an Agenda. Am J Psychiatry. 2018 Jan 1;175(1):15-27. doi: 10.1176/appi.ajp.2017.17030283. Epub 2017 Oct 3. PMID: 28969442; PMCID: PMC5756100.
- Sumner, C. J., Kolb, S. J., Harmison, G. G., Jeffries, N. O., Schadt, K., Finkel, R. S., Dreyfuss, G. & Fischbeck, K. H. 2006. SMN mRNA and protein levels in peripheral blood: biomarkers for SMA clinical trials. *Neurology,* 66**,** 1067-73.
- Surmeier, D.J., Obeso, J.A., Halliday, G.M., (2017). Selective neuronal vulnerability in Parkinson disease. Nature Rev. Neurosci., 18, 101–113.
- Szabo et al, (2020). A human iPSC-astroglia neurodevelopmental model reveals divergent transcriptomic patterns in schizophrenia. Biorxiv,. https:// doi.org/10.1101/2020.11.07.372839.

- Takahashi, K. et al, (2007). Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell, 131, 861–872.
- Takahashi, N., Sakurai, T., Davis, K. L., & Buxbaum, J. D. (2011). Linking oligodendrocyte and myelin dysfunction to neurocircuitry abnormalities in schizophrenia. Progress in neurobiology, 93(1), 13–24. https://doi.org/10.1016/j.pneurobio.2010.09.004
- Takenawa, T., and H. Miki. 2001. "WASP and WAVE Family Proteins: Key Molecules for Rapid Rearrangement of Cortical Actin Filaments and Cell Movement." Journal of Cell Science 114 (Pt 10): 1801–9.
- Tanaka, S. et al, (2013). DPP6 as a candidate gene for neuroleptic-induced tardive dyskinesia. Pharmacogenom. J., 13 (1), 27–34.
- Tang, X. et al, (2016). KCC2 rescues functional deficits in human neurons derived from patients with Rett syndrome. Proc. Natl. Acad. Sci., 113, 751–756.
- Tchieu, J. et al, (2010). Female human iPSCs retain an inactive X chromosome. Cell Stem Cell, 7 (3), 329–342.
- Tcw, J. et al. An Efficient Platform for Astrocyte Differentiation from Human Induced Pluripotent Stem Cells. Stem Cell Reports 9, 600-614 (2017). https://doi.org:10.1016/j.stemcr.2017.06.018
- Tegtmeyer, M., & Nehme, R. (2022). Leveraging the Genetic Diversity of Human Stem Cells in Therapeutic Approaches. Journal of molecular biology, 434(3), 167221. https://doi.org/10.1016/j.jmb.2021.167221
- Tegtmeyer, Matthew, and Ralda Nehme. 2022. "Leveraging the Genetic Diversity of Human Stem Cells in Therapeutic Approaches." Journal of Molecular Biology 434 (3): 167221.
- Thomson, J. A. et al. Embryonic stem cell lines derived from human blastocysts. Science 282, 1145-1147 (1998).
- Thunander Sundbom, L. et al, (2017). Are men under-treated and women over-treated with antidepressants? Findings from a cross-sectional survey in Sweden. BJPsych. Bull., 41 (3), 145–150.
- Tian, Run-Hui, Yang Bai, Jing-Yang Li, and Kai-Min Guo. 2019. "Reducing PRLR Expression and JAK2 Activity Results in an Increase in BDNF Expression and Inhibits the Apoptosis of CA3 Hippocampal Neurons in a Chronic Mild Stress Model of Depression." Brain Research 1725 (December): 146472.
- Toma, C., Pierce, K. D., Shaw, A. D., Heath, A., Mitchell, P. B., Schofield, P. R., & Fullerton, J. M. (2018). Comprehensive cross-disorder analyses of CNTNAP2 suggest it is unlikely to be a primary risk gene for psychiatric disorders. PLoS genetics, 14(12), e1007535. https://doi.org/10.1371/journal.pgen.1007535
- Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C. Y., Dennison, C. A., Hall, L. S., Lam, M., Watanabe, K., Frei, O., Ge, T., Harwood, J. C., Koopmans, F., Magnusson, S., Richards, A. L., Sidorenko, J., Wu, Y.,

… Schizophrenia Working Group of the Psychiatric Genomics Consortium (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature, 604(7906), 502–508. https://doi.org/10.1038/s41586-022-04434-5

