AWARD NUMBER: W81XWH-16-1-0130

TITLE: Single-Cell Dissection of Human Pancreatic Islet Dysfunction in Diabetes

PRINCIPAL INVESTIGATOR: Michael L. Stitzel, Ph.D.

CONTRACTING ORGANIZATION: The Jackson Laboratory
Bar Harbor, ME 04609

REPORT DATE: MARCH 2019

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MARCH 2019	Final	1JUNE2016 - 30NOV2018

4. TITLE AND SUBTITLE
Single-Cell Dissection of Human Pancreatic Islet Dysfunction in Diabetes

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-16-1-0130

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Michael L. Stitzel, Ph.D.

E-Mail: Michael.Stitzel@jax.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Jackson Laboratory for
Genomic Medicine
10 Discovery Drive
Farmington, CT 06032-2374

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Islets are composed of ≥5 endocrine cell types that perform complementary functions to maintain proper glucose homeostasis. This cellular heterogeneity impedes our ability to understand the precise transcriptional repertoire and regulatory landscape of each cell type and to determine how these programs in each cell type are perturbed in type 2 diabetes (T2D). The overarching goal of this project is to determine, with single cell resolution, changes in cellular composition and cell-specific gene expression programs elicited by T2D in human islets using innovative single cell transcriptomic (scRNA-seq; Aim 1) and epigenomic (scATAC-seq; Aim 2) technologies.

15. SUBJECT TERMS
Single cell; epigenome; scATAC-seq; scRNA-seq; transcriptome; human islet; alpha; beta; delta; pancreatic polypeptide (PP); gamma; epsilon; endocrine.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
Unclassified	Unclassified	Unclassified	UU	65	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Page

1. INTRODUCTION:

Islets are composed of ≥5 endocrine cell types that perform complementary functions to maintain proper glucose homeostasis. This cellular heterogeneity impedes our ability to understand the precise transcriptional repertoire and regulatory landscape of each cell type and to determine how these programs in each cell type are perturbed in type 2 diabetes (T2D). The overarching goal of this project is to determine, with single cell resolution, changes in cellular composition and cell-specific gene expression programs elicited by T2D in human islets using innovative single cell transcriptomic (scRNA-seq; Aim 1) and epigenomic (scATAC-seq; Aim 2) technologies.

2. KEYWORDS:

Single cell; epigenome; scATAC-seq; scRNA-seq; transcriptome; human islet; alpha; beta; delta; pancreatic polypeptide (PP); gamma; epsilon; endocrine

3. ACCOMPLISHMENTS:

What were the major goals of the project?

Major goals of the project:

Aim 1: Islet single cell transcriptomes

1a: Non-diabetic (ND) islet single cell transcriptomes

Milestone: ~1000 single cell transcriptome profiles from 5 ND islets

Achieved: 21,370 single cell transcriptomes from 5 ND islets sequencing results (~2000% of cells; 100% of samples)

1b: Type 2 diabetic (T2D) islet single cell transcriptomes

Milestone: ~1000 single cell transcriptomes from 5 T2D islets

Achieved: 19,683 single cell transcriptome profiles from 5 T2D islets (~2000 % of cells, 100% of samples)

1c: Determine cell type transcriptome signatures in ND and T2D samples

Milestone: Comprehensive analysis of islet transcriptomes and identification of cell type-specific transcriptomes / “signature” genes

Achieved: In total, we identified 188 cell-specific genes; 80 for alpha, 57 for beta, 25 for delta, and 19 for gamma/PP cells.

1d: Identify cell type-specific expression differences between ND and T2D samples

Milestone: Identification of cell type-specific differential expression in T2D vs. ND samples

Achieved: Comparative analysis of each cell-type transcriptome profiles in T2D and ND individuals revealed a total of 58 genes with altered expression (FDR < 5%).

Aim 2: Islet single cell epigenomes

2a: Non-diabetic (ND) islet single cell epigenomes

Milestone: ~1000 single cell epigenome (scATAC-seq) profiles

Achieved: 459 single cell epigenome profiles (46%)

2b: Type 2 diabetic (T2D) islet single cell epigenomes

Milestone: ~1000 single cell epigenome (scATAC-seq) profiles

Achieved: 886 single cell epigenome profiles (88%) from 9 individuals (90%)

2c: Determine cell type epigenome signatures in ND and T2D samples

Milestone: Comprehensive analysis of islet epigenomes and identification of cell type- specific regulatory element use/epigenome signatures

Achieved: Established and evaluated scATAC-seq analysis pipelines; QC and unsupervised clustering analyses of 1012 single cell ATAC-seq profiles (553 cells from 5 ND and 459 cells from 5 T2D islet donors); scATAC-seq data did not cluster into distinct cell types despite use of multiple approaches and algorithms

2d: Identify cell type-specific epigenomic differences between ND and T2D samples

Milestone: Identification of cell type-specific differences in regulatory element use/epigenome signatures in T2D and ND states

Achieved: scATAC-seq profiles did not cluster into distinct cell types, so we were unable to determine cell type-specific differences in regulatory element use from these data.

What was accomplished under these goals?

1) *Major activities*: We completed single cell transcriptome and epigenome profiling of islets from 5 T2D and 5 matched ND donors. In total, we have generated and analyzed 41,053 single cell transcriptomes (median n=4034 cells per donor) and 886 single cell open chromatin profiles (median n=123 cells per donor). With single cell transcriptome data, we have clustered data to identify cell populations and "signature genes" that are distinctly expressed in each islet cell type, and differences in islet cell proportions between T2D and ND donors, and identified differentially expressed genes relative islet cell proportion differences and completed analyses to identify differentially expressed genes in each islet cell type between the T2D and ND donors. Single cell ATAC-seq was of high quality, but the single cell profiles were either not sufficiently deep or dynamic or were to similar between islet cell types to identify distinct endocrine cell type clusters.

2) *Specific objectives*: The specific objectives in this period was to complete profiling and analyses of single cell transcriptomes and epigenomes from T2D and ND individuals to 1) identify cell type-specific gene expression and *cis*-regulatory element use; and 2) identify altered gene expression and *cis*-regulatory element use in T2D islet cell types compared to ND islet cell types.

3) *Significant results:*

Aim 1: Single cell transcriptomes

Over the total project period, we completed single cell transcriptome profiling of 41,053 cells from 10 islet donors (19,683 from five T2D donors and 21,370 from five ND donors). Using Seurat[1], we identified 29,101 endocrine cells forming distinct alpha, beta, delta, and gamma/PP clusters and 11,952 exocrine cell type (stellate, ductal, acinar) or endothelia clusters (**Figure 1**), which we annotated based on expression of classic marker genes for each cell type (*GCG, INS, SST, PPY, COL1A1, KRT19, PRSS1,* and *ESAM*, respectively). Importantly, cells from both T2D and ND clustered together, and single cell transcriptomes from different individuals were evenly mixed within a given cell population, suggesting that individual-specific effects were not major confounders of these and subsequent analyses.

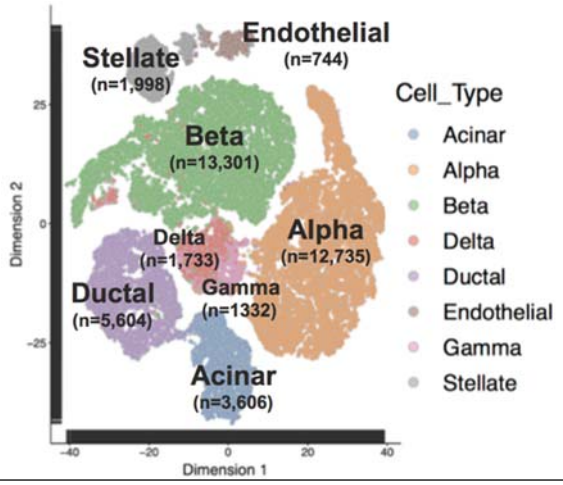


Figure 1. Single cell transcriptomes of 41,053 cells from five T2D and ND islet donors cluster into distinct endocrine (alpha, beta, delta, gamma/PP), exocrine (ductal, stellate, acinar) and endothelial cell types.

Cell type-specific genes:

In total, we identified 188 cell-specific genes; 80 for alpha, 57 for beta, 25 for delta, and 19 for gamma/PP cells. Cell-specific genes were identified using a Wilcoxon rank sum test and

genes with FDR < 0.05 were regarded as significant after completing a "one vs. all" comparison approach (e.g., beta cells vs. all other cell types). These genes were detected in their respective islet cell type from both T2D and ND donors. **Figure 2** shows the top 10 genes for each cell type ranked by log fold-change. Cell-specific marker genes included potential cell surface proteins, which we are currently exploring for their ability to isolate pure islet cell types, particularly delta and gamma cells, as part of a new project funded by the American Diabetes Association Pathway to Stop Diabetes award program.

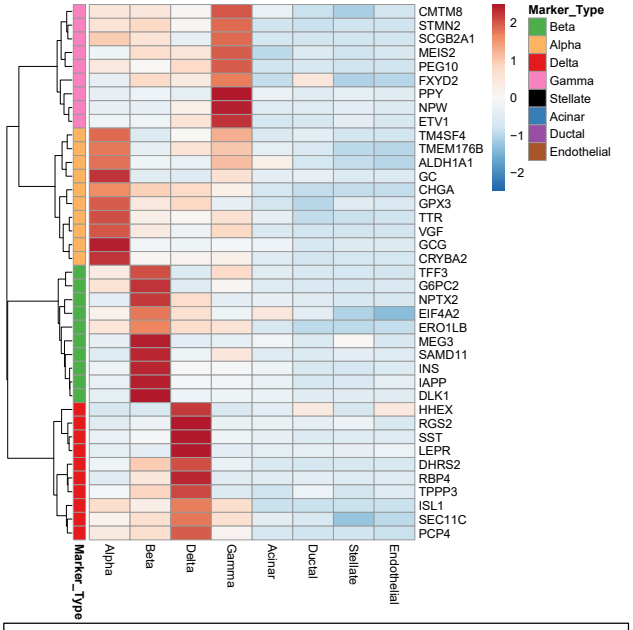


Figure 2. Heatmap of top 10 genes exhibiting specific expression in islet alpha (orange), beta (green), delta (red) and gamma (pink) endocrine cell types.

Identification of islet alpha and beta cell type subpopulations :

Separate clustering of 13,301 beta and 12,735 alpha cells, respectively, (using the top 500 most variably expressed genes in each case) uncovered both alpha and beta cell subpopulations (**Figure 3;** n=1,222 beta cells; n=1,162 alpha cells). Alpha and beta subgroups had enriched expression of 78 and 109 genes, respectively, many of which (e.g., *PDIA3*, etc.) are implicated in endoplasmic reticulum (ER) stress response pathways. 67 of these 78 and 109 genes were shared between the cell types. Gene ontology (GO) analysis revealed that these genes contribute to immune activation, ER stress, and unfolded protein responses. Similar ER-stressed subpopulations have been recently reported[2,3]. As shown in **Table 1**, these alpha and beta cell subpopulations are observed in all donors, both from T2D and ND individuals, at largely similar frequencies.

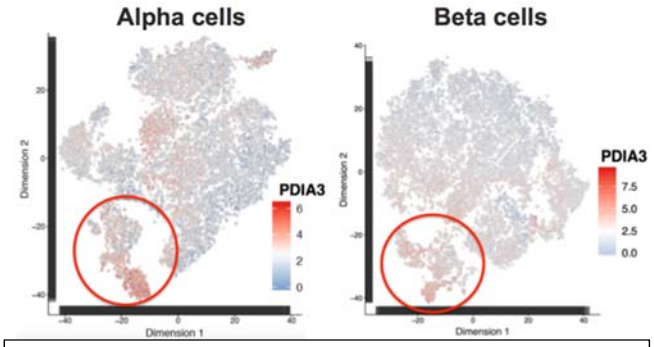


Figure 3. Identification of a common "stressed" subpopulation in both alpha and beta cells. Expression levels of one of the marker genes (*PDIA3*) in each cell type is shown. Red circle denotes the stressed cell subpopulation.

Table 1. Proportion of alpha and beta cell subpopulations per islet donor

	T2D 1	T2D 2	T2D 3	T2D 4	T2D 5	ND 1	ND 2	ND 3	ND 4	ND 5
Beta	**19.88**	3.98	5.26	2.24	1.73	5.6	8.63	4.49	7.4	**23.35**
Alpha	**30.12**	2.14	3.28	3.26	2.26	4.37	4.75	5.02	8.24	**24.19**

Reduced beta cell proportions in T2D islets compared to ND islets:

We compared the proportions of beta cells captured from T2D and ND donor islets. As shown in **Figure 4,** beta cells consistently comprised a smaller proportion of total islet endocrine cell types sampled from T2D donors compared to ND donors. Indeed, we detected a statistically significant 18% reduction in average beta cell proportions in T2D islet donors ($p = 0.018$). These data support previous studies indicating that beta cell mass is reduced in type 2 diabetic individuals[4,5].

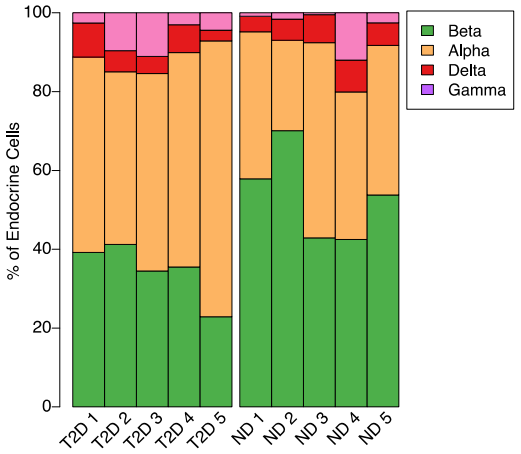


Figure 4. Relative endocrine cell composition of each donor islet profiled. Fewer beta cells were sampled in each T2D donor islet compared to ND donor islets.

Differentially expressed genes in each T2D islet cell type

We compared gene expression between each islet cell type from T2D vs. ND individuals. In total, we identified 58 genes exhibiting significant (FDR<5%) differential expression in at least one islet cell type (**Figure 5**). Approximately 1/3 of these genes were differentially expressed in two or more cell types (n= 19; e.g., *c19orf77, RPS26, MTRNR2L8*). The remaining 2/3 (n=39) were differentially expressed in a single cell type. Notable genes with increased expression in T2D beta cells include *GPX3*, an oxidative stress response gene that is upregulated in diabetic compared to non-diabetic mice[6], and the autophagy regulator *c10orf10/DEPP1*[7]. Surprisingly, we identified increased expression of a suite of genes (*PPY, S100A6,* and *NPY*) recently described in a mouse model of chronic beta cell depolarization[8]. Thus, it appears T2D beta cells in human islets may exhibit this chronic depolarization signature. *PPP1R1A,* a gene whose expression correlates with insulin secretion[9] and was also detected as down-regulated in T2D islet samples from the IMIDIA biobank[10], was among the genes detected with reduced expression in this study. We presented these data at the 20th Anniversary Servier-International Group of Insulin Secretion (IGIS) meeting, where we discussed with Jesper Gromada the opportunity to combine our data with theirs to

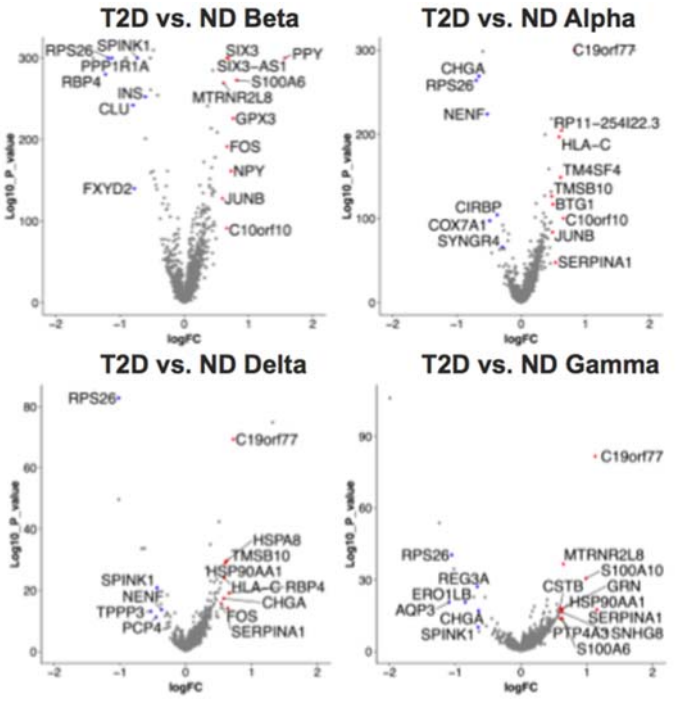


Figure 5. Differentially expressed genes in T2D beta, alpha, delta, and gamma/PP islet cells.

increase sample size and power to detect additional, robust steady state gene expression differences between T2D and ND islet cell types. We are presently setting up this collaborative effort, with the intention to co-author a manuscript on the meta-analysis results within the next 3-6 months. Moreover, we continue to expand upon this cohort in our ADA Pathway to Stop Diabetes Accelerator Award project.

Stated goals not met: All Aim 1 goals were met or exceeded.

Aim 2: Single cell epigenomes

Over the total project period, we completed open chromatin profiling of 886 islet cells from 9 of the islet donors using single cell ATAC-seq (scATAC-seq). Fluidigm C1 capture for one donor sample failed, and capture rates of single cells using this platform ranged from 42-70%. As shown in **Figure 6,** only 22 cells showed low enrichment of cut sites at promoters/transcription start sites (TSS) using QC analytics at (https://github.com/ParkerLab/ataqv). These poor-quality samples were discarded, and we proceeded with analysis of the remaining 864 scATAC-seq profiles. We applied the ChromVar pipeline (https://github.com/GreenleafLab/chromVAR) as previously described[11] to identify distinct clusters of scATAC-seq profiles corresponding to the different islet cell types. Briefly, fragment counts for each open chromatin region (OCR) and each cell were determined. For each OCR, we then determined which transcription factor binding sequence motifs mapped to each

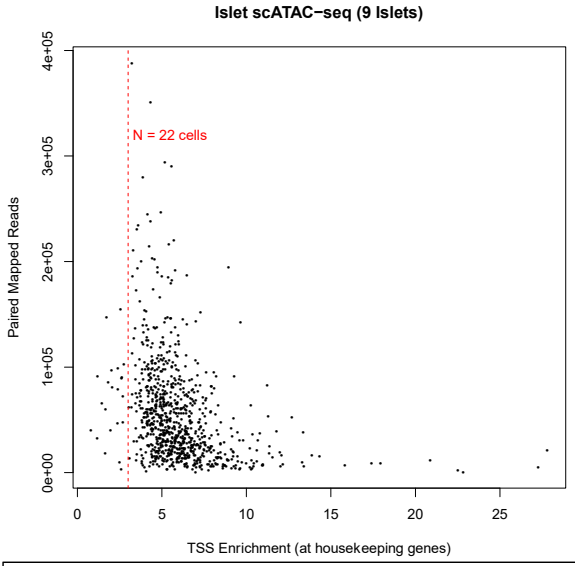


Figure 6. Quality control of scATAC-seq profiles. We used ataqv to distinguish high vs. low quality scATAC-seq profiles. Profiles from 22 cells with a TSS enrichment at housekeeping genes <3 (dashed red line) were discarded.

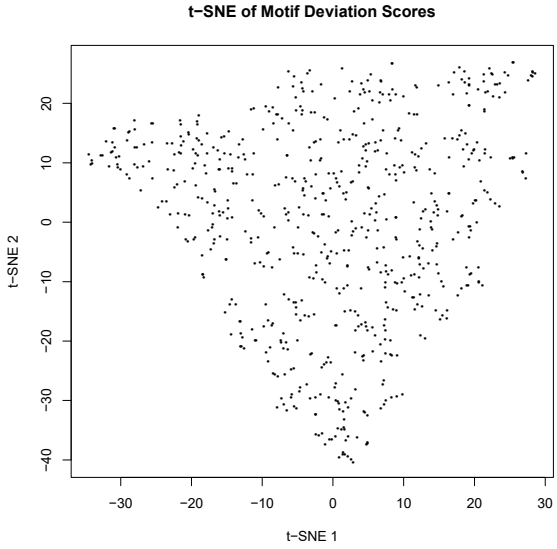


Figure 7. Islet scATAC-seq profiles do not assemble into distinct clusters. ChromVar was used to assess cell type clustering of 886 islet scATAC-seq profiles based on variability of transcription factor binding site motifs.

peak using a list of 386 transcription factor motifs from the HOMER database. Fragment counts in all peaks that mapped to a single motif were summed and used to calculate a deviation score across the different cells. Principal Components Analyses (PCA) and t-SNE was then used to cluster the single cells based on their motif deviation scores. Unfortunately, in contrast to the islet single cell transcriptomes, the single cell epigenomes did not assemble into discrete clusters

(**Figure 7**). This is despite the fact that, as we reported in the previous progress report, aggregation of single cell profiles into a "synthetic bulk islet" profile empirically resembled those of intact islets and identified approximately 113,931 open chromatin regions (OCRs) at FDR <0.0001. We suspect two issues may be confounding our analyses: dynamic range of ATAC-seq and the number of cells profiled. First, dynamic range of ATAC-seq data is much lower than transcriptomes, i.e., the number of cuts and insertions the transposase can make at a given chromosomal location in each cell's nucleus ranges from 0-2, not up to hundreds or thousands as can be observed in scRNA-seq. As we discussed in the potential pitfalls and alternative approaches section of the proposal, we sought to assign each scATAC-seq profile to a given islet cell type in a supervised manner based on cumulative transposase insertions in promoters/TSSs or nearby putative regulatory elements of the "signature genes" exhibiting cell type-specific expression according to scRNA-seq. Sadly, this also did not definitively identify distinct cell type epigenomes. It is possible that approaches to sample hundreds to thousands of single cells, such as those described by Shendure and colleagues[12] or recently developed by 10X Genomics, might produce a large enough number of scATAC-seq profiles to facilitate clustering. Moving forward, we largely favor an approach to profile enriched populations of each cell type, using antibodies against new cell surface proteins we have discovered or others recently described. Indeed, we used these scATAC-seq data and challenges as rationale in our successful ADA Pathway to Stop Diabetes Accelerator Award application and are currently employing this sorted cell approach.

Statement of Goals not met: Despite generating high quality single cell ATAC-seq profiles according to established metrics, we were unable to identify islet cell type-specific *cis*-regulatory element use and therefore unable to determine T2D differences in these profiles. This may reflect technical issues, such as the low dynamic range of the single cell ATAC-seq data, or biological truth, i.e., that each islet cell type uses largely the same *cis*-regulatory elements. We are exploring these possibilities in and independently funded project by comparing sorted cell ATAC-seq profiles to these and newly-generated single cell ATAC-seq profiles.

Bibliography:

1. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
2. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346-360.e4 (2016).
3. Xin, Y. *et al.* Pseudotime Ordering of Single Human β-Cells Reveals States of Insulin Production and Unfolded Protein Response. *Diabetes* **67**, 1783–1794 (2018).
4. Lawlor, N. *et al.* Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27**, 208–222 (2017).
5. Lawlor, N., Khetan, S., Ucar, D. & Stitzel, M. L. Genomics of Islet (Dys)function and Type 2 Diabetes. *Trends Genet.* **33**, 244–255 (2017).
6. Iwata, K., Nishinaka, T., Matsuno, K. & Yabe-Nishimura, C. Increased gene expression of glutathione peroxidase-3 in diabetic mouse heart. *Biol. Pharm. Bull.* **29**, 1042–1045 (2006).
7. Li, W. *et al.* DEPP/DEPP1/C10ORF10 regulates hepatic glucose and fat metabolism partly via ROS-induced FGF21. *FASEB J.* **32**, 5459–5469 (2018).

8. Stancill, J. S. *et al.* Chronic β-Cell Depolarization Impairs β-Cell Identity by Disrupting a Network of Ca2+-Regulated Genes. *Diabetes* **66**, 2175–2187 (2017).
9. Taneera, J. *et al.* Identification of novel genes for glucose metabolism based upon expression pattern in human islets and effect on insulin secretion and glycemia. *Hum. Mol. Genet.* **24**, 1945–1955 (2015).
10. Solimena, M. *et al.* Systems biology of the IMIDIA biobank from organ donors and pancreatectomised patients defines a novel transcriptomic signature of islets from individuals with type 2 diabetes. *Diabetologia* **61**, 641–657 (2018).
11. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
12. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).

What opportunities for training and professional development has the project provided?

Nathan Lawlor, Romy Kursawe, and I attended the Boston Ithaca Islet Club Meeting in Worcester, MA on April 28- 29, 2018. Nathan and I also recently attended the 20th Anniversary Servier-IGIS Symposium on March 28-31, 2019, where I presented an invited talk and Nathan presented the results of this single cell genomics study. Both generated great interest from other investigators, and I anticipate our discussions there will likely result in 3-4 new collaborative research efforts with leaders in the beta cell biology field from Yale University, University of Toronto, Joslin Diabetes Center in Boston, and an investigator formerly with Regeneron.

How were the results disseminated to communities of interest?

I have presented portions of these results at the Boston Ithaca Islet Club Meeting in Worcester, MA, April 28-29, 2018 and at the 78th Scientific Sessions of the American Diabetes Association in Orlando, FL, June 21-26, 2018. Nathan Lawlor recently presented the final data at the 20th Anniversary Servier-International Group on Insulin Secretion Symposium in St. Jean Cap Ferrat, France, March 28-31, 2019, where it was met with strong enthusiasm and interest. We will combine these data with those we are generating from newly funded studies, and I will present the combined results in an invited talk at the 79th Scientific Sessions of the American Diabetes Association in June 2019

4. IMPACT: Describe distinctive contributions, major accomplishments, innovations, successes, or any change in practice or behavior that has come about as a result of the project relative to:

What was the impact on the development of the principal discipline(s) of the project?

Our data and presentations are contributing to the growing appreciation and debate surrounding the existence and identities of alpha and beta cell subpopulations and to identifying differentially expressed genes in each diabetic cell type.

What was the impact on other disciplines?

Our single cell transcriptome studies and datasets led to some collaborative discussions with Dr. Zhijin Wu at Brown University. She is developing novel approaches to analyze single cell data to reveal new and unique insights into the dynamics and mechanisms of gene expression at the single cell level. The results of this collaboration have yielded co- authorship on a manuscript describing their methodology, which was accepted to *Bioinformatics*. Additionally, our single cell datasets have been used by additional groups developing tools and analysis algorithms for single cell transcriptome analysis.

What was the impact on technology transfer?

Nothing to Report

What was the impact on society beyond science and technology?

Data from this study were shared in the Community Health Discussion seminar series, two community outreach initiatives involving The Jackson Laboratory and The Children's Museum in West Hartford. In this educational presentation for the general public, I presented initial results from our single cell islet transcriptome analyses and explained how they can reshape our understanding of pathogenic events/processes contributing to diabetes and how they may shift our approach to preventing and treating diabetes. I also presented a TED-style talk entitled "Targeting Type 2 Diabetes: Precision Approaches to a Global Disease" for the JAXtaposition lecture series describing our single cell transcriptome studies of islet cell "communities".

6. PRODUCTS: List any products resulting from the project during the reporting period. If there is nothing to report under a particular item, state "Nothing to Report."

- **Publications, conference papers, and presentations**
  Report only the major publication(s) resulting from the work under this award.

  Journal publications.

  Khetan S, Kursawe R, Youn A, Lawlor N, Marquez E, Ucar D, and Stitzel ML. Chromatin accessibility profiling uncovers genetic- and T2D disease state-associated changes in cis-regulatory element use in human islets. *Diabetes.* 2018. Sep 4. PMID: 30181159 Acknowledgment of federal support: Yes

  Wu Z, Zhang Y, Stitzel ML, and Wu H. Two-phase differential expression analysis for single cell RNA-seq. *Bioinformatics*. 2018 Apr 24. PMID: 29688282. Acknowledgment of federal support: Yes

  Lawlor N, Khetan S, Ucar D, and Stitzel ML. Genomics of Islet (Dys)function and Type 2 Diabetes. *Trends Genet*. 2017 Apr;33(4):244-255. PMID28245910. Acknowledgment of federal support: Yes

  Lawlor N, George J, Bolisetty M, Kursawe R, Sun L, V S, Kycia I, Robson P, Stitzel ML. Single cell transcriptomes identify human islet cell signatures and reveal cell-type-

specific expression changes in type 2 diabetes. 2017. *Genome Res.* Feb; 27(2):208-222. PMID: 27864352. Acknowledgment of federal support: Yes

Please see Appendix

Books or other non-periodical, one-time publications.

Nothing to Report

Other publications, conference papers, and presentations.

Boston Ithaca Islet Club, Worcester, MA, April 28-29, 2018 (national, talk)
78th Scientific Sessions of the American Diabetes Association, Orlando FL June 21-26, 2018 (international, talk)

20th Servier-IGIS Symposium, St Jean Cap-Ferrat, France March 28-31, 2019 (international, poster; *please see Appendix)*

- **Website(s) or other Internet site(s)** Nothing to Report
- **Technologies or techniques** Nothing to Report
- **Inventions, patent applications, and/or licenses** Nothing to Report

• **Other Products** Nothing to Report

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS What individuals have worked on the project?

Name: Michael L. Stitzel
Project Role: Initiating PI
Researcher Identifier: http://orcid.org/0000-0001-5630-559X
Nearest person month worked this period: 1 CM
Nearest person month worked entire project period: 3 CM
Contribution to Project: Dr. Stitzel has managed the project, directing both the experiments and analyses completed by Drs. Kursawe and Mr. Lawlor

Name: Romy Kursawe
Project Role: Research Assistant IV/Lab manager
Researcher Identifier:
Nearest person month worked this period: 1 CM
Nearest person month worked entire project period: 8 CM
Contribution to Project: Dr. Kursawe has completed all of the experiments for the project, including processing islets, preparing single cell suspensions, preparing RNA, transposing nuclei, and preparing ATAC-seq libraries

Name: Joshy George
Project Role: Computational Scientist

Researcher Identifier:
Nearest person month worked this period: 0 CM
Nearest person month worked entire project period: 1 CM
Contribution to Project: Dr. George established components of the single cell transcriptome profiling platform and trained Mr. Lawlor in these and other analytical pipelines for the project.

Name: Nathan Lawlor
Project Role: Data Analyst
Research Identifier: http://orcid.org/0000-0003-3263-6057
Nearest person month worked this period: 1 CM* [*0.3 CM was rounded up to 1 to indicate Mr. Lawlor actively contributed to this project in the final reporting period]
Nearest person month worked entire project period: 3 CM
Contribution to Project: Mr. Lawlor has implemented and runs the published analysis pipelines for single cell ATAC-seq and single cell RNA-seq and internal pipelines established by Dr. George. In addition, he has completed differential analyses between cell (sub)populations and has participated in analyses and content for peer-reviewed publications and presentation of work supported by this funding.

What other organizations were involved as partners?

Nothing to Report

8. SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS:

QUAD CHARTS:

9. APPENDICES:

a. Khetan S, Kursawe R, Youn A, Lawlor N, Marquez E, Ucar D, and Stitzel ML. Chromatin accessibility profiling uncovers genetic- and T2D disease state-associated changes in cis-regulatory element use in human islets. *Diabetes.* 2018. Sep 4. PMID: 30181159 Acknowledgment of federal support: Yes

b. Wu Z, Zhang Y, Stitzel ML, and Wu H. Two-phase differential expression analysis for single cell RNA-seq. *Bioinformatics*. 2018 Apr 24. PMID: 29688282. Acknowledgment of federal support: Yes

c. Lawlor N, Khetan S, Ucar D, and Stitzel ML. Genomics of Islet (Dys)function and Type 2 Diabetes. *Trends Genet*. 2017 Apr;33(4):244-255. PMID28245910. Acknowledgment of federal support: Yes

d. Lawlor N, George J, Bolisetty M, Kursawe R, Sun L, V S, Kycia I, Robson P, Stitzel ML.

Single cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. 2017. *Genome Res.* Feb; 27(2):208-222. PMID: 27864352. Acknowledgment of federal support: Yes

e. Abstract for 20th Anniversary Servier-IGIS Symposium presentation

Type 2 Diabetes–Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets

Shubham Khetan,[1,2] Romy Kursawe,[1] Ahrim Youn,[1] Nathan Lawlor,[1] Alexandria Jillette,[1] Eladio J. Marquez,[1] Duygu Ucar,[1,2,3] and Michael L. Stitzel[1,2,3]

Diabetes 2018;67:2466–2477 | https://doi.org/10.2337/db18-0393

Type 2 diabetes (T2D) is a complex disorder in which both genetic and environmental risk factors contribute to islet dysfunction and failure. Genome-wide association studies (GWAS) have linked single nucleotide polymorphisms (SNPs), most of which are noncoding, in >200 loci to islet dysfunction and T2D. Identification of the putative causal variants and their target genes and whether they lead to gain or loss of function remains challenging. Here, we profiled chromatin accessibility in pancreatic islet samples from 19 genotyped individuals and identified 2,949 SNPs associated with in vivo *cis*-regulatory element use (i.e., chromatin accessibility quantitative trait loci [caQTL]). Among the caQTLs tested (*n* = 13) using luciferase reporter assays in MIN6 β-cells, more than half exhibited effects on enhancer activity that were consistent with in vivo chromatin accessibility changes. Importantly, islet caQTL analysis nominated putative causal SNPs in 13 T2D-associated GWAS loci, linking 7 and 6 T2D risk alleles, respectively, to gain or loss of in vivo chromatin accessibility. By investigating the effect of genetic variants on chromatin accessibility in islets, this study is an important step forward in translating T2D-associated GWAS SNP into functional molecular consequences.

