STEM CELL REPORTS
Report
ISSCR
OPEN ACCESS

Genetics of growth rate in induced pluripotent stem cells

Brian N. Lee,[1,15] Henry J. Taylor,[1,2,3,15] Filippo Cipriani,[4,15] Narisu Narisu,[1] Catherine C. Robertson,[1,5] Amy J. Swift,[1] Neelam Sinha,[1] Tingfen Yan,[1] Lori L. Bonnycastle,[1] Nathan Dale,[4] Annie Butt,[4] Hemant Parsaud,[4] Stefan Semrau,[4] NYSCF Global Stem Cell Array Team, GENESiPS Consortium, iPSCORE Consortium, Joshua W. Knowles,[6,7,8,9] Ivan Carcamo-Orive,[10,11] Agnieszka D'Antonio-Chronowska,[12] Kelly A. Frazer,[13,14] Leslie G. Biesecker,[1] Scott Noggle,[4] Michael R. Erdos,[1] Daniel Paull,[4,16,17] Francis S. Collins,[16,17] and D. Leland Taylor[1,16,17,18,*]

[1]Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
[2]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK
[3]Heart and Lung Research Institute, University of Cambridge, Cambridge CB1 8RN, UK
[4]New York Stem Cell Foundation Research Institute, New York, NY 10019, USA
[5]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[6]Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA 94304, USA
[7]Stanford Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA 94304, USA
[8]Stanford Diabetes Research Center, Stanford University School of Medicine, Stanford, CA 94304, USA
[9]Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, CA 94304, USA
[10]IKERBASQUE, Basque Foundation for Science, 48009 Bilbao, Spain
[11]Department of Endocrinology, Metabolism, Nutrition, and Kidney Disease, Biobizkaia Health Research Institute, 48903 Cruces, Spain
[12]Center for Epigenomics, University of California, San Diego, School of Medicine, La Jolla, CA 92093, USA
[13]Institute of Genomic Medicine, University of California, San Diego, La Jolla, CA 92093, USA
[14]Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA
[15]These authors contributed equally
[16]These authors contributed equally
[17]Senior author
[18]Lead contact
*Correspondence: leland.taylor@nih.gov
https://doi.org/10.1016/j.stemcr.2026.102928

SUMMARY

Induced pluripotent stem cells (iPSCs) enabled the generation of diverse cell types; however, certain fundamental biological properties, such as the genetic and epigenetic determinants of proliferation, remain poorly characterized. We quantified proliferation across 602 unique donors with a time-lapse imaging-based growth area under the curve (gAUC) phenotype and correlated gAUC with cell line gene expression and genotype. We identified 3,091 differentially expressed genes and found that rare deleterious variants in *WDR54*, *TMEM250*, and *C2orf81* were associated with reduced iPSC growth. Notably, *WDR54* was differentially expressed with respect to gAUC. Although no common variants were associated, common genetic variation explained 71%–75% of the variance. These results indicate a complex genetic architecture of iPSC growth rates, where rare, large-effect variants in important growth regulators are layered onto a highly polygenic background. These findings can impact the design of pooled iPSC-based studies and disease models, which may be confounded by intrinsic growth differences.

INTRODUCTION

The development of human induced pluripotent stem cells (iPSCs) has transformed many aspects of biomedical research. The first iPSCs were described 18 years ago, derived from human fibroblasts reprogrammed into a pluripotent state (Park et al., 2008). Since then, numerous protocols have been established to differentiate iPSCs into a wide variety of cell types, including those difficult to obtain as primary human tissues (Veres et al., 2019; Walczak et al., 2016). Given their differentiation potential and ease of *in vitro* manipulation, iPSCs enable diverse basic and translational applications including drug and cellular therapy testing (Chehelgerdi et al., 2023) and autotransplantation (Melton 2021).

Many basic biological questions that are difficult to address *in vivo* can be assessed with iPSCs. For example, genome-wide association studies (GWASs) often implicate disease-associated loci in non-coding regions, complicating the identification of effector gene(s) and cellular mechanisms (Ward and Kellis 2012). Gene expression studies across various tissues have not resolved all such signals, suggesting that some arise from hard-to-capture cell types or states (Umans et al., 2021). Large-scale genetic studies using iPSCs offer one way to investigate the effects of natural human variation and disease-associated variants in otherwise inaccessible cell types and contexts (Neavin et al., 2023). As iPSC repositories grow, such studies will become routine, requiring a detailed understanding of iPSC biology to optimize study designs and mitigate confounding.

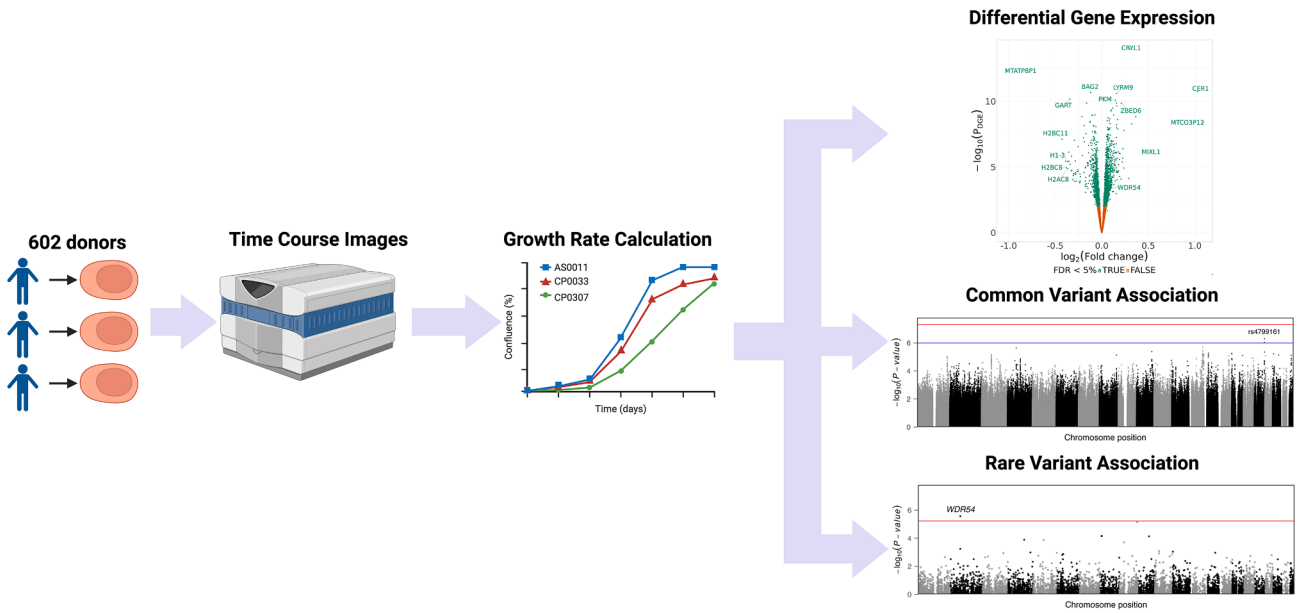

Figure 1. Graphical overview of this study

Human iPSCs from multiple cohorts were cultured in 96-well plates until confluence. Daily cell density measurements were obtained via imaging cytometry. Upon reaching confluence, cells were harvested for bulk RNA sequencing, genotyping, and genome sequencing, enabling differential expression analysis and both common- and rare-variant association studies.

In this study (Figure 1), we considered one such variable—growth rates of undifferentiated iPSCs. Growth rates can be a critical factor in pooled study designs where different proliferation rates may cause certain lines to dominate a pool (Mitchell et al., 2020). Across 602 lines, we measured rates of cell growth from daily images, paired with RNA sequencing on a subset of 333 lines. Using these data, we performed an association analysis of iPSC growth rate with gene expression, common genetic variation, and rare genetic variation. Our results support a polygenic model of cell proliferation, whereby common variants contribute small effects and rare variants may exert larger influences on cellular growth dynamics.

RESULTS

Cohort characterization and overview of genetic data

We obtained iPSC lines from 602 donors (ESM Table 1). These lines included both male (278 lines; 46.2%) and female (324 lines; 53.8%) donors. Lines were grown in two types of media, either mTeSR Plus (n = 347) or Freedom (n = 255), which we controlled for in subsequent analyses as a covariate (STAR Methods).

We genotyped all 602 donors and imputed variants to GRCh38 (STAR Methods). Focusing on common single nucleotide polymorphisms (SNPs; minor allele frequency [MAF] > 5%), we retained 5,985,990 variants after quality control. We grouped participants according to their genetic similarity to 1,000 Genomes genetic ancestry groups, including participants from European (1KG-EUR-like, n = 446), African (1KG-AFR-like, n = 39), American (1KG-AMR-like, n = 44), East Asian (1KG-EAS-like, n = 51), and South Asian (1KG-SAS-like, n = 22) ancestry groups (STAR Methods; Figure S1). While genotype imputation is accurate for common SNPs, rare genetic variants are often missed. To identify rare genetic variants, we used genome sequencing (GS) or exome sequencing (ES) data to call exonic variants in 516 donors (STAR Methods). For the majority of the cell lines, DNA originated from the iPSC lines; for a subset (n = 195), we used DNA obtained directly from donor cells, either blood or fibroblasts. On average, we generated approximately 656.0 million reads per cell line for a depth of 31×. After quality control, we retained 178,745 exonic single nucleotide or short insertion-deletion variants for downstream analyses.