- Tsiouda, T. et al, (2020). Sex Differences and Adverse Effects between Chemotherapy and Immunotherapy for Non-Small Cell Lung Cancer. J. Cancer, 11, 3407–3415.
- Tsurusaki Y, Okamoto N, Ohashi H, Kosho T, Imai Y, Hibi-Ko Y, Kaname T, Naritomi K, Kawame H, Wakui K, Fukushima Y, Homma T, Kato M, Hiraki Y, Yamagata T, Yano S, Mizuno S, Sakazume S, Ishii T, Nagai T, Shiina M, Ogata K, Ohta T, Niikawa N, Miyatake S, Okada I, Mizuguchi T, Doi H, Saitsu H, Miyake N, Matsumoto N. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. Nat Genet. 2012 Mar 18;44(4):376-8. doi: 10.1038/ng.2219. PMID: 22426308.
- Turovsky, E. A. et al. Mechanosensory Signaling in Astrocytes. J Neurosci 40, 9364-9371 (2020). https://doi.org:10.1523/JNEUROSCI.1249-20.2020
- UCT Stem cell research initiative. http://www.cellbiology. uct.ac.za/stem-cell-initiative.
- Uher, R. et al, (2010). Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. Am. J. Psychiatry, 167 (5), 555–564.
- Uranova, N. A., Vostrikov, V. M., Vikhreva, O. V., Zimina, I. S., Kolomeets, N. S., & Orlovskaya, D. D. (2007). The role of oligodendrocyte pathology in schizophrenia. The international journal of neuropsychopharmacology, 10(4), 537–545. https://doi.org/10.1017/S1461145707007626
- Valencia AM, Collings CK, Dao HT, St Pierre R, Cheng YC, Huang J, Sun ZY, Seo HS, Mashtalir N, Comstock DE, Bolonduro O, Vangos NE, Yeoh ZC, Dornon MK, Hermawan C, Barrett L, Dhe-Paganon S, Woolf CJ, Muir TW, Kadoch C. Recurrent SMARCB1 Mutations Reveal a Nucleosome Acidic Patch Interaction Site That Potentiates mSWI/SNF Complex Chromatin Remodeling. Cell. 2019 Nov 27;179(6):1342-1356.e23. doi: 10.1016/j.cell.2019.10.044. Epub 2019 Nov 20. PMID: 31759698; PMCID: PMC7175411.
- van Erp, T. G., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., Agartz, I., Westlye, L. T., Haukvik, U. K., Dale, A. M., Melle, I., Hartberg, C. B., Gruber, O., Kraemer, B., Zilles, D., Donohoe, G., Kelly, S., McDonald, C., Morris, D. W., Cannon, D. M., … Turner, J. A. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. Molecular psychiatry, 21(4), 547–553. https://doi.org/10.1038/mp.2015.63
- Velasco, S. et al, (2019). Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. Nature, 570 (7762), 523–527.
- Velasco, S., Paulsen, B., Arlotta, P., (2021). 3D Brain Organoids: Studying Brain Development and Disease Outside the Embryo. Annu. Rev. Neurosci.,. https://doi. org/10.1146/annurev-neuro-070918-050154.