Type 2 diabetes (T2D) is a complex disease resulting from the combined effects of an individual's genetic predisposition and environmental exposures (1,2). It ultimately manifests when islets cannot secrete sufficient insulin to compensate for insulin resistance in peripheral tissues (3). Genome-wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) in >200 loci that confer genetic susceptibility to T2D and/or alter quantitative measures of islet (dys)function (4,5). These SNPs are predominantly noncoding (~90%) and enriched within islet-specific *cis*-regulatory elements (*cis*-REs) (6–9), implicating perturbed islet transcription in T2D etiology (2). However, identifying the causal variants in each T2D-associated GWAS locus, their molecular effects, and the genes and pathways they affect remains critical to translate genetic associations into mechanistic understanding and treatments.

Quantitative trait locus (QTL) analyses have linked common genetic variants to in vivo gene expression changes (eQTL) for multiple cell types (10), including islets (8,11,12). However, eQTLs cannot pinpoint the causal variants among the multiple SNPs in linkage disequilibrium (LD) with each other. QTL approaches have recently been applied in cell lines to link genetic variation to epigenomic changes, such as DNaseI sensitivity (13), chromatin accessibility (caQTLs) (14–16), and histone modification levels (17). However, little is known about how genetic variation affects epigenomes of clinically relevant primary tissues such as islets.

In this study, we used the Assay for Transposase-Accessible Chromatin-sequencing (ATAC-seq) (18) to profile genome-wide chromatin accessibility in islets from 19 individuals (14 without diabetes [ND] and 5 with T2D). Using caQTL analysis, we identified genetic variants altering in vivo chromatin accessibility in islets and exhibiting concordant effects on in vitro luciferase reporter activity. Finally, we identified putative causal variants altering islet chromatin accessibility in 13 distinct T2D-associated GWAS loci. Together, this study provides a road map for

[1]The Jackson Laboratory for Genomic Medicine, Farmington, CT
[2]Department of Genetics and Genome Sciences, University of Connecticut, Farmington, CT
[3]Institute of Systems Genomics, University of Connecticut, Farmington, CT
Corresponding authors: Michael L. Stitzel, michael.stitzel@jax.org, and Duygu Ucar, duygu.ucar@jax.org.

Received 5 April 2018 and accepted 22 August 2018.
This article contains Supplementary Data online at http://diabetes.diabetesjournals.org/lookup/suppl/doi:10.2337/db18-0393/-/DC1.

© 2018 by the American Diabetes Association. Readers may use this article as long as the work is properly cited, the use is educational and not for profit, and the work is not altered. More information is available at http://www.diabetesjournals.org/content/license.

translating T2D-associated GWAS SNPs into functional molecular effects.

RESEARCH DESIGN AND METHODS

Study Subjects and Islet Culture

Fresh human cadaveric pancreatic islets were procured from Prodo Laboratories or the Integrated Islet Distribution Program (Supplementary Table 1) and processed according to institutional review board–approved protocols. Upon receipt, cells were transferred into PIM(S) media supplemented with PIM(ABS) and PIM(G) (Prodo Laboratories) and incubated overnight in a T-150 nontissue culture–treated flask (VWR) at 37°C and 5% CO_2 overnight. The following day, nuclei and total RNA were isolated for ATAC-seq and RNA-seq library preparation and analysis (8). Genomic DNA was isolated from islet explant cultures using Qiagen DNeasy Blood & Tissue Kit as previously described (8).

ATAC-seq Profiling

Islet ATAC-seq libraries were prepared as previously described (8) from 22 donors. Per donor, three replicates, each consisting of 50–100 islet equivalents (50,000–100,000 cells), were transposed. Libraries were barcoded, pooled into three-donor batches (corresponding to nine barcoded transposition reactions), and sequenced using 2 × 75 bp Illumina NextSeq 500 to an average depth of 62.6 (± 18.6) million paired-end reads per donor (Supplementary Table 2). Low quality portions of reads were trimmed using Trimmomatic (19) and aligned to the hg19 human genome assembly using Burrows-Wheeler Aligner-MEM (20). For each replicate, reads were shifted as previously described and duplicate reads were removed (21,22). Technical replicates were merged using SAMtools after confirming high correlation between them. Open chromatin regions (OCRs) were called for each islet sample using MACS2 (23) (with parameters -callpeak --nomodel -f BAMPE) at a false discovery rate (FDR) of 1%. Islet ATAC-seq samples with less than 30,000 OCRs were excluded from further analyses, yielding data for 19 individuals. OCRs on sex chromosomes and those overlapping low mapability regions (blacklisted regions available from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/) were excluded, resulting in 154,437 autosomal OCRs detected in at least one individual using R package DiffBind (24). deepTools was used to generate bedgraph files for UCSC Genome Browser sessions (with parameters --normalizeUsingRPKM --centerReads --scaleFactor = 1 -bs = 25).

OCR Chromatin State Annotations

Previously described chromatin states for islets (8), ENCODE, and National Institutes of Health Roadmap Epigenomics (25) cells/tissues were used to annotate islet OCRs and visualized using ggplot2 (26). OCRs overlapping ≥2 different chromatin states were assigned a single state using the following hierarchy: Active transcription start site (TSS) > Bivalent TSS > Weak TSS > Flanking TSS > Active Enhancer-1 > Active Enhancer-2 > Weak Enhancer > Genic Enhancer > Strong Transcription > Weak Transcription > Repressed Polycomb > Weak Repressed Polycomb > Quiescent. Previously described stretch enhancer (SE) regions (6,8) were overlapped with islet OCRs and tested for enrichment using the Fisher exact test. For each tissue-specific test, the background set comprised SEs from all other tissues (n = 30).

Genotyping, Imputation, and caQTL Analysis

Each islet donor was genotyped using Illumina Infinium Omni2.5Exome (n = 11) or Omni5Exome (n = 8) BeadChip arrays (Supplementary Table 1). We mapped Illumina array probe sequences to the hg19 genome assembly using Burrows-Wheeler Aligner. SNPs with ambiguous probe alignments, 1000 Genomes (1000G) phase 1 variants with minor allele frequency of ≥1% within 7 bp of the 3′ end of probes, or call rates <95% were excluded. All alleles were oriented relative to the reference. Genotype calls were merged using vcftools/0.1.12a suite (vcf-merge command). After removing SNPs with missing data (--max-missing 1), ~2.4 million SNPs were used for imputation (1000G phase 3 version 5 [27]) and phasing (Eagle version 2.3 [28]) using the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html) (29).

VerifyBamID (30) was used to match ATAC-seq bam files to individuals' genotypes to eliminate the possibility of a sample swap. Islet OCRs overlapping only monomorphic SNPs were removed from caQTL analyses, yielding 84,499 OCRs. Allele-specific counts were obtained for 195,207 SNPs within these OCRs, and caQTLs were detected using RASQUAL (15). To minimize confounding factors such as batch effects, we adopted the strategy described in Kumasaka et al. (15) and used the first five principal components as covariates in the RASQUAL model. Significant caQTLs were identified using a two-stage multiple hypothesis testing correction (15): *1*) correcting for the multiple SNPs tested within each OCR using Bonferroni correction, and *2*) then correcting for the number of OCRs tested genome wide by controlling FDR at 10% using RASQUAL's permutation test ("--random-permutation") 50 times.

To visualize chromatin accessibility patterns at caQTLs, first we calculated the number of ATAC-seq reads (normalized with respect to library size) spanning each base pair for all 19 samples using BEDTools ("genomecov"). Next, islet donors were grouped based on their genotypes for each displayed caQTL; average read counts were calculated for each genotype group and plotted using the "polygon" function in R.

Differential OCR Analyses

Differential chromatin accessibility analyses were conducted between islet ATAC-seq profiles of five T2D and five ND individuals with the most comparable demographics (Supplementary Table 3). To identify statistically

significant T2D disease state–associated chromatin accessibility changes, only OCRs meeting the following criteria were used for differential analyses (n = 52,387): *1*) present in ≥3 islet donors and *2*) present in ≥1 T2D and ≥2 ND (or ≥1 ND and ≥2 T2D) individuals. Race, sex, and significant surrogate variables (n = 2) from surrogate variable analysis (SVA) (31) were used as covariates to minimize confounding factors. edgeR (32) R package was used to identify differentially accessible OCRs.

GWAS SNP Enrichment in Islet caQTLs

The NHGRI-EBI GWAS Catalog of GWAS index SNPs for 184 diseases/traits was retrieved on 19 January 2017 (https://www.ebi.ac.uk/gwas/) and LD-pruned using PLINK (33) version 1.9 (parameters --maf 0.05 --clump --clump-p1 0.0001 --clump-p2 0.01 --clump-r2 0.2 --clump-kb 1000) to avoid testing enrichment for multiple SNPs representing the same genetic association signal/locus per trait. For index SNPs exhibiting pairwise correlation r^2 >0.2, only the SNP with the more significant P value was retained. We used GREGOR (34) on this LD-pruned list to determine whether GWAS index or linked SNPs (r^2 >0.8, LD window size = 1 Mb, minimum neighbor number = 500) were enriched in islet caQTLs or differentially accessible OCRs.

Transcription Factor Motif Enrichments

Homer (35) findMotifsGenome.pl script (parameters: hg19, –size given) was used to identify transcription factor (TF) motifs enriched in islet OCRs. We compared motifs in OCRs that are accessible only in islets (n = 40,271 islet-specific OCRs) to motifs in OCRs that are also accessible in adipose, $CD4^+$ T, GM12878, or peripheral blood mononuclear cells (PBMCs) (n = 41,639 shared/background OCRs) (Fig. 1*C*). Motifs enriched in caQTL-containing OCRs (Fig. 2*D*) were identified by comparing caQTL OCRs (n = 2,949) to all islet OCRs (n = 154,437). TFs were clustered based on the similarity of their position weight matrices (PWMs) using Kullback-Leibler divergence method implemented in TFBSTools (36). Motif enrichments for differential OCRs (n = 1,515) were calculated against all OCRs used in differential analysis (n = 52,387).

RNA-seq Profiling

Total RNA was isolated from each islet sample using TRIzol (8). Stranded RNA-seq libraries were prepared from total RNA using the TruSeq Stranded mRNA kit (Illumina) for the 19 individuals with high-quality ATAC-seq data; External RNA Controls Consortium (ERCC) Mix 1 or Mix 2 spike-ins were randomly added to each sample (Thermo Fisher, catalog #4456740) (Supplementary Table 4). RNA-seq from 10 islet samples used in differential analyses were sequenced together on an Illumina NextSeq 500 to minimize batch effects, whereas the remaining nine samples were sequenced on Illumina HiSeq 2500, each to an average sequencing depth of 87.2 (±27.8) million paired-end reads (Supplementary Table 4). Paired-end RNA-seq reads were trimmed to remove low-quality base calls using Trimmomatic (19). Bowtie2 (37) and RSEM (38) (rsem-calculate-expression) were used to determine fragments per kilobase of transcript per million mapped reads (FPKM) and expected read counts for all Ensembl hg19 Release 70 transcripts.

Differential Gene Expression Analyses

RNA-seq data from 10 islet samples (Supplementary Table 3) were used for differential expression analysis. Expected read counts for autosomal genes with FPKM >5 in ≥3 RNA-seq samples (n = 10,116) were used in differential analyses based on edgeR (32) models (FDR 10%). Race, sex, ERCC spike-in, and significant surrogate variables (n = 1) from SVA were used as covariates to minimize the impact of confounding factors on T2D disease state–associated gene expression changes.

eQTL Analysis

RSEM expected read counts (38) for 9,656 expressed genes (median FPKM >5) were used to identify islet eQTLs from 19 donors using RASQUAL (15). Only SNPs within the gene body or within 50 kb flanking the gene body were tested. To minimize potential batch effects, we adopted the strategy described in Kumasaka et al. (15) and used the first four principal components, in addition to age, sex, race, T2D status, and sequencing date as covariates in the RASQUAL model. A two-stage multiple hypothesis testing correction (15) was used to determine significant eQTLs similar to caQTLs, where only 10 permutation tests were used in step two.

Islet caQTL-eQTL Overlaps

Quantile-quantile (QQ) plots for caQTL P values were generated against the expectation of a uniform P value distribution between 0 and 1. The QQ plot was generated for islet eQTL SNPs from 112 individuals (8) and caQTL SNPs from 19 individuals by conditioning on lead caQTL SNPs that were either statistically significant at FDR 10% or background sets of randomly selected nonsignificant ones. Random sets of nonsignificant SNPs (n = 2,545) were generated 10 times to eliminate sampling bias; a representative result from one random set is shown in Fig. 2*F* and Supplementary Fig. 2*G*.

Gateway Cloning of Selected Islet caQTL Sequences and Alleles

Islet genomic DNA from individuals homozygous for the reference or alternate allele was used as templates to PCR amplify sequences containing each allele for 13 islet caQTLs (Supplementary Table 5). The corresponding 26 PCR amplicons were cloned into the pDONR201 vector using BP Clonase (Invitrogen). Sequences were validated by Sanger sequencing. Each islet caQTL sequence was transferred from pDONR201 into the Gateway-modified pGL4.23F plasmid (39) with LR Clonase.

Luciferase Reporter Assays

MIN6 cells were seeded in 96-well plates at a density of 60,000 cells per well 24 h prior to transfection as previously

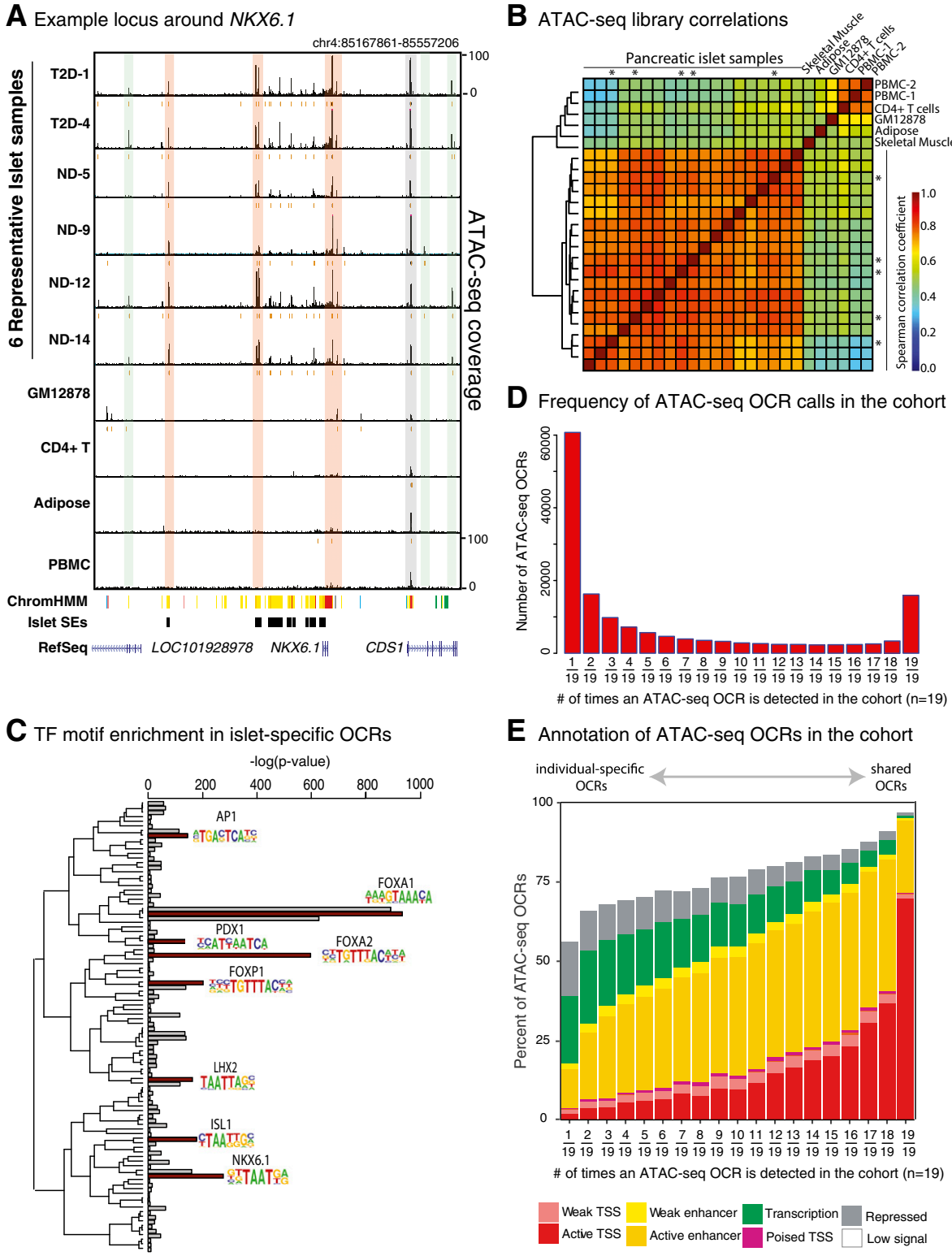


Figure 1—Human pancreatic islet chromatin accessibility profiles from 19 donors. *A*: UCSC Genome Browser view of ATAC-seq profiles at the *NKX6.1* locus from six representative islet samples (ND and T2D individuals), the lymphoblastoid cell line GM12878, CD4$^+$ T cells, adipose tissue, and PBMCs (data from two individuals). Orange and gray rectangles denote islet-specific or ubiquitously accessible regions, respectively, among cell types/tissues profiled. Green rectangles highlight regions showing variable accessibility between islet samples in the cohort. All chromatin accessibility profiles are normalized to their respective library size and have the same scale. Islet ChromHMM chromatin state annotations of these accessible sites (color code key found in Fig. 1*E*), islet SEs, and RefSeq gene models are also shown. *B*: Heatmap of Spearman correlation coefficients between ATAC-seq profiles from 19 islet samples and other cell types. Asterisks mark islet ATAC-seq samples from T2D donors ($n = 5$). *C*: TF motif enrichments in OCRs unique to islet samples ($n = 40{,}271$) compared with islet OCRs that are also detected in skeletal muscle, adipose tissue, GM12878, CD4$^+$ T cells, or PBMCs ($n = 41{,}639$). TFs are clustered with respect to the similarity of

described (39). Gateway-modified Firefly (0.072 pmol) (pGL4.23, Promega) plasmid containing each islet caQTL sequence (Supplementary Table 5) and 2 ng Renilla (pRL-TK, Promega) plasmid were cotransfected in triplicate using Lipofectamine 2000 Transfection reagent (Life Technologies). The Dual Luciferase Reporter Assay system (Promega) was used to determine Firefly and Renilla luciferase activity in each sample. Cells were lysed with 1× passive lysis buffer 38–40 h after transfection. Luminescence was measured on a Synergy 2 Microplate Reader (BioTek). Firefly values were normalized to Renilla to control for differences in cell number or transfection efficiency.

RESULTS

Human Pancreatic Islet Chromatin Accessibility Maps

To determine the genome-wide location of *cis*-REs in human islets, we generated high-quality ATAC-seq profiles from 19 islet donors (Supplementary Fig. 1*A*, Supplementary Tables 1 and 2). Investigating chromatin accessibility near the *NKX6.1* locus, a well-characterized β-cell–specific TF, revealed both OCRs unique to islet samples (Fig. 1*A*, orange and green rectangles) and OCRs shared with other cell types (22,40) (Fig. 1*A*, gray rectangle). Overall, chromatin accessibility profiles from 19 islets were highly correlated to each other and to those from sorted islet α- and β-cells (Fig. 1*B* and Supplementary Fig. 1*B*) (41). Notably, ATAC-seq profiles from T2D donors (n = 5; Fig. 1*B*, asterisks) did not cluster separately from ND donors, suggesting that the T2D disease state does not lead to global remodeling of human islet chromatin accessibility.

Collectively, we identified 154,437 islet OCRs accessible in at least one of the 19 individuals (Supplementary Table 6). Comparison with reported chromatin state annotations in human islets (6,8) assigned 12.9% and 23.14% of these OCRs as putative promoters and enhancers, respectively (Supplementary Fig. 1*C*). Putative promoter OCRs were shared with several of 30 tissues profiled by the National Institutes of Health Roadmap Epigenomics project (25). Putative enhancer OCRs were more specific to islets, consistent with previous observations of cell type specificity of enhancers (42). To further assess the islet specificity of detected OCRs, we compared them to SEs, which are long (>3 kb) contiguous enhancer chromatin states that govern cell-specific functions and often harbor disease-associated SNPs relevant to the cognate cell type (6). The majority (90%) of islet SEs overlapped islet OCRs (Supplementary Fig. 1*D*), significantly greater than overlaps observed between islet OCRs and SEs in other tissues (Fisher exact test $P < 2.2 \times 10^{-16}$). As anticipated, DNA sequence binding motifs of islet-specific TFs, such as PDX1 and NKX6.1, were significantly enriched in OCRs that are accessible in islet samples and not in GM12878, PBMCs, $CD4^+$ T cells, skeletal muscle, or adipose tissues (Fig. 1*C*). Together, these observations indicate that high-quality chromatin accessibility maps of islets from multiple individuals reveal *cis*-REs (OCRs) important for islet-specific gene regulation.

Only 10% (n = 15,917) of islet OCRs were detected in all 19 individuals (Fig. 1*D*), which were overwhelmingly annotated as promoters (Fig. 1*E*, red bars). In contrast, 39.3% (n = 60,713) of OCRs were detected in only 1 out of 19 individuals (Fig. 1*D*) and were found predominantly (45%) in quiescent/low signal chromatin states (Fig. 1*E*, white bars). Though we cannot eliminate the possibility of false positives in OCR detection, these might also represent individual-specific enhancers missed in reference islet chromatin states, as references were based on data from 2–3 individuals. OCRs detected in 2–18 individuals (Fig. 1*D*) were mostly enhancers (Fig. 1*E*, orange/yellow bars), suggesting that genetic differences (i.e., SNPs) between individuals may alter the chromatin accessibility, and therefore the activity, of human islet enhancers.

Genetics of Chromatin Accessibility in Human Islets

To identify genetic variants (SNPs and small insertions/deletions) that alter chromatin accessibility of islet OCRs in which they reside (Fig. 2*A*), we genotyped islet samples and conducted caQTL analysis using RASQUAL (15), a method that can discover QTLs from small sample sizes. Using RASQUAL, we identified 2,949 SNPs associated with increased or decreased chromatin accessibility at FDR 10% (Supplementary Fig. 2*A*, Supplementary Table 7) from 19 islet samples. For example, the rs488797 C allele was associated with reduced OCR accessibility in an islet SE in the intron of *CELF4* (Fig. 2*B*), a gene selectively expressed in islets (8,40). CC homozygous islet donors exhibited dramatically lower accessibility than CT or TT genotypes (Fig. 2*B*, compare blue CC, pink CT, and green TT profiles). Moreover, ATAC-seq sequences overlapping rs488797 in CT heterozygous samples almost exclusively contained the T allele (Fig. 2*B*, inset), reinforcing genetics as a strong determinant of chromatin accessibility at this OCR.

The rs488797 C allele is predicted to disrupt FOXA2 binding (Fig. 2*B*, compare top sequence between brackets to FOXA2 binding motif below). To test this, we analyzed FOXA2 ChIP-seq data from two islet donors (HI101 and HI32) (7) (Fig. 2*C*). We leveraged FOXA2 ChIP-seq reads and genetic LD to infer genotypes of these individuals in this region. As the caQTL SNP rs488797 alters in vivo islet chromatin accessibility, we imputed its genotype using a linked SNP rs648005 (T/C) (r^2 = 0.99 with rs488797).

their PWMs. Motif logos are shown for TFs highlighted with maroon bars. *D*: Histogram of the number of times an ATAC-seq OCR is detected in the cohort, ranging from individual-specific OCRs (n = 1) to shared OCRs (n = 19). *E*: Stacked bar plot showing islet ChromHMM chromatin state annotations of OCRs, binned according to the number of times an ATAC-seq OCR is detected in the cohort. Note that common OCRs predominantly overlap promoter states, whereas individual-specific OCRs overlap mostly unannotated (i.e., quiescent/low signal) regions.

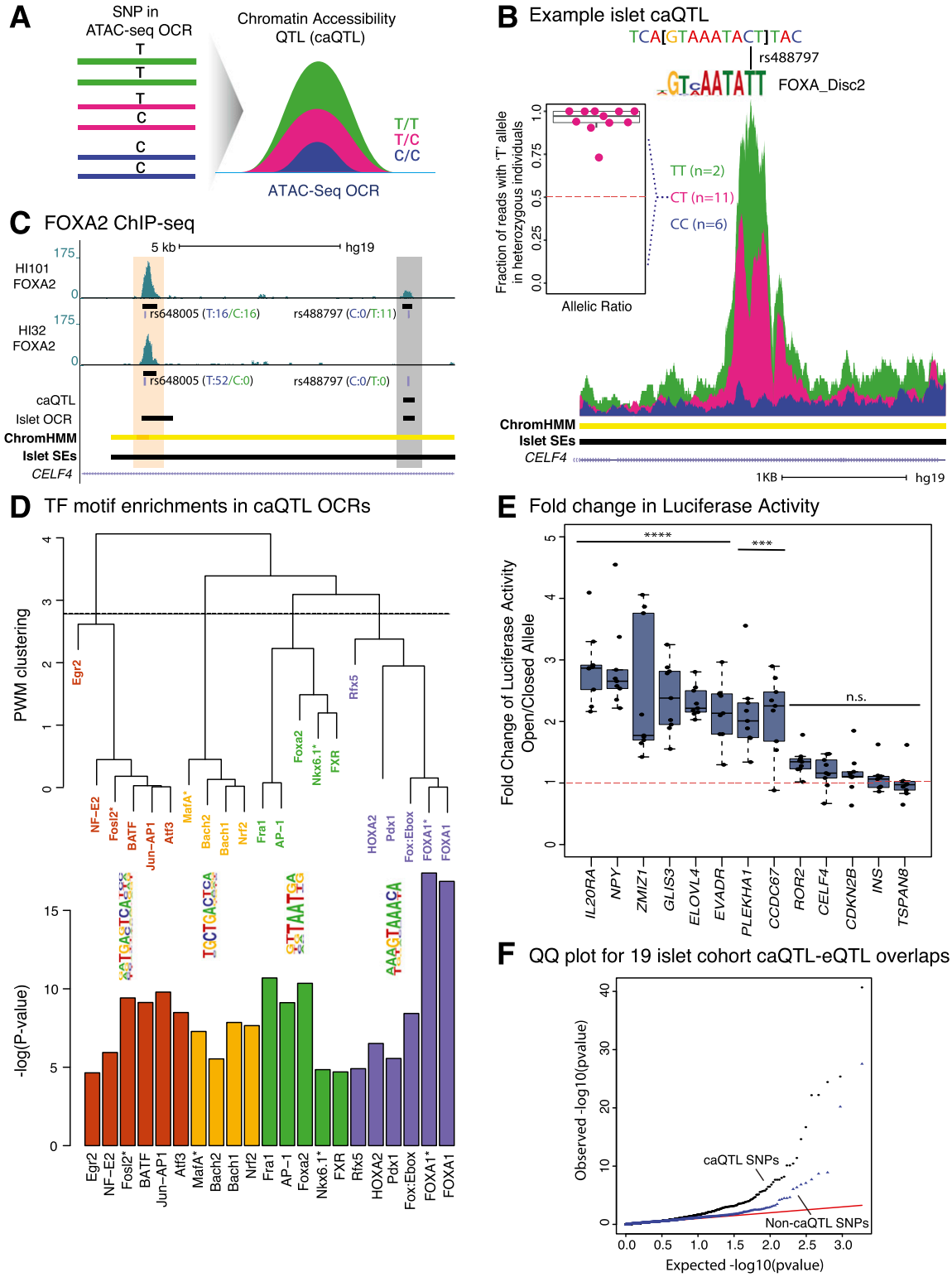


Figure 2—caQTL analysis identifies genetic variants affecting human islet *cis*-RE use. *A*: Schematic depicting genotype effects on chromatin accessibility detected by caQTL analyses. *B*: Average ATAC-seq read counts of islet samples with CC (blue), CT (pink), or TT (green) genotypes at rs488797, an islet caQTL overlapping an islet SE within an intron of *CELF4*. The fraction of ATAC-seq reads overlapping rs488797 that contain the opening T allele in CT heterozygous islet samples ($n = 11$) is plotted in the inset. The rs488797 C allele is predicted to disrupt a FOXA2 binding motif (logo shown below the reference genome sequence), which is consistent with reduced chromatin accessibility observed for the C allele. Average read counts for islet samples with the TT genotype is 50.5 at the OCR summit. Islet samples with CT or CC genotype exhibited maximum average read counts of 32.36 and 6.5, respectively. Islet ChromHMM chromatin states, islet SEs, and RefSeq gene models are displayed as in Fig. 1. hg19 chromosome coordinates: chr18:34969218–34972156. *C*: UCSC Genome Browser view of FOXA2 ChIP-seq profiles (7) at the *CELF4* locus for islets from two individuals (HI101, HI32). FOXA2 ChIP-seq read pileups are shown for the islet caQTL SNP (rs488797, gray rectangle) and a nearby SNP (rs648005, orange rectangle) in high LD ($r^2 = 0.99$), suggesting that the

rs648005 overlaps a distinct OCR and a FOXA2 binding site 8,178 nucleotides away but is neither an islet caQTL nor is predicted to disrupt a FOXA2 motif. In HI101, FOXA2 ChIP-seq reads overlapping rs648005 contained both C and T alleles (Fig. 2*C*, top), indicating that HI101 is heterozygous at rs648005 and, by extension, at rs488797 with high probability. However, FOXA2 ChIP-seq reads overlapping the caQTL SNP rs488797 exclusively contained the T allele, consistent with the islet caQTL analysis and supporting FOXA2 motif disruption predictions. In HI32, FOXA2 ChIP-seq reads at rs648005 contained only the T allele, suggesting that this individual is a TT homozygote at rs648005, and therefore a CC homozygote at rs488797 with high probability. Notably, no FOXA2 binding is observed at rs488797 for HI32, providing further support that the C allele disrupts FOXA2 binding. Supplementary Table 8 provides predicted motif disruptions from HaploReg (43) for all islet caQTL including rs488797.

Islet caQTLs were uniformly distributed across the autosomal chromosomes (Supplementary Fig. 2*B*), and the majority (>98%) were located within 200 kb flanking the TSS of the nearest islet-expressed gene (Supplementary Fig. 2*C*). Twelve percent of islet caQTLs were in promoters, whereas 30% overlapped enhancers (Supplementary Fig. 2*D*). Islet caQTLs were exclusively enriched in islet SEs compared with SEs in other tissues (Supplementary Fig. 2*E*). Finally, sequence motifs for islet-specific TFs, such as FOXA2, NKX6.1, and PDX1, were enriched in caQTL OCRs (Fig. 2*D*). To validate this, we overlapped caQTL OCRs with ChIP-seq data from human islets for islet-specific TFs and ubiquitous CTCF (7). We found that FOXA2, NKX6.1, and PDX1 binding (i.e., ChIP-seq peaks) were enriched at caQTLs (Supplementary Fig. 2*F*), in contrast to CTCF, whose binding sites were not enriched at islet caQTLs. Together, these results suggest that motif enrichment analyses likely reflect actual binding of these TFs at caQTL OCRs. Surprisingly, sequence motifs of oxidative stress-responsive TFs, such as BACH1, BACH2, and NRF2, were also enriched in caQTL OCRs, suggesting that some caQTLs may modulate stress/stimulus-responsive *cis*-RE activity.

To determine if caQTL alleles altering in vivo chromatin accessibility elicit concordant effects on in vitro enhancer activity, we selected a subset of caQTLs (n = 13) that were nearby genes exhibiting islet-specific expression (8) (e.g., Fig. 2*B*). We cloned DNA sequences containing each islet caQTL allele (Supplementary Table 5) and measured their enhancer activity using luciferase reporter assays in MIN6 mouse β-cells. We observed allelic effects on luciferase activity for 8 of the 13 caQTLs tested (Fig. 2*E*). Importantly, for all 8 caQTLs, the allele that increased in vivo chromatin accessibility also increased in vitro enhancer activity (Fig. 2*E*). Finally, we studied whether caQTL variants were also associated with variability in islet gene expression levels using islet eQTL data from this cohort (n = 19). As shown in Fig. 2*F*, caQTL variants exhibited more significant allelic effects on islet gene expression than randomly selected variants in OCRs (noncaQTLs). We observed the same trend comparing these caQTLs to eQTLs detected in a larger independent cohort (n = 112) (Supplementary Fig. 2*G*) (8). Importantly, for 84% of caQTL-eQTL pairs in our cohort (37/44) (Supplementary Fig. 2*H*), we observed a concordant direction of effect (Pearson r = 0.691), i.e., higher chromatin accessibility is associated with increased gene expression and vice versa (Supplementary Fig. 2*H*; Q1 and Q3), linking chromatin accessibility effects of these variants to downstream changes in islet gene expression.