Derivation of growth rate from cell images

To measure cell growth rate, we plated cells from the iPSC lines of all 602 donors (ESM Table 1) on the New York Stem Cell Foundation (NYSCF) robotic culture system, performed at least one passage, seeded cells at a standard density on fresh plates, and cultured cells until wells reached ~90% confluency (STAR Methods; Figure S3A). We used a Celigo imaging instrument to image each well once a day

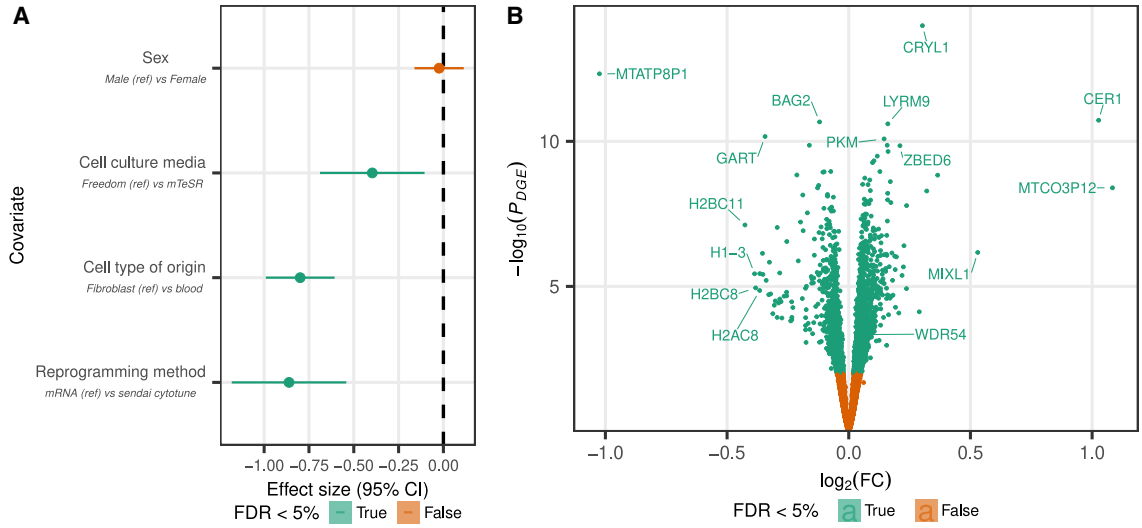

Figure 2. Covariate associations with growth area under the curve (gAUC) and differential gene expression analyses

(A) Effect size (*y* axis) with 95% confidence intervals (CIs; lines) of iPSC line characteristics (*x* axis). Color denotes FDR <5%.

(B) Volcano plots of log2(fold change) (*x* axis) and $-\log_{10}$(P) (*y* axis) from differential gene expression analysis. Color denotes associated genes (FDR <5%).

from which we measured well coverage (expressed as a percent). In total, we had 139,246 timestamped well coverage measurements across 26,066 wells spanning all 602 lines (average 43.3 wells per cell line). We excluded wells with incomplete coverage measurements or that were highly dissimilar to replicate wells, observing an average coverage >71% across all lines (STAR Methods; Figure S3B). For each remaining well, we used the series of well coverage measurements to derive a growth area under the curve (gAUC) to represent growth rate for downstream analyses (STAR Methods). After quality control procedures (STAR Methods), we retained on average 27.0 wells with gAUC estimates per line (25.7 standard deviation [SD]) and observed high reproducibility across the replicate wells (intraclass correlation coefficient [ICC] = 0.96, $p = 1.11 \times 10^{-16}$, F test; Figure S3C). We also compared gAUC in 16 twin participants (eight pairs) and found them to be correlated (ICC = 0.65; $p = 0.02$, F test). These observations suggest a genetic component of gAUC, though cell-specific effects or technical artifacts may also contribute (e.g., epigenomic events that take place during the development of pluripotency from a differentiated cell). To explore the effect of known technical or biological factors on gAUC, we tested for gAUC association with a variety of iPSC attributes including genetically imputed donor sex, pluripotency reprogramming method, starting cell type for reprogramming, and cell culture media used (STAR Methods). We found three terms to be associated with gAUC: cell culture media ($p = 0.008$), starting cell type ($p = 1.68 \times 10^{-15}$), and pluripotency reprogramming method ($p = 1.80 \times 10^{-7}$), where the Freedom cell culture media, fibroblast origin cell type, and mRNA reprogramming method were associated with increased iPSC growth, respectively (Figure 2A).

Growth-associated differential gene expression analysis

We sought to characterize the transcriptional profile of iPSC growth rates and performed RNA sequencing on 333 lines (STAR Methods). On average, we generated 235 million reads. After quality control and gene expression quantification, we retained expression data for 15,603 genes (STAR Methods).

We tested for associations of gene expression with gAUC and identified 3,091 differentially expressed genes (DEGs; false discovery rate [FDR] < 5%; Figure 2B; STAR Methods). Many of the genes that we found are well-established cell growth and proliferation genes, confirming the gAUC phenotype: *KDM4C* (Zhu et al., 2024), *FGF8* (Liu et al., 2015), *L1CAM* (Kiefel et al., 2012), *COBL* (Haag et al., 2012), *NCAPG2* (Zhan et al., 2017), *PIK3CA* (Samuels et al., 2005), *RNF43* (Koo et al., 2012), and *KRT17* (Mikami et al., 2017). To gain a better understanding of the biological programs underlying these associations, we tested for enriched gene sets from the "biological process" gene ontology (GO) database (STAR Methods). We found 111 enriched gene sets (FDR <5%), which clustered into eight groups of functionally related terms based on semantic similarity (Figure S2). Notably, these clusters defined key biological processes regulating

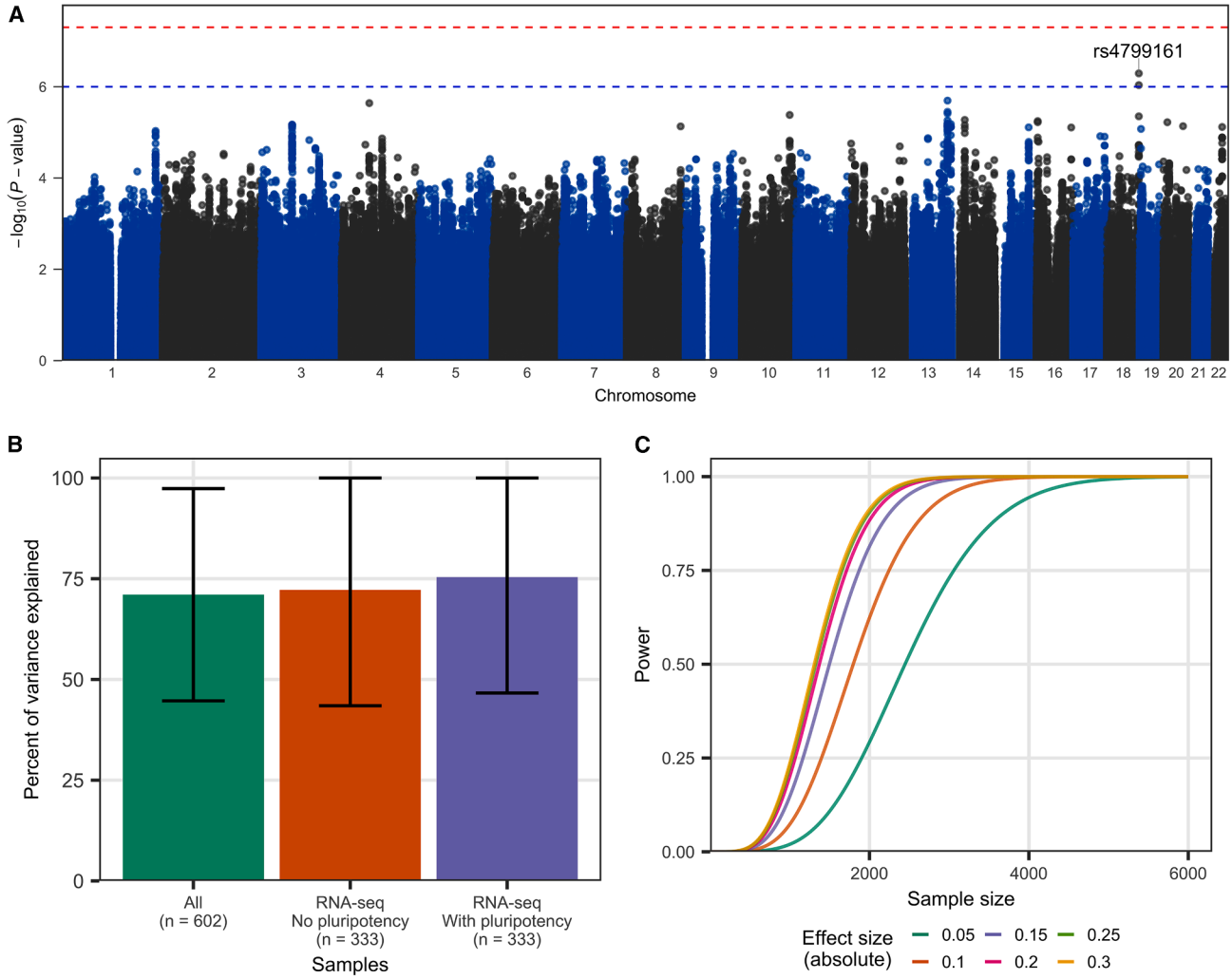

Figure 3. Common-variant association study

(A) Manhattan plot of common variant associations (points) with gAUC. Nominal association threshold ($p < 1 \times 10^{-6}$) is indicated by the blue line.

Genome-wide association threshold ($p < 5 \times 10^{-8}$) is indicated by the red line.

(B) Percent of growth rate variance explained by the genetic relatedness matrix (*y* axis) across different models (*x* axis). The horizontal axis and colors indicate the model—either all samples (green), samples with RNA-seq (orange), or RNA-seq samples with pluripotency covariates (purple). The vertical error bars indicate 95% confidence intervals (CIs).

(C) Power estimates (*y* axis) by sample size (*x* axis). Color indicates absolute effect size modeled.

cellular growth and homeostasis, including glycolysis, chromatin stability, nucleotide metabolism, and stress response.

Contribution of common genetic variants to growth rate

We performed a genome-wide common variant association study (CVAS) to test for common variant associations with gAUC using all 602 iPSC lines, controlling for culture media type, starting cell type, reprogramming method, and the first 10 genetic principal components (PCs) as covariates (STAR Methods). We did not identify any genome-wide associated signals ($p < 5 \times 10^{-8}$). However, we found one locus with suggestive evidence of association ($p < 1 \times 10^{-6}$; Figure 3A), tagged by rs4799161. The rs4799161 variant is located within the *RP11-383C5.4* lncRNA and is approximately 451 kb upstream of *SALL3*, the nearest protein-coding gene. *SALL3* has been found to influence methylation of Wnt-signaling-related genes in iPSCs and, therefore, may play an important role in both cell proliferation and stem cell differentiation pathways (Kuroda et al., 2019).