- Veugelers, M., B. De Cat, H. Ceulemans, A. M. Bruystens, C. Coomans, J. Dürr, J. Vermeesch, P. Marynen, and G. David. 1999. "Glypican-6, a New Member of the Glypican Family of Cell Surface Heparan Sulfate Proteoglycans." The Journal of Biological Chemistry 274 (38): 26968–77.
- Viengchareun, Say, Nathalie Servel, Bruno Fève, Michael Freemark, Marc Lombès, and Nadine Binart. 2008. "Prolactin Receptor Signaling Is Essential for Perinatal Brown Adipocyte Function: A Role for Insulin-like Growth Factor-2." PloS One 3 (2): e1535.
- Vigilante, Alessandra, Anna Laddach, Nathalie Moens, Ruta Meleckyte, Andreas Leha, Arsham Ghahramani, Oliver J. Culley, et al. 2019. "Identifying Extrinsic versus Intrinsic Drivers of Variation in Cell Behavior in Human iPSC Lines from Healthy Donors." Cell Reports 26 (8): 2078–87.e3.
- Vijzelaar, R., Snetselaar, R., Clausen, M., Mason, A. G., Rinsma, M., Zegers, M., Molleman, N., Boschloo, R., Yilmaz, R., Kuilboer, R., Lens, S., Sulchan, S. & Schouten, J. 2019. The frequency of SMN gene variants lacking exon 7 and 8 is highly population dependent. *PLoS One,* 14**,** e0220211.
- Vikhreva, O. V., Rakhmanova, V. I., Orlovskaya, D. D., & Uranova, N. A. (2016). Ultrastructural alterations of oligodendrocytes in prefrontal white matter in schizophrenia: A post-mortem morphometric study. Schizophrenia research, 177(1-3), 28–36. https://doi.org/10.1016/j.schres.2016.04.023
- Villa, C., Combi, R., Conconi, D., Lavitrano, M., (2021). Patient-Derived Induced Pluripotent Stem Cells (iPSCs) and Cerebral Organoids for Drug Screening and Development in Autism Spectrum Disorder: Opportunities and Challenges. Pharm, 13, 280.
- Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, Diekhans M, Sulovari A, Munson KM, Lewis AP, Hoekzema K, Porubsky D, Li R, Nurk S, Koren S, Miga KH, Phillippy AM, Timp W, Ventura M, Eichler EE. Segmental duplications and their variation in a complete human genome. Science. 2022 Apr;376(6588):eabj6965. doi: 10.1126/science.abj6965. Epub 2022 Apr 1. PMID: 35357917; PMCID: PMC8979283.
- Voulgaris, D., Nikolakopoulou, P. & Herland, A. Generation of Human iPSC-Derived Astrocytes with a mature star-shaped phenotype for CNS modeling. Stem Cell Rev Rep (2022). https://doi.org:10.1007/s12015-022-10376-2
- W. Skiller et al., Ketamine-Induced Toxicity in Neurons Differentiated from Neural Stem Cells, Mol. Neurobiol. (n. d.). https://doi.org/10.1007/s12035-015-9248-5.
- Wadhawan, S., Runz, H., Buchard, J., (2001). The genome as pharmacopeia: association of genetic dose with phenotypic response. Biochem. Pharmacol.,. https://doi.org/10.1016/j.bcp.2015.02.005.
- Wang X, Lee RS, Alver BH, Haswell JR, Wang S, Mieczkowski J, Drier Y, Gillespie SM, Archer TC, Wu JN, Tzvetkov EP, Troisi EC, Pomeroy SL, Biegel JA, Tolstorukov MY, Bernstein BE, Park PJ, Roberts CW. SMARCB1-mediated SWI/SNF complex function is

essential for enhancer regulation. Nat Genet. 2017 Feb;49(2):289-295. doi: 10.1038/ng.3746. Epub 2016 Dec 12. PMID: 27941797; PMCID: PMC5285474.

- Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. 2014. Genetic screens in human cells using the CRISPR-Cas9 system. *Science,* 343**,** 80-4.
- Warren, Curtis R., John F. O'Sullivan, Max Friesen, Caroline E. Becker, Xiaoling Zhang, Poching Liu, Yoshiyuki Wakabayashi, et al. 2017. "Induced Pluripotent Stem Cell Differentiation Enables Functional Validation of GWAS Variants in Metabolic Disease." Cell Stem Cell 20 (4): 547–57.e7.
- Weick, J. P. et al. Deficits in human trisomy 21 iPSCs and neurons. Proc Natl Acad Sci U S A 110, 9962-9967 (2013). https://doi.org:10.1073/pnas.1216575110
- Weiner DJ, Ling E, Erdin S, Tai DJC, Yadav R, Grove J, Fu JM, Nadig A, Carey CE, Baya N, Bybjerg-Grauholm J; iPSYCH Consortium; ASD Working Group of the Psychiatric Genomics Consortium; ADHD Working Group of the Psychiatric Genomics Consortium; Berretta S, Macosko EZ, Sebat J, O'Connor LJ, Hougaard DM, Børglum AD, Talkowski ME, McCarroll SA, Robinson EB. Statistical and functional convergence of common and rare genetic influences on autism at chromosome 16p. Nat Genet. 2022 Nov;54(11):1630-1639. doi: 10.1038/s41588-022-01203-y. Epub 2022 Oct 24. PMID: 36280734; PMCID: PMC9649437.
- Weiner DJ, Nadig A, Jagadeesh KA, Dey KK, Neale BM, Robinson EB, Karczewski KJ, O'Connor LJ. Polygenic architecture of rare coding variation across 394,783 exomes. Nature. 2023 Feb;614(7948):492-499. doi: 10.1038/s41586-022-05684-z. Epub 2023 Feb 8. PMID: 36755099.
- Welch, J. D. et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell 177, 1873-1887 e1817 (2019). https://doi.org:10.1016/j.cell.2019.05.006
- Wells, M. et al, (2018). Genome-wide screens in accelerated human stem cell-derived neural progenitor cells identify Zika virus host factors and drivers of proliferation. Biorxiv,. https://doi.org/10.1101/476440.
- Wells, M. F. et al. Natural variation in gene expression and viral susceptibility revealed by neural progenitor cell villages. Cell Stem Cell (2023). https://doi.org:10.1016/j.stem.2023.01.010
- Wells, Michael F., James Nemesh, Sulagna Ghosh, Jana M. Mitchell, Curtis J. Mello, Daniel Meyer, Kavya Raghunathan, et al. 2021. "Natural Variation in Gene Expression and Zika Virus Susceptibility Revealed by Villages of Neural Progenitor Cells." bioRxiv. https://doi.org/10.1101/2021.11.08.467815.
- Wilson HL, Crolla JA, Walker D, Artifoni L, Dallapiccola B, Takano T, Vasudevan P, Huang S, Maloney V, Yobb T, Quarrell O, McDermid HE. Interstitial 22q13 deletions: genes other than SHANK3 have major effects on cognitive and language development. Eur J