Chromatin Accessibility Changes in T2D Versus ND Islet Samples

To assess potential environmental effects of T2D disease state on the islet epigenome, we compared chromatin accessibility between five T2D donors and five ND donors with comparable demographics, e.g., age, race, sex) (Supplementary Fig. 3*A*, Supplementary Table 3). After completing SVA (31) to remove unwanted variation in the data, e.g., batch effect, sex, postmortem interval, drug treatment (Supplementary Fig. 3*B*), we identified 1,515 of 52,387 (2.8%) OCRs that were differentially accessible between T2D and ND islet samples at FDR 10% (see RESEARCH DESIGN AND METHODS, Fig. 3*A*, and Supplementary Fig. 3*C*; 609 at FDR 5%, and 79 at FDR 1%), where 714 have increased (opening OCRs) and 801 have decreased (closing OCRs) accessibility in T2D compared with ND samples (Fig. 3*A*, Supplementary Table 9). There was a remarkable difference in the chromatin state annotation of opening and closing OCRs. Closing OCRs, e.g., the one highlighted near

rs648005/rs488797 genotypes are TC/CT for HI101 and TT/CC for HI32. No FOXA2 binding is observed at rs488797 in HI32, whose CC genotype is predicted to disrupt the FOXA2 binding motif on both parental chromosomes. In HI101, who is heterozygous at rs488797, all FOXA2 ChIP-seq reads contained the T allele, supporting predictions that the C allele disrupts FOXA2 binding. *D*: TF motifs significantly enriched in islet caQTL OCRs. TFs are clustered based on their PWM similarity using hierarchical clustering, resulting in four major TF groups. Bar plots of *P* values are color coded according to this clustering. A representative motif logo is shown for each cluster. Asterisks mark the TF that corresponds to depicted motif logos. *E*: Luciferase reporter activity in MIN6 β-cells of sequences containing human islet caQTL alleles at selected loci. Plots show the ratio of luciferase activity of the more accessible, open allele relative to the less accessible, closed allele. Dashed red line indicates balanced activity of caQTL alleles. Error bars are SEM. $^{****}P < 0.0001$; $^{***}P < 0.001$, according to two-sided Mann-Whitney test. Three plasmid preparations were tested for each sequence on three separate occasions. *F*: QQ plot of observed (y-axis) vs. expected (x-axis) islet eQTL (eQTLs from 19 individuals in this study) association *P* values for islet caQTL SNPs (black) or randomly selected noncaQTL SNPs (blue). Higher enrichment of eQTLs among statistically significant caQTLs links regulation of chromatin accessibility to gene expression. Red line denotes the line of equality ($y = x$).

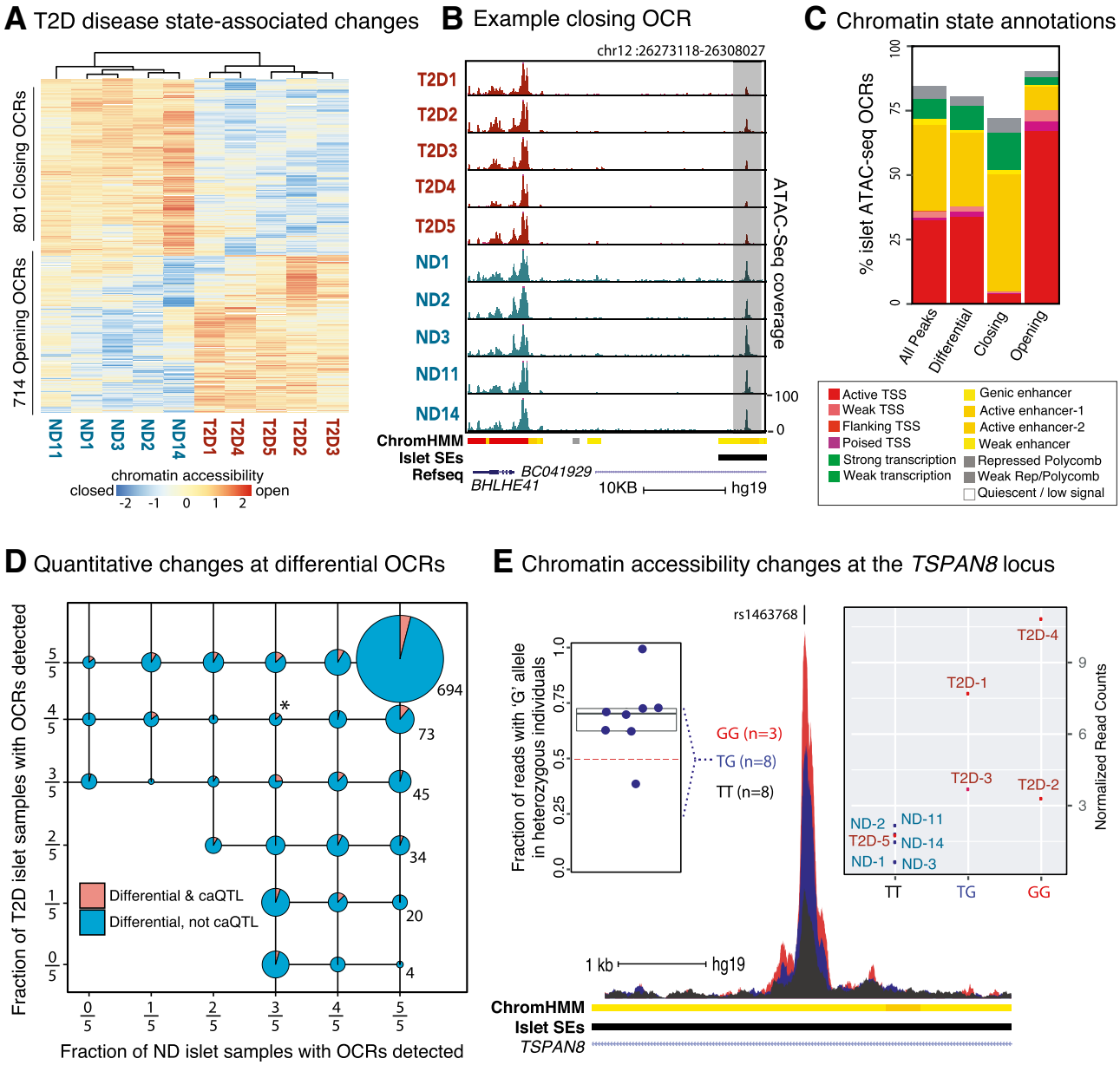


Figure 3—Chromatin accessibility differences between T2D and ND islet samples. *A*: T2D disease state–associated chromatin accessibility changes. Heatmap represents normalized chromatin accessibility levels at differentially accessible sites (FDR 10%). *B*: UCSC Genome Browser view around the *BHLHE41* locus, highlighting an example of closing OCR in T2D islet samples. *C*: Islet ChromHMM chromatin state annotations of all islet OCRs (n = 52,387) and differentially accessible OCRs (n = 1,515), further separated into closing (n = 801) or opening (n = 714) OCRs. Note that closing and opening OCRs predominantly overlap islet enhancer and promoter states, respectively. *D*: Plot showing the fraction of ND (*x*-axis) and T2D (*y*-axis) islet samples that have OCRs detected at differentially accessible regions, demonstrating that the majority of accessibility changes in T2D islet samples are quantitative in nature. The size of each pie represents the number of differential OCRs for that category. Pie sizes are listed for the rightmost column. Pink wedges indicate the proportion of T2D disease state–associated differential OCRs that are also islet caQTLs. Asterisk denotes the group that contains the opening OCR shown in panel *E*. *E*: T2D opening OCR that is also an islet caQTL. Average chromatin accessibility of all 19 islet samples at the T2D-associated *TSPAN8* locus, stratified by rs1463768 genotype. Average read counts for islet samples with the GG genotype is 97.33 at the OCR summit. Islet samples with TG or TT genotypes exhibited maximum average read counts of 67.75 and 29.125, respectively. Left inset shows the fraction of ATAC-seq reads containing the G allele for each of the heterozygous islet samples (n = 8). Right inset shows chromatin accessibility of the five ND and five T2D islet samples used in the differential OCR analysis, stratified by rs1463768 genotype. hg19 coordinates: chr12:71586245–71591030.

BHLHE41 (Fig. 3*B*, gray rectangle), mostly overlapped enhancers (48%), whereas opening OCRs extensively overlapped promoters (70%) (Fig. 3*C*). This difference was also reflected in TF motif enrichments, where opening and closing OCRs were enriched for distinct motifs (Supplementary Fig. 3*D*). Interestingly, motifs for PDX1 and TFs that regulate stress responses, such as ATF3/JUNB, AP-1, and BACH1, were enriched in closing OCRs, which may

represent epigenomic signatures of previously described molecular perturbations in dysfunctional and T2D islets, including PDX1 export from the nucleus (44), perturbation of oxidative stress responses (45,46), and inactivation of β-cell survival pathways (47).

The overwhelming majority (>99%) of T2D disease state–associated changes in chromatin accessibility were quantitative, not qualitative, i.e., OCRs did not completely appear/disappear with T2D disease state (Fig. 3*D*). A total of 654 genes were associated with opening OCRs (predominantly enhancers), and 622 genes were associated with closing OCRs (predominantly promoters). Differentially accessible OCRs at gene promoters exhibited modest positive correlation with the corresponding gene's expression (Supplementary Fig. 3*E*). T2D disease state–associated OCRs were not enriched for any GO terms or KEGG/Wiki pathways. Differential gene expression analyses from the same ND and T2D samples revealed few significant changes (Supplementary Table 9), where only 120 and 54 genes were up- or downregulated, respectively, with T2D disease state at FDR 10% (Supplementary Table 10).

Finally, given the significant impact of genetics on islet chromatin accessibility, we asked which T2D disease state–associated chromatin accessibility changes may be driven by genetic differences. Interestingly, 6% of the differentially accessible OCRs overlapped islet caQTLs (39 opening OCRs, 51 closing OCRs) (Fig. 3*D*, Supplementary Fig. 3*F*), including the opening OCR that contains the caQTL variant rs1463768. Four of five T2D islet samples were heterozygous or homozygous for opening G allele for this variant, whereas all five ND donors were homozygous for the closing T allele (Fig. 3*E*). rs1463768-containing sequences did not show allelic differences in luciferase reporter activity in MIN6 cells (Fig. 2*E*). Therefore, it remains uncertain whether genotype, environment (i.e., T2D disease state), or genotype–environment interactions are responsible for islet chromatin accessibility changes at this and other overlapping loci.

T2D-Associated GWAS SNPs Altering Islet Chromatin Accessibility

The vast majority (>90%) of GWAS SNPs associated with T2D (4,48) and metabolic measures of islet (dys)function (49,50) are noncoding and overlap islet SEs (6,7). To test whether T2D- and islet (dys)function–associated GWAS SNPs alter chromatin accessibility in islets, we assessed overlaps of GWAS index and linked SNPs (see RESEARCH DESIGN AND METHODS) with islet caQTLs. Among 184 diverse trait- and disease-associated SNP sets tested, only those associated with T2D (2.97 fold), fasting plasma glucose (13.46 fold), and BMI-adjusted fasting glucose–related (7.43 fold) traits were significantly enriched in islet caQTLs (Fig. 4*A*; $P < 5.43 \times 10^{-04}$, FDR 5%). In contrast, DNaseI sensitivity QTLs (13) in lymphoblastoid cell lines were enriched for mostly autoimmune disease–associated GWAS SNPs (Supplementary Fig. 4*A*), emphasizing the specificity of T2D-associated GWAS SNP enrichments in islet caQTLs.

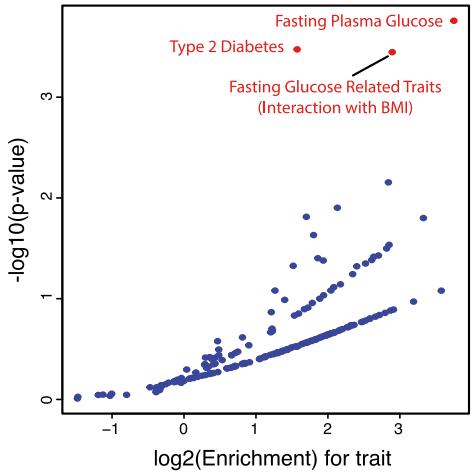
We identified SNPs in 13 T2D-associated loci that alter islet chromatin accessibility, thereby nominating these as putative causal/functional SNPs (Fig. 4*B*, Supplementary Fig. 4*B*). caQTL SNP alleles for 4 of 13 T2D-associated loci (*ADCY5, ZMIZ1, MTNR1B, RNF6*) were previously linked to altered in vitro enhancer activity (51), in vivo chromatin accessibility (52), or in vivo steady state gene expression in islets (8,11,12,53). Importantly, T2D-associated risk alleles for these four loci exhibit concordant effects on chromatin accessibility and gene expression in islets, i.e., same direction of effect (Fig. 4*B*). For 6 of 13 T2D-associated caQTLs, the risk allele decreased chromatin accessibility, designated as loss of function (Fig. 4*B*). This included the T2D-associated caQTL SNP rs11708067 in the third intron of *ADCY5*, which overlaps an islet SE. The risk allele for this variant is associated with reduced chromatin accessibility (Fig. 4*C*), consistent with recent reports linking the rs11708067 risk allele to decreased transcriptional reporter activity in rodent β-cells (MIN6 and 832/13) and to decreased *ADCY5* expression in ND human islets in vivo (12,51). The T2D risk allele was associated with increased chromatin accessibility for the remaining 7 of 13 islet caQTLs (Fig. 4*B*), designated as gain of function. For example, the T2D risk allele A at rs6937795 increased in vivo islet chromatin accessibility in the *IL20RA* locus (Fig. 4*D*) and conferred 2.5-fold higher transcriptional reporter activity than the nonrisk C allele (Fig. 2*E*). Although targeted approaches are required to establish causality, our analyses nominate rs6937795 as a strong candidate for causal SNP in the T2D-associated *IL20RA* locus.

DISCUSSION

In this study, we integrated ATAC-seq data and genotypes from 19 islet donors to link 2,949 SNPs with altered in vivo chromatin accessibility. Allelic effects on in vivo chromatin accessibility correlated well with effects on in vitro enhancer activity; 8 of 13 caQTLs tested showed concordant allelic effects in luciferase reporter assays. Although we cannot eliminate the possibility of false-positive associations for the remaining 5 caQTLs, these loci may also represent *1*) α-cell–specific, *2*) species-specific, or *3*) poised/primed *cis*-REs (16), which need to be tested in future studies in human cells.

The data suggest that islet caQTL variants modulate regulatory programs important for islet cell identity and function. They were enriched in islet-specific TF motifs, TF ChIP-seq peaks (7), and islet SEs (6). They were specific to islets, as only 2.3% (68/2,949) of the islet caQTL variants altered chromatin accessibility in induced pluripotent stem cells or macrophages (data not shown) (16,54). Islet caQTL SNPs were linked to more significant effects on islet gene expression levels than variants that do not significantly impact chromatin accessibility (i.e., noncaQTL SNPs). Increasing the cohort size and separating islet cell types in future studies should lead to increased convergence between islet caQTLs and islet eQTLs. Furthermore, studying

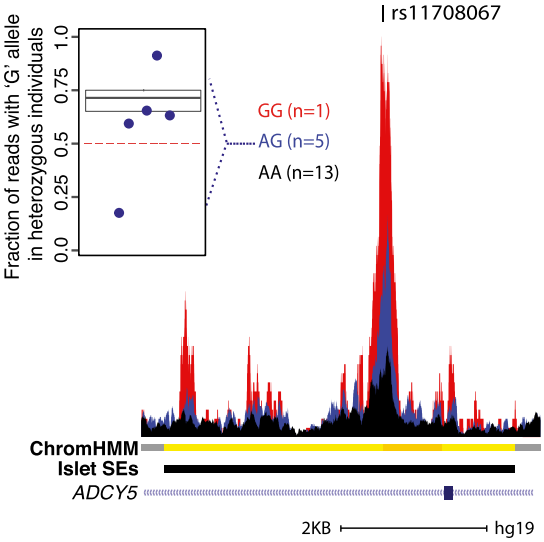
A GWAS Disease SNP enrichments at caQTLs



B caQTLs linked to T2D-associated SNPs

	Gene	Index SNP	caQTL SNP	r^2	caQTL Alleles	Risk Allele	Accessible Allele	eQTL Allele
Loss-of-function	*ADCY5*	rs11717195	rs11708067	0.94	A/G	A	G	G
	*PLEKHA1**	rs10510110	rs2421016	0.99	C/T	T	C	NA
	IGF2BP2	rs1470579	rs10428126	1	T/C	C	T	NA
	LOC100507477	rs642858	rs9376483	0.92	C/T	C	T	NA
	*ROR2**	rs1873747	rs7855529	1	T/C	C	T	NA
	*EVADR**	rs1048886	rs16869158	0.81	A/T	T	A	NA
Gain-of-function	*IL20RA**	rs6937795	rs6937795	1	A/C	A	A	NA
	*INS**	rs3842770	rs115420895	1	G/A	A	A	NA
	RNF6	rs10507349	rs34584161	0.93	A/G	A	A	A
	*CDKN2B**	rs2383208	rs10811661	0.95	T/C	T	T	NA
	*ZMIZ1**	rs12571751	rs703977	0.98	T/G	T	T	T
	MTNR1B	rs10830963	rs10830963	1	C/G	G	G	G
	IRS1	rs2943640	rs2943656	0.85	A/G	G	G	NA

C Loss-of-function T2D caQTL



D Gain-of-function T2D caQTL

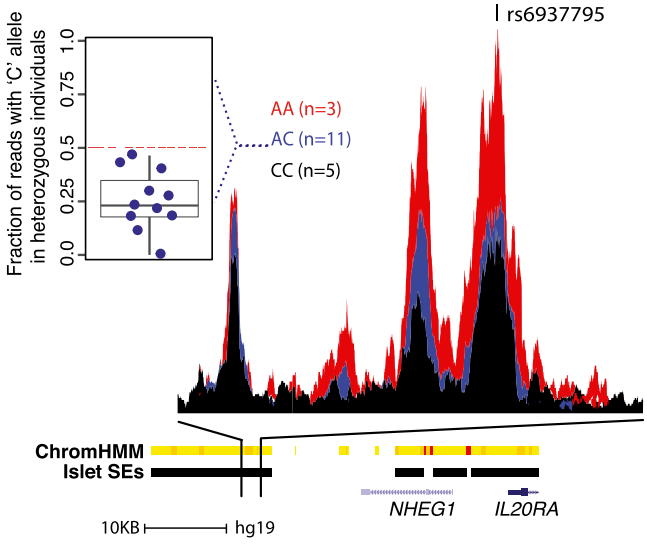


Figure 4—GWAS SNP enrichment in islet caQTLs. *A*: Disease- and trait-associated GWAS SNP enrichment in islet caQTLs. Enrichment (*x*-axis, observed/expected number of disease SNPs) and significance (*y*-axis) of GWAS SNP islet caQTL overlaps are plotted. Red dots indicate significantly enriched diseases/traits at FDR 5% (after correcting for multiple hypothesis testing; 184 GWAS catalog diseases and traits tested). *B*: Table showing the T2D-associated GWAS index or linked ($r^2 > 0.8$) SNP overlapping islet caQTLs. Asterisks mark sequence variants tested for allelic effects on luciferase activity shown in Fig. 2*E*. The eQTL allele refers to the allele linked to higher gene expression in islets (8,11,12). Reported pairwise SNP correlations (r^2 values) are based on European populations. *C*: Average chromatin accessibility in islet samples stratified by genotype at rs11708067 in the *ADCY5* locus. The inset shows the fraction of ATAC-seq reads containing the rs11708067 G allele in each of the heterozygous islet samples ($n = 5$). This is a putative loss-of-function T2D-associated caQTL, in which the T2D risk allele A at rs11708067 is associated with lower chromatin accessibility in islets and lower gene expression levels. Average read counts for islet samples with the GG genotype is 44 at the OCR summit. Islet samples with AG or AA genotypes exhibit maximum average read counts of 24.4 or 10.08, respectively. hg19 coordinates: chr3:123062482–123067947. *D*: Average chromatin accessibility in islet samples stratified by genotype at rs6937795 in the *IL20RA* locus. The fraction of ATAC-seq reads containing the C allele in each of the heterozygous islet samples ($n = 11$) is plotted in the inset. This is an example of a gain-of-function T2D-associated caQTL, in which the T2D risk allele is associated with higher chromatin accessibility at this OCR. Average read counts for islet samples with the AA genotype is 40.33 at the OCR summit. Islet samples with AC or CC genotypes exhibited maximum average read counts of 22.81 or 19.6, respectively. hg19 coordinates for zoomed-in view of ATAC-seq average read counts: chr6:137289071–137292315; hg19 coordinates for ChromHMM chromatin state, islet SE, and RefSeq Gene models: chr6:137277485–137324778.

islets under stress conditions could identify and link primed enhancers and response eQTLs, which have been reported in other cell types (16,55).

T2D disease state–associated changes in chromatin accessibility were limited and quantitative, i.e., few OCRs completely lost or gained accessibility with T2D, suggesting that the T2D disease state does not lead to extensive remodeling of steady-state chromatin accessibility in islets. However, we acknowledge that T2D disease state–associated epigenetic changes may be masked by multiple factors,

including *1*) relatively low HbA_{1c} values for T2D donors (5.3–7.4%); *2*) cell type–specific changes hidden by whole-islet measurements; *3*) steady-state, normoglycemic culture conditions of the islets that may mask changes elicited by the diabetic state; and *4*) limited power due to cohort size (n = 10) and genetic diversity. We found that 6% of differentially accessible OCRs associated with T2D disease state overlapped caQTLs. Future studies integrating genotype and environment and their interaction in larger, genetically stratified cohorts should contribute to more precise understanding of epigenomic changes associated with T2D disease state.

This study demonstrates the utility of using islet caQTL analyses to identify and prioritize putative functional variants among hundreds of linked, "credible set" T2D-associated SNPs (4,9,48). Even with a relatively small cohort (n = 19), we identified putative causal variants at 13 T2D GWAS loci, based on their chromatin accessibility effects. These include four loci (*ADCY5, MTNR1B, RNF6,* and *ZMIZ1*) in which the same or linked ($r^2 > 0.8$) SNP functions as an islet eQTL (8,11,12,53). Importantly, the risk allele exhibited concordant effects on islet chromatin accessibility and gene expression for each locus. Finally, we identified allelic effects on both in vivo and in vitro islet enhancer activity for multiple new loci, such as rs6937795 in the *IL20RA* locus, and linked the risk alleles at each locus to increased or decreased activity. This study provides new understanding of genetic variant effects on islet chromatin accessibility and enumerates targets for site-specific and hypothesis-driven investigation.

Acknowledgments. The authors are indebted to the anonymous pancreatic islet organ donors and their families, without whom this entire study would not be possible. A subset of human pancreatic islets was provided by the National Institute of Diabetes and Digestive and Kidney Diseases–funded Integrated Islet Distribution Program at City of Hope (grant 2UC4DK098085). The authors thank Jane Cha, Jackson Laboratory for Genomic Medicine, for help generating artwork for the figures; members of the Stitzel and Ucar laboratories for helpful discussion and critiques during study design and execution; and Taneli Helenius, Jackson Laboratory for Genomic Medicine, and anonymous reviewers, whose comments, questions, and suggested edits greatly improved the quality and clarity of the manuscript.

Funding. This study was made possible by generous financial support from The Jackson Laboratory startup funds (to M.L.S. and D.U.); the Doug Coleman Research Fund at The Jackson Laboratory; the National Institute of Diabetes and Digestive and Kidney Diseases under award number DK092251 (to M.L.S.); the Assistant Secretary of Defense for Health Affairs, through the Peer Reviewed Medical Research Program under award number W81XWH-16-1-0130 (to M.L.S.); and the American Diabetes Association Pathway to Stop Diabetes Accelerator Award (1-18-ACE-15) (to M.L.S.).

Opinions, interpretations, conclusions, and recommendations are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health, Department of Defense, or American Diabetes Association.

Duality of Interest. No other potential conflicts of interest relevant to this article were reported.

Author Contributions. S.K., D.U., and M.L.S. conceived the study and designed experiments. R.K. and M.L.S. collected and prepared each islet sample for genotyping and sequencing. S.K. analyzed the data. A.Y., N.L., and E.J.M. contributed to bioinformatics and statistical analyses of the data. S.K. and A.J. cloned and tested caQTL allelic effects using luciferase reporters. S.K., D.U., and M.L.S. wrote the manuscript. All authors reviewed and edited the final manuscript. M.L.S. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Data Availability. The accession number for human islet ATAC-seq and RNA-seq data reported in this article is NCBI Sequence Read Archive: SRP117935.

Prior Presentation. Parts of this study were presented at the Pancreatic Diseases Gordon Research Conference, Waterville Valley, NH, 18–23 June 2017, and the Boston Ithaca Islet Club Meeting, Worcester, MA, 28–29 April 2018.

References

1. Halban PA, Polonsky KS, Bowden DW, et al. β-Cell failure in type 2 diabetes: postulated mechanisms and prospects for prevention and treatment. Diabetes Care 2014;37:1751–1758
2. Lawlor N, Khetan S, Ucar D, Stitzel ML. Genomics of islet (dys)function and type 2 diabetes. Trends Genet 2017;33:244–255
3. Ashcroft FM, Rorsman P. Diabetes mellitus and the β cell: the last ten years. Cell 2012;148:1160–1171
4. Fuchsberger C, Flannick J, Teslovich TM, et al. The genetic architecture of type 2 diabetes. Nature 2016;536:41–47
5. Mohlke KL, Boehnke M. Recent advances in understanding the genetic architecture of type 2 diabetes. Hum Mol Genet 2015;24(R1):R85–R92
6. Parker SCJ, Stitzel ML, Taylor DL, et al.; NISC Comparative Sequencing Program; National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program Authors; NISC Comparative Sequencing Program Authors. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci U S A 2013;110:17921–17926
7. Pasquali L, Gaulton KJ, Rodríguez-Seguí SA, et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet 2014;46:136–143
8. Varshney A, Scott LJ, Welch RP, et al.; NISC Comparative Sequencing Program. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc Natl Acad Sci U S A 2017;114:2301–2306
9. Gaulton KJ, Ferreira T, Lee Y, et al.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nat Genet 2015;47:1415–1425
10. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348:648–660
11. Fadista J, Vikman P, Laakso EO, et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc Natl Acad Sci U S A 2014;111:13924–13929
12. van de Bunt M, Manning Fox JE, Dai X, et al. Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. PLoS Genet 2015;11:e1005694
13. Degner JF, Pai AA, Pique-Regi R, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 2012;482:390–394
14. Gaffney DJ, McVicker G, Pai AA, et al. Controls of nucleosome positioning in the human genome. PLoS Genet 2012;8:e1003036
15. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet 2016;48:206–213
16. Alasoo K, Rodrigues J, Mukhopadhyay S, et al.; HIPSCI Consortium. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet 2018;50:424–431
17. McVicker G, van de Geijn B, Degner JF, et al. Identification of genetic variants that affect histone modifications in human cells. Science 2013;342:747–749
18. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin,

DNA-binding proteins and nucleosome position. Nat Methods 2013;10:1213–1218
19. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–2120
20. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–1760
21. Lawlor N, Youn A, Kursawe R, Ucar D, Stitzel ML. Alpha TC1 and Beta-TC-6 genomic profiling uncovers both shared and distinct transcriptional regulatory features with their primary islet counterparts. Sci Rep 2017;7:11959
22. Ucar D, Márquez EJ, Chung C-H, et al. The chromatin accessibility signature of human immune aging stems from $CD8^+$ T cells. J Exp Med 2017;214:3123–3144
23. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9:R137
24. Ross-Innes CS, Stark R, Teschendorff AE, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature 2012;481:389–393
25. Kundaje A, Meuleman W, Ernst J, et al.; Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317–330
26. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. New York, Springer, 2009
27. Auton A, Brooks LD, Durbin RM, et al.; 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015;526:68–74
28. Loh P-R, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. Nat Genet 2016;48:811–816
29. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. Nat Genet 2016;48:1284–1287
30. Jun G, Flickinger M, Hetrick KN, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 2012;91:839–848
31. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012;28:882–883
32. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–140
33. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575
34. Schmidt EM, Zhang J, Zhou W, et al. GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics 2015;31:2601–2606
35. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 2010;38:576–589
36. Tan G, Lenhard B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinformatics 2016;32:1555–1556
37. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–359
38. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011;12:323
39. Stitzel ML, Sethupathy P, Pearson DS, et al.; NISC Comparative Sequencing Program. Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. Cell Metab 2010;12:443–455
40. Scott LJ, Erdos MR, Huyghe JR, et al. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat Commun 2016;7:11764
41. Ackermann AM, Wang Z, Schug J, Naji A, Kaestner KH. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Mol Metab 2016;5:233–244
42. Heintzman ND, Hon GC, Hawkins RD, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 2009;459:108–112
43. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res 2016;44:D877–D881
44. Guo S, Dai C, Guo M, et al. Inactivation of specific β cell transcription factors in type 2 diabetes. J Clin Invest 2013;123:3305–3316
45. Abebe T, Mahadevan J, Bogachus L, et al. Nrf2/antioxidant pathway mediates β cell self-repair after damage by high-fat diet-induced oxidative stress. JCI Insight 2017;2:92854
46. Kondo K, Ishigaki Y, Gao J, et al. Bach1 deficiency protects pancreatic β-cells from oxidative stress injury. Am J Physiol Endocrinol Metab 2013;305:E641–E648
47. Gurzov EN, Barthson J, Marhfour I, et al. Pancreatic β-cells activate a JunB/ATF3-dependent survival pathway during inflammation. Oncogene 2012;31:1723–1732
48. Scott RA, Scott LJ, Mägi R, et al. An expanded genome-wide association study of type 2 diabetes in Europeans. Diabetes 2017;66:2888–2902
49. Wood AR, Jonsson A, Jackson AU, et al. A genome-wide association study of IVGTT-based measures of first phase insulin secretion refines the underlying physiology of type 2 diabetes variants. Diabetes 2017;66:2296–2309
50. Dimas AS, Lagou V, Barker A, et al.; MAGIC Investigators. Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. Diabetes 2014;63:2158–2171
51. Roman TS, Cannon ME, Vadlamudi S, et al. A type 2 diabetes-associated functional regulatory variant in a pancreatic islet enhancer at the Adcy5 locus. Diabetes 2017;66:2521–2530
52. Thurner M, van de Bunt M, Torres JM, et al. Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. eLife 2018;7:7
53. Lyssenko V, Nagorny CLF, Erdos MR, et al. Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. Nat Genet 2009;41:82–88
54. Banovich NE, Li YI, Raj A, et al. Impact of regulatory variation across human iPSCs and differentiated cells. Genome Res 2018;28:122–131
55. Nédélec Y, Sanz J, Baharian G, et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. Cell 2016;167:657–669.e21

Bioinformatics, 2018, 1–9
doi: 10.1093/bioinformatics/bty329
Advance Access Publication Date: 24 April 2018
Original Paper

OXFORD

Gene expression

Two-phase differential expression analysis for single cell RNA-seq

Zhijin Wu[1,2,3,*], Yi Zhang[1], Michael L. Stitzel[4,5,6] and Hao Wu[7,*]

[1]Department of Biostatistics, [2]Center for Statistical Sciences and [3]Center for Computational Molecular Biology, Brown University, Providence, RI 02806, USA, [4]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, [5]Institute for Systems Genomics and [6]Department of Genetics & Genome Sciences, University of Connecticut, Farmington, CT 06032, USA and [7]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on October 23, 2017; revised on February 19, 2018; editorial decision on April 19, 2018; accepted on April 21, 2018

Abstract

Motivation: Single-cell RNA-sequencing (scRNA-seq) has brought the study of the transcriptome to higher resolution and makes it possible for scientists to provide answers with more clarity to the question of 'differential expression'. However, most computational methods still stick with the old mentality of viewing differential expression as a simple 'up or down' phenomenon. We advocate that we should fully embrace the features of single cell data, which allows us to observe binary (from Off to On) as well as continuous (the amount of expression) regulations.
Results: We develop a method, termed SC2P, that first identifies the phase of expression a gene is in, by taking into account of both cell- and gene-specific contexts, in a model-based and data-driven fashion. We then identify two forms of transcription regulation: phase transition, and magnitude tuning. We demonstrate that compared with existing methods, SC2P provides substantial improvement in sensitivity without sacrificing the control of false discovery, as well as better robustness. Furthermore, the analysis provides better interpretation of the nature of regulation types in different genes.
Availability and implementation: SC2P is implemented as an open source R package publicly available at https://github.com/haowulab/SC2P.
Contact: zhijin_wu@brown.edu or hao.wu@emory.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Studies of transcriptome have been arguably the most active field in genomics research. Traditionally, gene expression is measured from 'bulk' samples pooling a large number (often in the scale of millions) of cells, thus the measurements reflect the average expression of a population of cells. For highly heterogeneous samples such as cancer or brain tissues, the bulk measurements fail to provide more detailed information for the transcriptomic variation. For example, bulk expression data cannot differentiate a '50% decrease in all cells' and a mixture of 'complete shut-down in half of the cells, while no change in the other half'.