Though we identified limited common variant associations, a substantial genetic component to gAUC may exist, consisting of highly polygenic, small effect sizes. To better understand the genetics of gAUC, we estimated the SNP-based, narrow-sense heritability modeling the genetic relatedness matrix (h_g^2) along with pluripotency of lines when applicable (STAR Methods). We analyzed the heritability in three ways: (1) across all samples, including media, starting cell type, and reprogramming method as covariates (n_{All} = 602), (2) across lines with RNA-seq using expression of *POU5F1*, *PODXL*, *NANOG*, and *SOX2* to control for the pluripotency (n_{RNA} = 333; only starting cell-type and reprogramming method as covariates since all of the RNA-seq lines were grown in mTeSR media), and (3) across lines with RNA-seq without controlling for pluripotency. Across all models, we found that the genetic component explained approximately 71%–75% of the variance in gAUC, indicating a strong heritable component to iPSC growth rates (Figure 3B). These findings are consistent with the high intraclass correlation coefficient identified within twin cell line pairs.

Finally, given our heritability findings, we hypothesized that the limited CVAS results were a consequence of an insufficient sample size. We performed calculations to determine an optimal sample size for investigating the effects of common variants on iPSC growth rates (STAR Methods). We estimated that identifying common variants associated with growth phenotypes at 80% power would require a minimum sample size of 1,720 individuals to detect effect sizes of 0.30 and up to 3,222 individuals to detect smaller effects of 0.05 (Figure 3C).

Contribution of rare genetic variants in protein coding regions to growth rate

For a highly polygenic phenotype, a CVAS often identifies loci with small effect sizes (O'Connor et al., 2019). However, rare deleterious or *de novo* mutations in coding regions may have large effects on cellular growth, as is evident in many cancers (Kandoth et al., 2013). To explore the role of rare variants (MAF <1%) in gAUC, we performed a rare variant association study (RVAS) across donors with GS or ES by aggregating variants across genes and testing for gene-based associations with gAUC (STAR Methods).

We first considered only variants with a predicted functional consequence of loss of function (pLoF) or putatively damaging missense with a CADD score ≥20. In total, after additional quality control filters (STAR Methods), we tested for associations of 8,480 genes with gAUC. We identified one association with decreased gAUC at *WDR54* ($P_{Bonferroni}$ < 0.05; Figure 4A), driven by five potentially damaging missense variants in *WDR54* (ESM Table 2). We did not detect any pLoF variants in *WDR54*. Because these missense variants occurred in DNA derived from both donor fibroblasts (n = 6) and iPSC lines themselves (n = 11), we cannot determine if they are somatic or germline. Given the difficulty in predicting the pathogenicity of missense variants, we conducted a sensitivity analysis by repeating the RVAS using a stricter CADD score threshold (≥25), resulting in 3,441 tests. Although the number of missense variants considered in *WDR54* decreased to two, the association with decreased gAUC remained ($P_{Bonferroni}$ < 0.05), supporting the robustness of this finding. We did not identify any additional associations in this restricted analysis.

While our initial analysis relied on conventional annotations such as CADD and variant effect predictor to assess variant deleteriousness, numerous newer methods now predict functional effects across coding and noncoding regions. To incorporate these newer annotation methods, we repeated the RVAS using DeepRVAT (Clarke et al., 2024), which uses deep set neural networks to aggregate predicted functional impacts across variant annotations (STAR Methods). With this approach, we replicated the *WDR54* association and identified two new associations with *TMEM250* and *C2orf81* ($P_{Bonferroni}$ < 0.05, corrected for 17,958 genes; Figure S4).

Finally, to further explore our rare variant associations, we queried *WDR54*, *TMEM250*, and *C2orf81* in the gAUC differential gene expression results. Although all three genes were tested, we only found an association with *WDR54* expression and gAUC (FDR <5%; $p = 4.58 \times 10^{-4}$; Figures 2B and 4C). Combined with the rare variant findings, these findings are consistent with previous reports of increased expression of *WDR54* and cell/tumor proliferation (Wei et al., 2023; Jeong et al., 2024).

DISCUSSION

In this study, we measured iPSC growth rates across 602 unique donors using an automated culture and imaging system. From the RNA-seq analysis, we identified thousands of DEGs associated with growth rate. Our rare variant association analysis identified *WDR54*, *TMEM250*, and *C2orf81* as having putatively damaging missense variants associated with reduced growth rate. This finding was also corroborated by our RNA-seq results, where increased expression of *WDR54* was associated with increased growth rate. Our common variant association analysis identified no associations with iPSC growth, suggesting limited statistical power.

We also considered the role of epigenetic factors in iPSC growth. We pursued three approaches to address this question: (1) we considered 16 cell lines from eight pairs of monozygotic twins and found their growth rates to be highly correlated (ICC = 0.65), though not as high as

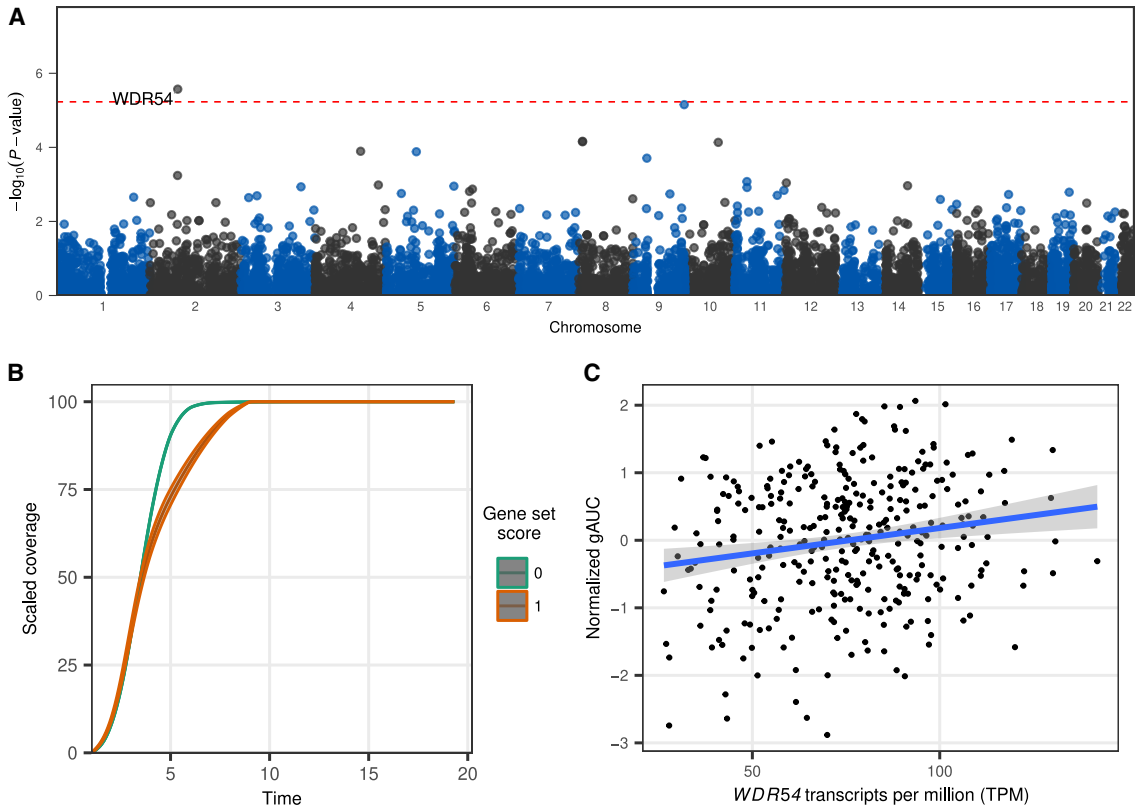

Figure 4. Rare-variant association study

(A) Manhattan plot of rare variant associations (points) with gAUC. The horizontal red line indicates Bonferroni *p* value threshold ($p < 5.90 \times 10^{-6}$).

(B) The mean confluence trajectories stratified by rare-variant burden scores (color) for *WDR54* with confluence percentage (*y* axis) over time (*x* axis).

(C) A scatterplot displaying the relationship between *WDR54* gene expression (*x* axis) and the average gAUC (*y* axis) per line (points). The shaded ribbon indicates the 95% CI.

replicates of the same cell line (ICC = 0.96). (2) We decomposed the variance explained by a genetic relatedness matrix and observed heritability of approximately 71%–75%. (3) We found differences in growth rates of iPSCs depending on the starting cell type (blood or skin), the pluripotency reprogramming method (mRNA-based or Sendai virus), and the media used for cell culture (mTeSR Plus or Freedom). Accordingly, we included these variables as covariates and demonstrated that a substantial portion of growth-rate variance remained attributable to genetics.

The differential gene expression analysis identified numerous genes with expression differences associated with growth rates. Many gAUC-associated genes are known oncogenes including *FGF8* (Liu et al., 2015), *L1CAM* (Kiefel et al., 2012), *KRT17* (Mikami et al., 2017), *NCAPG2* (Zhan et al., 2017), *PIK3CA* (Samuels et al., 2005), and *RNF43* (Koo et al., 2012). The differential expression of genes associated with chromatin remodeling (i.e., *KDM4C*) (Zhu et al., 2024), cytoskeletal reorganization (*COBL*) (Haag et al., 2012), and DNA damage repair (*KDM4C*) (Zhu et al., 2024) further implicates cellular features that modulate cell growth.

Though the functions of both *TMEM250* and *C2orf81* have not yet been extensively studied, the association between *WDR54* and iPSC growth rates comports with prior studies of the function of this gene. *WDR54* is a prominent oncogene within colorectal cancer, bladder cancer, and T cell acute lymphoblastic leukemia and has also been proposed as a therapeutic target for head and neck squamous cell carcinoma (Yuan et al., 2019; Wei et al., 2023; Li et al., 2023). Knockdown of *WDR54* causes reduced cell growth and increased apoptosis (Yuan et al., 2019; Wei et al., 2023). Mechanistic studies show that *WDR54* enhances extracellular-signal-regulated kinase activation—a process that ultimately regulates cell growth, proliferation, and differentiation—by sustaining epidermal growth factor receptor signaling at the plasma membrane (Maeda et al., 2019). These established roles of *WDR54* in cellular

proliferation reinforce the biological context for the gene expression and rare-variant association finding.