Hum Genet. 2008 Nov;16(11):1301-10. doi: 10.1038/ejhg.2008.107. Epub 2008 Jun 4. PMID: 18523453.

- Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D, Lee C, Eskin E, Voineagu I, Ernst J, Geschwind DH. Chromosome conformation elucidates regulatory relationships in developing human brain. Nature. 2016 Oct 27;538(7626):523-527. doi: 10.1038/nature19847. Epub 2016 Oct 19. PMID: 27760116; PMCID: PMC5358922.
- Won H, Huguet G, Jacquemont S. Rare and common autism risk variants converge across 16p. Nat Genet. 2022 Nov;54(11):1587-1588. doi: 10.1038/s41588-022-01219-4. PMID: 36303073.
- Worsdorfer, P. et al, (2019). Generation of complex human organoid models including vascular networks by incorporation of mesodermal progenitor cells. Sci. Rep., 9 (1), 15663.
- Wouters, O.J. et al, (2020). Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009–2018. JAMA, 323 (9), 844–853.
- Xie, H. G., Kim, R. B., Wood, A. J., & Stein, C. M. (2001). Molecular basis of ethnic differences in drug disposition and response. Annual review of pharmacology and toxicology, 41, 815–850. https://doi.org/10.1146/annurev.pharmtox.41.1.815
- Xiong, Y. et al. A Comparison of mRNA Sequencing with Random Primed and 3'-Directed Libraries. Sci Rep 7, 14626 (2017). https://doi.org:10.1038/s41598-017-14892-x
- Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011 Oct 16;43(11):1059-65. doi: 10.1038/ng.947. PMID: 22001755.
- Yagi H, Furutani Y, Hamada H, Sasaki T, Asakawa S, Minoshima S, Ichida F, Joo K, Kimura M, Imamura S, Kamatani N, Momma K, Takao A, Nakazawa M, Shimizu N, Matsuoka R. Role of TBX1 in human del22q11.2 syndrome. Lancet. 2003 Oct 25;362(9393):1366-73. doi: 10.1016/s0140-6736(03)14632-6. PMID: 14585638.
- Yamanaka, S., (2020). Pluripotent Stem Cell-Based Cell Therapy—Promise and Challenges. Cell Stem Cell, 27, 523–531.
- Yamasaki, A.E., Panopoulos, A.D., Belmonte, J.C.I., (2017). Understanding the genetics behind complex human disease with large-scale iPSC collections. Genome Biol., 18, 135.
- Yamasaki, Amanda E., Athanasia D. Panopoulos, and Juan Carlos Izpisua Belmonte. 2017. "Understanding the Genetics behind Complex Human Disease with Large-Scale iPSC Collections." Genome Biology 18 (1): 135.
- Yang, N. et al, (2017). Generation of pure GABAergic neurons by transcription factor programming. Nature Methods, 14 (6), 621–628.
- Yao, L. et al, (2021). Genetic Imaging of Neuroinflammation in Parkinson's Disease: Recent Advancements. Front. Cell Dev. Biol., 9, 655819