Single-cell RNA-sequencing (scRNA-seq) emerged recently as a powerful technology to investigate transcriptomic variation and regulation at the individual cell level (Buettner *et al.*, 2015; Patel *et al.*, 2014; Picelli *et al.*, 2013; Ramsköld *et al.*, 2012; Shalek *et al.*, 2014; Tang *et al.*, 2009; Usoskin *et al.*, 2015). It is in scRNA-seq that we finally observe evidence of binary status of transcription (Shalek *et al.*, 2013; Wills *et al.*, 2013), which we refer to as 'phases' in transcription. Phase I corresponds to low level non-specific transcription (for example, as a result of random initiation), and Phase II corresponds to targeted specific transcription. The regulation of transcription includes a phase transition between Phase I and Phase II, as well as continuous regulation within Phase II.

Even though the analysis of scRNA-seq data is multifaceted, including cell clustering (Kiselev *et al.*, 2017; Ntranos *et al.*, 2016),

© The Author(s) 2018. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty329/4984507
by Jackson Laboratory user
on 29 June 2018

pseudo-time construction (Trapnell *et al.*, 2014) and rare cell type identification (Grün *et al.*, 2015; Jiang *et al.*, 2016), differential expression (DE) remains the most fundamental question to be answered. The scRNA-seq technology makes it possible for scientists to provide answers with more clarity even to the simple question of DE. Due to the special characteristics of scRNA-seq data, including excessive zero counts for both biological and technical reasons, higher variability and multi-modal distribution that cannot be attributed to the zero counts (Bacher and Kendziorski, 2016), DE methods developed for bulk RNA-seq cannot be directly applied. We illustrate some of these characteristics in Figure 1, where histograms of the log counts in three cells are presented. A spike of zero counts is observed in all three cells, most obvious in Cell A and to a less extent in Cells B and C. A substantial fraction of genes have non-zero but very low counts (with log2 counts less than 3). Another group of genes reach counts that are orders-of-magnitudes higher, sometimes forming a second mode, which is most obvious in Cell B. Cell B appears to have a greater proportion of genes with high expression level, though it also has more than twice as many genes with zero counts as seen in Cell C. These examples suggest that non-zero count is not a reliable reflection of expression activity and to dichotomize genes into on/off categories by one arbitrary cutoff may lead to systematic biases between cells.

Recently, a few methods were developed specifically for scRNA-seq DE. SCDE (Kharchenko *et al.*, 2014) uses a mixture of Poisson and negative binomial distributions to capture the two phases, and then identifies DE when the gene is on. Monocle (Trapnell *et al.*, 2014) uses a generalized additive model (GAM) to test the differences in marginal mean expressions; BPSC (Vu *et al.*, 2016) uses a beta-Poisson mixture model to capture the bimodality in the expression, and then implements a generalized linear model (GLM) for DE test for, again, the differences in marginal mean expressions.

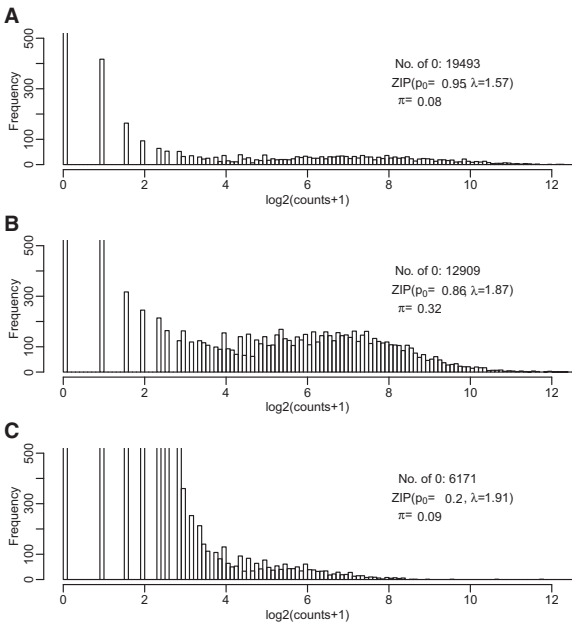


Fig. 1. Histogram of three cells in the human brain dataset. The y axis is trimmed at 500 to allow the visualization of lower frequencies. The parameters are described in the Methods section. π is the estimated prior proportion of genes in Phase II. (**A**) A cell with extremely high zero inflation, and a small fraction (8%) of genes in Phase II expression. (**B**) A cell with high zero inflation, but also a high proportion (32%) of genes in Phase II expression. (**C**) A cell with low zero inflation, but also small fraction (9%) in Phase II expression

Even though these methods have noticed and mentioned the phenomenon of two-phase transcription from scRNA-seq data, they dismissed the importance of the phase transition. Genes in Phase I are often considered technical 'dropouts' that failed to be detected, and the DE analyses are mostly focused on the marginal changes or within the Phase II, e.g. when the gene is 'on'. Even when phase transition is considered in some methods, it is not recognized as an important form of DE in its own right. For example, MAST (Finak *et al.*, 2015) includes a test for phase change but only declares a gene DE if '*the estimated fold-change is greater than 1.5*' in addition to low FDR. D^3E (Delmans and Hemberg, 2016) is a method based on a bursting model that explicitly considers 'On' and 'Off' status of gene expression. The detection of DE, however, is marginal: the method uses non-parametric or likelihood ratio tests to test a null hypothesis that the distributions of expression across two groups are identical. When the null is rejected, it does not infer the reason being a change in bursting rate or in burst size. Korthauer *et al.* (2016) also considers the possibility of multi-modal distribution of a gene's expression, and presents a Bayesian modeling framework (scDD) that identifies differential distribution (DD) across conditions. The transition between phases is not directly inferred. Instead, genes that are identified as showing DD are subsequently classified by their patterns of difference, including mean shift, differential proportion of the same components, differential modality or a combinations of these.

We advocate that the lower mode in the distribution of gene expression corresponds to a phase of inactivity, and phase transition is the first important step in transcription regulation, hence it is essential to the understanding of the regulation mechanism. Thus a principled, data-driven approach rather than arbitrary cutoff for determining phase is necessary. We observe, in multiple biological systems, that DE can take the form of phase transition or magnitude tuning, and a combination of these two. Most interestingly, we observe examples of 'compensation' (presented in the Results section): a population of cells may have a lower percentage expressing a particular gene, but the cells expressing that gene do so at a higher level. In such cases, the average expression level may remain the same and be completely unidentifiable in bulk RNA-seq.

In this work, we present a statistical method, termed **SC2P**, that identifies the phase for each gene in each cell, given the context (both biological and technical) of each cell sample and gene-specific profile. With this latent phase inferred, we identify genes that go through different forms of DE. This includes genes that are turned on with different frequencies in different populations (**Form I**), as well as genes that are transcribed at different rates (**Form II**). These different forms of DE reveal, potentially, different mechanisms in the regulation of transcription, such as initiation versus elongation speed (Jonkers and Lis, 2015), bursting frequency versus bursting size (Dar *et al.*, 2012; Raj *et al.*, 2006), or different half-life of RNA transcripts. Being able to distinguish the forms of DE between cell types, or over time, will also elucidate the relationship between expression and genomic/epigenomic elements: some markers may be associated with the probability of expression while others may be associated with the amount of expression.

2 Materials and methods

2.1 Data model

We begin with the expression measured as sequence read counts for G genes and C cells in a $G \times C$ matrix $\mathbf{Y}$. For a particular gene, we use a two-component mixture model to describe its expression from

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty329/4984507
by Jackson Laboratory user
on 29 June 2018

individual cells. This characterizes the phenomenon observed in multiple publicly available datasets (Darmanis *et al.*, 2015; Shekhar *et al.*, 2016) as well as our data that many genes demonstrate a bimodal distribution: one component corresponds to very low counts with an excess of zero, consistent with a background, inactive transcription; the other component corresponds to higher counts with a long right tail that are approximately normal in log scale. We refer to these as the two 'phases' of transcription. A key difference separating our model from those described in existing methods is our treatment of the first component, by allowing cell-specific parameters. The status of each gene in each cell, i.e. which component the observed count is generated from, is latent, but inferable given the observed count and the gene-cell contexts.

We use a zero-inflated Poisson (ZIP) distribution to model Phase I (inactive transcription), and a lognormal-Poisson (LNP) model for Phase II (targeted specific transcription). Specifically, let Y_{gi} denote the count observed on gene g in cell i, and Z_{gi} denote the binary latent expression state ($Z_{gi}=1$ for Phase II). We model Phase I with a ZIP distribution $Y_{gi}|Z_{gi}=0 \sim ZIP(p_i, \lambda_i)$, where p_i is for the extra point mass at 0 to account for zero-inflation, and λ_i is the Poisson rate. Both the zero-inflation and the Poisson parameter are cell specific, reflecting the heterogeneity in low counts among cells. In scRNA-seq data, each sample is a single cell. Thus the parameters p_i and λ_i reflect both cell effects and sample preparation effects, which are not separately identifiable.

Conditioning on a gene in Phase II, or the 'on' phase, the observed count is modeled by log-normal Poisson (LNP) mixture distribution, with θ_{gi} denote the mean expression rate:

$$\theta_{gi}|Z_{gi}=1 \sim LN\left(\mu_g, \sigma_g^2\right), \;\; Y_{gi}|\theta_{gi} \sim Poisson\left(\theta_{gi} S_i\right)$$

Here S_i is the size factor representing the sequencing depth in cell i. We use lognormal-Poisson distribution instead of the often-used negative binomial (gamma-Poisson) distribution for two reasons. First, the heterogeneity between samples in scRNA-seq data are much greater than that in the bulk data, making the gamma model no longer flexible enough (more detailed discussions are provided in the Supplementary Material). Second, the log normal model offers more convenience in downstream DE testing procedure, since we can use existing methods for linear models on log transformed data. The LNP model for the phase II distribution is cell- and gene-specific, capturing the expression heterogeneity among cells and genes. Marginally, the model gives

$$P\left(Y_{gi}=y_{gi}\right) = (1-\pi_i) ZIP\left(y_{gi}|p_i, \lambda_i\right) + \pi_i LNP\left(y_{gi}|\mu_g, \sigma_g^2\right).$$

where π_i represents the prior probability for gene g in cell i to be in the specific transcription phase. The parameters for the ZIP model could vary between genes, but we choose the simplification by assuming the same parameters p_i and λ_i for all genes within a cell for better model identification and easier parameter estimation. With this simplification, the cell's profile provides information about the inactive transcription as well as the technical issues such as extraction and counting efficiency for the sample. The gene's profile across cells provides information about a gene's expected expression when it enters the active transcription phase. We estimate cell specific parameters for the ZIP and gene specific parameters for the LNP distributions (detailed below). Given these hyper parameters and the observed count, the posterior probability of each gene in Phase II is computed. Most existing methods determine the phases by applying an arbitrary threshold to all genes and all cells (Kharchenko *et al.*, 2014; Shalek *et al.*, 2013), which fails to consider the cell- and gene-specific characteristics. MAST attempts to derive gene-specific thresholds by implementing an *ad hoc* 'adaptive thresholding' to estimate thresholds based on average expression level of genes. It takes TPM (transcripts per million) as inputs to normalize out one particular cell-specific characteristics: the total sequencing depth. However, MAST ignores the differences in expression distributions from different cells. Our proposed method achieves cell- and gene-specific inference for phases based on a rigorous statistical model. This leads to a data-driven determination of transcription phase for genes, and subsequently better DE detection results.

2.2 Parameter estimation

2.2.1 Estimation of ZIP parameters

Cell-specific ZIP parameters p_i and λ_i are estimated for each cell separately. We developed a robust and efficient ZIP estimation method, which takes advantage of the linearity of log transformed probability mass in a Poisson or a ZIP random variable. Specifically, for a Poisson random variable Y, $\log P(Y=k) = -\lambda - \log(k!) + k\log(\lambda)$. Define the expected frequency as $D_k \equiv P(Y=k)$, we see that $\log D_k + \log(k!)$ has a linear relationship with k with slope $\log\lambda$. This linear relationship remains even when there is zero inflation, except for $k=0$. Given the observed frequencies $d_k \equiv \sum_{i=1}^{n} I(y_i=k)/n$, we regress $\log d_k + \log(k!)$ on k to estimate λ, with decreasing weights for higher k to enforce robustness. With λ estimated, we use the zero frequency exceeding expectation ($\exp(-\widehat{\lambda})$) to estimate the inflation. If the observed zero counts does not exceed ($\exp(-\widehat{\lambda})$), we set the inflation as zero (i.e. the possibility of zero-depletion is not considered). Specifically, given $\widehat{\lambda}$, we estimate the zero inflation as

$$\widehat{p} = \max\left(0,\; d_0 - P\left(Y=0|\widehat{\lambda}\right)\right) = \max\left(0, d_0 - \exp\left(-\widehat{\lambda}\right)\right).$$

2.2.2 Estimation of LNP parameters

With ZIP parameters estimated, we use the 99th quantile of the estimated ZIP distribution as initial threshold to filter out low-count genes, that is, genes with counts greater than the 99th quantile of ZIP are considered as in phase II in the initial round. This step will provide more accurate and stable foreground estimation. Note that the thresholds established here are not used as naive cutoffs to distinguish the two components, which was a common approach taken by some previous single-cell analyses (Kharchenko *et al.*, 2014; Shalek *et al.*, 2013). Instead, the counts passing the threshold are used to estimate the Phase II parameters μ_g and σ_g via empirical Bayesian shrinkage methods (Smyth, 2004). In detail, we log transform the counts and feed them into the shrinkage estimation procedure, by posing a common prior $\mu_g \sim N\left(\mu_0, \sigma_0^2\right)$ and $\sigma_g^2 \sim Inv\text{-}\chi^2\left(\nu_0, \tau^2\right)$ and borrow information across genes, to obtain estimates μ_g and σ_g^2. For genes that rarely enter the Phase II, the shrinkage procedure stabilizes the estimates. For genes with many high counts, there will be less shrinkage. We then plug in these estimates to obtain the posterior probability of being in phase II (π_{gi}) given each gene's counts in each cell as

$$\begin{aligned}\widehat{\pi}_{gi} &= P\left(Z_{gi}=1|Y_{gi}=y_{gi}\right) \\ &= \frac{\widehat{\pi}_i LNP\left(y_{gi}|\widehat{\mu}_g, \widehat{\sigma}_g^2\right)}{\widehat{\pi}_i LNP\left(y_{gi}|\widehat{\mu}_g, \widehat{\sigma}_g^2\right) + (1-\widehat{\pi}_i) ZIP\left(y_{gi}|\widehat{\lambda}_i, \widehat{p}_i\right)}.\end{aligned}$$

Here $\widehat{\pi}_i$ is the estimated mixture probability for Phase II. We initialize π_i as the proportion of genes exceeding the 99th percentile of the $ZIP(p_i, \lambda_i)$. We could iteratively estimate $\widehat{\pi}_i$ and LNP parameters μ_g

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty329/4984507
by Jackson Laboratory user
on 29 June 2018

and σ_g based on an EM algorithm. In practice, we found that extra iterations do not significantly alter the final result. Thus we skip the iterative procedure for computational efficiency.

The probability mass function (PMF) of LNP distribution does not have close-form. It can be efficiently and accurately approximated by

$$LNP(y|\mu,\sigma^2) \approx \Phi(\log_2(y+0.5)|\mu,\sigma^2) - \Phi(\log_2 \max(0,y-0.5)|\mu,\sigma^2).$$

where $\Phi(.|\mu,\sigma^2)$ is the cumulative distribution function (CDF) of Gaussian distribution with mean μ and variance σ^2. Numerical comparison with Monte-Carlo approximation method confirms that the Gaussian CDF approximation achieves excellent accuracy (Supplementary Fig. S2). We use this approximation in our implementation for computing efficiency.

2.3 Two-phase differential expression tests

With the inferred latent phase status of each gene in each cell, we propose a single-cell two-phase testing procedure (SC2P) that identify genes with DE in either the frequency or the magnitude of expression in Phase II. The first class of DE includes genes that are turned on to Phase II with different frequencies between cell populations. We dichotomize the each gene's phase based on the posterior probability (Phase II if $\widehat{\pi}_{gi} > 0.99$ by default, though the user may choose a different cutoff). A logistic regression model of $\widehat{Z}_{gi}$ is used to detect DE in this class. The second class of DE includes genes that show a difference in the magnitude of expression level given these are in Phase II. For each gene, the log2-transformed counts in cells with $\widehat{Z}_{gi} = 1$ are used as input data, and the test is conducted using LIMMA (Smyth, 2004). In both phases, false discovery rate (Benjamini and Hochberg, 1995) is used to control type I error.

3 Results

We demonstrate the benefit of SC2P on two independent datasets. In the first dataset (referred to as 'human brain data'), single cell sequencing data on 466 cells from human cortical tissue are obtained from GEO under accession number GSE67835. The libraries were prepared with Nextera XT DNA Sample Preparation Kit (Illumina), and sequenced by Illumina NextSeq instrument using 2 × 75 paired-end read (details are available in the appendix of Darmanis *et al.* (2015)). Cell-specific markers are identified from bulk sequencing of purified cell types in the mouse brain (Zhang *et al.*, 2014), as described in Darmanis *et al.* (2015). These cell-type-defining markers were then used to classify single cells from human brain into predefined cell types: oligodentrocytes ($n = 38$), astrocytes ($n = 62$), microglia ($n = 16$), neurons ($n = 131$), endothelial ($n = 20$), oligodendrocyte precursor cells ($n = 18$), fetal quiescent ($n = 110$) and fetal replicating cells ($n = 25$). There are also 46 cells classified as 'hybrid'.

In the second dataset (referred to as 'T2D data'), 978 cells from human pancreatic islet are profiled (Lawlor *et al.*, 2017). Cells were processed on the C1 Single Cell Autoprep System. Multiplexed single cell libraries were prepared with Nextera XT reagent, and All sequencing was performed on a NextSeq500 (Illumina). Raw sequence data is under accession SRP075970 in NCBI Sequence Read Archive (SRA). The processed dataset is available at Gene Expression Omnibus (GEO) with accession number GSE86473. Cell types are classified using known marker genes as described in Lawlor *et al.* (2017).

3.1 Data exploration

We illustrate the typical characteristics of scRNA-seq data that motivated our model using the *human brain* dataset. Figure 1 shows the distribution of expressions from all genes for three different cells. There is extremely high number of zeros in cell A, but we still observe about 8% of genes in Phase II, and these genes reach high counts. Figure 1B shows a cell that appear to have much greater fraction of the genes in Phase II, with an estimated π (proportion of genes in Phase II) at 32%, though still with substantial zero inflation ($p_0 = 0.86$). Figure 1C shows a cell with little zero inflation (less than a third of zero counts compared to Cell A), but also a low fraction of genes in Phase II (9%), similarly to Cell A. In addition, the expression level tends to be lower in this cell compared with cell A. These examples demonstrate that 'non-zero count' is not a reliable reflection of expression activity, and that the zero inflation is a sample specific feature. The proportion of genes with 'detected' expression, if defined as any none-zero count or counts above a universal cutoff, is a poor reflection of overall expression in a cell. To dichotomize genes into on/off categories by one arbitrary cutoff will also lead to systematic biases between cells.

3.2 Data-driven determination of phases

Our method estimates cell specific Phase I parameters, as well as gene specific Phase II parameters. Given an observed count of a specific gene in a particular cell, the conditional Phase II probability is computed given both the cell and gene context. Figure 2A is an example of all genes in a cell, from the T2D dataset. Their probabilities of being in Phase II increase as counts increase, and essentially approach 1 for genes with counts greater than 20. There is a great deal of variability among genes as we do not make a simple cutoff for all genes in a cell. A gene (red circled) with a count as high as 18 is inferred to be most likely in Phase I, while another gene (green circled) as low as 6 is inferred to be probably in Phase II. This may appear counter intuitive, but Figure 2B explains the difference. The red gene is observed to have counts over several hundred in general, making the observation of 18 an extreme outlier. In contrast, the green gene has much lower expression. Figure 2C shows the Phase II probability for these two genes against observed counts across cells. Again, there is not a perfectly monotonic relationship because different cells have different Phase I parameters.

3.3 Examples of different forms of DE

With latent phases of a gene's expression inferred, we are able to detect DE in different forms: a difference in the Phase II proportion between conditions, or a different level of expression, or a combination of both? Here we show DE detection examples from comparing alpha cells between Type II diabetic patients and controls in the T2D dataset.

Figure 3 illustrates examples of four forms of DE identified. Figure 3A shows a gene that is more prevalent in T2D cells (78% off in non T2D cells versus 36% off in T2D cells), but among the cells that do express the gene, the mean expression level and spread are similar in both populations. Figure 3B shows a gene is expressed in the majority of cells regardless of disease status, but the mean expression level is higher in T2D cells. Figure 3C shows a gene that is up-regulated in T2D cells in both types of DE regulation: the gene is more likely to be turned on in T2D cells, and when it is turned on the magnitude of expression tends to be higher. These three forms of DE lead to a difference in average expression between two cell populations, which can potentially be detected by bulk RNA-seq as well, though the mechanism of regulation would not be identifiable.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty329/4984507
by Jackson Laboratory user
on 29 June 2018

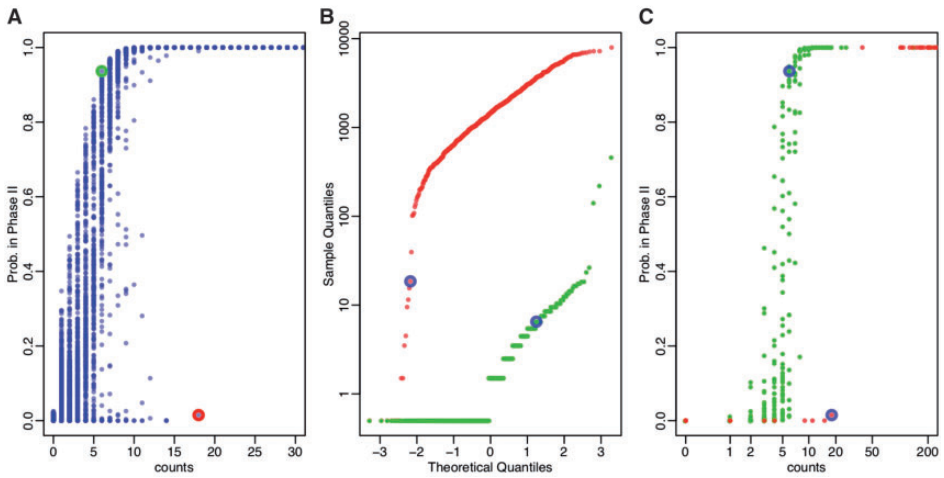


Fig. 2. Cell- and gene-specific phase determination. Data are from the in the T2D dataset. (**A**) Estimated probabilities of being in Phase II given observed counts, for all genes in one cell. (**B**) For two genes highlighted in Panel A, normal quantile–quantile plot of their counts across all cells. Their counts in the cell shown in Panel A are circled. (**C**) The estimated probability of Phase II for the same two genes plotted against observed counts across different cells

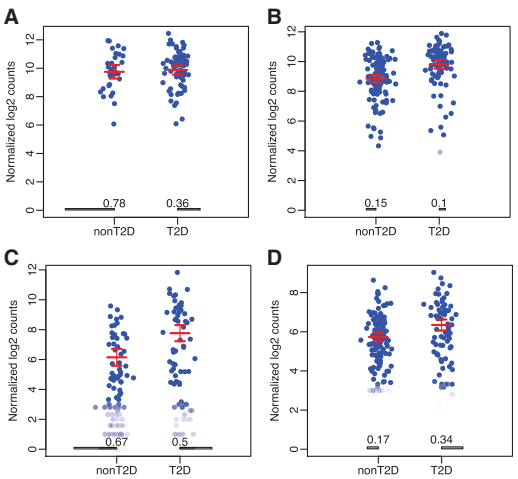


Fig. 3. Examples of different forms of differential expression from the T2D dataset: (**A**) phase transition alone (Form I *P*-value = 4.41e-11, Form II *P*-value = 0.98); (**B**) magnitude regulation only (Form I *P*-value = 0.22, Form II *P*-value = 4.9e-06); (**C**) phase transition and magnitude regulation in concordant manner (Form I *P*-value = 8.42e-03, Form II *P*-value = 4.04e-05); (**D**) expression compensation (Form I *P*-value = 3.91e-03, Form II *P*-value = 1.61e-05). Each figure plots the expressions for a particular gene from all cells. The bars at the bottom of the figures represent the proportions of cells not expressing the gene (in Phase I). Each dot represents the log expression values for this gene from a cell. (*P*-values are from the proposed SC2P method)

Most interestingly, we also observe a form of DE that achieves a 'compensation' effect in expression. Figure 3D shows a gene that is turned on in a smaller proportion of T2D cells (83% non-T2D cells have the gene on, versus 68% of T2D cells), but among those cells that do express the gene, the expression level is higher on average in T2D cells. Genes that undergo DE in this form may end up with similar level of average expression between cell populations, and may not be identified by bulk RNA-seq, or any analysis that only seeks marginal differences.

These examples demonstrate the importance of identifying DE in both forms in order to gain a full understanding of the mechanism of DE. From our proposed method, SC2P reports the estimated proportions of cells in phase I/II, the fold change in phase II, and false discovery rate (FDR) associated with either type of DE, thus provides more comprehensive information for DE detection.

3.4 DE detection comparison with existing methods

We compare the DE detection performance of SC2P with several existing methods: SCDE (Kharchenko *et al.*, 2014), BPSC (Vu *et al.*, 2016) and MAST (Finak *et al.*, 2015).

3.4.1 SC2P has higher sensitivity without sacrificing false discovery control

First, we validated the ability to identify known DE genes. There is a lack of gold standard for true positives in data, but more than 20 cell type marker genes are given in the human brain dataset. These marker genes are identified by comparing purified cell types via bulk RNA-seq (Darmanis *et al.*, 2015; Zhang *et al.*, 2014). They provide a partial list of true positives with strong signal, thus the ability to recover these genes among the top genes declared as DE is a reasonable validation of sensitivity.

Figure 4 shows the results from human brain dataset, comparing astrocytes and oligodendrocytes cells. Figure 4A compares the ability to recover known marker genes from the top ranked DE genes reported by four methods. Overall, SCDE, MAST and SC2P provide comparable overall results, and BPSC performs unfavorably. In addition, there are many more marker genes belonging to the Form I DE than Form II, indicating that the phase transition is more prevalent than magnitude adjustment between cell types. Even though SCDE reports these genes as DE, this mechanism is not revealed. The results for recovering DE in known markers in other comparisons are provided in Supplementary Material (neuron versus oligodendrocyte in Supplementary Fig. S3, and astrocyte versus neurons in Supplementary Fig. S4), and they lead to the same conclusion.

We focus on the comparison with MAST hereafter since it is the only other method in the group that also provides the functionality of testing DE in two phases. Figure 4B shows MAST and SC2P identify many genes in common for both forms of DE, with SC2P being much more sensitive, when both methods control FDR at 0.05. To ensure that the high sensitivity of SC2P is not achieved by sacrificing the control of false discoveries, we performed the following permutation test to assess the type I error control from DE tests.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty329/4984507
by Jackson Laboratory user
on 29 June 2018

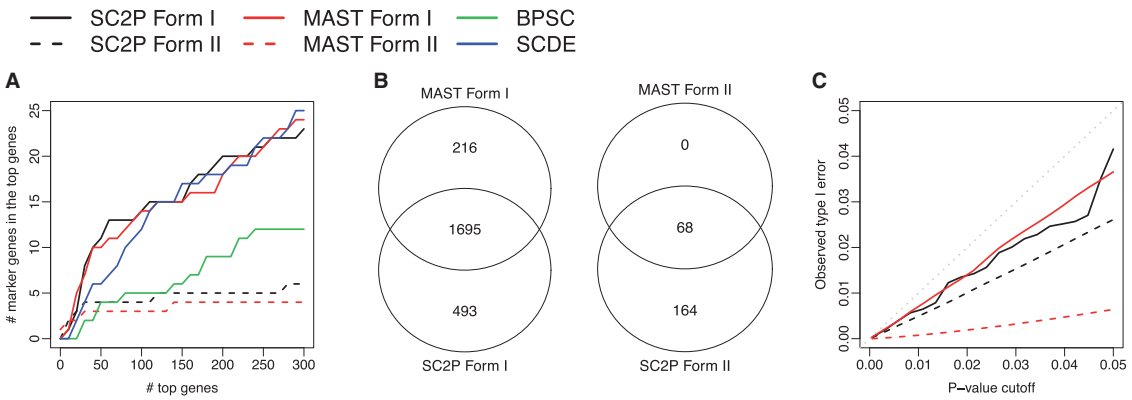


Fig. 4. DE detection in human brain data, for astrocytes and oligodendrocytes comparison. Genes with FDR <0.05 from the statistical tests are deemed DE. (**A**) Recovery of known marker genes among top ranked DE genes; (**B**) Overlaps of DE genes in both forms from MAST and SC2P; (**C**) Assessment of type I error control based from permutation

We randomly shuffle the cells among two conditions, and then perform DE test on the shuffled dataset. All DE genes detected from the shuffled dataset should be false positives, and the resulting *P*-values from the DE test on shuffled dataset should follow uniform distribution. We then compute the observed type I error rate for a given *P*-value threshold, and compare that with the nominal *P*-value to evaluate the type I error control from the statistical test. Figure 4C shows that the observed type I errors based on a permutation test for both SC2P and MAST are well controlled and below the nominal type I error for both forms of DE detection, validating that the higher sensitivity of SC2P is not from inflated type I error. Comparison between other human brain cell types and the T2D data (Supplementary Figs S3–S6) lead to the same conclusion. These results show SC2P has better sensitivity than MAST, while having the comparable type I error control and accuracy in ranking genes.

We obtained the lists of different DE genes called by SC2P and MAST. The heatmap of these gene's expression data are presented in Supplementary Figures S7 and S8. These figures provide a direct visualization of the raw data, thus are not obscured by the choice of modeling or processing, though they do not provide quantitative assessment of performance.

3.4.2 Robustness of SC2P

One critical property of any DE detection method is the robustness: that the discoveries are not sensitive to a few outliers. When we declare that a gene is differentially expressed between two cell populations, this result should not be driven by only a few cells. In other words, the analysis should be highly independent on the inclusion or exclusion a few cells, which are random samples from cell populations we study.

We compared the *P*-values obtained from the full dataset with the *P*-values from reduced dataset obtained by randomly removing 5% of the cells, from a population of 100 cells total in the neurons versus oligodendrocytes comparison. The panels in the second row of Figure 5 show excellent concordance between the two sets of *P*-values in DE detection from SC2P, in the testing of both forms of DE. In contrast, the panels in the first row of Figure 5 show such comparison results from MAST, which present substantial difference between results from the full data and reduced data. Most strikingly, we observe qualitatively different answers between the two sets of *P*-values: there are non-trivial amount of genes reported to have extreme statistical significance (with $\log_{10}$ *P*-value ≤ -10) when using all cell, but become non-significant (with log *P*-value

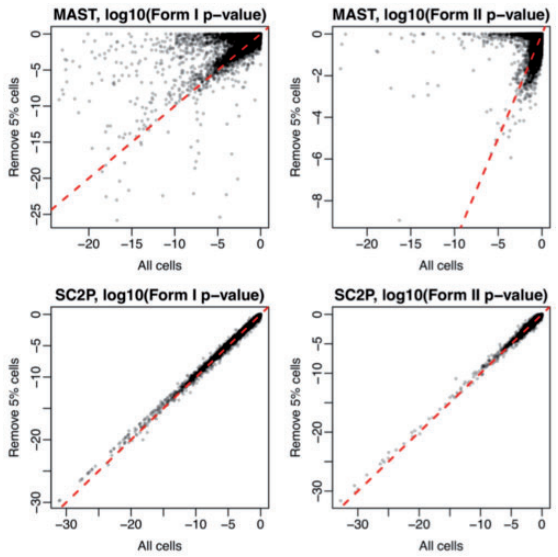


Fig. 5. Robustness of DE detection. Figures show comparison of *P*-values from testing DE using all cells in the dataset or a subset with 5% cells randomly removed, in the human brain data (neurons versus oligodendrocytes). Form 1 (phase transition) and Form 2 (magnitude difference in phase II) DE are compared separately

near 0) when 5% cells are excluded. This contrast in robustness is observed in both DE in phase transition and in expression level within Phase II, in both datasets.