Strengths of the study include the large number of lines considered, an automated system to precisely measure cell growth over time, and a broad characterization of iPSC growth contributors through SNP genotyping, DNA sequencing, and bulk RNA sequencing. One limitation is the inability to distinguish germline from somatic mutations in these iPSC lines; future studies may benefit by identifying somatic mutations through comparison of iPSC sequence with donor blood or skin DNA sequence (Ji et al., 2012). A second limitation is the inability to detect common variants associated with growth rate, suggesting future studies may need to be several times larger to have appropriate power. A third limitation is our dependence on computational methods to identify pathogenic coding variants, which advancements in artificial intelligence models will likely improve.

Awareness of the genetic drivers of iPSC growth can provide important insights for future research designs. For instance, in cell village association studies, lines with aggressive growth and enhanced differentiation characteristics can overwhelm a heterogeneous pool of cells, potentially confounding experimental outcomes (Mitchell et al., 2020; Neavin et al., 2023). As large-scale genetic studies of iPSCs become more common, study designs should consider differences in starting cell type, type of media, and pluripotency reprogramming method. These studies may also benefit from genetic screening of individual lines for both somatic and germline rare variants in relevant genes, including *WDR54*, *TMEM250*, and *C2orf81*.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Leland Taylor (leland.taylor@nih.gov).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Raw data (individual-level genotype and RNA-seq) for the FUSION, iPSCORE, GENESiPS, and HipSci cell lines are available in either via database of Genotypes and Phenotypes (dbGaP) accession number phs004457.v1.p1 or European Genome-phenome Archive (EGA) accession number EGAS00001008440.
- NYSCF Global Stem Cell Array (GSCA) data are available upon request from the NYSCF Research Institute.
- Cell growth data for non-GSCA cell lines, summary statistics, and code have been deposited to GitHub at https://github.com/NHGRI/lee_etal_2026-ipsc_growth.

ACKNOWLEDGMENTS

Funding: This research was supported by the New York Stem Cell Foundation (NYSCF), an NYSCF Druckenmiller Fellowship (to D.L.T.); the US National Institutes of Health grants ZIAHG000024 (to F.S.C. and L.G.B.), K99DK13917501 (to D.L.T.), U01HL107388 (to the GENESiPS Consortium), R01DK106236 (to J.W.K.), P30DK116074 (to the GENESiPS Consortium), R01DK116750 (to J.W.K.), R01DK120565 (to J.W.K.), R01DK137889 (to J.W.K.), DP3DK112155 (to K.A.F.), and U01HL107442 (to K.A.F.); the Ikerbasque Research Fellowship sponsored by European Union (EU) grant H2020-MSCA-COFUND-2020-101034228-WOLFRAM2 (to I.C.O.); the MCIU/AEI/FEDER-EU grant PID2023-148986OB-I00 (to I.C.O.); the Gates Cambridge Trust (to H.J.T.); and the NIH Oxford-Cambridge scholars' program (to H.J.T.). The data from the FUSION samples used for the research were obtained from THL Biobank (study number THLBB2025_5). We thank all study participants for their generous participation in Finnish health research. We also acknowledge Wellcome Trust Sanger Institute as the source of the HipSci human induced pluripotent cell lines, which was generated under the Human Induced Pluripotent Stem Cell Initiative funded by a grant from the Wellcome Trust and Medical Research Council, supported by the Wellcome Trust (WT098051), where clinical samples were collected by Prof Andrew Hattersley and colleagues from the University of Exeter supported by the Wellcome Trust (098395), and acknowledges Life Science Technologies Corporation as the provider of CytoTune-iPS.

AUTHOR CONTRIBUTIONS

B.N.L., H.J.T., F.C., M.R.E., D.P., F.S.C., and D.L.T. designed the research. F.S.C., A.J.S., L.L.B., N.D., A.B., H.P., S.N., J.W.K., I.C.-O., A.D.-C., K.A.F., N.S., M.R.E., F.S.C., and D.L.T. performed research. B.N.L., H.J.T., N.N., C.C.R., N.S., T.Y., S.S., D.P., and D.L.T. analyzed data. L.G.B., F.S.C., and D.L.T. supervised the study. B.N.L., H.J.T., F.C., N.N., A.J.S., L.G.B., D.P., F.S.C., and D.L.T. wrote the paper.

DECLARATION OF INTERESTS

L.G.B. is a member of the Illumina Medical Ethics advisory board, receives research support from Merck, and royalties from Wolters-Kluwer.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Summary of induced pluripotent stem cells (iPSCs)
  - iPSC thawing and processing
  - iPSC culture, expansion, and passaging
  - iPSC karyotyping
  - Ethics statement
- METHOD DETAILS
  - RNA extraction and sequencing
  - DNA extraction, genotyping, and sequencing

- Cellular imaging of individual wells

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Growth rate derivation from cellular images and replicate concordance
  - SNP genotype quality control procedures and imputation
  - Variant calling and quality control procedures in genomic sequencing
  - RNA-seq data processing and quality control procedures
  - Determination of covariates for downstream analyses
  - Differential gene expression and gene ontology enrichment analyses
  - Genetic relatedness matrix calculation
  - Variance decomposition analyses
  - Genetic principal component and ancestry assignment
  - Common variant association testing
  - Power analyses for common variant associations
  - Rare variant annotation, aggregation, and association testing

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.stemcr.2026.102928.

Received: July 18, 2025
Revised: April 21, 2026
Accepted: April 22, 2026
Published: May 21, 2026

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nat. Protoc. *5*, 1564–1573.

Arthur, T.D., Nguyen, J.P., Henson, B.A., D'Antonio-Chronowska, A., Jaureguy, J., Silva, N., Arias, A.D., Benaglio, P., Berggren, W.T., D'Antonio, M., et al. (2025). Multiomic QTL mapping reveals phenotypic complexity of GWAS loci and prioritizes putative causal variants. Cell Genom. *5*, 100775.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B *57*, 289–300.

Carcamo-Orive, I., Hoffman, G.E., Cundiff, P., Beckmann, N.D., D'Souza, S.L., Knowles, J.W., Patel, A., Hendry, C., Papatsenko, D., Abbasi, F., et al. (2017). Analysis of Transcriptional Variability in a Large Human iPSC Library Reveals Genetic and Non-genetic Determinants of Heterogeneity. Cell Stem Cell *20*, 518–532.e9.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.

Chehelgerdi, M., Behdarvand Dehkordi, F., Chehelgerdi, M., Kabiri, H., Salehian-Dehkordi, H., Abdolvand, M., Salmanizadeh, S., Rashidi, M., Niazmand, A., Ahmadi, S., et al. (2023). Exploring the promising potential of induced pluripotent stem cells in cancer research and therapy. Mol. Cancer *22*, 189.

Chen, H., Huffman, J.E., Brody, J.A., Wang, C., Lee, S., Li, Z., Gogarten, S.M., Sofer, T., Bielak, L.F., Bis, J.C., et al. (2019). Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. Am. J. Hum. Genet. *104*, 260–274.

Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Celedón, J.C., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am. J. Hum. Genet. *98*, 653–666.

Clarke, B., Holtkamp, E., Öztürk, H., Mück, M., Wahlberg, M., Meyer, K., Munzlinger, F., Brechtmann, F., Hölzlwimmer, F.R., Lindner, J., et al. (2024). Integration of variant annotations using deep set networks boosts rare variant association testing. Nat. Genet. *56*, 2271–2280.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.

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

Dong, X., Li, X., Chang, T.-W., Scherzer, C.R., Weiss, S.T., and Qiu, W. (2021). powerEQTL: an R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis. Bioinformatics *37*, 4269–4271.

Dyer, S.C., Austine-Orimoloye, O., Azov, A.G., Barba, M., Barnes, I., Barrera-Enriquez, V.P., Becker, A., Bennett, R., Beracochea, M., Berry, A., et al. (2025). Ensembl 2025. Nucleic Acids Res. *53*, D948–D957.

Grenko, C.M., Taylor, H.J., Bonnycastle, L.L., Xue, D., Lee, B.N., Weiss, Z., Yan, T., Swift, A.J., Mansell, E.C., Lee, A., et al. (2024). Single-cell transcriptomic profiling of human pancreatic islets reveals genes responsive to glucose exposure over 24 h. Diabetologia *67*, 2246–2259.

Haag, N., Schwintzer, L., Ahuja, R., Koch, N., Grimm, J., Heuer, H., Qualmann, B., and Kessels, M.M. (2012). The actin nucleator Cobl is crucial for Purkinje cell development and works in close conjunction with the F-actin binding protein Abp1. J. Neurosci. *32*, 17842–17856.

Jeong, E.-J., Kim, E., and Kim, Y.S. (2024). Identification of novel therapeutic targets for head and neck squamous cell carcinoma through bioinformatics analysis. Sci. Rep. *14*, 32102.

Ji, J., Ng, S.H., Sharma, V., Neculai, D., Hussein, S., Sam, M., Trinh, Q., Church, G.M., McPherson, J.D., Nagy, A., and Batada, N.N. (2012). Elevated coding mutation rate during the reprogramming of human somatic cells into induced pluripotent stem cells. Stem Cell. *30*, 435–440.

Jun, G., Wing, M.K., Abecasis, G.R., and Kang, H.M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. *25*, 918–925.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Kiefel, H., Bondong, S., Hazin, J., Ridinger, J., Schirmer, U., Riedle, S., and Altevogt, P. (2012). L1CAM: a major driver for tumor cell invasion and motility. Cell Adh. Migr. *6*, 374–384.

Koo, B.-K., Spit, M., Jordens, I., Low, T.Y., Stange, D.E., van de Wetering, M., van Es, J.H., Mohammed, S., Heck, A.J.R., Maurice, M.M., and Clevers, H. (2012). Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of Wnt receptors. Nature *488*, 665–669.