- Yazawa, M. et al, (2011). Using induced pluripotent stem cells to investigate cardiac phenotypes in Timothy syndrome. Nature, 471 (7337), 230–234.
- Yoon, S. J., Lyoo, I. K., Haws, C., Kim, T. S., Cohen, B. M., & Renshaw, P. F. (2009). Decreased glutamate/glutamine levels may mediate cytidine's efficacy in treating bipolar depression: a longitudinal proton magnetic resonance spectroscopy study. Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology, 34(7), 1810–1818. https://doi.org/10.1038/npp.2009.2
- Yu, C., Mannan, A. M., Yvone, G. M., Ross, K. N., Zhang, Y. L., Marton, M. A., Taylor, B. R., Crenshaw, A., Gould, J. Z., Tamayo, P., Weir, B. A., Tsherniak, A., Wong, B., Garraway, L. A., Shamji, A. F., Palmer, M. A., Foley, M. A., Winckler, W., Schreiber, S. L., Kung, A. L. & Golub, T. R. 2016. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat Biotechnol*, 34**,** 419-23.
- Yu,J. et al,(2009).Human Induced Pluripotent Stem Cells Free of Vector and Transgene Sequences. Science, 324, 797–801.
- Yurov, Y.B. et al, (2014). X chromosome aneuploidy in the Alzheimer's disease brain. Mol. Cytogenet., 7 (1), 20.
- Zhang, W. et al, (2012). Evolution of iPSC disease models. Protein Cell, 3 (1), 1–4.
- Zhang, Y. et al, (2013). Rapid single-step induction of functional neurons from human pluripotent stem cells. Neuron, 78 (5), 785–798.
- Zhang, Y. et al, (2013). Rapid Single-Step Induction of Functional Neurons from Human Pluripotent Stem Cells. Neuron, 78, 785–798.
- Zhang, Y. et al. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. Neuron 89, 37-53 (2016). https://doi.org:10.1016/j.neuron.2015.11.013
- Zhang, Y. et al. Rapid single-step induction of functional neurons from human pluripotent stem cells. Neuron 78, 785-798 (2013). https://doi.org:10.1016/j.neuron.2013.05.029
- Zhang, Y., Pak, C., Han, Y., Ahlenius, H., Zhang, Z., Chanda, S., Marro, S., Patzke, C., Acuna, C., Covy, J., Xu, W., Yang, N., Danko, T., Chen, L., Wernig, M. & Sudhof, T. C. 2013. Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron*, 78**,** 785-98.
- Zhao, W.N. et al, (2012). A high-throughput screen for Wnt/beta-catenin signaling pathway modulators in human iPSC-derived neural progenitors. J. Biomol. Screen., 17 (9), 1252–1263.
- Zhou, B., Zuo, Y. X., & Jiang, R. T. (2019). Astrocyte morphology: Diversity, plasticity, and role in neurological diseases. CNS neuroscience & therapeutics, 25(6), 665–673. https://doi.org/10.1111/cns.13123
- Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 10, 1523 (2019). https://doi.org:10.1038/s41467-019-09234-6

- Zimmermann H. (1994). Signalling via ATP in the nervous system. Trends in neurosciences, 17(10), 420–426. https://doi.org/10.1016/0166-2236(94)90016-7
- Zucker, I., Prendergast, B.J., (2020). Sex differences in pharmacokinetics predict adverse drug reactions in women. Biol. Sex Differ., 11 (1), 32.

Appendix

Appendix 1 - Data resources generated during this thesis

Data type	Cell type	# of Samples	# of Cells	Publicly available
Cell Painting	iPSCs	365	>6M	Yes
Cell Painting	NPCs	84	>2M	Yes
Cell Painting	Neurons	84	>2M	Yes
Cell Painting	Astrocytes	50	>1M	Yes
scRNA-seq	iPSCs	45	~500,000	Yes (soon)
scRNA-seq	NPCs	45	~500,000	Yes (soon)
scRNA-seq	Neurons	45	~500,000	Yes (soon)
scRNA-seq	Astrocytes	45	~500,000	Yes (soon)