We ran such analyses for 10 times, each time randomly selected 5% of the cells are removed. We observe that at least 5 out of the 10 times, the *P*-values from MAST show significant discordant. On the other hand, SC2P shows great consistence in all cases. The scatterplots for all 10 runs are provided as Supplementary Figure S9. We further perform additional analyses by removing 10, 20 and 50% cells. Each analysis is run 10 times. In each scenario, we compute the Pearson's correlation coefficients of *P*-values before and after removing cells. The distributions of the correlation coefficients from MAST and SC2P are presented in Supplementary Figure S10. In all scenarios, SC2P has much higher correlations than MAST, indicating better robustness. These results show that compared to MAST, SC2P is much more robust to outlier cells, benefited from our method for estimating transcription phases.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty329/4984507
by Jackson Laboratory user
on 29 June 2018

Table 1. Time (in seconds) required for MAST and SC2P

# cells	100	200	500	1000	2000	5000
MAST	211.8	297.4	476.9	756.6	1214.6	2897.9
SC2P	63.3	85.7	160.0	285.3	574.4	1704.4

3.4.3 Simulation

We conduct simulation studies to further compare the DE detection performances from MAST and SC2P. The simulated data are generated based on the human brain data so that they mimic the real data characteristics. The detailed simulation procedures and results are presented in Supplementary Material Section 8 and Supplementary Figures S13–S15. Overall, the simulation results are consistent with the real data results: SC2P are MAST provide comparable gene ranks, but SC2P is more sensitive due to better statistical inference.

3.4.4 Comparison with DESeq2

DESeq2 (Love *et al.*, 2014) is a very popular tool for detecting DE genes in bulk RNA-seq data. Though it is not specifically designed for scRNA-seq, it is worth exploring its performance in scRNA-seq DE detection. We ran DESeq2 on the *brain* and *T2D* datasets and compared its performance with other methods. The results are presented in Supplementary Material Section 9 and Supplementary Figures S16–S20. In terms of recovering known marker genes, DESeq2 fell below the group of better performers (SCDE, MAST and SC2P) but was better than BPSC. DESeq2 tended to identify many more genes as DE at any FDR cutoff ranging from 1% to 20%, at a cost of inflated type I error. Though DESeq2 could discover many genes that were identified by SC2P or MAST or both, its observed type I error was much greater than nominal type I error, meaning it identified many more false positives than expected. In addition, since DESeq2 tests for the mean expression difference between groups, it does not reveal whether the form of DE involves phase transition. Overall, these drawbacks make DESeq2 undesirable for DE analysis for scRNA-seq data.

3.5 Computational performance

SC2P provides excellent computational performance. We profiled the times required for different methods to run DE analyses. All profiling was done on a MacBook pro laptop with i7 2.7 GHz CPU and 16 G RAM. When there are 100 cells in each group, SC2P takes 63.3 s, MAST takes 211.8 s and BPSC takes 3167.6 s. SCDE recommends to run on multiple cores. On a single core, it did not finish after five hours. So we focus on the comparison between SC2P and MAST. Table 1 summarizes the times (in seconds) required for different numbers of cells. Overall, SC2P is 2-3 times faster than MAST.

4 Discussion

Transcription is a complex process that is usually divided into three phases, including initiation (in higher eukaryotes, this is followed by the pause and release from pause of RNA Pol II), elongation and termination (Venkatesh and Workman, 2015). These steps are under regulation in various extent. The initiation, for example, involves intricate cooperation of multiple complexes in the disassembly of nucleosomes that creates a nucleosome-depleted region (NDR) which makes the DNA accessible to Pol II. Maintaining the NDR also allows multiple rounds of transcription to take place. Once initiated (and released from the pause), multiple factors affect the elongation speed, hence the production rate of RNA transcripts. The number of transcripts of a particular gene depends on both the production and degradation rate. Real-time measurements of transcription activity, taken from fluorescence in situ hybridization (FISH) in individual cells, indicate that genes transition between inactive and active states of transcription (Dar *et al.*, 2012; Raj *et al.*, 2006). The transition from inactive to active state leads to pulsatile expression patterns often referred to as bursting. As a result, we observe, in scRNA-seq data, gene expression counts that exemplify two modes of regulation: one mode that accounts for a binary transition from an inactive phase (Phase I) into an active, high expression phase (Phase II) and another mode that accounts for a regulation of the expression level within Phase II.

With bulk RNA-seq, the average expression of a large population of cells is measured, masking the heterogeneities among cells. scRNA-seq makes it possible to understand the transcriptional variation at the single cell level, providing evidence of bimodal expression regulation. However, the detection of DE has either remained as a comparison of the mean expression (Trapnell *et al.*, 2014; Vu *et al.*, 2016), or with arbitrary cut off for expression phases. In this work, we advocate that the DE test in scRNA-seq should be performed in both modes: phase transition and magnitude tuning. To achieve that, a vital first step is to accurately estimate the phases of expression for all genes in all cells. We present evidence that there are differences in overall detection rate among cells, and this is positively correlated with but different from the non-zero percentage (Supplementary Figs S10 and S11). This simple but effective method provides DE identification with increased sensitivity without sacrificing specificity, as well as greatly improved robustness. Furthermore, the results provide better interpretation of the DE regulation mechanism.

The excess of zero counts in scRNA-seq data is observed widely, though the source of these zero counts is debated. Some treat zero as unexpressed (Finak *et al.*, 2015), others consider the zeros as technical dropouts and use imputation to recover the unobserved expression (Huang *et al.*, 2017; Lin *et al.*, 2017; Zhu *et al.*, 2016). There are definitely technical dropouts, especially in low depth sequencing. On the other hand, the genome accessibility varies among cells (Buenrostro *et al.*, 2015; Thurman *et al.*, 2012) and transcription is certainly not active throughout the entire genome in a given cell. Therefore, we believe that both biological and technical reasons contribute to observed zero counts. Since scRNA-seq measures the quantity of RNA in a cell, not the transcription activity itself, even in inactive phase, there are RNA molecules already transcribed and not completely degraded. This is consistent with data from FISH experiments, in which cells without active transcription sites have fewer but non-zero reporter mRNA (Raj *et al.*, 2006). Thus we argue that zero counts (as well as very low counts) are 'lack of evidence' for active transcription.

Existing threshold-based methods for phase determination fail to properly account for important data characteristics, including the variation of Phase I counts across cells. A major contribution of our method is providing data-driven thresholds that account for technical and biological factors, and both the cell- and gene-specific characteristics in the determination of expression status. Our current model only considers gene-specific factors in Phase II, while treating the Phase I parameters as if they were the same across genes. This is a choice for computational simplicity, as the variability due to Poisson counting at low counts makes it difficult to identify small difference in the Poisson rate. However, as public data accumulates, we will be able to observe a gene's expression over a wide variety of conditions and in very large sample sizes. With multi-experiment

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty329/4984507
by Jackson Laboratory user
on 29 June 2018

databases we will be able to extend the model to estimate gene-specific patterns in both phases. Our long term goal is to establish gene-specific priors for both phases to accurately infer DE in cell- and gene-specific context. Such work on large scale databases have been presented on microarray platforms (McCall *et al.*, 2011, 2014), which we predict will be greatly improved by single cell level data.

We model Phase II expression with a LNP distribution, instead of Gamma-Poisson (negative binomial), which is the most common choice for bulk RNA-seq data (Anders and Huber, 2010; Love *et al.*, 2014; Robinson and Smyth, 2007; Wu *et al.*, 2013). The Gamma distribution is often a choice of mathematical convenience and it is very similar to lognormal when the dispersion parameter is small, which is usually the case in bulk RNA-seq, since the expression level is an average over a large collection of cells. When the dispersion is small, both the dispersion parameter in the Gamma distribution and the parameter σ^2 in the lognormal distribution correspond to the square of the biological coefficient of variation (BCV) (Wu *et al.*, 2013). However, when the CV is large and often exceeding 1 (Supplementary Material, Section 1), it would force the Gamma distribution to be extremely skewed and have a mode at 0, and lose its flexibility in shape. Using lognormal distribution to model the true expression rate not only allows better flexibility, but also allows easy extension to accommodate more complex study designs, such as mixed effects and nested design, by using existing methods for linear models on the log transformed data.

The datasets we tested do not use unique molecular identifiers (UMIs) (Islam *et al.*, 2014; Kivioja *et al.*, 2012), which are additional barcodes added to RNA transcripts before amplification. In UMI data, reads that map to a gene and share the same UMI are counted as originating from the same transcript, thus UMI data have much lower counts. Additional error correction of the UMIs in preprocessing may be necessary, and different normalization strategy is recommended (Stegle *et al.*, 2015). These factors complicate the assessment of DE detection, and we have not included such comparison in this paper. The lower count level in UMI data makes it more difficult to decompose the two latent phases. At this stage, the current methods including SC2P may not work well for UMI data, or data with low depth sequencing, in terms of detecting DE in the form of phase changes.

Funding

This work was partially supported by NIH/NIGMS award R01GM122083 for HW, by R01GM122083, NSF DBI1054905, P20GM109035 for ZW. Single cell transcriptome work in the Stitzel Lab is supported by the Assistant Secretary of Defense for Health Affairs, through the Peer Reviewed Medical Research Program under Award No. W81XWH-16-1-0130. Opinions, interpretations, conclusions and recommendations are solely the responsibility of the authors, are not necessarily endorsed by the Department of Defense.

Conflict of Interest: none declared.

References

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

Bacher,R. and Kendziorski,C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, **17**, 63.

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

Buenrostro,J.D. *et al.* (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486–490.

Buettner,F. *et al.* (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.

Dar,R.D. *et al.* (2012) Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci. USA*, **109**, 17454–17459.

Darmanis,S. *et al.* (2015) A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. USA*, **112**, 7285–7290.

Delmans,M. and Hemberg,M. (2016) Discrete distributional differential expression (d 3 e)—a tool for gene expression analysis of single-cell rna-seq data. *BMC Bioinformatics*, **17**, 110.

Finak,G. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.

Grün,D. *et al.* (2015) Single-cell messenger RNA sequencing reveals rare *intestinal* cell types. *Nature*, **525**, 251–255.

Huang,M. *et al.* (2017) Gene expression recovery for single cell RNA sequencing. doi: https://doi.org/10.1101/138677.

Islam,S. *et al.* (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, **11**, 163.

Jiang,L. *et al.* (2016) GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol.*, **17**, 144.

Jonkers,I. and Lis,J.T. (2015) Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.*, **16**, 167–177.

Kharchenko,P.V. *et al.* (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.

Kiselev,V.Y. *et al.* (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, **14**, 483–486.

Kivioja,T. *et al.* (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72.

Korthauer,K.D. *et al.* (2016) A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.*, **17**, 222.

Lawlor,N. *et al.* (2017) Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome Res.*, **27**, 208–222.

Lin,P. *et al.* (2017) CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.*, **18**, 59.

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

McCall,M.N. *et al.* (2011) The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res.*, **39**, D1011–D1015.

McCall,M.N. *et al.* (2014) The gene expression barcode 3.0: improved data processing and mining tools. *Nucleic Acids Res.*, **42**, D938–D943.

Ntranos,V. *et al.* (2016) Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol.*, **17**, 112.

Patel,A.P. *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.

Picelli,S. *et al.* (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, **10**, 1096–1098.

Raj,A. *et al.* (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.*, **4**, e309.

Ramsköld,D. *et al.* (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**, 777–782.

Robinson,M. and Smyth,G. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.

Shalek,A.K. *et al.* (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, **498**, 236–240.

Shalek,A.K. *et al.* (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, **510**, 363–369.

Shekhar,K. *et al.* (2016) Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, **166**, 1308–1323.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 1.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty329/4984507
by Jackson Laboratory user
on 29 June 2018

Stegle,O. *et al*. (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet*., **16**, 133.

Tang,F. *et al*. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.

Thurman,R.E. *et al*. (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75.

Trapnell,C. *et al*. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol*., **32**, 381–386.

Usoskin,D. *et al*. (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci*., **18**, 145–153.

Venkatesh,S. and Workman,J.L. (2015) Histone exchange, chromatin structure and the regulation of transcription. *Nat. Rev. Mol. Cell Biol*., **16**, 178.

Vu,T.N. *et al*. (2016) Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics*, **32**, 2128–2135.

Wills,Q.F. *et al*. (2013) Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol*., **31**, 748–752.

Wu,H. *et al*. (2013) A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, **14**, 232–243.

Zhang,Y. *et al*. (2014) An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci*., **34**, 11929–11947.

Zhu,L. *et al*. (2016) A unified statistical framework for RNA sequence data from individual cells and tissue. *arXiv preprint arXiv: 1609.08028*.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty329/4984507
by Jackson Laboratory user
on 29 June 2018

CellPress

Review

Genomics of Islet (Dys) function and Type 2 Diabetes

Nathan Lawlor,[1] Shubham Khetan,[1,2] Duygu Ucar,[1,3] and Michael L. Stitzel[1,2,3,*]

Pancreatic islet dysfunction and beta cell failure are hallmarks of type 2 diabetes mellitus (T2DM) pathogenesis. In this review, we discuss how genome-wide association studies (GWASs) and recent developments in islet (epi)genome and transcriptome profiling (particularly single cell analyses) are providing novel insights into the genetic, environmental, and cellular contributions to islet (dys) function and T2DM pathogenesis. Moving forward, study designs that interrogate and model genetic variation [e.g., allelic profiling and (epi)genome editing] will be critical to dissect the molecular genetics of T2DM pathogenesis, to build next-generation cellular and animal models, and to develop precision medicine approaches to detect, treat, and prevent islet (dys)function and T2DM.

Lay of the Land: (Functional) Genomic Landscape of Islets and T2DM

T2DM is a complex metabolic disorder with both genetic and environmental components. It results from the dysfunction and loss of insulin-secreting beta cells in the endocrine pancreas (Islets of Langerhans) as they work to secrete more insulin to counteract insulin resistance in peripheral tissues (adipose, skeletal muscle, and liver). Ultimately, T2DM manifests as uncontrolled elevations in blood glucose levels. **GWAS** (see Glossary) have systematically identified hundreds of **single nucleotide variants** (SNVs), representing >150 regions of the genome (loci) [1], that are associated with T2DM risk and differences in T2DM-related quantitative metabolic traits, such as insulin, proinsulin, and glucose levels. Most (>90%) of these SNVs reside in noncoding regions of the genome. In parallel, functional (epi)genomics approaches to map open chromatin using **DNase I hypersensitive site sequencing** (DNase-seq), **assay for transposase-accessible chromatin sequencing** (ATAC-seq), and histone modification and transcription factor (TF)-binding patterns using **chromatin immunoprecipitation sequencing** (ChIP-seq) have identified genome-wide location of regulatory elements (REs), such as promoters, enhancers, and insulators, in >150 human cell types and tissues. T2DM SNVs are significantly and specifically enriched in islet-specific REs [2–7], suggesting that changes in islet RE activity and target gene expression are a common mechanism underlying the molecular genetics of islet dysfunction and T2DM [8] (Figure 1A). Indeed, recent studies have identified putative factors binding these REs and have detected allelic effects on their binding and target gene expression [9–11].

In this review, we discuss how recent studies are improving our understanding of how islet REs are perturbed by SNVs contributing to T2DM risk [1,12–19] and are elucidating the transcriptional underpinnings of islet responses to (patho)physiological environmental changes, such as aging, circadian rhythms, Western diet and lifestyle, as well as oxidative, endoplasmic reticulum (ER), and inflammatory stress responses [20–25]. We explore how studies applying next-generation sequencing (NGS) to profile individual cells are improving our comprehension of islet biology and reshaping our view of T2DM pathogenesis. Finally, we examine similarities and differences between mice and humans in the 'omics of islet function and T2DM (summarized in

Trends

T2DM is a multi-tissue metabolic disorder that results when pancreatic islets fail to compensate for insulin resistance in peripheral tissues.

Recent studies reaffirm the common variant origins of T2DM genetic risk. Variants overlap noncoding genomic regions, implicating regulatory defects in T2DM etiology.

Environmental stressors are associated with changes in gene expression programs leading to T2DM progression.

Single cell sequencing technologies permit investigation of islet cell type transcriptomes and epigenomes with single cell resolution and/or precision. Such methods provide greater insight into cell type-specific perturbations and their roles in T2DM.

Recent studies suggest that other cells (alpha, delta, and PP/gamma) in the islet have important roles in islet/and/or beta cell function, resilience, and T2DM pathogenesis.

[1]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA
[2]Department of Genetics & Genome Sciences, University of Connecticut, Farmington, CT 06032, USA
[3]Institute for Systems Genomics, University of Connecticut, Farmington, CT 06032, USA

*Correspondence:
michael.stitzel@jax.org (M.L. Stitzel).

© 2017 Elsevier Ltd. All rights reserved.

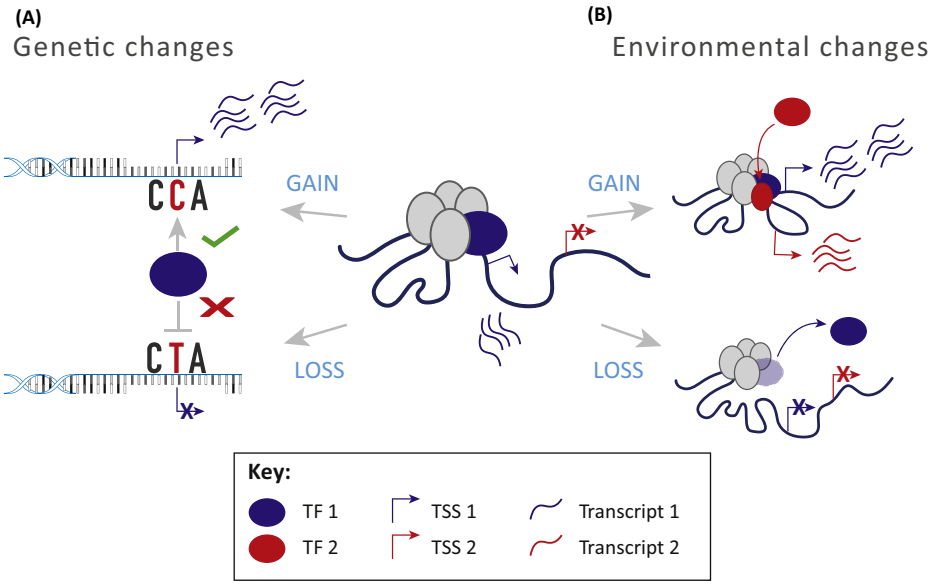


Figure 1. Genomic Effects of Genetic and Environmental Perturbations Contributing to Pancreatic Islet Dysfunction and Type 2 Diabetes Mellitus (T2DM). (A) DNA single nucleotide variants (SNVs) may enhance (gain-of-function) or diminish (loss-of-function) transcription element (e.g., enhancer) activity and islet gene expression. Most T2DM-associated SNVs reside in noncoding regions of the genome and overlap islet regulatory elements (REs) [2,3,12,14,15,32,47], implicating disruptions in gene regulatory network components as a central molecular feature in disease pathogenesis. A subset of SNVs has been linked to changes in basal islet gene expression [11,31]. (B) Environmental factors, such as inflammation, diet, aging, circadian rhythms, and stress, may also influence RE activity, resulting in altered and/or novel transcription of genes essential for islet function [20–25,48–50,57,58]. Abbreviations: TF, transcription factor; TSS, transcription start site

Figure 2, Key Figure). Throughout, we highlight future challenges and opportunities and offer perspectives on how these recent developments set the stage for precision medicine approaches to understand, treat, and prevent T2DM.

Homing in on T2DM Genetic Risk and Architecture

Since initial T2DM GWAS reports in 2007 [26–29], the list of genomic loci in which sequence variation contributes to T2DM risk and variability in quantitative measures of pancreatic islet function has grown to over 150 [1,14,30]. Associated SNVs at each locus contribute modestly to increased T2DM risk [odds ratios (OR) 1.05–1.75]. Together, these loci only explain a fraction of T2DM heritability [13,14]. Genetic consortia continue to dissect the genetic architecture of T2DM using larger cohorts with increasing ethnic diversity and/or representation. Recent efforts have reported [12,14,30,99] fewer 'new' T2DM loci (N=10) than previous studies. Importantly, however, they are refining the genetic signals at known (previously associated) T2DM loci to define **'credible sets' of single nucleotide polymorphisms** (SNPs) that are the most probable causal and/or functional SNPs driving the association and, consequently, the resulting molecular and/or phenotypic consequences.

The GOT2D and T2D-GENES consortia sought to identify less common SNVs (0.1%<MAF<5%) with larger effect size that may underlie common variant associations or may account for some of the T2DM 'missing heritability' using a combined whole-genome sequencing (WGS), exome sequencing, and genotype imputation approach [14]. These efforts identified protein-coding variants and/or mutations that are the most likely causative variant or effector transcripts for 12 out of 78 GWAS loci, confirming five nominated in previous studies (*PPARG, KCNJ11-ABCC8, SLC30A8, GCKR,* and *PAM* loci) and identifying seven

Glossary

Assay for transposase-accessible chromatin sequencing (ATAC-seq): a technique used to profile regions of open chromatin from small cell numbers.
Chromatin immunoprecipitation sequencing (ChIP-seq): a method used to study DNA–protein interactions.
Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET): a method used to study 3D chromatin interactions genome wide.
CpG sites: areas of DNA containing a cytosine nucleotide directly linked to a single phosphate group and guanine nucleotide. These sites are often methylated and influence transcription.
Credible sets of SNPs: lists of sequence variants with 95% posterior probability of containing a/the causal disease-associated SNP (s) [99].
Deconvolution: a statistical framework to resolve a heterogeneous mixture into its constituent elements.
Dedifferentiation: the process in which a mature differentiated cell type reverts to an earlier developmental and/or precursor state.
DNA methylation: molecular process wherein a methyl group is covalently attached to a DNA base without altering the DNA sequence.
DNase I hypersensitive site sequencing (DNase-seq): a method used to characterize regulatory and open chromatin regions of the genome.
Expression quantitative trait loci (eQTL): approach to link sequence variation at a position in the genome to expression of target gene(s).
Genome-wide association study (GWAS): statistical association of sequence variation with disease risk or variability in a measurable phenotypic trait and/or feature.
Glycated hemoglobin (HbA1C): a type of hemoglobin modification that is measured to determine plasma glucose concentration.
RNA-sequencing (RNA-seq): measures the amount of RNA in a sample at a given time.
Single nucleotide polymorphism (SNP): nucleotide variation at a specific location in the genome that exists with >5% frequency in the population.

new ones (*FES, TM6SF2,* and *RREB1* in the *PRC1, CILP2,* and *SSR1* loci, respectively, and *TSPAN8, THADA, HNF1A,* and *HNF4A*). For the remaining loci, noncoding SNVs constitute the putative causal SNVs. Comparison of multiple genetic models with the empirical data generated in this study suggest that a long tail of common variants with lower effect sizes may comprise the missing heritability and reaffirms the importance of common, regulatory variation in the genetic architecture of T2DM (see Outstanding Questions). Perhaps most importantly, this immense effort has narrowed the list of putative causal SNVs to a handful for five loci and by 50% on average for the 78 T2DM-associated autosomal loci investigated [14]. Similar themes and reductions in credible sets were reported for fasting glucose- and insulin-associated loci [30].

Ongoing islet epigenomic and transcriptomic analyses are progressively defining the regulatory potential of variant loci, identifying SNV-RE overlaps, and nominating potential target genes, whose dysfunction is likely to contribute to T2DM [2,3,11,12,14,15,30–32]. Open chromatin (DNase-seq, ATAC-seq) and histone modification and/or TF-binding profiling (ChIP-seq) indicate that T2DM and related trait-associated SNVs are especially prominent in islet distal REs and **stretch/super enhancers** [2,3,5,33,34]. Due to the long distances over which REs might act, additional work to elucidate the target genes of T2DM SNV-containing REs is needed. Chromosome conformation capture techniques, such as 3C, 4C, 5C [35], Hi-C [36], **chromatin interaction analysis by paired-end tag sequencing** (ChIA-PET) [37], and HiCHIP [38] will be important components to effectively map interactions between REs and their target genes (see Outstanding Questions). In two separate studies, **RNA-sequencing** (RNA-seq) of 89 [31] and 118 [11] human islet samples identified 616 and 2341 **expression quantitative trait loci** (eQTLs), respectively. These analyses were the first studies linking SNVs to gene expression changes in islets to define the putative genetic control of islet function and failure. However, of the 216 eQTLs common to both studies, only 14 overlapped with T2DM-associated loci [11]. This may be due to power limitations and an inability to detect eQTLs beyond their primary signal. Alternatively, this relatively low overlap could suggest that T2DM SNVs affect islet physiological or pathophysiological responses, not just basal expression, as has been measured to date. Indeed, a recent study suggested that several putative T2DM GWAS genes are regulated by NFAT, a TF involved in calcineurin signaling responses [39]. Alternatively, the detection of eQTLs overlapping T2DM-associated SNVs in peripheral tissues, such as skeletal muscle [40] and adipose [41] tissue, reminds us that these other metabolic tissues should not be ignored in the T2DM molecular genetics and pathogenesis, and warrant further investigation of genomic variation in these tissues.

Recent islet studies suggest that regulatory noncoding RNAs (ncRNAs) contribute to diabetes progression and beta cell (dys)function [31,42,43]. Aberrant expression of 17 long noncoding (lncRNAs) has been associated with **glycated hemoglobin** (HbA1c) levels [31]. This study identified eQTLs for two of these transcripts (*LOC283177* and *SNHG5*), but the eQTL SNVs did not overlap with T2DM SNVs [31]. Similarly, a study by Morán and colleagues identified nine out of 55 T2DM-associated loci that contained lncRNAs located within 150 kb of, but not directly overlapping, the reported lead SNVs [42]. In the *KCNQ1* locus, T2DM risk SNVs overlap both *KCNQ1* and *KCNQ1OT1* [43,44], a long intergenic noncoding RNA (lincRNA) also found to be significantly induced in T2DM islets [42]. We anticipate that additional links will emerge in the coming years. Other studies suggest that islet lncRNA alterations could also contribute to type 1 diabetes mellitus (T1DM), because a T1DM GWAS SNV (rs941576) was identified in the *MEG3* lincRNA locus [43,45]. Functional analyses in human islets and rodent models will clarify the roles of these ncRNAs in islet development, (dys)function, and diabetes.

DNA methylation studies of nondiabetic (ND) and T2DM islets have suggested that epigenetic dysregulation promotes T2DM development [46,47]. DNA methylation profiling of 15 T2DM and

Single nucleotide variant (SNV): changes in a given nucleotide sequence in the genome.
Stretch/super enhancers: extended (>3 kb) regions of the genome marked by enhancer chromatin states; enriched near genes that are important for cell type identity and cell type-specific functions.
Subpopulation: a subset of cells within a tissue distinguished by the expression of specific marker genes and/or proteins.
***Trans*-differentiation:** the process in which a mature cell type converts into another mature cell type.

Key Figure

Converging and Diverging Genetic, Environmental, and Cellular Aspects of Islet (Dys)function and Type 2 Diabetes Mellitus (T2DM) in Mice and Humans

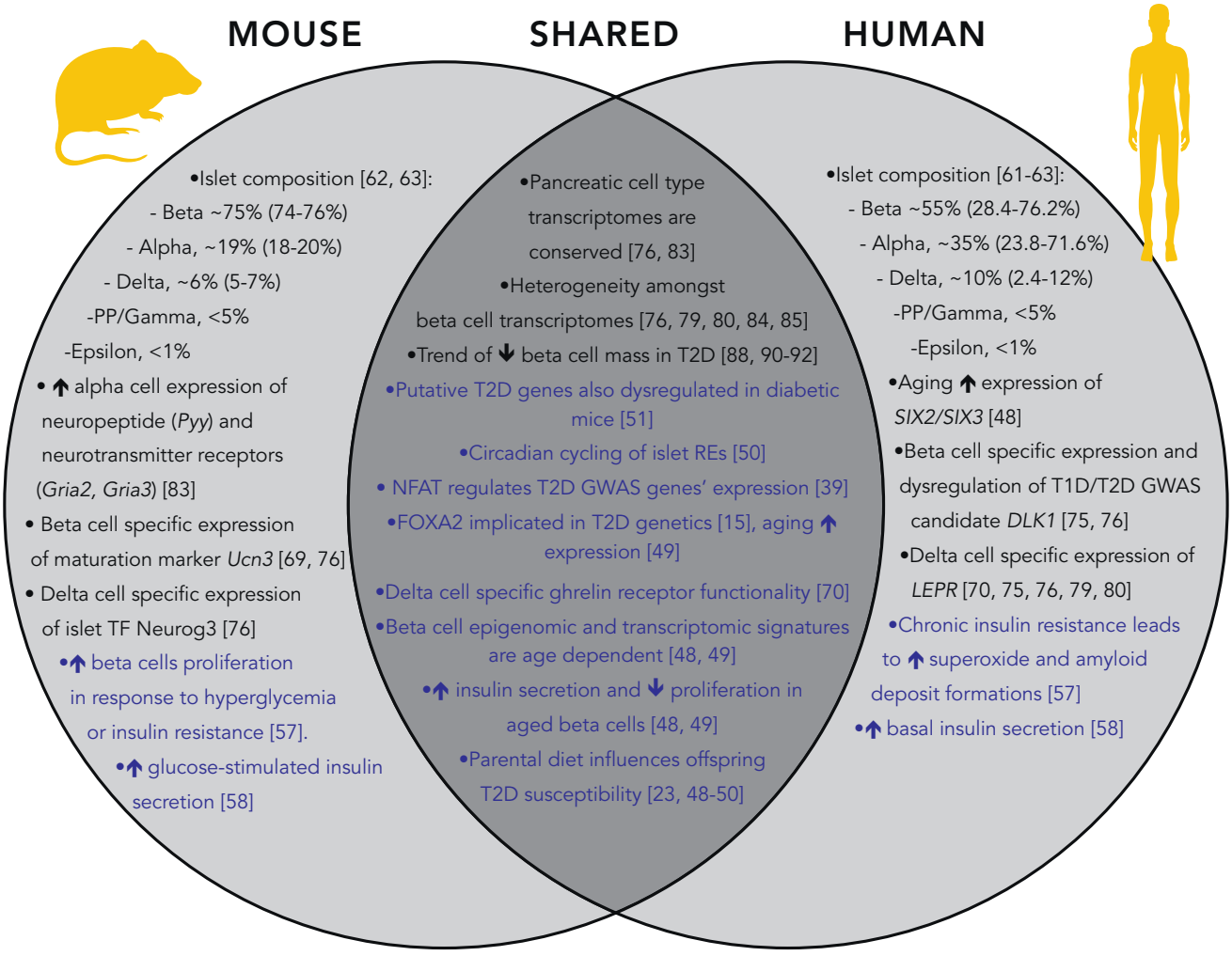


Figure 2. Parallel analyses of human and mouse islets are revealing important similarities [15,23,39,48–51,70,76,79,80,83–85] (A) and differences [48,57–58,61–63,69-70,75-76,79–80,83] (B,C) between molecular features of islet identity and (dys)function in mice and humans. Black text highlights significant findings regarding islet cellular composition and identity. Blue text highlights longitudinal and/or comparative analyses of genome-wide molecular data sets and environmental effects on islet (dys)function. These features reaffirm the value of modeling T2DM in mice to delineate important species-specific differences in islet biology that may reflect distinct T2DM causative mechanisms. Abbreviations: ↑, increase; ↓, decrease; GWAS, genome-wide association study RE, regulatory element; TF, transcription factor; T1DM, type 1 diabetes mellitus

34 ND islets using the Illumina 450BeadChip identified 1649 differentially methylated **CpG sites** for 853 genes, 17 of which reside in T2DM-associated loci [46]. Surprisingly, most (97%) of these CpG sites were hypomethylated in T2DM islets, suggesting that they suffer from decreased methyl donor levels or decreased activity of DNA methyltransferases.

Genomics of Islet Responses to Environmental Changes and T2DM Pathogenesis

Intrinsic and extrinsic environmental changes, such as aging, and Western diet and/or lifestyle, respectively, are linked to islet dysfunction and T2DM risk [23,48–50] (Figure 1B). Multiple groups have begun to characterize the genomic effects of these environmental inputs and insults on islets. Transcriptome profiling of adult and juvenile islet beta cells identified 565 (209 up, 356 down) and 6123 (2083 up, 4040 down) differentially expressed genes in humans and mice, respectively [48,49]. Signatures of decreased proliferative capacity in aged islets and/or beta cells were apparent in both species, perhaps best illustrated by increased *CDKN2A/B* expression, a gene cluster with established cellular senescence functions and implicated as 'Type 2 Diabetogenes' for a T2DM GWAS signal on 9p21 [48,49,51]. Unexpectedly, transcriptome and epigenome signatures suggested superior insulin secretory capacity of adult islets, which was confirmed functionally by glucose-stimulated insulin secretion (GSIS) assays [48,49]. DNA methylation and histone profiling indicated that these expression differences were largely mediated by chromatin remodeling and epigenetic modification of distal Res, such as enhancers. Using whole-genome bisulfite sequencing (WGBS), Avrahami and colleagues identified approximately 14 368 aging-related differentially methylated regions (DMRs) between the beta cells of juvenile and adult mice. DMRs overlapping distal REs outnumbered those overlapping promoters 3:1, and exhibited larger changes in magnitude of methylation. Distal DMRs that lost methylation with aging were enriched for binding sites of important islet TFs, such as Foxa2, Neurod1, and Pdx1, suggesting these factors mediate the expression differences and improved functionality in adult islets. Finally, genes showing differential expression in adult islets were accompanied by differential methylation at nearby distal REs more often than at their promoters. These data suggest that, in addition to their importance in T2DM genetic risk, enhancers also govern important transcriptional regulatory changes accompanying or mediated by aging.