Kuroda, T., Yasuda, S., Tachi, S., Matsuyama, S., Kusakawa, S., Tano, K., Miura, T., Matsuyama, A., and Sato, Y. (2019). SALL3 expression balance underlies lineage biases in human induced pluripotent stem cell differentiation. Nat. Commun. *10*, 2175.

Kuznetsova, A., Brockhoff, P.B., and Christensen, R.H.B. (2017). lmertest package: tests in linear mixed effects models. J. Stat. Softw. *82*, 1–26.

Liu, R., Huang, S., Lei, Y., Zhang, T., Wang, K., Liu, B., Nice, E.C., Xiang, R., Xie, K., Li, J., and Huang, C. (2015). FGF8 promotes colorectal cancer growth and metastasis by activating YAP1. Oncotarget *6*, 935–952.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Zhang, D., Fu, Q., Wang, S., Zhang, X., Lin, Z., Wang, Z., Song, J., Su, Z., Xue, V., et al. (2023). WDR54 exerts oncogenic roles in T-cell acute lymphoblastic leukemia. Cancer Sci. *114*, 3318–3329.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. https://doi.org/10.48550/arXiv.1303.3997.

Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448.

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

Lowry, W.E., Richter, L., Yachechko, R., Pyle, A.D., Tchieu, J., Sridharan, R., Clark, A.T., and Plath, K. (2008). Generation of human induced pluripotent stem cells from dermal fibroblasts. Proc. Natl. Acad. Sci. USA *105*, 2883–2888.

Maeda, A., Nishino, T., Matsunaga, R., Yokoyama, A., Suga, H., Yagi, T., and Konishi, H. (2019). Transglutaminase-mediated cross-linking of WDR54 regulates EGF receptor-signaling. Biochim. Biophys. Acta Mol. Cell Res. *1866*, 285–295.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

Melton, D. (2021). The promise of stem cell-derived islet replacement therapy. Diabetologia *64*, 1030–1036.

Mikami, Y., Fujii, S., Nagata, K., Wada, H., Hasegawa, K., Abe, M., Yoshimoto, R.U., Kawano, S., Nakamura, S., and Kiyoshima, T. (2017). GLI-mediated Keratin 17 expression promotes tumor cell growth through the anti-apoptotic function in oral squamous cell carcinomas. J. Cancer Res. Clin. Oncol. *143*, 1381–1393.

Mitchell, J.M., Nemesh, J., Ghosh, S., Handsaker, R.E., Mello, C., Meyer, D., Raghunathan, K., de Rivera, H., Tegtmeyer, M., Hawes, D., et al. (2020). Mapping genetic effects on cellular phenotypes with "cell villages". Preprint at bioRxiv. https://doi.org/10.1101/2020.06.29.174383.

Neavin, D.R., Steinmann, A.M., Farbehi, N., Chiu, H.S., Daniszewski, M.S., Arora, H., Bermudez, Y., Moutinho, C., Chan, C.-L., Bax, M., et al. (2023). A village in a dish model system for population-scale hiPSC studies. Nat. Commun. *14*, 3240.

O'Connor, L.J., Schoech, A.P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A.L. (2019). Extreme polygenicity of complex traits is explained by negative selection. Am. J. Hum. Genet. *105*, 456–476.

Pagliuca, F.W., Millman, J.R., Gürtler, M., Segel, M., Van Dervort, A., Ryu, J.H., Peterson, Q.P., Greiner, D., and Melton, D.A. (2014). Generation of functional human pancreatic β cells in vitro. Cell *159*, 428–439.

Panopoulos, A.D., D'Antonio, M., Benaglio, P., Williams, R., Hashem, S.I., Schuldt, B.M., DeBoever, C., Arias, A.D., Garcia, M., Nelson, B.C., et al. (2017). iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. Stem Cell Rep. *8*, 1086–1100.

Park, I.-H., Zhao, R., West, J.A., Yabuuchi, A., Huo, H., Ince, T.A., Lerou, P.H., Lensch, M.W., and Daley, G.Q. (2008). Reprogramming of human somatic cells to pluripotency with defined factors. Nature *451*, 141–146.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Paull, D., Sevilla, A., Zhou, H., Hahn, A.K., Kim, H., Napolitano, C., Tsankov, A., Shang, L., Krumholz, K., Jagadeesan, P., et al. (2015). Automated, high-throughput derivation, characterization and differentiation of induced pluripotent stem cells. Nat. Methods *12*, 885–892.

Pemberton, T.J., Wang, C., Li, J.Z., and Rosenberg, N.A. (2010). Inference of unexpected genetic relatedness among individuals in HapMap Phase III. Am. J. Hum. Genet. *87*, 457–464.

Ramos-Carreño, C., Torrecilla, J.L., Hong, Y., and Suárez, A. (2022). scikit-fda: Computational Tools for Machine Learning with Functional Data. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*. 2022 IEEE 34th International

Conference on Tools with Artificial Intelligence (ICTAI) (IEEE), pp. 213–218.

Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894.

Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902.

Samuels, Y., Diaz, L.A., Schmidt-Kittler, O., Cummins, J.M., Delong, L., Cheong, I., Rago, C., Huso, D.L., Lengauer, C., Kinzler, K.W., et al. (2005). Mutant PIK3CA promotes cell growth and invasion of human cancer cells. Cancer Cell *7*, 561–573.

Scott, L.J., Erdos, M.R., Huyghe, J.R., Welch, R.P., Beck, A.T., Wolford, B.N., Chines, P.S., Didion, J.P., Narisu, N., Stringham, H.M., et al. (2016). The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat. Commun. *7*, 11764.

Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference*. Proceedings of the python in science conference. Python in Science Conference (SciPy), pp. 92–96.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Soneson, C., Love, M.I., and Robinson, M.D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. [version 2; peer review: 2 approved]. F1000Res. *4*, 1521.

Sprouffske, K., and Wagner, A. (2016). Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. BMC Bioinf. *17*, 172.

Streeter, I., Harrison, P.W., Faulconbridge, A., Flicek, P., Parkinson, H., Clarke, L., and Clarke, L. (2017). The human-induced pluripotent stem cell initiative-data resources for cellular genetics. Nucleic Acids Res. *45*, D691–D697.

Taylor, H.J., Hung, Y.-H., Narisu, N., Erdos, M.R., Kanke, M., Yan, T., Grenko, C.M., Swift, A.J., Bonnycastle, L.L., Sethupathy, P., et al. (2023). Human pancreatic islet microRNAs implicated in diabetes and related traits by large-scale genetic analysis. Proc. Natl. Acad. Sci. USA *120*, e2206797120.

Umans, B.D., Battle, A., and Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? Trends Genet. *37*, 109–124.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *11*, 11.10.1–11.10.33.

Veres, A., Faust, A.L., Bushnell, H.L., Engquist, E.N., Kenty, J.H.-R., Harb, G., Poh, Y.-C., Sintov, E., Gürtler, M., Pagliuca, F.W., et al. (2019). Charting cellular identity during human in vitro β-cell differentiation. Nature *569*, 368–373.

Walczak, M.P., Drozd, A.M., Stoczynska-Fidelus, E., Rieske, P., and Grzela, D.P. (2016). Directed differentiation of human iPSC into insulin producing cells is improved by induced expression of PDX1 and NKX6.1 factors in IPC progenitors. J. Transl. Med. *14*, 341.

Ward, L.D., and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. *30*, 1095–1106.

Wei, X., Wang, B., Wu, Z., Yang, X., Guo, Y., Yang, Y., Fang, Z., Yi, C., Zhang, L., Fan, X., et al. (2023). WD repeat protein 54-mediator of ErbB2-driven cell motility 1 axis promotes bladder cancer tumorigenesis and metastasis and impairs chemosensitivity. Cancer Lett. *556*, 216058.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation *2*, 100141.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.

Yuan, Y., Qi, G., Shen, H., Guo, A., Cao, F., Zhu, Y., Xiao, C., Chang, W., and Zheng, S. (2019). Clinical significance and biological function of WD repeat domain 54 as an oncogene in colorectal cancer. Int. J. Cancer *144*, 1584–1595.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics *26*, 976–978.

Yu, G. (2026). enrichplot: Visualization of Functional Enrichment Result. R package version 1.32.0. https://doi.org/10.18129/B9.bioc.enrichplot.

Zhang, D., Dey, R., and Lee, S. (2020). Fast and robust ancestry prediction using principal component analysis. Bioinformatics *36*, 3439–3446.

Zhan, P., Xi, G.-M., Zhang, B., Wu, Y., Liu, H.-B., Liu, Y.-F., Xu, W.-J., Zhu, Q., Cai, F., Zhou, Z.-J., et al. (2017). NCAPG2 promotes tumour proliferation by regulating G2/M phase and associates with poor prognosis in lung adenocarcinoma. J. Cell Mol. Med. *21*, 665–676.

Zhu, K., Zhang, H., Luan, Y., Hu, B., Shen, T., Ma, B., Zhang, Z., and Zheng, X. (2024). KDM4C promotes mouse hippocampal neural stem cell proliferation through modulating ApoE expression. FASEB J. *38*, e23511.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
mTeSR1 medium	STEMCELL Technologies	Cat: #85850
CloneR	STEMCELL Technologies	Cat: #05888
MycoAlert® Mycoplasma PLUS Detection Kit	Lonza	Cat: #LT07-710
penicillin-streptomycin	Thermo Fisher	Cat: #15140122
Accutase	Sigma Aldrich	Cat: #A6964
QubitTM 1X dsDNA Broad Range Assay Kit	Thermo Fisher	Cat. #Q33266
QiaAMP Mini kit	Qiagen	Cat. #51304
NEBNext Poly(A) mRNA Magnetic Isolation Module	New England Biolabs	Cat. #E7490
RNA 6000 Nano Kit	Agilent	Cat. #5067-1511
TRIzol™	Thermo Fisher	Cat: #15596026
Y27632 dihydrochloride	AbMole	Cat: #M1817
Deposited data		
iPSCORE iPSC lines genome sequencing data	dbGaP	phs001325.v6.p1
FUSION iPSC lines data	EGA	EGAS00001008440
Experimental models: Cell lines		
GENESiPS iPSC lines	Carcamo-Orive et al. (2017)	See publication
iPSCORE iPSC lines	Panopoulos et al. (2017) Arthur et al. (2025)	See publication
HipSci iPSC lines	Streeter et al. (2017)	See publication
NYSCF global stem cell array iPSC lines	Paull et al. (2015)	See publication
Software and algorithms		
scikit-FDA	Ramos-Carreño et al. (2022)	v0.10.1
Growthcurver	Sprouffske et al. (2016)	v0.3.1
Salmon	Patro et al. (2017)	v1.10.3
verifyBamId	Jun et al. (2015)	v1.1.1
KING	Manichaikul et al. (2010)	N/A
DESeq2	Love et al. (2014)	v1.36.0
RUVSeq	Risso et al. (2014)	v1.30.0
GCTA	Yang et al. (2011)	v1.94.1
TopMED imputation server	Das et al. (2016)	N/A
FRAPOSA	Zhang et al. (2020)	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GMMAT	Chen et al. (2016)	v1.4.2
powerEQTL	Dong et al. (2021)	v0.3.6
GATK Best Practices pipeline	Van der Auwera et al. (2013)	N/A
Variant Effect Predictor	McLaren et al. (2016)	v111
CADD	Rentzsch et al. (2019)	v1.6
DeepRVAT	Clarke et al. (2024)	v1.1
Other		
Opera Phenix High-Content Screening System	PerkinElmer	N/A
Geltrex coated plates	Thermo Fisher	Cat: #A1413301
NovaSeq 6000 platform	Illumina	N/A
NovaSeqX platform	Illumina	N/A
Celigo imaging instrument	Revvity Inc.	Cat. #200-BFFL-5C
GRCh38 human reference transcriptome	Ensembl	v111

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Summary of induced pluripotent stem cells (iPSCs)

We used existing lines and newly generated lines. We obtained existing lines from four primary studies: the Human Induced Pluripotent Stem Cells Initiative (HipSci; n = 33), GENEticS of Insulin Sensitivity (GENESiPS; n = 119), iPSC Collection for Omic REsearch (iPSCORE; n = 183), and NYSCF Global Stem Cell Array (GSA; n = 255) (Paull et al., 2015; Streeter et al., 2017; Carcamo-Orive et al., 2017; Panopoulos et al., 2017; Arthur et al., 2025). In addition, we generated 12 iPSC lines from human fibroblasts of consented donors from the Finland-United States Investigation of NIDDM (FUSION) study using the NYSCF automated, high-throughput iPSC characterization and differentiation platform as described previously (Paull et al., 2015; Scott et al., 2016). In total, this study considered 610 lines, which included sixteen twin participants (eight twin pairs). Except for the twin concordance analysis and genetic relatedness matrix (GRM) calculation, we considered only one of the lines from each twin pair for downstream analyses, retaining 602 lines.

iPSC thawing and processing

As iPSCs were delivered to NYSCF in various non-standard, non-uniform cryovials, they were not immediately available for thawing. As such, we used manual processes to thaw iPSCs not stored in Matrix tubes. We removed cryovials containing frozen iPSCs from liquid nitrogen storage and transferred them on dry ice before thawing. We placed vials into a 37°C water bath for ~90 s before resuspending them in prewarmed mTeSR1 medium (STEMCELL Technologies, Cat: #85850, Vancouver, BC) containing 10% CloneR (STEMCELL Technologies, Cat: #05888) or Freedom media (Life Technologies custom media #A14577SA) supplemented with 10 μM of thiazovivin (TOCRIS Bioscience). We transferred cells into a 15 mL conical tube and added additional mTeSR1 with 10% CloneR or Freedom with 10 μM of thiazovivin dropwise to the conical tube to ensure complete neutralization of the cryopreservative medium for a total volume of 15 mL per conical tube. We took a small aliquot (10 μL) of each cell suspension for cell counting using a dead/total count application on an Opera Phenix High-Content Screening System (herein referred to as Opera Phenix; PerkinElmer, Waltham, MA). We centrifuged cell suspensions at 800 revolutions per minute for four minutes and aspirated the supernatant without disturbing the cell pellet. We resuspended cells in either mTeSR1 with 10% CloneR or Freedom with 10 μM of thiazovivin media solution and plated cells on a 12-well plate, 24-well plate, or 96-well Geltrex coated plate (Thermo Fisher, Cat: #A1413301), depending on their respective live cell counts as follows. If we counted fewer than 100,000 live cells for a given vial, we resuspended cells in 200 μL of either mTeSR with 10% CloneR or Freedom with 10 μM of thiazovivin media and seeded cells into one well of a 96-well plate; if we counted fewer than 250,000 but more than 100,000 live cells, we

resuspended cells in 500 μL of media solution and seeded cells into one well of a 24-well plate; and if we counted over 250,000 live cells, we resuspended cells in 1.0 mL of media solution seeded into one well of a 12-well plate.

iPSC culture, expansion, and passaging

After thawing and plating, we cultured cells for 48 h at 37°C in 5% CO_2 and either mTeSR1 with 10% CloneR or Freedom with 10 μM of thiazovivin media without changing media and without penicillin-streptomycin. We fed cells daily with mTeSR1 or Freedom. We collected supernatant from each well and tested for mycoplasma contamination with the MycoAlert Mycoplasma PLUS Detection Kit (Lonza, Cat: #LT07-710, Rockville, MD) and the accompanying MycoAlert Assay Control Set (Lonza, Cat: #LT02-518) as per the manufacturer's instructions. After we tested for mycoplasma contamination, we transferred mycoplasma-free cells into the automated workcells and fed them either mTeSR1 or Freedom supplemented with penicillin-streptomycin (pen-strep; Thermo Fisher, Cat: #15140122). Upon most wells reaching 60–80% confluence, we passaged cells on the NYSCF robotic culture system into a Geltrex coated 24-well plate (Thermo Fisher) at a density of 50,000 cells per well. We performed all liquid handling steps using an OT-II liquid handler (OpenTrons, Long Island City, NY) and took measurements on a Synergy HT BioTek plate reader (BioTek, Winooski, VT).

We aspirated culture medium without disturbing the cell layer and washed wells with Accutase (Sigma-Aldrich, Cat: #A6964, St. Louis, MO). We dissociated cells in this same reagent before performing additional washes with 250 μL of Accutase. We incubated cells for 20 min at 37°C in 5% CO_2. Once cells were fully dissociated, we neutralized Accutase by adding prewarmed mTeSR1 or Freedom containing 10 μM Y27632 (Accutase-neutralization media ratio: 1:1). We added 100 μL of media for a 96-well plate; 250 μL for a 24-well plate, and 500 μL for a 12-well plate, where the media was either mTeSR1 or Freedom. We transferred cell suspensions into an intermediate 96-well round-bottom plate and centrifuged at 120 g for five minutes. After centrifugation, we aspirated the supernatant and resuspended cells in 1.0 mL of prewarmed mTeSR1 or Freedom containing 10 μM of Y27632. We took 10 μL of each cell suspension for a dead/total cell count on the Opera Phenix. We transferred cells from the intermediate plate to the destination plate(s) at the appropriate cell density (for 24-well plates, 50,000 live cells per well; for 96-well plates, 10,000–20,000 live cells per well). We then returned destination plates to the incubator for a minimum of 24 h.

After 24 h, we performed a half-feed followed by daily media changes. Upon most wells reaching 60–80% confluence, we passaged cells into Geltrex coated 96-well plates (Thermo Fisher) with 10,000–20,000 cells per well in mTeSR1 or Freedom with pen-strep and 10 μM Y27632 dihydrochloride (AbMole, Cat: #M1817, Houston, TX). After 24 h, we performed a half-feed before subsequent daily feeds with mTeSR1 or Freedom. Upon most wells reaching ∼90% confluency, we dissociated and counted cells before we transferred the cell suspension into Matrix tubes and centrifuged at 600×g for 5 min. We aspirated the supernatant without disturbing the cell pellets. We resuspended cells intended for bulk RNA sequencing in 100 μL of TRIzol (Thermo Fisher; Cat: #15596026) per Matrix tube and stored at −80°C. We left cells intended for SNP genotyping and genome/exome sequencing as dry pellets.

iPSC karyotyping

We karyotyped all iPSC lines. Briefly, DNA was extracted from the iPSCs using an epMotion (Eppendorf) and ReliaPrep 96 gDNA Miniprep HT System (Promega) kit. The DNA was tested using the Illumina Global Screening Array or the Illumina CoreExome 24 platform, with data analyzed via Genome Studio (Illumina) and the Copy Number Variation (CNV) analysis, cnvPartition 3.2.0. The line was considered passing (normal) in the absence of major CNVs (>2.5 Mb). For one line, we detected trisomy 8. We considered all downstream analyses with and without this line and found no meaningful effect on the results. We retained this line in the reported results.

Ethics statement

All research in this study was conducted under category 1A as defined by the International Society for Stem Cell Research guidelines. All tissue samples were sourced under IRB #: OH95-HG-N030. Study subjects provided tissues under full consent and agreement to generate and use iPSC lines for broad research. All samples are anonymized such that personally identifiable information is not accessible.

METHOD DETAILS

RNA extraction and sequencing

We performed bulk RNA sequencing for all cell lines in-house at the NIH Intramural Sequencing Core (NISC). We extracted total RNA by transferring the 100 μL TRIzol cell suspension to a 1.5 mL microfuge tube and adding 900 μL of

additional TRIzol following the manufacturer's instructions. We eluted the RNA in 20 μL RNAse-free water. We assessed total RNA using the RNA 6000 Nano Kit (Agilent, Cat. #5067-1511, Santa Clara, CA) and quantitated using the QubitTM RNA Broad Range Assay kit (Thermo Fisher, Cat. #Q10211). We used 1.0 μg of total RNA with an RNA Integrity Number (RIN) $\geq$ 7 for library preparation with the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, Cat. #E7490, Ipswich, MA). We sequenced libraries on the Illumina NovaSeq 6000 platform using an S4 flow cell with 151 bp paired-end reads.