Circadian rhythm links behavior and metabolism to day–night cycles. Notably, insulin secretion oscillates with a circadian periodicity. Analysis of mouse islet transcriptomes revealed that approximately 27% of the beta cell transcriptome (*N*=3905 genes) demonstrated circadian oscillation, including genes responsible for insulin synthesis, transport, and stimulated exocytosis [50]. The human orthologs of 481 of these genes exhibited circadian oscillations in human islets. ChIP-seq identified 742 out of 3905 of these oscillatory genes as direct targets of the circadian clock TFs CLOCK and BMAL1. As with aging, most differential sites were at distal REs. Beta cell-specific deletion of *Bmal1* resulted in islet failure and diabetes in mice. This study demonstrates the importance of circadian rhythms in islet function and suggests that genetic or environmental perturbation of this program contribute to T2DM risk and pathophysiology. GWAS results suggest that this could be the case, because SNVs in the *CRY2* locus, a component of the circadian machinery, and *MTNR1B,* a gene encoding a melatonin receptor, are associated with altered islet function and T2DM [1,52]. It will be interesting to see whether genetic perturbations in circadian clock TFs or their binding sites emerge as one of the molecular mechanisms underlying T2DM GWAS.

Maternal nutrition and *in utero* stresses have been linked to T2DM risk for offspring in humans and rodents [23,53–55]. Although changes in fetal nutrition are suggested to influence offspring metabolism via epigenetic modifications [23,56], the genome-wide effects on the islet (epi) genome have not been determined. Similarly, stress responses to elevated oxidative and/or ER stress lead to islet failure, impaired insulin secretion, and T2DM susceptibility [57–59]. Ultimately, these responses converge on the nucleus and involve the redistribution or covalent modifications of master TFs (MAFB, NKX6-1, and PDX1) or stress response factors (FOXO1, ATF4, and HIF1alpha) [20,22,53,57,58]. (Epi)genomic analyses of these stress responses are warranted and may reveal important connections between T2DM SNVs and altered islet stress

responses. Moving forward, it will be crucial to understand the extent to which genetic and epigenetic changes interact in T2DM pathogenesis (see Outstanding Questions). Response QTL (reQTL) and epigenome-wide association studies (EWAS) [56] should provide these important links (see Outstanding Questions). Indeed, studies of SNV effects on immune cell responses identified 121 reQTLs, 38 of which overlapped autoimmune disease-associated SNVs [60]. Specific factor(s) and pathway(s) activated by insulin resistance appear to differ between mouse and human islets [57,58] (Figure 2); thus, we emphasize that caution must be taken in study design and interpretation to interrogate this and possibly other islet responses.

Deconstructing Pancreatic Islet Cellular and/or Functional Heterogeneity

Islets comprise 1–5% of the pancreas and consist of at least five endocrine cell types performing coordinated but distinct functions and each producing a unique hormone in the islet: beta (insulin), alpha (glucagon), delta (somatostatin), gamma (pancreatic polypeptide), and epsilon (ghrelin) cells [61–64]. Precise understanding of islet molecular changes during T2DM development is likely complicated by variability in islet cell type composition. On average, islets comprise 55% beta cells, 35% alpha, 10% delta, and less than 5% and 1% gamma/PP and epsilon cells, respectively [61–63]. However, this can vary considerably between donors, with ranges of 28.4–76.2%, 23.8–71.6%, and 2.4–12% for beta, alpha, and delta cell compositions, respectively [61] (Figure 2). This cellular heterogeneity, combined with donor-to-donor variability, masks the molecular repertoire of each cell type and impedes clear understanding of the molecular programs perturbed in each cell type by T2DM pathogenesis.

Until recently, most studies had focused on epigenetic and transcriptional analyses of whole islets or, at the expense of other cell types, beta cells. However, recent studies demonstrating roles for alpha [65–67] and delta cells [68–71] in modulating beta cell function and/or resilience and in T2DM pathogenesis are fueling renewed interest in these cell types. First attempts to overcome these obstacles and understand the molecular repertoire of each islet cell type focused on transcriptomic analyses of sorted and enriched cell type populations [61,72–74]. However, such methods were unable to effectively isolate and enrich the less abundant nonbeta cells [75], leaving much of the functional genomic landscape of islets imprecisely assigned and/or classified or, in the case of rarer islet cell types, undefined.

Within the past year, multiple groups have applied single cell transcriptome profiling to islets to begin to address questions about islet physiology [75–83] (see Outstanding Questions) with single cell resolution, such as: (i) what is the gene repertoire of each islet cell type? (ii) Does the gene repertoire reveal any new and/or unexpected roles for each cell type in islet (patho) physiology? (iii) Are there novel cell types or unappreciated **subpopulations** in islets? These studies are providing new appreciation of the repertoire of both islet beta and nonbeta cells. Given that much of the beta cell transcriptional repertoire has been extensively studied [61,72–74], several features have been validated, including genes involved in cell survival and/or maturation (*PDX1*), regulation of insulin secretion (*RGS16, SYT13,* and *ENTPD3*), and diabetes-associated genes (*DLK1, MEG3,* and *SLC2A2*) [75,76,78–81,83]. Unique expression of genes encoding TFs (*IRX2*), membrane glycoproteins (*DPP4*), and hormone transporters (*TTR*) were also validated in alpha cells. Analysis of single alpha cell transcriptomes uncovered signatures involved in wound healing (*FAP*), blood clotting (*F10*), and tissue biogenesis (*LOXL4*) [75,76,78–81,83], suggesting that they share functions akin to pancreatic fibroblast and/or mesenchymal cells.

Single cell profiling has provided new views of the roles of delta and PP/gamma cells in islet physiology and the molecular genetics of islet failure and diabetes. For example, these studies revealed that delta cells uniquely express appetite-suppressing leptin (*LEPR*) and appetite-stimulating ghrelin (*GHSR*) hormone receptors [75,79,80], implicating them as the integrators

and regulators of these pathways in the islet. *GHSR* functionality has been demonstrated in both human and mouse delta cells [70]. *LEPR* expression is unique to human delta cells, suggesting that these cells uniquely mediate the leptin response in human islets [70,75,76,79,80] (Figure 2). Expression of genes associated with congenital hyperinsulinemia (CHI) (*UCP2* and *HADH*) in delta cells further implicates this cell type in the molecular genetics of CHI [75]. PP/gamma cell transcriptomes exhibited enrichment of genes involved in neuronal development (*MEIS2* and *FEV*) [75,78–80] and serotonin catalysis and reuptake (*TPH1* and *SLC6A4*) [75,79,80,83]. Together, these findings suggest that delta and PP/gamma cells act as the 'brains' of the pancreatic islets, capable of receiving and integrating various neuronal signals to coordinate islet function. Due to their scarcity in human pancreatic islets (<1% of islet volume), our knowledge of the epsilon cell repertoire and its putative function(s) remain speculative. Nonetheless, the insights gleaned from these initial studies undoubtedly motivate follow-up studies that continue transitioning from whole-islet to functional constituent cell studies. Identification of genes encoding cell type-specific surface markers (beta, *LRRTM3* and *CASR*; alpha, *DPP4* and *PLCE1*; delta, *LEPR*, *GHSR*, and *ERBB4*; PP/gamma, *SLC6A4* and *PTGFR*; and epsilon, *ANXA13*) [75,79] provide new targets that may be exploited for more accurate purification of each islet cell type and analysis of its specific responses to genetic and environmental stressors.

Islet Subpopulations and Cell Type Heterogeneity

Detection of heterogeneous beta cell subpopulations was reported for enriched cell and single cell studies. These include four subpopulations with differing expression of *ST8SIA1* and *CD9* [84], five subpopulations defined by *RBP4, FFAR4/GPR120, ID1, ID2*, and *ID3* expression [80], and subpopulations characterized by ER stress-associated [76] and oxidative stress-associated genes [79]. *Fltp/CFAP126* expression has been reported to distinguish proliferating and mature beta cell subpopulations in mice [85], but single cell transcriptome analyses failed to detect this distinction in human beta cells [75,83]. However, proliferative and mature human beta cells were identified by single cell mass cytometry analysis [86], suggesting that mice and humans make use of distinct cell growth pathways. Given that each study detected distinct beta cell subpopulations with different gene signatures, it remains difficult to distinguish whether these subpopulations are functionally distinct cells or the result of technical confounders, such as the time to sort and enrich in a harsh cell sorting environment, time for cell capture, or cell and transcript capture efficiency [87].

Single Cell Dissection of Islet Dysfunction and T2DM

Single cell transcriptome analyses provide a fresh and agnostic opportunity to investigate the putative mechanisms underlying islet dysfunction in T2DM. To date, single cell transcriptome profiling has been completed for a total of 1831 and 1970 islet cells from 26 ND and 15 T2DM donors, respectively [75,80,81,83]. Comparison of T2DM and ND single cell transcriptomes suggest that specific alterations in islet cell type transcriptomes underlie T2DM pathogenesis (Figure 3A). However, changes in cell proportions (Figure 3B), identity, and plasticity (Figure 3C, D) have also been regarded as potential contributors to T2DM [72,88–92]. Specifically, decreases in diabetic beta cell mass were suggested to be caused by reversion to endocrine progenitor (hormone-negative) cells (Figure 3C) or different islet cell types (Figure 3D) rather than to apoptosis. The model of transformed beta cell identity remains controversial. A recent study concluded that the observed magnitude of decline in beta cell numbers in T2DM islets is not accompanied by proportionate increases in cells exhibiting ***trans*-differentiation** markers or increases in other islet cell types [93]. Rather, the presence of endocrine progenitor-like cells in T2DM islets may represent newly forming endocrine cells [93]. Single cell profiling also did not identify transcriptomic evidence of **dedifferentiated** or *trans*-differentiated cells in T2DM islets (Figure 3C,D) [75,80,83]. Similar trends were observed in whole-islet RNA-seq data upon **deconvolution**, where cell type proportions did not significantly vary between hypoglycemic

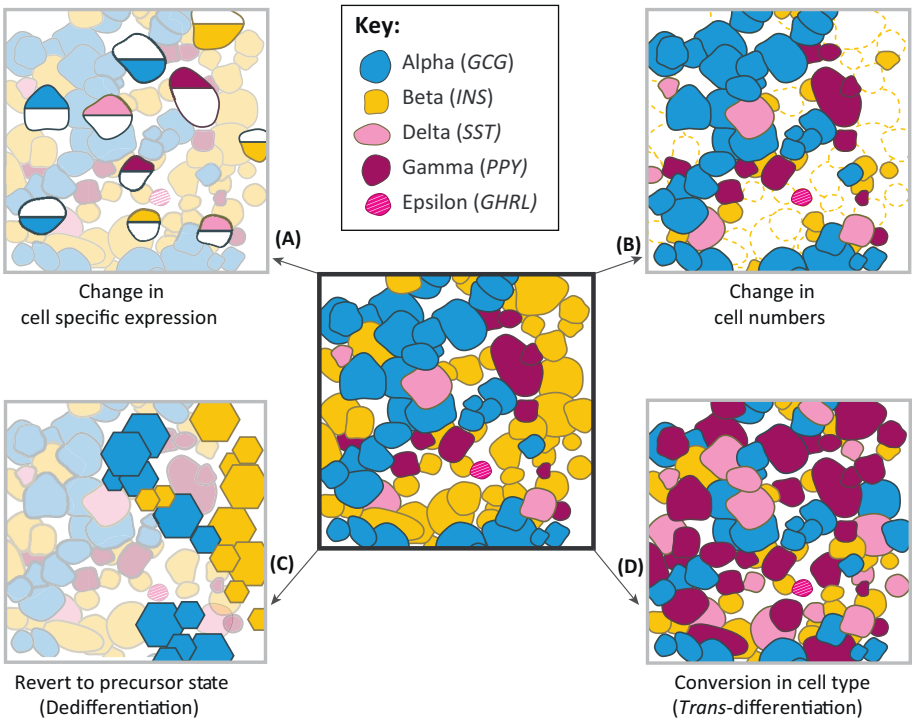


Figure 3. Proposed Cellular Mechanisms Contributing to Type 2 Diabetes Mellitus (T2DM) Development. (Center) Cartoon representation of human islet cellular composition. Studies have described the following phenomena: (A) Islet single cell transcriptomic studies [75,80,83] suggest that cell type-specific changes in gene expression (depicted as half-shaded cells) contribute to T2DM pathogenesis. These studies suggest that potential pathogenic expression changes occur in each islet cell type, not just beta cells. (B) Decreases in beta cell (in yellow) numbers [25,92,100,101], thought to precede islet dysfunction and development of insulin resistance. (C) Alterations in islet cellular identity may also account for islet failure. Dedifferentiation of islet cell types to precursor cell types and/or states (hexagons) has been proposed to underlie the loss of beta cell mass and function in T2DM [88,90–92]. (D) Similarly, *trans*-differentiation of islet cell types may lead to imbalances in islet cell proportions and improper function [72,88,89]

and hyperglycemic islets [76]. Thus, the transcriptome data to date do not provide supporting evidence of dedifferentiation in T2DM islets.

Transcriptomes of each cell type from ND and T2DM donors exhibited remarkable correlation overall. However, specific changes in gene expression were reported in T2DM beta cells, including reduced expression of *INS* [75,80], genes important for insulin secretion (*STX1A*) [75] and beta cell proliferation (*FXYD2*) [80,83], as well as elevated expression of genes implicated in T2DM GWAS (*DLK* and , *DGKB*) [75]. Transcriptional differences were also identified in T2DM alpha cells, including expression of *CD36* [75,80], a crucial activator of the NLRP3 inflammasome [94], and *RGS4,* a negative regulator of GSIS [80]. Several genes were dysregulated in T2DM delta cell transcriptomes [75,83]. However, the underlying biology of these candidates remains undefined, with no association with islet growth or function [83]. Aside from these encouraging examples, these single cell studies have not reached consensus regarding differentially expressed genes between T2DM and ND cell types. Differences in islet donor variability, islet isolation and/or transport, and single cell dissociation and/or sequencing protocols may explain these inconsistencies across studies. We expect that sampling thousands of single cells each from hundreds of individuals for large-scale meta-analyses will provide a more convergent list of cell type-specific genes and pathways disrupted in T2DM islets. It will also be important for future studies to profile cells from individuals at different points

along the T2DM pathogenesis spectrum, such as prediabetic individuals (5.5<$HbA1_c$<6.0) to identify and distinguish primary from secondary genomic changes that may be the cause or consequence of progression to T2DM.

Concluding Remarks and Future Directions

The past few years have marked exciting developments in our understanding of the underlying genomic, environmental, and cellular components driving T2DM pathogenesis. Numerous common (and only few rare) genetic SNVs have been implicated in T2DM progression [13,14]. It is unclear whether the 'missing T2DM heritability' is explained by a larger distribution of common SNVs with minimal effect sizes, whether current methods have missed critical rare SNVs, or whether it will be captured by gene–gene and gene–environment interactions (such as detected by reQTL). Thus far, most catalogued T2DM-SNVs occur in, and disrupt, islet RE function; however, the causal connections between the two remain challenging to decipher. eQTL and chromatin accessibility QTL (caQTL) [95,96] studies have been, and will continue to be, essential for linking genetic variants to molecular phenotypes. A subsequent challenge will be to link these molecular effects to pathways [39] and (patho)physiological phenotypes [97].

Functional genomic studies have identified minimal overlap between islet eQTLs and T2DM-SNVs [11,31], suggesting that responses to environmental stress factors are key mediators of T2DM pathogenesis. Mouse models have been instrumental in elucidating the genetic and molecular regulation of these responses and how environmental stressors influence islet (dys) function. However, observed differences between mice and humans in islet morphology, composition, expression, and function remind us to exercise caution when extrapolating findings in mice to human T2DM. Studies comparing the genomic features of human islets and models are essential to define conserved features and those that require modification to determine what aspects of islet dysfunction and T2DM we can model effectively and to decide how and/or where we should manipulate or humanize the mouse (epi)genome to better model human T2DM. (Epi)genome editing technologies, such as CRISPR/Cas9, can then be applied to develop new cellular and animal models to more effectively study islet phenotypic changes resulting from genetic and environmental variation. We anticipate that these integrative genomic studies and techniques will also serve as valuable resources to determine the underlying genetic changes and mechanisms of beta cell dysfunction that lead to T1DM [98].

Rapid developments in single cell NGS technologies have renewed interest in the less-studied islet cell types. Deconstructing the major molecular changes that occur in each cell type during T2DM progression has proven challenging, yielding inconsistent results between studies due to patient donor variability and technical sequencing artifacts. This is also likely the result of limited statistical power. In the future, it will be interesting to perform meta-analyses of available transcriptomic data to maximize our confidence of changes in cell specific expression programs. Together, the innovative new genomic technologies of the past few years will allow us to more precisely define, model, and manipulate the genes and pathways that have gone awry in T2DM, with the ultimate goal of designing novel therapeutic approaches.

Outstanding Questions

T2DM-associated GWAS variants explain only a small portion of T2DM heritability, with rare variants showing minimal contribution. Does a long tail of common variants with small effect sizes explain this missing heritability? Or are we simply 'underpowered' to detect rare variants and their contribution to T2DM heritability?

What are the genes targeted by T2DM GWAS sequence variant (SV)-containing regulatory elements? Are these links context specific? Does the risk allele enhance (gain-of-function) or repress (loss-of-function) RE function?

How do the transcriptomes and/or epigenomes of islets and islet cell types change when subjected to variable environmental stressors (oxidative stress, inflammation, diet, etc.)? How are they changed by intrinsic (aging, circadian rhythms, etc.) environmental factors? Which SVs regulate and alter these islet responses?

What are the precise cellular and molecular pathophysiological changes in each cell type that lead to T2DM? Are the major pathological changes beta cell specific or do they involve other islet cell types and/or non-islet cell types?

How many islet and single cell samples must be obtained to effectively capture combined cell type heterogeneity while controlling for technical and experimental confounders? How many samples are needed to observe genetic and/or epigenetic differences between T2DM and ND states? Would stratification of islets by T2DM risk genotype improve cell type-specific T2DM signatures?

Acknowledgments

Work in the Stitzel Lab is supported by the Assistant Secretary of Defense for Health Affairs, through the Peer Reviewed Medical Research Program under Award No. W81XWH-16-1-0130 and by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) under award number R00DK092251. Opinions, interpretations, conclusions, and recommendations are solely the responsibility of the authors, are not necessarily endorsed by the Department of Defense, and do not necessarily represent the official views of the National Institutes of Health. We thank Jane Cha for her expertise and help in figure design and artwork. We gratefully acknowledge members of the Stitzel and Ucar labs for helpful discussion and feedback on this work, and we thank the anonymous reviewers for their helpful suggestions to improve the perspective and content of this manuscript.

Supplemental Information

Supplemental Information associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.tig.2017.01.010

References

1. Mohlke, K.L. and Boehnke, M. (2015) Recent advances in understanding the genetic architecture of type 2 diabetes. *Hum. Mol. Genet.* 24, R85–R92
2. Parker, S.C.J. *et al.* (2013) Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U. S. A.* 110, 17921–17926
3. Pasquali, L. *et al.* (2014) Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* 46, 136–143
4. Roadmap Epigenomics Consortium *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330
5. Quang, D.X. *et al.* (2015) Motif signatures in stretch enhancers are enriched for disease-associated genetic variants. *Epigenetics Chromatin* 8, 23
6. Trynka, G. *et al.* (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* 45, 124–130
7. Finucane, H.K. *et al.* (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235
8. Stitzel, M.L. *et al.* (2015) Transcriptional regulation of the pancreatic islet: implications for islet function. *Curr. Diab. Rep.* 15, 66
9. Kulzer, J.R. *et al.* (2014) A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *Am. J. Hum. Genet.* 94, 186–197
10. Fogarty, M.P. *et al.* (2014) Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. *PLOS Genet.* 10, e1004633
11. van de Bunt, M. *et al.* (2015) Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. *PLoS Genet.* 11, e1005694
12. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium *et al.* (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* 46, 234–244
13. Morris, A.P. *et al.* (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* 44, 981–990
14. Fuchsberger, C. *et al.* (2016) The genetic architecture of type 2 diabetes. *Nature* 536, 41–47
15. Gaulton, K.J. *et al.* (2015) Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* 47, 1415–1425
16. Agarwala, V. *et al.* (2013) Evaluating empirical bounds on complex disease genetic architecture. *Nat. Genet.* 45, 1418–1427
17. Strawbridge, R.J. (2011) Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* 60, 2624–2634
18. Steinthorsdottir, V. (2014) Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* 46, 294–298
19. Hara, K. *et al.* (2014) Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum. Mol. Genet.* 23, 239–246
20. Aouacheri, O. *et al.* (2015) The investigation of the oxidative stress-related parameters in type 2 diabetes mellitus. *Can. J. Diabetes* 39, 44–49
21. Chaudhari, N. *et al.* (2014) A molecular web: endoplasmic reticulum stress, inflammation, and oxidative stress. *Front. Cell. Neurosci.* 8, 213
22. Gorasia, D.G. *et al.* (2015) Pancreatic beta cells are highly susceptible to oxidative and ER stresses during the development of diabetes. *J. Proteome Res.* 14, 688–699
23. Rando, O.J. and Simmons, R.A. (2015) I'm eating for two: parental dietary effects on offspring metabolism. *Cell* 161, 93–105
24. Zephy, D. and Ahmad, J. (2015) Type 2 diabetes mellitus: role of melatonin and oxidative stress. *Diabetes Metab. Syndr. Clin. Res. Rev.* 9, 127–131
25. Laybutt, D.R. *et al.* (2007) Endoplasmic reticulum stress contributes to beta cell apoptosis in type 2 diabetes. *Diabetologia* 50, 752–763
26. Scott, L.J. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316, 1341–1345
27. Sladek, R. *et al.* (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445, 881–885
28. Burton, P.R. *et al.* (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678
29. Diabetes Genetics Initiative of the Broad Institute of Harvard *et al.* (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316, 1331–1336
30. Liu, C.-T. *et al.* (2016) Trans-ethnic meta-analysis and functional annotation illuminates the genetic architecture of fasting glucose and insulin. *Am. J. Hum. Genet.* 99, 56–75
31. Fadista, J. *et al.* (2014) Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. U. S. A.* 111, 13924–13929
32. Stitzel, M.L. *et al.* (2010) Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab.* 12, 443–455
33. Ackermann, A.M. *et al.* (2016) Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol. Metab.* 5, 233–244
34. Hnisz, D. *et al.* (2013) Super-enhancers in the control of cell identity and disease. *Cell* 155, 934–947
35. Dekker, J. and Mirny, L. (2016) The 3D genome as moderator of chromosomal communication. *Cell* 164, 1110–1121
36. Lieberman-Aiden, E. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293
37. Li, G. *et al.* (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.* 11, R22
38. Mumbach, M.R. *et al.* (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* 13, 919–922
39. Keller, M.P. *et al.* (2016) The transcription factor Nfatc2 regulates β-cell proliferation and genes associated with type 2 diabetes in mouse and human islets. *PLOS Genet.* 12, e1006466
40. Scott, L.J. *et al.* (2016) The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat. Commun.* 7, 11764
41. Small, K.S. *et al.* (2011) Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat. Genet.* 43, 561–564
42. Morán, I. *et al.* (2012) Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab.* 16, 435–448
43. Arnes, L. and Sussel, L. (2015) Epigenetic modifications and long noncoding RNAs influence pancreas development and function. *Trends Genet.* 31, 290–299
44. Voight, B.F. *et al.* (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* 42, 579–589

45. Wallace, C. *et al.* (2010) The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nat. Genet.* 42, 68–71
46. Dayeh, T. *et al.* (2014) Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. *PLoS Genet.* 10, e1004160
47. Volkmar, M. *et al.* (2012) DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. *EMBO J.* 31, 1405–1426
48. Arda, H.E. *et al.* (2016) Age-dependent pancreatic gene regulation reveals mechanisms governing human β cell function. *Cell Metab.* 23, 909–920
49. Avrahami, D. *et al.* (2015) Aging-dependent demethylation of regulatory elements correlates with chromatin state and improved β cell function. *Cell Metab.* 22, 619–632
50. Perelis, M. *et al.* (2015) Pancreatic β cell enhancers regulate rhythmic transcription of genes controlling insulin secretion. *Science* 350, aac4250
51. Kluth, O. *et al.* (2014) Differential transcriptome analysis of diabetes-resistant and -sensitive mouse islets reveals significant overlap with human diabetes susceptibility genes. *Diabetes* 63, 4230–4238
52. Persaud, S.J. and Jones, P.M. (2016) A wake-up call for type 2 diabetes? *N. Engl. J. Med.* 375, 1090–1092
53. Halban, P.A. *et al.* (2014) β-cell failure in type 2 diabetes: postulated mechanisms and prospects for prevention and treatment. *Diabetes Care* 37, 1751–1758
54. Schulz, L.C. (2010) The Dutch Hunger Winter and the developmental origins of health and disease. *Proc. Natl. Acad. Sci.* 107, 16757–16758
55. Li, J. *et al.* (2017) Prenatal exposure to famine and the development of hyperglycemia and type 2 diabetes in adulthood across consecutive generations: a population-based cohort study of families in Suihua, China. *Am. J. Clin. Nutr.* 105, 221–227
56. Rakyan, V.K. *et al.* (2011) Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* 12, 529–541
57. Dai, C. *et al.* (2016) Stress-impaired transcription factor expression and insulin secretion in transplanted human islets. *J. Clin. Invest.* 126, 1857–1870
58. Dai, C. *et al.* (2012) Islet-enriched gene expression and glucose-induced insulin secretion in human and mouse islets. *Diabetologia* 55, 707–718
59. Guo, S. *et al.* (2013) Inactivation of specific β cell transcription factors in type 2 diabetes. *J. Clin. Invest.* 123, 3305–3316
60. Lee, M.N. *et al.* (2014) Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* 343, 1246980
61. Blodgett, D.M. *et al.* (2015) Novel observations from next-generation RNA sequencing of highly purified human adult and fetal islet cell subsets. *Diabetes* 64, 3172–3181
62. Brissova, M. *et al.* (2005) Assessment of human pancreatic islet architecture and composition by laser scanning confocal microscopy. *J. Histochem. Cytochem.* 53, 1087–1097
63. Cabrera, O. *et al.* (2006) The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proc. Natl. Acad. Sci. U. S. A.* 103, 2334–2339
64. Ionescu-Tirgoviste, C. (2015) A 3D map of the islet routes throughout the healthy human pancreas. *Sci. Rep.* 5, 14634
65. Stanojevic, V. and Habener, J.F. (2015) Evolving function and potential of pancreatic alpha cells. *Best Pract. Res. Clin. Endocrinol. Metab.* 29, 859–871
66. Jamison, R.A. *et al.* (2011) Hyperglucagonemia precedes a decline in insulin secretion and causes hyperglycemia in chronically glucose-infused rats. *Am. J. Physiol. Endocrinol. Metab.* 301, E1174–E1183
67. Rodriguez-Diaz, R. (2011) Alpha cells secrete acetylcholine as a non-neuronal paracrine signal priming beta cell function in humans. *Nat. Med.* 17, 888–892
68. Hauge-Evans, A.C. (2009) Somatostatin secreted by islet delta-cells fulfills multiple roles as a paracrine regulator of islet function. *Diabetes* 58, 403–411
69. van der Meulen, T. (2015) Urocortin3 mediates somatostatin-dependent negative feedback control of insulin secretion. *Nat. Med.* 21, 769–776
70. DiGruccio, M.R. *et al.* (2016) Comprehensive alpha, beta and delta cell transcriptomes reveal that ghrelin selectively activates delta cells and promotes somatostatin release from pancreatic islets. *Mol. Metab.* 5, 449–458
71. Molina, J. *et al.* (2014) Control of insulin secretion by cholinergic signaling in the human pancreatic islet. *Diabetes* 63, 2714–2726
72. Bramswig, N.C. *et al.* (2013) Epigenomic plasticity enables human pancreatic α to β cell reprogramming. *J. Clin. Invest.* 123, 1275–1284
73. Dorrell, C. *et al.* (2011) Transcriptomes of the major human pancreatic cell types. *Diabetologia* 54, 2832–2844
74. Nica, A.C. *et al.* (2013) Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome. *Genome Res.* 23, 1554–1562
75. Lawlor, N. *et al.* (2017) Single cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* 27, 208–222
76. Baron, M. *et al.* (2016) A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* 3, 346–360
77. Grün, D. *et al.* (2016) De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* 19, 266–277
78. Li, J. *et al.* (2016) Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Rep.* 17, 178–187
79. Muraro, M.J. *et al.* (2016) A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* 3, 385–394
80. Segerstolpe, Å. *et al.* (2016) Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* 24, 593–607
81. Wang, Y.J. *et al.* (2016) Single cell transcriptomics of the human endocrine pancreas. *Diabetes* 65, 3028–3038
82. Xin, Y. *et al.* (2016) Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. *Proc. Natl. Acad. Sci. U. S. A.* 113, 3293–3298
83. Xin, Y. *et al.* (2016) RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.* 24, 608–615
84. Dorrell, C. *et al.* (2016) Human islets contain four distinct subtypes of β cells. *Nat. Commun.* 7, 11756
85. Bader, E. *et al.* (2016) Identification of proliferative and mature β-cells in the islets of Langerhans. *Nature* 535, 430–434
86. Wang, Y.J. *et al.* (2016) Single-cell mass cytometry analysis of the human endocrine pancreas. *Cell Metab.* 24, 616–626
87. Liu, S. and Trapnell, C. (2016) Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research* 5, 182
88. Cinti, F. *et al.* (2016) Evidence of β-cell dedifferentiation in human type 2 diabetes. *J. Clin. Endocrinol. Metab.* 101, 1044–1054
89. Lu, J. *et al.* (2014) Transdifferentiation of pancreatic α-cells into insulin-secreting cells: from experimental models to underlying mechanisms. *World J. Diabetes* 5, 847–853
90. Talchai, C. *et al.* (2012) Pancreatic β-cell dedifferentiation as mechanism of diabetic β-cell failure. *Cell* 150, 1223–1234
91. Wang, Z. *et al.* (2014) Pancreatic β-cell dedifferentiation in diabetes and re-differentiation following insulin therapy. *Cell Metab.* 19, 872–882
92. Butler, A.E. *et al.* (2003) Beta-cell deficit and increased beta-cell apoptosis in humans with type 2 diabetes. *Diabetes* 52, 102–110
93. Butler, A.E. *et al.* (2016) β-cell deficit in obese type 2 diabetes, a minor role of β-cell dedifferentiation and degranulation. *J. Clin. Endocrinol. Metab.* 101, 523–532
94. Sheedy, F.J. *et al.* (2013) CD36 coordinates NLRP3 inflammasome activation by facilitating the intracellular nucleation

from soluble to particulate ligands in sterile inflammation. *Nat. Immunol.* 14, 812–820

95. Kumasaka, N. *et al.* (2016) Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* 48, 206–213
96. Degner, J.F. *et al.* (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–394
97. Thomsen, S.K. *et al.* (2016) Systematic functional characterization of candidate causal genes for type 2 diabetes risk variants. *Diabetes* 65, 3805–3811
98. Soleimanpour, S.A. and Stoffers, D.A. (2013) The pancreatic β cell and type 1 diabetes: innocent bystander or active participant? *Trends Endocrinol. Metab.* 24, 324–331
99. Maller, J.B. *et al.* (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* 44, 1294–1301
100. Cerf, M.E. (2013) Beta cell dysfunction and insulin resistance. *Front. Endocrinol.* 4, 37
101. Halban, P.A. *et al.* (2014) β-cell failure in type 2 diabetes: postulated mechanisms and prospects for prevention and treatment. *Diabetes Care* 37, 1751–1758

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes

Nathan Lawlor,[1,4] Joshy George,[1,4] Mohan Bolisetty,[1] Romy Kursawe,[1] Lili Sun,[1] V. Sivakamasundari,[1] Ina Kycia,[1] Paul Robson,[1,2,3] and Michael L. Stitzel[1,2,3]

[1] *The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut 06032, USA;* [2] *Institute for Systems Genomics, University of Connecticut, Farmington, Connecticut 06032, USA;* [3] *Department of Genetics & Genome Sciences, University of Connecticut, Farmington, Connecticut 06032, USA*

Blood glucose levels are tightly controlled by the coordinated action of at least four cell types constituting pancreatic islets. Changes in the proportion and/or function of these cells are associated with genetic and molecular pathophysiology of monogenic, type I, and type 2 (T2D) diabetes. Cellular heterogeneity impedes precise understanding of the molecular components of each islet cell type that govern islet (dys)function, particularly the less abundant delta and gamma/pancreatic polypeptide (PP) cells. Here, we report single-cell transcriptomes for 638 cells from nondiabetic (ND) and T2D human islet samples. Analyses of ND single-cell transcriptomes identified distinct alpha, beta, delta, and PP/gamma cell-type signatures. Genes linked to rare and common forms of islet dysfunction and diabetes were expressed in the delta and PP/gamma cell types. Moreover, this study revealed that delta cells specifically express receptors that receive and coordinate systemic cues from the leptin, ghrelin, and dopamine signaling pathways implicating them as integrators of central and peripheral metabolic signals into the pancreatic islet. Finally, single-cell transcriptome profiling revealed genes differentially regulated between T2D and ND alpha, beta, and delta cells that were undetectable in paired whole islet analyses. This study thus identifies fundamental cell-type–specific features of pancreatic islet (dys)function and provides a critical resource for comprehensive understanding of islet biology and diabetes pathogenesis.