DNA extraction, genotyping, and sequencing

AKESOgen (Tempus, Chicago, IL) performed SNP genotyping of all NYSCF GSA iPSC lines. We extracted DNA using the Promega Maxwell 96 gDNA Miniprep HT System following the manufacturer's recommendations (Promega, Cat. #A2670, Madison, WI). We transferred DNA to AKESOgen for sequencing using the Illumina Global Screening Array and processed raw data using Illumina Genome Studio with reference genome GRCh37. We used three Illumina BeadChip arrays: InfiniumOmni2-5Exome-8v1-3 v1.3 (Illumina, Cat. #20031813, San Diego, CA), Infinium Global Screening Array-24 Kit (Illumina, Cat. #20030770), and Infinium HumanCore-24 v1.2 (Illumina, Cat. #20024566).

We performed SNP genotyping of all other iPSC lines (i.e., from the FUSION, HipSci, iPSCORE, and GENESiPS cohorts) at the NHGRI Genomics Core. We resuspended DNA pellets in 100 μL 1X PBS (Thermo Fisher, Cat. #10010031) and transferred to a 1.5 mL microfuge tube (Fisher Scientific, Cat. #14-222-169). We added 100 μL of additional 1X PBS to wash the Matrix tubes and combined the volumes. We extracted DNA using the QiaAMP Mini kit (Qiagen, Cat. #51304, Hilden, Germany) following manufacturer instructions. We resuspended DNA in 50 μL Buffer AE and quantitated using QubitTM 1X dsDNA Broad Range Assay Kit (Thermo Fisher, Cat. #Q33266). We genotyped 250 ng of DNA using the Illumina Omni 2.5 Exome-8 v1.5 beadchip array. We analyzed data in GenomeStudio with the v1.5_A1_GRCh37 manifest and the v1.5_A1 cluster file (Illumina).

Genosity (Invitae Corporation, South San Francisco, CA) performed exome sequencing (ES) of NYSCF GSA iPSC lines. We extracted DNA using the Promega Maxwell 96 gDNA Miniprep HT System following manufacturer's instructions (Promega). We transferred DNA to Genosity for sequencing using Twist Human Core Exome EF Multiplex Complete Kit (Twist Bioscience, Cat. #PN 100803, South San Francisco, CA). Genosity sequenced at a minimum coverage of 25X of 90% of bases for 183 samples and returned the trimmed FASTQ files for downstream analyses.

We performed genome sequencing (GS) of iPSC lines from the FUSION, HipSci, and GENESiPS cohorts at NISC. We resuspended DNA pellets in 100 μL 1X PBS (Thermo Fisher) and transferred the contents to a 1.5 mL microfuge tube (Fisher Scientific). We added 100 μL of additional 1X PBS to wash the matrix tubes and combined the volumes. We extracted DNA using the QiaAMP Mini kit (Qiagen). We resuspended DNA in 50 μL Buffer AE and quantitated using the Qubit 1X dsDNA Broad Range Assay Kit (Thermo Fisher, Cat. #Q33266). We submitted 1.0 μg of DNA to NISC for library preparation with the Illumina TruSeq PCR-free Sample Preparation kit (Illumina, Cat. #20015962). We sequenced libraries on the Illumina NovaSeqX platform using 25B flow cells.

We downloaded genome sequencing data sourced from donor blood or fibroblasts from iPSCORE (dbGaP accession number phs001325.v6.p1) (Arthur et al., 2025; Panopoulos et al., 2017).

Cellular imaging of individual wells

To derive a growth rate phenotype, we used daily measurements of well coverage (expressed as a percent) for each well. We imaged all plates stored within the automated system nightly using Celigo imagers during culture (Revvity Inc., Cat. #200-BFFL-5C, Waltham, MA). We used the default confluence analysis parameters from the Celigo software to measure well coverage.

QUANTIFICATION AND STATISTICAL ANALYSIS

Growth rate derivation from cellular images and replicate concordance

We only considered wells that contained at least three well coverage measurements between 30 and 70%. We removed outlier wells using magnitude-shape plot outlier detection in scikit-FDA v0.10.1 (Ramos-Carreño et al., 2022). Because most donors had multiple replicate wells, we identified and removed wells with low concordance by calculating the pairwise Pearson correlation coefficients among all replicate wells for each donor. Specifically, we rounded all well coverage measurement time points to the nearest day and aligned the resulting time series of well coverage measurements for each well pair based on shared time points. We calculated Pearson correlation coefficients using all complete, aligned observations. We excluded any well that exhibited a Pearson correlation coefficient (r) < 0.75 with any other replicate well from

the same donor. From these filtered data, we min-max normalized the individual per-well time series measurements and fitted each to a standard logistic curve (time measured in days versus normalized well coverage) using Growthcurver v0.3.1 (Sprouffske and Wagner 2016). We further excluded replicate wells that either failed to fit a growth curve or showed a residual sum of squares from the logistic curve fit exceeding three standard deviations above the mean error. Finally, for each remaining well, we derived the growth Area-Under-the-Curve (gAUC) metric from the fitted curves.

To assess concordance of gAUC across replicate wells corresponding to each donor and within the eight twin pairs, we calculated the intraclass correlation coefficient (ICC) using a one-way random-effects model for single measurements (ICC(1,1)) as implemented in the Python statsmodels package v0.14.4 (Seabold and Perktold 2010).

SNP genotype quality control procedures and imputation

After SNP chip genotyping, we aligned array probe sequences to the GRCh37 reference genome using the Burrows-Wheeler Aligner (BWA) v0.17.7. (Li and Durbin 2009) Pre-imputation quality control procedures included removing multi-mapping probes, probes containing known 3′ variants, and non-SNP variants. We further filtered out variants with per-variant missingness >5% and Hardy-Weinberg equilibrium (HWE) $p < 10^{-6}$, and removed samples exceeding 5% missingness. Using the 1000 Genomes Project Phase 3 data as a reference, we verified and corrected strand orientation, allele assignments, genomic positions, and reference/alternate designations (1000 Genomes Project Consortium et al., 2015). We excluded biallelic variants with minor allele frequencies (MAF) > 40%, variants absent from the reference panel, and variants with discordant allele assignments. We merged genotyping data from individual arrays prior to lifting over and imputation.

We conducted lift over from GRCh37 to GRCh38, pre-imputation variant-level quality control procedures, and genotype imputation using the TOPMed Imputation Server (Das et al., 2016). We used Eagle v2.4 to perform SNV pre-phasing and minimac4 to impute SNV genotype dosages with the TOPMed imputation panel v3 (Das et al., 2016; Loh et al., 2016). After imputation, we retained biallelic SNPs meeting the following criteria: MAF >5%, imputation quality $r^2 > 0.3$, and HWE $p > 10^{-10}$. We merged the quality-controlled imputed data across studies and used PLINK v2.00a5 with the "–king-cutoff" flag to remove genetically identical samples (KING relatedness coefficient >0.354) (Manichaikul et al., 2010; Chang et al., 2015). Our final study set contained 602 unique individuals. To infer donor sex for all 602 individuals based on genotype calls, we used bcftools v1.17 with the "guess-ploidy" plugin considering genotypes in the non-pseudoautosomal region of chromosome X from quality-controlled common variant data (Danecek et al., 2021).

We also generated linkage disequilibrium (LD)-pruned genotypes for downstream analyses, including genetic principal component (PC) calculation and power analyses. We retained only biallelic SNPs with MAF >5%, HWE $p > 10^{-10}$, and imputation quality $r^2 > 0.8$. We excluded high-LD regions defined by Anderson et al., (2010). We applied LD pruning using a sliding window of 1 kb, a step size of 0.8, and an LD threshold of $r^2 < 0.1$, retaining 66,917 pruned variants for downstream analysis.

Variant calling and quality control procedures in genomic sequencing

After GS and ES, we processed reads according to the GATK Best Practices pipeline for variant discovery (Van der Auwera et al., 2013). Following adapter trimming, we aligned paired-end reads to the GRCh38 reference genome using BWA MEM v0.17.7 with default parameters (Li and Durbin 2009). We subsequently filtered per-lane BAM files to retain high-quality, primary alignments, merged, and sorted using samtools v1.22 (Li 2013). To ensure accurate variant calling, we recalibrated base quality scores using the Genome Analysis Toolkit (GATK) v4.0.5.1 BaseRecalibrator, incorporating dbSNP v151 as well as the Mills and 1000 Genomes gold standard indels to correct systematic biases in base quality scores (Van der Auwera et al., 2013; Sherry et al., 2001; McKenna et al., 2010). We performed variant calling per sample in GVCF mode using the GATK HaplotypeCaller (McKenna et al., 2010). We calibrated variants (indels and SNPs) via Variant Quality Score Recalibration (VQSR), which was trained using the Mills and 1000 Genomes gold standard indels, HapMap v3.3, sites found to be polymorphic on the Illumina Omni 2.5M SNP array, and dbSNP v151 to finalize high-quality, recalibrated SNP and indel calls (Sherry et al., 2001; Pemberton et al., 2010; Van der Auwera et al., 2013; McKenna et al., 2010).

We used VerifyBamID v1.1.1 to compare the sequencing data to quality-controlled genotypes and identified no sample swaps (Jun et al., 2015). For downstream analyses, we retained biallelic variants (SNPs and indels with length $\leq$5 bp) meeting the following criteria: MAF <1% and HWE $p > 10^{-10}$. We kept genetically distinct samples (KING relatedness coefficient <0.354) as determined from SNP genotype data (n = 516).

RNA-seq data processing and quality control procedures

After RNA sequencing, we performed transcript quantification using Salmon v1.10.3 in quasi-mapping mode with default parameters, aligning reads to the GRCh38 human reference transcriptome (Ensembl v111) (Dyer et al., 2025; Patro et al., 2017). We used tximport v1.32.0 to convert transcript-level abundance estimates to gene-level counts for differential gene expression analysis via the lengthScaledTPM function (Soneson et al., 2015).

As an additional quality control procedure, we used STAR v2.7.11a to align reads to the GRCh38 reference genome in a paired-end fashion (Dobin et al., 2013). We identified sample swaps and contaminated samples by comparing the allelic RNA-seq read count distribution to the finalized genotypes of the 602 donors using VerifyBamID v1.1.1 (Jun et al., 2015). We identified and resolved 38 pairs of sample swaps. We removed 25 contaminated samples from the analyses (FREEMIX or CHIPMIX >0.055). After these procedures, 333 iPSC lines remained for final analyses. For downstream analyses, we considered only genes with transcripts per million (TPM) > 0.5 in at least 50% of samples, resulting in 15,603 features.