[Supplemental material is available for this article.]

Pancreatic islets of Langerhans are clusters of at least four different hormone-secreting endocrine cell types that elicit coordinated—but distinct—responses to maintain glucose homeostasis. As such, they are central to diabetes pathophysiology. On average, human islets consist mostly of beta (54%), alpha (35%), and delta (11%) cells; up to a few percent gamma/pancreatic polypeptide (PP) cells; and very few epsilon cells (Brissova et al. 2005; Cabrera et al. 2006; Blodgett et al. 2015). Human islet composition is neither uniform nor static but varies between individuals and across regions of the pancreas (Brissova et al. 2005; Cabrera et al. 2006; Blodgett et al. 2015). Cellular heterogeneity complicates molecular studies of whole human islets and may mask important role(s) for less common cells in the population (Dorrell et al. 2011b; Bramswig et al. 2013; Nica et al. 2013; Blodgett et al. 2015; Liu and Trapnell 2016). Moreover, it complicates attempts to identify epigenetic and transcriptional signatures distinguishing diabetic from nondiabetic (ND) islets, leading to inconsistent reports of genes and pathways affected (Gunton et al. 2005; Marselli et al. 2010; Taneera et al. 2012; Dayeh et al. 2014). Conventional sorting and enrichment techniques are unable to specifically purify each human islet cell type (Dorrell et al. 2008; Nica et al. 2013; Bramswig et al. 2013; Hrvatin et al. 2014; Blodgett et al. 2015), thus a precise understanding of the transcriptional repertoire governing each cell type's identity and function is lacking. Identifying the cell-type–specific expression programs that contribute to islet dysfunction and type 2 diabetes (T2D) should reveal novel targets and approaches to prevent, monitor, and treat T2D.

In this study, we sought to decipher the transcriptional repertoire of each islet cell type in an agnostic and precise manner by capturing and profiling pancreatic single cells from ND and T2D individuals. From these profiles, we identified transcripts uniquely important for each islet cell type's identity and function. Finally, we compared T2D and ND individuals to identify islet cell-type–specific expression changes that were otherwise masked by islet cellular heterogeneity. The insights and data from this study provide an important foundation to guide future genomics-based interrogation of islet dysfunction and diabetes.

Results

Islet single-cell transcriptomes accurately recapitulate those of intact islets

Pancreatic islets (>85% purity and >90% viability) were obtained from eight human cadaveric organ donors (five ND, three T2D) (Fig. 1A; Supplemental Table S1). Each islet sample was processed to generate single-cell RNA-seq libraries (Fig. 1A; single cell) and paired bulk RNA-seq libraries at three different stages of islet processing (Fig. 1A; baseline, intact, and dissociated). All RNA-seq

[4] **These authors contributed equally to this work.**
Corresponding author: michael.stitzel@jax.org
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.212720.116. Freely available online through the *Genome Research* Open Access option.

© 2017 Lawlor et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at http://creativecommons.org/licenses/by/4.0/.

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

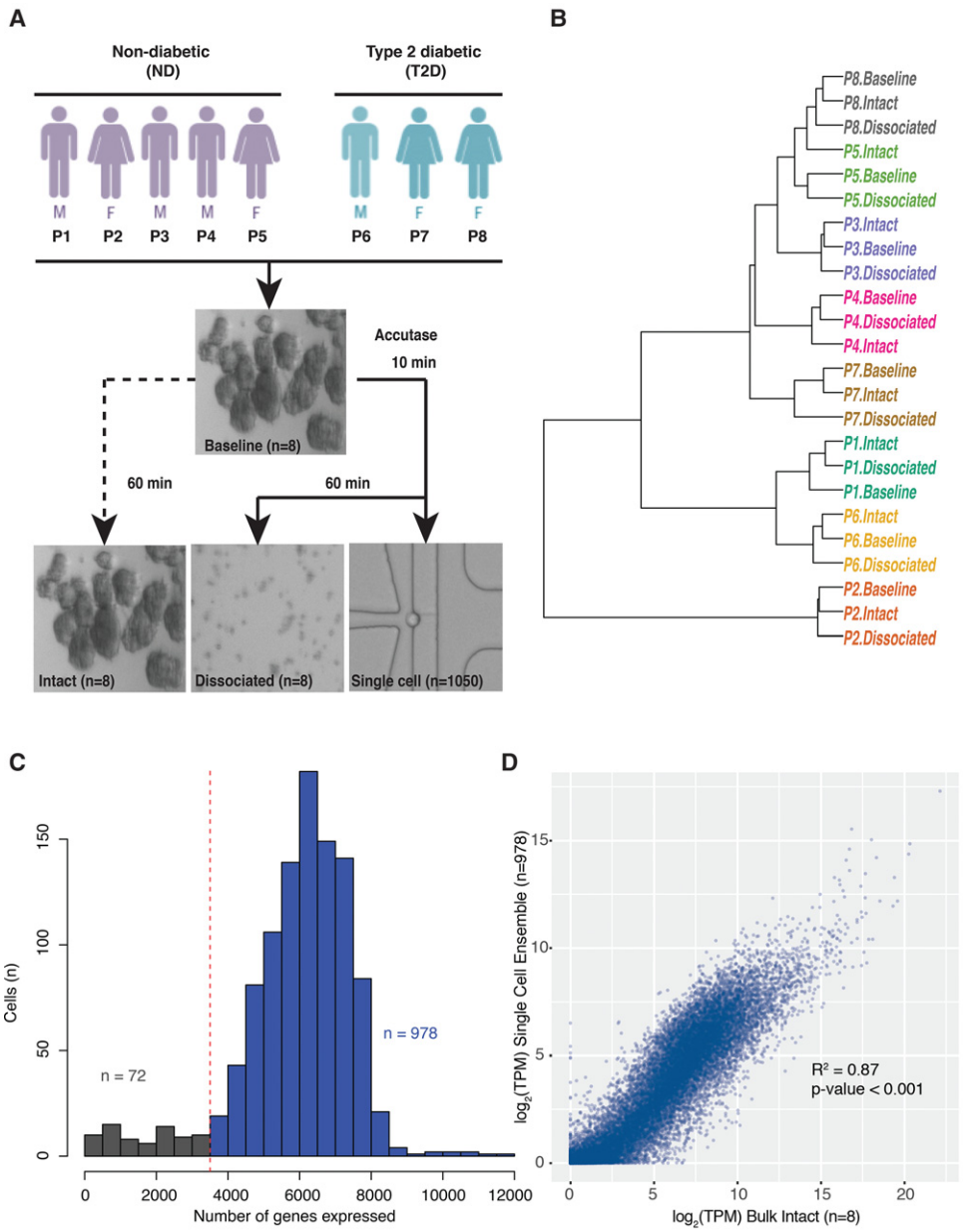


Figure 1. Single-cell transcriptomes reflect those of paired intact islets. (*A*) Schematic of experimental workflow. Islets from each donor sample ($n = 8$ individuals) were dissociated using Accutase, and single-cell transcriptomes were synthesized from 1050 cells captured using 11 Fluidigm C1 chips. In parallel, "bulk" RNA-seq libraries were prepared from remaining dissociated single cells (dissociated) and from intact islets either flash frozen (baseline) or incubated/processed (intact). (*B*) Unsupervised hierarchical clustering of baseline, intact, and dissociated islet transcriptomes demonstrates clustering by person and not by processing/experimental condition. (*C*) Histogram demonstrating the number of genes detected in each single cell. Cells expressing less than 3500 genes ($n = 72$) were removed from downstream analyses. (*D*) Scatter plot comparing intact islet bulk RNA-seq ($n = 8$) and ensemble single-cell RNA-seq ($n = 978$) data demonstrates high correlation. (R^2) Pearson's *R*-squared; (TPM) transcripts per million; (P) person.

methods employed SMARTer chemistry (Methods), and bulk islet cDNA libraries were sequenced to an average approximate depth of 34 million reads (Supplemental Table S2). Baseline, intact, and dissociated transcriptomes from each person were highly correlated (Supplemental Fig. S1). Transcriptomes clustered by donor and not by processing condition or incubation time (Fig. 1B), strongly suggesting that islet processing did not significantly alter islet transcriptomes.

A total of 1050 islet cells (622 ND and 428 T2D) were captured on 11 Fluidigm C1 chips. cDNA libraries were constructed from captured cells and barcoded, fragmented, pooled, and sequenced to an average depth of 3 million reads (Supplemental Table S2). Two separate library preparations from the same amplified cDNA for each of 83 single cells demonstrated remarkable correlation, suggesting minimal batch effects resulting from the cDNA processing and sequencing steps. Resequenced samples are highlighted in Supplemental Table S2 but were not included in subsequent analyses. Transcript coverage is indicated in Supplemental Figure S2. Approximately 81% (21,484/26,616) of protein-coding genes and long intergenic noncoding RNAs (lincRNAs) were detected

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

in at least one cell from the collection. On average, each single cell expressed 5944 genes (Fig. 1C). Cells expressing less than 3500 genes ($n = 72$) also exhibited high mitochondrial alignment rates and other reported transcriptional metrics of cell death and/or poor quality (Ilicic et al. 2016; Xin et al. 2016) and were removed from subsequent analyses (Fig. 1C).

We next assessed the extent to which the remaining 978 single-cell transcriptomes represent the expression patterns observed in intact islets. Single-cell transcriptome ensembles from each person were highly correlated (Pearson's R^2 ranged from 0.91–0.98) (Supplemental Fig. S3), regardless of disease state. Pearson's R^2 values between individuals' single-cell ensembles and corresponding "bulk" transcriptomes ranged from 0.75–0.86 (Supplemental Fig. S4) and did not differ substantially between ND ($R^2 = 0.87$) and T2D ($R^2 = 0.85$) samples (Supplemental Fig. S5). Overall, ensemble/aggregate single-cell transcriptome profiles correlated well with those of pooled bulk islet transcriptomes from all individuals (Fig. 1D, $R^2 = 0.87$). These results suggest that the islet single-cell transcriptomes are high quality and effectively reflect bulk islet transcriptomes.

Single-cell profiling captures transcriptomes of major and minor pancreatic endocrine and exocrine cell types

Five islet endocrine cell types have been assigned based on exclusive and robust expression of the peptide hormone genes *INS* (beta), *GCG* (alpha), *SST* (delta), *PPY* (PP/gamma), and *GHRL* (epsilon) (Baetens et al. 1979; Nussey and Whitehead 2001; Zhao et al. 2008; Li et al. 2016; Xin et al. 2016; Wang et al. 2016). The pancreas also contains three exocrine cell types—acinar, stellate, and ductal—that critically support digestion through synthesis and transport of digestive enzymes (Pandol 2011; Reichert and Rustgi 2011). Each also has been identified by specific marker gene expression, including the serine peptidase gene *PRSS1* (acinar) (Dabbs 2013), the extracellular matrix protein gene *COL1A1* (stellate) (Mathison et al. 2010), and the structural keratin gene *KRT19* (ductal) (Dorrell et al. 2008, 2011a,b; Reichert and Rustgi 2011). We used these marker genes to determine the representation of each islet cell type among our 978 profiled single cells.

Density plots (Fig. 2A) revealed bimodal expression of each marker gene across the population of single cells. Therefore, we employed Gaussian mixture modeling (GMM) to classify the cells unambiguously (Fig. 2B). Approximate $\log_2$ counts per million (CPM) thresholds for each marker gene used to classify cell types are provided in Supplemental Table S3. This approach identified 617 single cells (~63%) from T2D and ND islets expressing a single marker gene representative of each major endocrine and exocrine cell type, examples of which are shown in Figure 2C. This included 239 alpha, 264 beta, 25 delta, and 18 PP/gamma cells (Table 1); the proportions of each cell type are in the ranges previously reported (Brissova et al. 2005; Cabrera et al. 2006; Blodgett et al. 2015). Only one cell expressing high levels ($\log_2$CPM > 15) of *GHRL* was identified, which we presume to be an exceedingly rare epsilon cell. Additionally, we obtained 19 stellate, 24 acinar, and 27 ductal cells (Table 1), presumably from exocrine contamination of the islet cell preparations. Only 21 cells (~2%) expressed none of the specified marker genes (Table 1). Approximately one-third (340/978) of cells expressed more than one marker gene; these were removed from subsequent analysis due to concerns that these represent two vertically stacked cells in a given capture site (for details, see Methods). Similar ratios of potential stacked cells have been reported in other studies using the Fluidigm C1 platform to capture mouse (Xin et al. 2016) and human islet cells (Wang et al. 2016). Collectively, these single-cell data capture transcriptome profiles representing each of the major and minor pancreatic endocrine and exocrine cell types. Genome Browser tracks representing aggregate single-cell expression for each islet cell type have been generated using HOMER (Heinz et al. 2010) and are made available (see Data Access) to facilitate their use and investigation by members of the islet biology and diabetes research communities.

Unsupervised analyses of islet single-cell transcriptomes identify discrete clusters corresponding to cell type

To determine if and how islet cell transcriptomes cluster, we completed unsupervised dimensionality reduction via t-distributed stochastic neighbor embedding (t-SNE) on 380 ND single-cell samples (excluding "multiple" labeled samples). t-SNE assembled single-cell transcriptomes into discrete clusters based upon 1824 highly expressed genes (see Methods; Supplemental Table S4); GMM-based marker gene analysis revealed that each cluster corresponded to a distinct endocrine and exocrine cell type (Fig. 3A; Supplemental Fig. S6). Unsupervised hierarchical clustering also grouped single-cell transcriptomes into discrete cell types (Fig. 3B). Despite being obtained from different individuals, 161/168 beta, 128/138 alpha, 15/16 delta, and 12/12 PP/gamma cell transcriptomes clustered into the same dendrogram branches, strongly suggesting that cell type encodes the greatest variation in the data. Exocrine cells and those expressing none of the specified marker genes ("none") clustered separately from the endocrine cell types. Importantly, this clustering was driven by neither sequencing depth (Supplemental Fig. S7B) nor expression of classic marker genes (*INS*, *GCG*, *SST*, *PPY*, *GHRL*, *COL1A1*, *PRSS1*, and *KRT19*), as cells continued to cluster into discrete cell types even when all marker genes were removed from the expression data sets (Supplemental Figs. S7C, S8). Recent studies have reported heterogeneity among beta cells. Specifically, Dorrell et al. characterized four subpopulations of human beta cells based on differing *ST8SIA1* and *CD9* expression (Dorrell et al. 2016). Similarly, Bader et al. 2016 distinguished two populations of proliferating (*Fltp*$^+$) and mature (*Fltp*$^-$) mouse beta cells. We did not find evidence of beta cell subpopulations (Supplemental Fig. S9), nor did we identify numerous proliferating cells (Supplemental Table S5). T2D single-cell transcriptomes ($n = 258$) also demonstrated clear clustering by cell type in unsupervised analyses (Supplemental Figs. S10–S14) based on 1908 highly expressed genes (Supplemental Table S4). Thus, each endocrine and exocrine pancreatic cell type exhibits a complex characteristic expression signature that uniquely identifies it.

Differential expression analyses reveal islet cell-type–specific transcriptional signatures

To identify gene signatures distinguishing each islet cell type, we completed a series of pairwise differential expression analyses (Supplemental Table S6) between each cell type (see Methods). After intersecting the results from each pairwise comparison, we identified a conservative collection of 154 islet endocrine cell-type "signature" genes (61 beta, 51 alpha, 17 delta, 25 gamma), as well as 202 exocrine genes (109 stellate, 31 acinar, 62 ductal) at 5% false-discovery rate (FDR) (Fig. 3C; Supplemental Table S7). Two genes exhibited overlap between the endocrine and exocrine signature lists: *FAP* (alpha and stellate cell overlap) and *TNS1* (beta and stellate cell overlap). Gene set enrichment analysis (GSEA) identified enrichment (FDR-adjusted P-value <0.05) of insulin

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

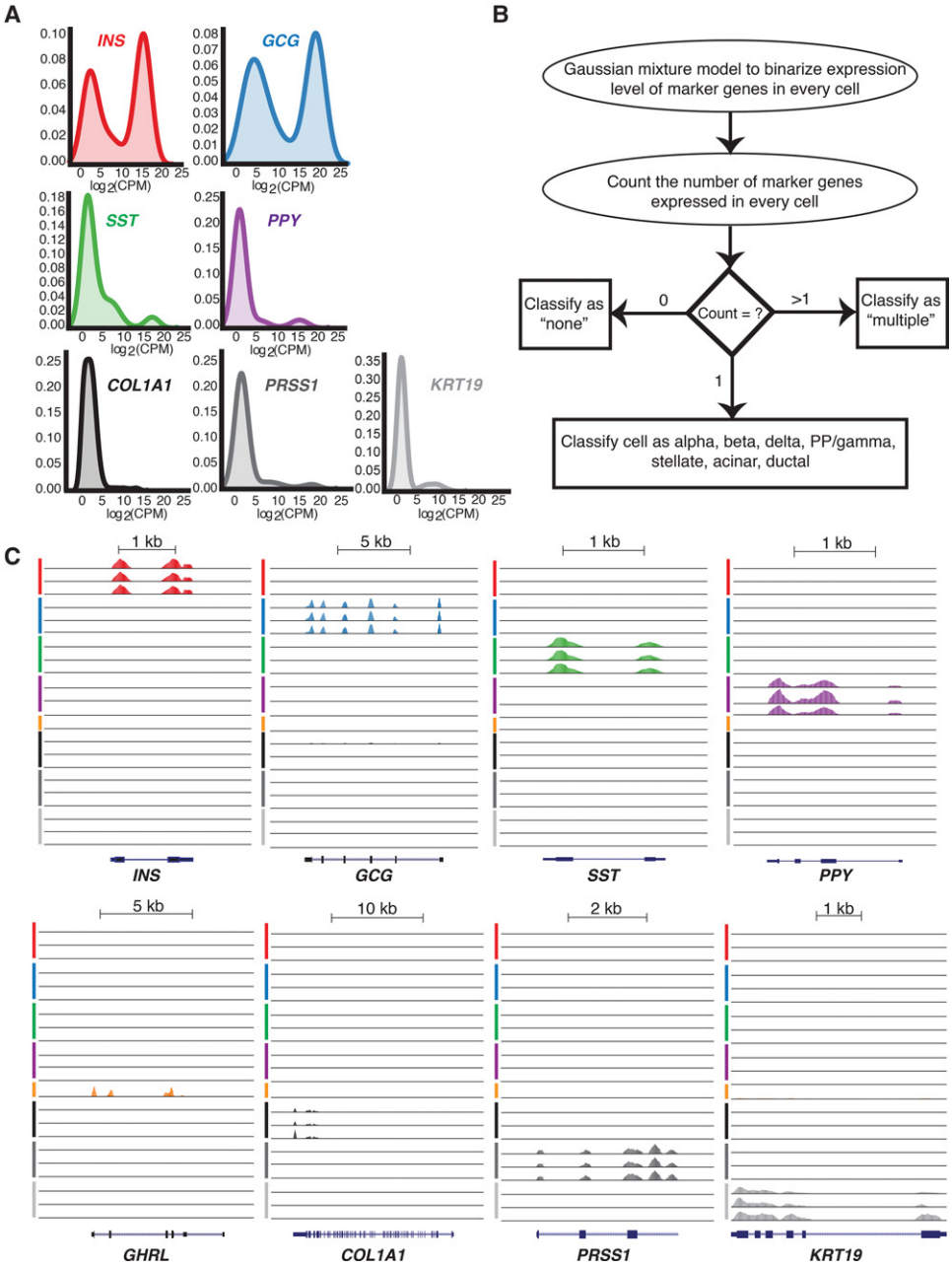


Figure 2. Cell-type classification based on marker gene expression. (*A*) Density plots demonstrating endocrine and exocrine marker gene expression across all single cells. (*B*) Schematic of the Gaussian mixture model method applied to assign cell-type identity based on marker gene expression. (*C*) UCSC Genome Browser views of representative single-cell expression profiles of *INS*, *GCG*, *SST*, *PPY*, and *GHRL* genes encoding beta, alpha, delta, PP/gamma, and epsilon cell hormones of the endocrine pancreas, respectively, and marker genes for stellate (*COL1A1*), acinar (*PRSS1*), and ductal (*KRT19*) cells of the exocrine pancreas. Line colors indicate putative beta (red), alpha (blue), delta (green), PP/gamma (purple), epsilon (orange), stellate (black), acinar (dark gray), and ductal cells (light gray). (PP) pancreatic polypeptide; (CPM) counts per million.

signaling, oxidative phosphorylation, maturity-onset diabetes of the young (MODY), and glycolysis/gluconeogenesis KEGG pathways in beta cells relative to the other endocrine cells (Supplemental Table S8).

Signature genes included previously reported beta-specific genes like *NKX6-1*, *DLK1*, and *ADCYAP1* (Fig. 3C, right) and alpha cell–specific genes like *IRX2*, *LOXL4*, and *DPP4*, a cell surface receptor and diabetes drug target (Dorrell et al. 2011a; Bramswig et al. 2013; Nica et al. 2013; Blodgett et al. 2015). Among delta cell signature genes, we detected exclusive expression of *HHEX*, a transcription factor reported to govern delta cell identity and function and linked to T2D GWAS (Zhang et al. 2014). Delta cells also specifically expressed *BCHE*, which encodes butyrylcholinesterase. BCHE catalyzes the breakdown of acetylcholine and ghrelin (Chen et al. 2015), thus providing a mechanism for delta cells to exert local inhibition of islet-influencing endocrine signals. PP/gamma cell–specific transcriptomes included *CTD-2008P7.8*, a lincRNA of unknown function; *CNTNAP5*, a member

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

Table 1. Number of profiled cells for each pancreatic cell type based on marker gene expression

Putative cell type (marker gene)	Cell ontology accession no.	Nondiabetic (ND)	Type 2 diabetic (T2D)
Alpha (*GCG*)	CL:0000171	138 (23.47%)	101 (25.9%)
Beta (*INS*)	CL:0000169	168 (28.57%)	96 (24.62%)
Delta (*SST*)	CL:0000173	16 (2.72%)	9 (2.31%)
PP/gamma (*PPY*)	CL:0002275	12 (2.04%)	6 (1.54%)
Epsilon (*GHRL*)	CL:0005019	1 (0.17%)	0
Stellate (*COL1A1*)	CL:0002410	9 (1.53%)	10 (2.56%)
Acinar (*PRSS1*)	CL:0002064	15 (2.55%)	9 (2.31%)
Ductal (*KRT19*)	CL:0002079	11 (1.87%)	16 (4.1%)
Multiple	—	208 (35.37%)	132 (33.85%)
None (other)	—	10 (1.7%)	11 (2.82%)
Total		588	390

of the neurexin family of cell adhesion molecules; and *ID4*, which encodes an inhibitor of DNA-binding protein. In addition to *DPP4*, we detected 30 islet signature genes whose proteins SWISSPROT predicts to localize to the cell surface (Supplemental Table S9). DPP4 antibodies have recently been used to isolate purer alpha cell populations from islets (Arda et al. 2016). Thus, antibodies against the other candidate cell-type–specific surface markers we have identified may be useful to purify other islet cell types.

Single-cell profiling identifies unexpected overlap in expression between minor and major islet cell types

Cell sorting and enrichment methods such as fluorescence-activated cell sorting (FACS) have been used to identify characteristic alpha and beta cell genes (Dorrell et al. 2011a,b; Bramswig et al. 2013; Nica et al. 2013; Blodgett et al. 2015). However, expression of *SST* or *PPY* in the reported alpha and beta cell gene sets suggests the presence of the less abundant delta and PP/gamma islet cell types in the enriched cell preparations. To distinguish genes exhibiting alpha- and beta-specific gene expression from those expressed also in delta and PP/gamma cells, we investigated the expression of previously reported alpha- and beta-specific genes (Supplemental Table S10; Supplemental Fig. S15) in our ND endocrine single-cell transcriptomes. Only 115/1683 previously reported beta-specific genes were expressed greater than fourfold higher in beta cells relative to the other endocrine cells (FDR < 0.05; one-way ANOVA followed by Tukey's honest significant difference [THSD]) (Fig. 3D). Similarly, 75/1853 reported alpha-specific genes were alpha cell enriched (Fig. 3E). Several genes previously reported to be enriched in the major islet cell types, such as *MAFA*, *SLC2A2*, *SIX3*, and *DLK1* in beta cells and *IRX2*, *DPP4*, and *ADORA2A* in alpha cells, were confirmed to be signature genes. Surprisingly, we found that 37 and 33 reported beta- and alpha-specific genes were also expressed in delta and PP/gamma cells, respectively (Fig. 3F; Supplemental Table S10). Notable examples included beta and delta cell expression of the congenital hyperinsulinemia (CHI) gene *HADH* and alpha and PP/gamma cell expression of the *ARX* transcription factor (Liu et al. 2011). *HADH* is typically associated with beta cell expression and, when mutated, leads to insulin hypersecretion and CHI (Kapoor et al. 2010; Pepin et al. 2010); these data implicate the delta cell in the molecular genetics of CHI. Misexpression of *ARX* has been shown to convey both alpha and PP/gamma cell features to cells (Collombat et al. 2007), suggesting that its expression in each cell type is important for identity and function.

Genes underpinning metabolic function, regulation of energy homeostasis, and satiety are specific to distinct islet cell types

Perturbations in genes involved in glucose sensing and proper maintenance of blood glucose levels contribute to T2D pathophysiology (Schuit et al. 2001; MacDonald et al. 2005). Beta cells regulate blood glucose through the secretion of insulin and are thus exquisitely sensitive to blood glucose levels. Glucose-stimulated insulin secretion (GSIS) is linked to universal basic pathways of cellular metabolism in beta cells. To gain insight into beta cell-type–specific transcriptomic features associated with GSIS, namely, glucose uptake and glycolysis, we examined the expression of relevant genes in our islet single-cell transcriptomes (Fig. 4A).

GSIS pathway genes associated with glucose sensing and uptake displayed highly beta cell–specific expression, including *SLC2A2*, which encodes the glucose transporter GLUT2; *G6PC2*, which encodes a subunit of glucose-6-phosphatase; and *PFKFB2*, which encodes an enzyme involved in regulation of glycolysis (Fig. 4A; Chen et al. 2008; Muller et al. 2015). While expressed in all cell types, the enzyme, ALDOA1, immediately downstream from PFK1 and associated with the glycerol phosphate (GP) shuttle, is enriched in beta cells, perhaps reflecting an additional point of GSIS control. Protein-coding genes for five subsequent glycolytic enzymatic steps from glyceraldehyde-3-phosphate to pyruvate were not significantly differentially expressed between cell types. Beta cells are known to be limited in their ability to produce lactate from pyruvate (Fridlyand and Philipson 2010); this is reflected by high *LDHB*/*LDHA* ratios that favor the lactate to pyruvate flux in beta cells.

The glycerol-3-phosphate shuttle allows NAD^+ regeneration in the cytosol to sustain glycolytic flux essential for GSIS. Cytoplasmic NAD^+ generation has been shown to be essential for GSIS (Eto et al. 1999). Both components of the glycerol-3-phosphate shuttle, cytoplasmic *GPD1* and mitochondrial *GPD2*, were expressed in beta cells, with the former representing a beta cell signature gene (Fig. 4A). Additionally, we identified the mitochondrial solute transporter *SLC25A34* as beta cell specific. While its transport specificities have yet to be determined, the closest yeast ortholog of *SLC25A34*, Oac1p/YKL120w (Palmieri et al. 1999; Marobbio et al. 2008), is thought to import oxaloacetate into the mitochondria. This is particularly intriguing considering our data and others (MacDonald et al. 2011) show the complete absence of pyruvate carboxylase (PC) expression in human beta cells, despite the essential role PC is known to play in rodent GSIS (Sugden and Holness 2011) through mitochondrial production of oxaloacetate. We hypothesize that *SLC25A34* may provide an alternate, cytoplasmic source for mitochondrial oxaloacetate in the human beta cell.

Single-cell profiling also allowed us to interrogate the transcriptional repertoire of less abundant delta and PP/gamma cell types, which have been elusive in both whole islet and sorted islet studies. While it is difficult to determine epsilon cell expression signatures with one ghrelin-positive cell, our ND data set includes 16 delta cells and 12 PP/gamma cells. Among the top 100 differentially expressed (FDR < 5%) genes in delta versus other islet endocrine cells are receptors for the appetite-regulating hormones leptin (*LEPR*) and ghrelin (*GHSR*), the growth factor neuregulin 4 (*ERBB4*), and the neurotransmitter dopamine (*DRD2*) (Fig. 4B). *GHSR* has recently been shown to be specifically expressed and

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

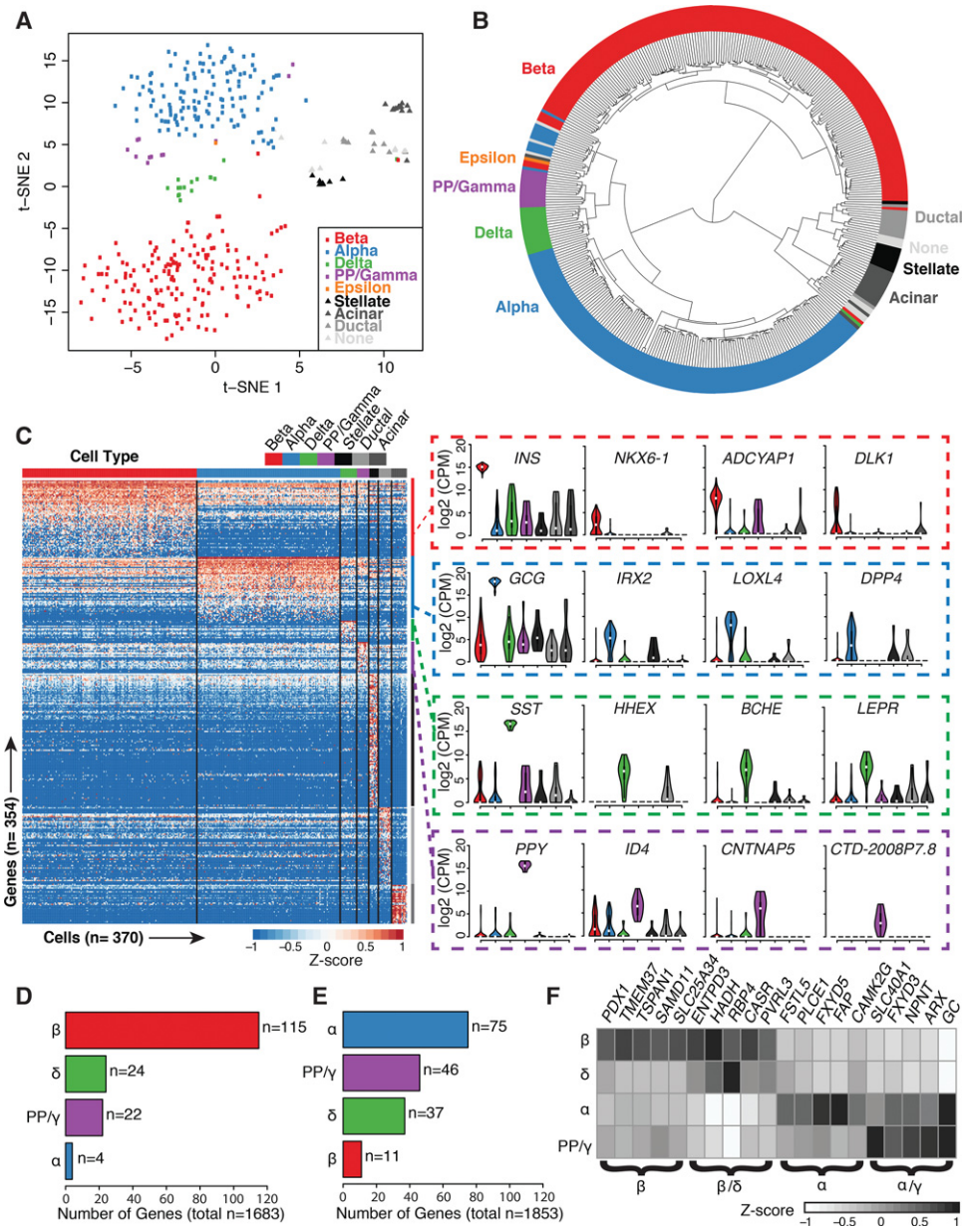


Figure 3. Statistical analysis of nondiabetic single-cell transcriptomes identifies cell-type–specific clusters and defines the signature genes of each islet cell type. (*A*) Unsupervised analysis of single-cell transcriptomes using t-distributed stochastic neighbor embedding (t-SNE) demonstrates grouping of single islet cell transcriptomes into the major constituent cell types. Respective cell labels and coloring were added after unsupervised analyses. (*B*) Unsupervised hierarchical clustering illustrates relationships of transcriptome profiles between respective endocrine and exocrine cells. (*C*) Supervised differential expression analysis of cell types determines cell-specific (signature) genes across all cells (see Methods). Values represent $\log_2$(CPM) expression after mean-centering and scaling between −1 and 1. Violin plots of selected signature gene expression are displayed to the *right* of the heatmap. (*D*,*E*) Bar plots depicting the numbers of previously reported beta-specific (*D*) and alpha-specific (*E*) genes (Dorrell et al. 2011b; Bramswig et al. 2013; Nica et al. 2013; Blodgett et al. 2015) confirmed to be expressed in each islet cell type after ANOVA and Tukey's honest significant difference (THSD) post-hoc analysis (Methods). (*F*) Several beta-specific genes demonstrate similar expression levels in delta cells, and alpha-specific genes demonstrate similar expression in PP/gamma cells. Values represent average $\log_2$(CPM) expression after mean-centering and scaling between −1 and 1. (β) Beta; (α) alpha; (δ) delta; (γ) PP/gamma cells.

functional in both human and mouse delta cells, reducing GSIS in human and mouse beta cells when induced (DiGruccio et al. 2016). *LEPR*, *DRD2*, and *ERBB4* expression is specific to human delta cells. In situ analyses (ViewRNA, Affymetrix) detected coexpression of *LEPR* in 79/102 (77%) of *SST*-expressing cells (Fig. 4D, arrowheads) in ND islets, confirming the delta cell–specific expression detected in Fluidigm C1 profiling. Thus, our data suggest intriguing roles for islet delta cells in the integration of metabolic signals via leptin, ghrelin, and dopamine signaling pathways.