Determination of covariates for downstream analyses

To determine covariates for subsequent models, we fit a linear mixed model implemented in lmerTest v3.1.358 to test if genetically-imputed donor sex, pluripotency reprogramming method, starting cell type, and cell culture media are associated (FDR <5%, Benjamini-Hochberg [BH] procedure) with gAUC, controlling for donor as a random effect (Kuznetsova et al., 2017; Benjamini and Hochberg 1995). We found pluripotency reprogramming method, starting cell type, and cell culture media were associated with gAUC and therefore included these three terms in subsequent association analyses. We did not include cell culture media type as a covariate in our differential expression analyses because all cell lines included in those analyses were cultured in mTeSR1 media.

Differential gene expression and gene ontology enrichment analyses

We normalized gAUC using inverse rank normalization and calculated the gAUC mean value for each RNA-seq sample. We performed differential expression analysis using a three-stage process as previously described in Taylor et al., (2023). First, we tested for differentially expressed genes across gAUC using DESeq2 v1.36.0 with default parameters and adjusting for starting cell type and pluripotency reprogramming method as fixed effect covariates (Love et al., 2014). Next, we used RUVSeq v1.30.0 with default parameters to derive a factor representing unwanted, latent variation using control gene expression, defined as genes with $p > 0.25$ from the initial differential expression analysis (Risso et al., 2014). Finally, we re-ran differential expression analysis and incorporated the calculated RUVSeq factor as an additional fixed-effect covariate. We performed multiple hypothesis correction using the BH procedure (Benjamini and Hochberg 1995).

Additionally, we performed gene ontology (GO) enrichment analysis similar to Grenko et al., (2024). Briefly, we calculated the enrichment of differentially expressed genes (FDR <5%) in GO terms from the biological process ontology using the compareCluster function from clusterProfiler v4.4.1 with options fun = "enrichGO" (Wu et al., 2021). We controlled for the number of tests performed using the BH procedure (Benjamini and Hochberg 1995). To create the network plots, we calculated the semantic similarity between enriched GO terms (FDR <0.05) using GOSemSim v2.22.0), constructed a similarity matrix using the pairwise_termsim function from enrichplot v1.26.1, and generated the plots using the emapplot function from enrichplot, where edges connecting GO terms (nodes) required a similarity ≥0.75 (Yu et al., 2010; Yu, 2026).

Genetic relatedness matrix calculation

We calculated a GRM across all donors with genotype data, including the eight monozygotic twins, using GCTA v1.94.1 and common genetic variants (MAF >5%) from the imputed genotype data (Yang et al., 2011). For the downstream analyses that required the GRM (i.e., variance decomposition, common variant genetic association testing, and rare variant genetic association testing), we subset the GRM to the relevant donors.

Variance decomposition analyses

We performed variance decomposition analyses using GCTA v1.94.1 at the cell line (i.e., donor) level with the mean gAUC value across all replicates per donor (Yang et al., 2011). We used three subsets of the available data: (i) all unique cell lines ($n = 602$) including cell culture media, starting cell type, and pluripotency reprogramming method as discrete covariates, (ii) all unique cell lines with bulk RNA-seq data incorporating starting cell type and pluripotency reprogramming method as discrete covariates ($n = 333$), and (iii) all unique cell lines with bulk RNA-seq data incorporating pluripotency markers as quantitative covariates and starting cell type and pluripotency reprogramming method as discrete covariates ($n = 333$). We modeled pluripotency with gene expression of four established pluripotency markers measured in TPM: *POU5F1*

(ENSG00000204531), *SOX2* (ENSG00000181449), *NANOG* (ENSG00000111704), and *PODXL* (ENSG00000128567) (Pagliuca et al., 2014; Lowry et al., 2008). For subsets (ii) and (iii), we did not include media as a discrete covariate since all cell lines with available RNA-seq data were cultured with mTeSR1 media.

We fit the following model:

$$Y = X \cdot \beta + g + \varepsilon$$

where Y represents the vector of sample-level mean phenotype values, X is the matrix of fixed-effect covariates (either media or gene expression of pluripotency markers), $g \sim N\left(0, K_g \sigma_g^2\right)$ is the random genetic effect modeled using the GRM K_g calculated as described above, and $\varepsilon \sim N\left(0, I\sigma_\varepsilon^2\right)$ represents the residual error term. Here, σ_g^2 and σ_ε^2 denote the genetic and residual variance components, respectively.

To quantify the proportion of phenotypic variance explained by genetic factors, we calculated the narrow-sense SNP heritability as:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2}$$

Genetic principal component and ancestry assignment

We calculated genetic PCs and assigned participants to genetic ancestry groups using Fast and Robust Ancestry Prediction by Online singular value decomposition and Shrinkage Adjustment (FRAPOSA) (Zhang et al., 2020). Briefly, FRAPOSA works by performing PC analysis in a diverse reference cohort with known genetic ancestry labels, systematically calculating PCs for each participant in a target cohort by comparing them to the reference, and grouping participants from the target cohort to the most similar genetic ancestry group from the reference. Using FRAPOSA with default parameters, LD-pruned genotypes as input, and 1000 Genomes Project Phase 3 data (1KG) as the reference cohort, we calculated the first 30 PCs and grouped participants into 1KG-AFR-like (African), 1KG-AMR-like (American), 1KG-EUR-like (European), 1KG-EAS-like (East Asian), or 1KG-SAS-like (South Asian) genetic ancestries (1000 Genomes Project Consortium et al., 2015).

Common variant association testing

We normalized gAUC growth phenotypes through rank inverse normalization. We conducted association tests using the Generalized Linear Mixed Model Association Tests (GMMAT) v1.4.2 framework with default parameters (Chen et al., 2016). The analyses proceeded in two steps: (i) fitting a null model using the GRM, and (ii) performing score tests to identify genetic associations. We included 10 genetic PCs, cell culture media type, starting cell type, and pluripotency reprogramming method as fixed-effect covariates. We considered variants with $p < 5 \times 10^{-8}$ as genome-wide associated and those with $p < 10^{-6}$ as nominally associated.

Power analyses for common variant associations

We estimated the sample size needed to detect common variant associations with gAUC using an unbalanced one-way ANOVA framework, implemented in the powerEQTL package v0.3.6 (Dong et al., 2021). Briefly, for a range of absolute effect size thresholds (absolute effect size <0.1, 0.2, 0.3, 0.4, 0.5), we calculated statistical power for up to 6,000 samples. Calculations assumed a genome-wide threshold of $p < 5 \times 10^{-8}$, a maximum MAF of 50%, and a residual standard deviation ($\sigma = 0.4584$) derived from variance decomposition across all 602 samples from the common variant association study. We estimated the effect size parameter "deltaVec" by averaging the observed differences in mean gAUC comparing heterozygotes to each homozygous genotype group for LD-pruned variants within each effect size bin.

Rare variant annotation, aggregation, and association testing

To investigate the role of rare variation on gAUC, we performed two complementary approaches to annotate the functional impact of rare variants and test for associations: (i) GMMAT association testing of damaging missense or predicted loss of function variants (pLoF) variants and (ii) DeepRVAT, which provides an end-to-end annotation to testing framework and utilizes 34 variant annotations (Clarke et al., 2024).

For approach (i), we annotated variants using Variant Effect Predictor (VEP, Ensembl v111) selecting the most severe consequence across all protein-coding transcripts (Ensembl v111 annotations) for each variant (Dyer et al., 2025; McLaren et al., 2016). We classified pLoF variants as those with consequences including transcript ablation, splice

acceptor variant, splice donor variant, stop gained, frameshift variant, stop lost, start lost, transcript amplification, feature elongation, or feature truncation. We also evaluated VEP-annotated missense variants for deleteriousness using CADD v1.6 and classified variants with scores $\geq$20 as damaging (Rentzsch et al., 2019).

For each gene, we aggregated coding variants with MAF <1% that were annotated as either pLoF or damaging missense. To be considered in downstream analyses, we required a gene to contain at least two qualifying variants and carry a non-zero burden score in at least 1% of the cohort, resulting in 8,480 genes.

We normalized gAUC growth phenotypes through rank inverse normalization. We tested for associations across all 516 lines with GS or ES data using GMMAT v1.4.2 with default parameters (Chen et al., 2019). The analyses proceeded in two steps: (i) fitting a null model using the GRM, (ii) performing hybrid set-based tests to identify gene-level rare variant associations using the provided variant-gene annotations. For the hybrid test, we used the SMMAT function implemented in GMMAT, which efficiently combines a burden test and an adjusted SKAT test. As covariates, we included 10 genetic PCs, cell culture media type, starting cell type, and pluripotency reprogramming method as fixed-effect covariates. We applied a Bonferroni-corrected threshold to account for the number of unique genes tested ($p < 5.90 \times 10^{-6}$, 8,480 tests). For the sensitivity analysis, we fit a regression in the same manner but using CADD $\geq$25 to define damaging missense variants, resulting in an adjusted Bonferroni-corrected threshold of $p < 1.45 \times 10^{-5}$ (3,441 tests) after gene aggregation.

For approach (ii), we used DeepRVAT v1.1 to generate gene burden scores for each line based on 34 variant annotations and tested for associations of the burden score with gAUC (Clarke et al., 2024). Briefly, DeepRVAT is a framework that uses deep set neural networks to aggregate trait-agnostic functional impact predictions for rare variants across an arbitrary number of variant annotations, producing a gene-level impairment score. Using the pre-trained gene impairment modules, we considered variants with MAF <1% and CADD $\geq$5 and computed gene burden scores across all lines with default parameters. We then tested the burden score for association with gAUC using the default score test implementation in DeepRVAT. Unlike GMMAT, this implementation does not account for relatedness or multiple measurements, so we restricted the analysis to the 504 unrelated individuals with GS or ES data, normalized gAUC using inverse rank normalization, and calculated the gAUC mean value for each line. As in approach (i), we controlled for the number of genes tested using Bonferroni correction.