PP/gamma, along with epsilon cells, are among the least studied islet cell types due to their scarcity in islets. Recent studies show that PP/gamma cells are crucial regulators of energy homeostasis (Yulyaningsih et al. 2014; Khandekar et al. 2015). In response to food intake, these cells secrete the anorexigenic hormone PPY to

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

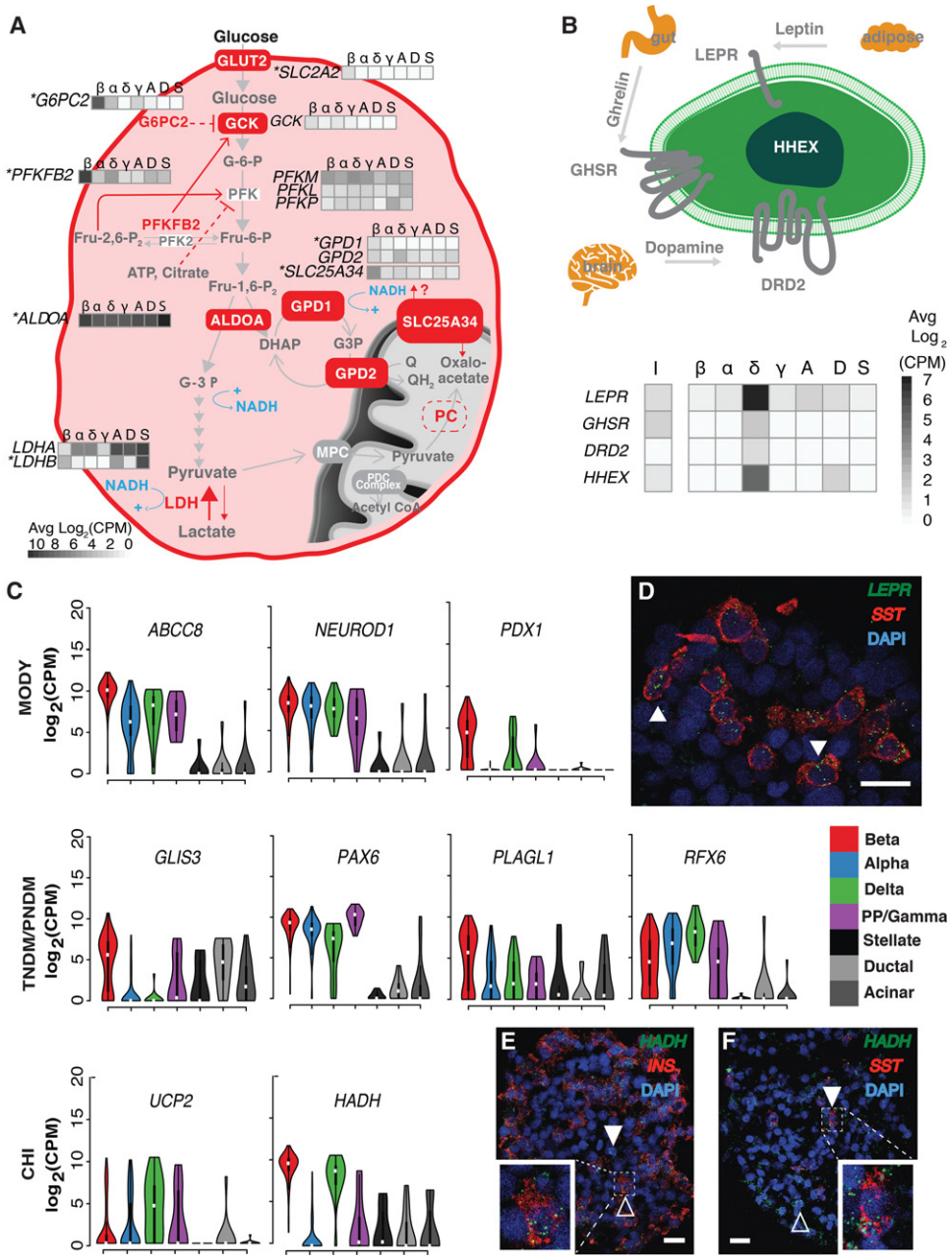


Figure 4. Cell-type–specific expression of metabolic, signaling, and diabetes trait genes. (*A*) Beta cell–specific expression of different isoforms of glycolytic and metabolic intermediate shuttles. Genes marked with an asterisk represent beta cell signature genes. (*B*) Delta cell–specific expression of neuroactive-ligand receptors and transcription factors. (I) Bulk intact islets; (β) beta; (α) alpha; (δ) delta; (γ) PP/gamma; (A) acinar; (D) ductal; (S) stellate cells. (*C*) Monogenic diabetes–associated genes and their cell-type–specific expression in islets. Violin plots show the $\log_2$(CPM) expression of each gene across cell types. (CHI) congenital hyperinsulinism; (MODY) maturity onset diabetes of the young; (TNDM) transient neonatal diabetes mellitus; (PNDM) permanent neonatal diabetes mellitus. (*D*) RNA in situ hybridization (ViewRNA, Affymetrix) of OCT-embedded islet sections from donor P3 labeling *SST* (red), *LEPR* (green), and nuclei (DAPI; blue). White arrowheads indicate $SST^+/LEPR^+$ cells. ViewRNA of OCT-embedded islet sections from donor P4 to detect the following: (*E*) *INS* (red), *HADH* (green), and nuclei (DAPI; blue) and (*F*) *SST* (red), *HADH* (green), and nuclei (DAPI; blue). White arrowheads highlight examples of $HADH^+/INS^-$ (*E*) and $HADH^+/SST^+$ (*F*) cells. Hollow arrowheads highlight $HADH^+/INS^+$ (*E*) and $HADH^+/SST^-$ (*F*) cells. In *D*–*F*, solid horizontal white lines indicate scale bars of 20 μm. In *E* and *F*, white dashed lines highlight a cell either co-expressing (*E*) *INS*/*HADH* or (*F*) *SST*/*HADH*. White squares in the *bottom left* of *E* and *bottom right* of *F* indicate magnified images of the cells highlighted in respective dashed white boxes.

facilitate vagal stimulation of neuropeptide Y receptors in the hypothalamus and induce satiety (Khandekar et al. 2015). Our data suggest interesting parallels in expression between PP/gamma cells and serotonergic neurons, a group of neurons that influence various cognitive and physiological processes including anxiety, mood, sleep, and satiety. We report expression of *FEV*, a serotonergic transcription factor and necessary driver of neuronal maturation previously reported in mouse beta cells (Ohta et al. 2011), in PP/gamma cells (average $\log_2$CPM of 2.172). Interestingly, FEV has also been implicated in beta cell differentiation, and *Fev*

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

−/− mice exhibit insulin production, insulin secretion, and glucose clearance defects (Ohta et al. 2011). Other related signature genes in PP/gamma cells include *TPH1*, encoding a tryptophan hydroxylase essential for the initial catalysis of serotonin, and *SLC6A4*, a serotonin reuptake transporter. Serotonin colocalizes with insulin in beta cells and promotes GSIS (Paulmann et al. 2009). Mice lacking *TPH1* are diabetic and exhibit impaired insulin secretion due to a lack of pancreatic serotonin (Paulmann et al. 2009). Elevated *FEV*, *TPH1*, and *SLC6A4* expression suggests PP/gamma cells share a suite of characteristic genes with serotonergic neurons that, in the pancreas, integrate central and peripheral hunger and satiety cues. We also observed high PP/gamma expression of muscarinic acetylcholine receptor M3, *CHRM3*, which stimulates exocrine pancreatic amylase (Gautam et al. 2005), insulin secretion (Kong and Tobin 2011; Molina et al. 2014), and smooth muscle contraction and gastric emptying (Eglen et al. 1994). These data implicate the less abundant delta and PP/gamma cell types as critical for islet function via the integration of systemic cues and warrant further studies to elucidate the function and health of these cells in normal and diabetogenic conditions.

Single-cell transcriptomes link rare and common diabetes genetic risk genes to islet cell types

We next sought to understand the cell type(s) involved in rare forms of diabetes, including transient/permanent neonatal diabetes (T/PNDM), CHI and MODY, as well as more common forms of islet dysfunction and diabetes (T1D/T2D). Monogenic diabetic disorders, including CHI, MODY, and neonatal diabetes, are characterized by mutations in a single gene, often resulting in beta cell dysfunction and death (Schwitzgebel 2014). Five monogenic diabetes risk genes (Supplemental Table S11; Hoffmann and Spengler 2012; Senniappan et al. 2013; Schwitzgebel 2014), were enriched in beta cells (i.e., greater than fourfold change in expression in specific islet cell type relative to other endocrine cells), including glucose transporter *SLC2A2* (data not shown), beta cell maturation transcription factor *PDX1*, and the sulfonylurea drug target *ABCC8* (Fig. 4C). *PDX1* expression has been reported in human (Li et al. 2016) and mouse (DiGruccio et al. 2016) beta and delta cells. Despite the modest number of delta cells sampled, our data also suggest moderate *PDX1* expression in delta cells (four of 16 delta cells with expression ≥16 CPM). Robust expression of *HADH* in both beta and delta cells (Fig. 4C) was confirmed by in situ (View RNA) analyses (Fig. 4E,F). Approximately 386/457 cells (84%) in *HADH* and *INS* labeled sections coexpressed both markers (shown in Fig. 4E). Adjacent *SST/HADH* colabeling yielded an approximately equal proportion ($n = 255/306$; 83%) of *SST*-negative/*HADH*-positive cells. Finally, 43/457 (9%) cells were *INS* negative/*HADH* positive, and 41/306 (13%) cells coexpressed *SST* and *HADH* (shown in Fig. 4F) in the respective serial sections. Another CHI-associated gene, *UCP2* (González-Barroso et al. 2008; Senniappan et al. 2013), which was reported to be highly expressed in human beta cells (Liu et al. 2013) and to suppress insulin secretion (Krauss et al. 2003), was enriched in delta cells (Fig. 4C). Delta cell expression of monogenic diabetes genes thus implicate this cell type in the molecular genetics of rare islet dysfunction and diabetes disorders, particularly CHI.

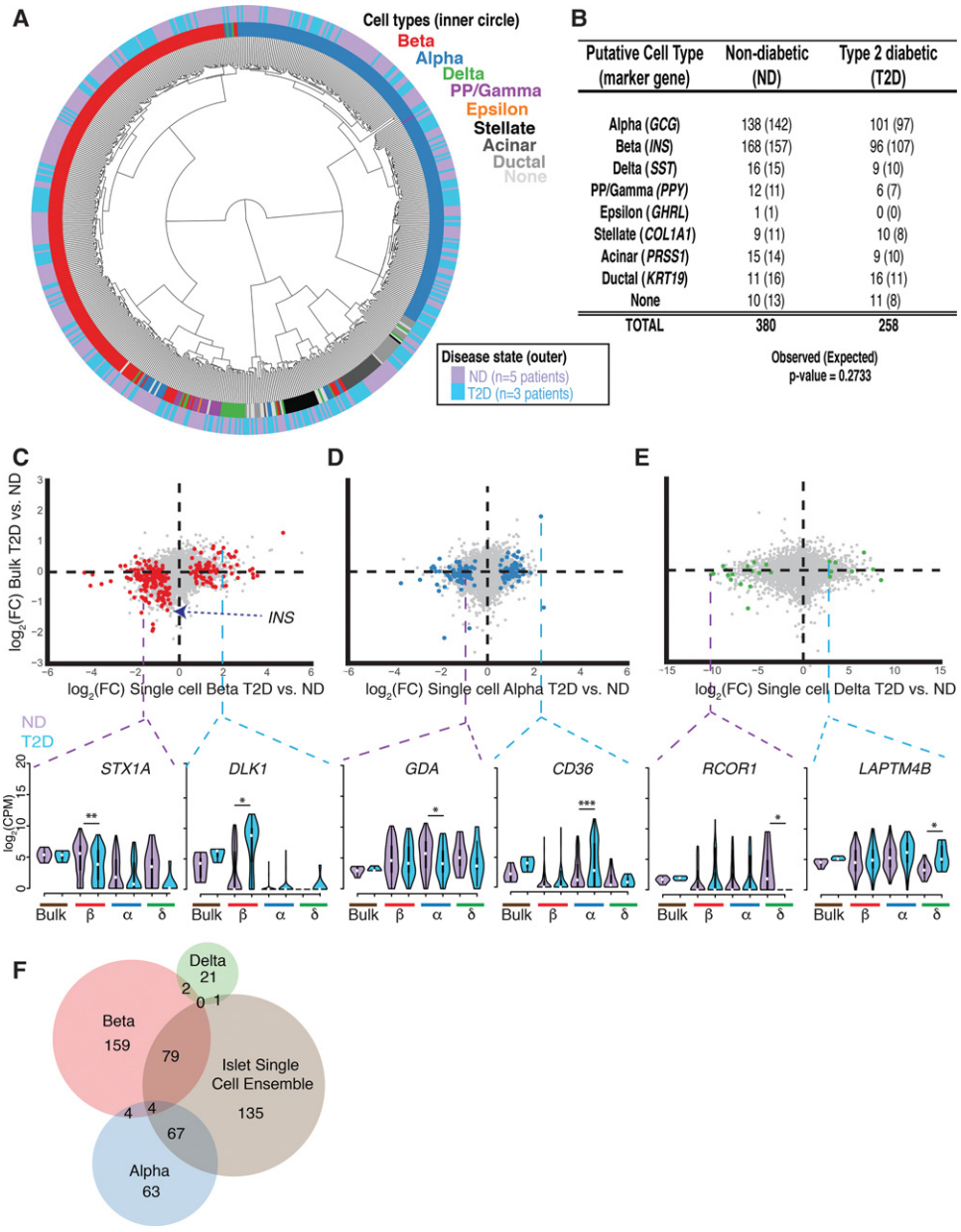
We also investigated cell type expression patterns of 536 islet expression quantitative trait loci (eQTL) target genes (Lyssenko et al. 2009; Dupuis et al. 2010; Dayeh et al. 2013; Fadista et al. 2014; Kulzer et al. 2014; van de Bunt et al. 2015). The majority of these genes ($n = 309$; Supplemental Table S11) were lowly expressed in both the endocrine islet single-cell transcriptomes and in the paired bulk islet transcriptomes (Supplemental Fig. S16A). One hundred fifty-nine additional genes did not exhibit a greater than or equal to fourfold expression change in any endocrine islet cell type. Of the remaining 68 eQTL genes, 54, 46, 51, and 43 were expressed in beta, alpha, delta, and PP/gamma cells, respectively. Surprisingly, beta and delta cells possessed the highest numbers of cell-type–specific eQTL genes (Supplemental Table S11).

Genome-wide association studies (GWAS) have identified more than 100 loci associated with T2D and related quantitative traits (Mohlke and Boehnke 2015). Because GWAS identify genetic variants associated with a disease, but not the specific gene(s) affected (Pearson and Manolio 2008; Manolio 2010), we took two approaches to assess cell-type expression of patterns of putative GWAS genes. First, we compiled and examined a list of 197 reported putative T1D and T2D GWAS genes (Bakay et al. 2013; Nica et al. 2013; Fadista et al. 2014; Marroqui et al. 2015; Mohlke and Boehnke 2015). Of these genes, 37 were expressed in beta, 24 in alpha, 28 in delta, and 22 in PP/gamma cells (Supplemental Table S11). Similarly, genes that were cell-type specific were expressed at higher levels in ND bulk intact islets compared with those genes without cell-type specificity (Supplemental Fig. S16B). Ten genes were uniquely expressed in beta cells, including *MEG3*, a type 1 diabetes (T1D)–associated lincRNA with reported expression in mouse beta cells and potential tumor suppressor activity (Modali et al. 2015), and *IAPP*, whose protein product, when aggregated, possesses cytotoxic properties that may contribute to beta cell death and dysfunction in T2D (Westermark et al. 2011). We also identified five putative T2D GWAS genes (including *HHEX*) to be uniquely expressed in delta cells. To conduct a more liberal analysis of putative GWAS genes, we identified all single-nucleotide polymorphisms (SNPs) associated with polygenic diabetes and related traits from the GWAS catalog (https://www.ebi.ac.uk/gwas/). For each reported SNP associated with T2D, T1D, fasting insulin, fasting glucose, and proinsulin, we examined the expression of all genes overlapping within one megabase of the chromosomal locus and identified 263 genes with cell-type–specific expression (Supplemental Table S12). Together, our observations of cell-type–specific expression of eQTL and monogenic and common (T2D GWAS) diabetes genes both confirm beta cell–specific expression of multiple diabetes-associated genes (*MEG3*, *DLK1*, *SLC2A2*, etc.) and implicate other cell types in the molecular genetic pathogenesis of diabetes. In light of recent studies (Zhang et al. 2014; DiGruccio et al. 2016) and our data, which suggest that delta cells may be critical regulators of glucose homeostasis and islet function, this provides a new avenue for investigation of T2D pathogenesis, as well as potentially new therapeutic targets and treatment options.

Comparison of T2D and ND single-cell transcriptomes uncovers cell-type–specific differences not detected in whole islets

Finally, we compared single-cell transcriptome profiles from T2D and ND donors to identify differentially regulated genes and obtain greater insight into the molecular genetic pathogenesis of diabetes. After unsupervised hierarchical clustering (Fig. 5A) and t-SNE analysis (Supplemental Figs. S17, S18) using 2754 of the most highly expressed genes (Supplemental Table S4), we observed that transcriptomes clustered by cell type regardless of disease state. As previously observed, clustering was not driven by marker gene expression (Supplemental Figs. S19, S20). For regions of the dendrogram (Fig. 5A) where samples appeared to cluster by disease

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press



Putative Cell Type (marker gene)	Non-diabetic (ND)	Type 2 diabetic (T2D)
Alpha (*GCG*)	138 (142)	101 (97)
Beta (*INS*)	168 (157)	96 (107)
Delta (*SST*)	16 (15)	9 (10)
PP/Gamma (*PPY*)	12 (11)	6 (7)
Epsilon (*GHRL*)	1 (1)	0 (0)
Stellate (*COL1A1*)	9 (11)	10 (8)
Acinar (*PRSS1*)	15 (14)	9 (10)
Ductal (*KRT19*)	11 (16)	16 (11)
None	10 (13)	11 (8)
TOTAL	380	258

Figure 5. Single-cell transcriptome analyses identify cell-type–specific expression changes in T2D islets. (*A*) T2D and ND single-cell transcriptomes cluster together by cell type after unsupervised hierarchical clustering. (*B*) Number of each ND and T2D cell type classified by marker gene expression as shown in Figure 2. The numbers of cells expected in each condition based on a χ^2 test are indicated in parentheses. (*C–E, top*) Scatter plots of $\log_2$ fold-change (FC) expression detected between T2D and ND samples from bulk intact RNA-seq (*y*-axis) and from Fluidigm C1 single-cell RNA-seq (*x*-axis) from beta cells (*left* plot; red), alpha cells (*middle* plot; blue), and delta cells (*right* plot; green). (*Bottom*) Violin plots highlight examples of differentially expressed genes in one single-cell type. Dashed purple lines represent repressed genes in the respective T2D cell type, while dashed blue lines represent induced genes. (*) FDR < 0.05, (**) FDR < 0.01, (***) FDR < 0.001. (*F*) Venn diagram showing the intersections of differentially expressed genes identified between T2D and ND transcriptomes at single-cell-type and islet single-cell ensemble resolution. The islet single-cell ensemble represents the pooled collection of beta, alpha, delta, and PP/gamma single cells.

state, we found that islet donor identity was an underlying factor that reflected sample subclustering (Supplemental Fig. S21). We obtained fewer beta cells among the T2D islet cells sampled compared with ND samples (Fig. 5B). However, observed differences in T2D and ND single-cell proportions did not differ significantly from expected cell-type proportions (Fig. 5B, χ^2 *P*-value = 0.2733), and none of the islets from these newly diagnosed T2D individuals exhibited as significant a decrease as previously reported (Butler et al. 2003; Cnop et al. 2005; Donath et al. 2005; Prentki and Nolan 2006).

Recent studies have reported features of beta cell de-differentiation under diabetogenic and stress conditions (Talchai et al. 2012; Wang et al. 2014; Cinti et al. 2016). However, we did not identify significant shifts in islet cell populations, increases in number of hormone-negative "none" cells, or appearances of new or more abundant populations of cells in T2D islets that clustered distinctly

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

from the known islet cell types in this study. Moreover, expression of reported de-differentiation genes including *FOXO1*, *NANOG*, and *POU5F1* (Talchai et al. 2012) did not differ significantly between T2D and ND islet cell types nor the paired bulk intact islet preparations (Supplemental Fig. S22). Finally, other de-differentiation markers such as *NEUROG3* and *MYCL* were not detected in our single-cell or bulk intact islet data. Thus, our analysis did not identify transcriptional evidence of de-differentiated cells in T2D islets.

Comparison of islet cell-type transcriptomes (e.g., T2D beta vs. ND beta) did, however, identify 410 genes that were differentially expressed (FDR < 5%) between T2D and ND donors (Supplemental Table S6) beta, (Fig. 5C, $n = 248$), alpha (Fig. 5D, $n = 138$), and delta cells (Fig. 5E, $n = 24$). We also identified differentially expressed genes in acinar ($n = 74$), ductal ($n = 35$), and stellate ($n = 28$) exocrine cell types (Supplemental Fig. S23; Supplemental Table S6). T2D beta cells exhibited a 1.4-fold decreased *INS* expression compared with ND beta cells (Fig. 5C). *STX1A* was significantly reduced ($\log_2$FC −1.5178) in T2D beta cells, consistent with reported decreases in STX1A protein levels in T2D beta cells (Andersson et al. 2012). STX1A combines with SNAP-25 and VAMP2 to form a tertiary SNARE protein complex important for insulin secretion in beta cells (Andersson et al. 2012), and *STX1A* inhibition drastically reduces GSIS and exocytosis (Vikman et al. 2006). Additionally, we detected elevated *DLK1* expression in T2D beta cells ($\log_2$FC 2.010), which has been implicated in T1D/T2D GWAS (Wallace et al. 2010) and is part of a dysregulated locus in T2D islets (Kameswaran et al. 2014). *Dlk1*$^{-/-}$ mice exhibit increased glucose sensitivity and insulin secretion (Abdallah et al. 2015), and high levels of serum DLK1 have been associated with insulin resistance in both rodents and humans (Chacón et al. 2008). Immunofluorescence indicates that DLK1 is beta cell specific in human but not mouse islets (Li et al. 2016), and FACS-enriched mouse beta cells show low expression of *Dlk1* in comparison to other sorted islet alpha and delta cells (DiGruccio et al. 2016), potentially implicating a unique role of this gene in human T2D progression. These findings suggest that perturbations in *STX1A* and *DLK1* expression may contribute to the beta cell dysfunction and impaired insulin secretion that is commonly observed in T2D pathogenesis.

Decreased beta cell function and mass are hallmarks of T2D pathophysiology (Cerf 2013; Halban et al. 2014). Our analyses suggest that transcriptional changes in nonbeta cells may also contribute to T2D pathogenesis. Specifically, we highlight increased expression of fatty acid translocase gene *CD36* ($\log_2$FC 2.296), as well as decreased expression of the guanine deaminase gene, *GDA* ($\log_2$FC −1.062), in T2D alpha cells. Soluble CD36 is a biomarker of T2D (Alkhatatbeh et al. 2013) and diabetic nephropathy (Shiju et al. 2015) and coordinates activation of the NLRP3 inflammasome, leading to proinflammatory cytokine release and reduced insulin secretion (Sheedy et al. 2013). Within T2D delta cell transcriptomes, we note increased *LAPTM4B* expression ($\log_2$FC 2.871) and drastically reduced *RCOR1* expression ($\log_2$FC −10.128). The underlying biological significance of these differentially regulated genes remains unclear and thus requires further investigation of their roles in nonbeta cell types and T2D pathology. We also compared the transcriptional differences between T2D and ND endocrine cells without first segregating them into islet cell types (334 ND and 212 T2D single-cell profiles). Approximately 66% of beta cell–specific ($n = 165/248$), 50% of alpha cell–specific ($n = 67/138$), and >90% of delta–specific ($n = 23/24$) changes in gene expression were missed when cell types were not defined and specifically compared (Fig. 5F). The decreased heterogeneity in the transcriptional profiles of cell-type–specific comparisons provides increased power to detect the transcriptomic differences and argues the importance of single-cell analysis in understanding the molecular basis of T2D.

Recent islet single-cell studies emerged while this study was under review. We therefore sought to validate our observed cell-type–specific differences in T2D islets using these independent data sets (Wang et al. 2016; Segerstolpe et al. 2016). We found that 54/77 genes and 32/171 were also significantly up- and down-regulated, respectively, in T2D beta cells in these studies ($P < 0.05$, two-sided Wilcoxon rank-sum test) (Supplemental Fig. S24A,B; Supplemental Table S13). Notably, *DLK1* consistently exhibited approximately fourfold induction in T2D beta cells in each study (Supplemental Fig. S24C,D) Similarly, 39/60 and 14/78 genes were significantly induced or repressed, respectively, in T2D alpha cells (Supplemental Fig. S24E,F). This included approximately twofold *CD36* induction in each study (Supplemental Fig. S24G,H). Validation rates for delta cells was notably lower, likely due to the relatively few cells profiled for comparison. However, we did note a significant increase ($\log_2$FC 1.203) in *LAPTM4B* in T2D delta cells from Segerstolpe et al. (2016), consistent with our data.

Discussion

In this study, we completed transcriptome profiling and analysis of 638 single islet cells from ND and T2D individuals. Single-cell RNA-seq protocols are often limited by their capture efficiency due to the fact that a limited proportion of each cell's total transcripts is represented in the final sequencing library (Liu and Trapnell 2016). Additionally, these approaches have difficulty detecting expression and changes in expression of low abundance transcripts. Despite these limitations, we observed a strong correlation between the transcriptomes of paired bulk islets and single cells, indicating these are high-quality and representative data sets. Based on single-cell transcriptome profiles, we have identified cells of each endocrine (alpha, beta, delta, PP/gamma, epsilon) and exocrine (stellate, ductal, acinar) type in the pancreas in an agnostic and data-driven manner.

This approach has defined expression signatures of each cell type with single-cell precision. Cell-type–specific expression patterns in our data such as *MAFA* in beta cells and *IRX2* in alpha cells are concordant with and extend those generated on a smaller set of cells and an independent platform (Li et al. 2016). Notably, our approach also uncovered important instances of shared expression between these cell types and the less common delta and PP/gamma islet populations, including genes mutated in CHI (*HADH*) and transcription factors regulating cell fate/identity (*ARX*). *HADH* encodes hydroxyacyl-CoA dehydrogenase, an important enzyme and negative regulator of glutamate dehydrogenase (GDH) and insulin secretion. Expression of *HADH* in islets has been shown to be beta cell specific (Kapoor et al. 2010; Pepin et al. 2010), and indeed, knockdown of *HADH* in rat 832/13 beta cells increases insulin secretion (Pepin et al. 2010). Surprisingly, our combined transcriptomic analyses and in situ (ViewRNA) validation of *HADH* revealed shared expression in beta and delta cells. These findings suggest delta cell dysfunction, in addition to beta cell dysfunction, as potential contributing factors to the development of monogenic diabetic disorders.

Most importantly, analysis of the delta and PP/gamma islet cell transcriptomes revealed cell-type–specific expression of multiple genes that suggest important roles for these cells in islet

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

physiology and the molecular genetics of islet dysfunction in rare (e.g., PNDM, TNDM, and MODY) and common (e.g., T2D) forms of diabetes. The novel transcriptome signatures uncovered for human delta and PP/gamma cells includes genes that strongly suggest important roles for each cell type in sensing and integrating specific systemic cues to govern islet (dys)function. This clearly warrants additional work to better understand the regulation and function of these cells in normal and diabetic states. New cell surface markers identified for each of these cell types could be used to specifically enrich and purify these populations for detailed functional analysis.

Finally, by comparing single-cell transcriptomes from T2D and ND islets, we were able to study quantitative changes in cell populations and cell-type–specific expression in T2D pathogenesis. Although not reaching statistical significance, we did observe a trend of decreased beta cells in T2D islets versus ND islets. This difference was not as pronounced as in previous reports, possibly due to the relatively modest number of cells sampled per individual. Alternatively, as most of the T2D islet single-cell transcriptomes came from newly diagnosed individuals, this difference may also reflect the shorter duration or decreased severity of T2D in these samples compared with other studies. Previous studies suggested that beta cell de-differentiation may underlie beta cell loss in T2D (Talchai et al. 2012; Wang et al. 2014; Cinti et al. 2016). However, a subsequent study comparing human islets from 14 T2D and 13 ND individuals did not identify clear evidence of this phenomenon (Butler et al. 2016). Similarly, our data do not provide transcriptome-based evidence of *trans*-differentiation or de-differentiation phenomena in T2D islets. We observed neither the appearance of new or distinct subpopulations among the T2D islet single cells nor significant changes of reported de-differentiation genes between T2D and ND cell types (e.g., T2D beta cells vs. ND beta cells), although it is possible that de-differentiated cells were simply not captured in our study. Overall, we identify 248, 138, and 24 genes exhibiting differential expression in T2D versus ND beta, alpha, and delta cells, respectively. Consistent with Simpson's paradox, approximately half of these genes in each major islet cell type (64% beta, 45% alpha) and ~90% of these in the less abundant delta cells were not detected in whole islet or single-cell islet transcriptomes when they were not stratified by cell type (Simpson 1951; Trapnell 2015). Each of these differentially regulated genes may represent important new candidate genes in T2D pathogenesis and therapeutic targeting.

Methods

Islet acquisition, processing, and dissociation

Islets were procured from ProdoLabs or the Integrated Islet Distribution Program (IIDP) and shipped in PIM(T) media (ProdoLabs) overnight on cold packs. Upon arrival, islets were washed and transferred into PIM(S) media with PIM(G) and PIM (ABS) supplements according to the manufacturer's instructions (ProdoLabs) and incubated at 37°C in a 5% CO_2 tissue culture incubator. Twenty-four hours after transfer, approximately 500 islet equivalents (IEq) were aliquoted and centrifuged at 180*g* for 3 min at room temperature (RT). One aliquot (100 IEq) was immediately flash frozen (Fig. 1A, baseline), one (200 IEq) was resuspended in 1 mL Prodo-media (Fig. 1A, intact), and one (200 IEq) was resuspended in 1 mL Accutase (Innovative Cell Technologies) (Fig. 1A, dissociated and single cell) and incubated for 10 min in a 37°C water bath, with pipetting every 2 min. Accutase-dissociated cells were filtered through a prewet cell strainer (BD) to collect single cells, rinsed with 9 mL prewarmed CMRL + 10%FBS media to stop the reaction, and centrifuged at 180*g* for 3 min at RT. Dissociated cells were resuspended in 300 µL CMRL + 10%FBS media. Cell size, number, and viability were assessed using Countess II FL (Thermo Fisher Scientific), and the cell suspension was diluted to a final concentration of 300 cells/µL. Total processing and handling time for each islet was ≤60 min.

Single-cell processing on the CI single-cell Autoprep system

After counting, cells were diluted to a final concentration range of 250–400 cells/µL and 5 µL loaded onto each C1 integrated fluidic circuit (IFC; 10- to 17-µm chip) for cell capture on the C1 single-cell Autoprep system. For each islet preparation, up to two microfluidic chips were used. After capture, cells were imaged within each capture nest with an EVOS FL auto microscope (Life Technologies). IFCs were subsequently loaded with additional reagents for subsequent cell lysis; SMARTer v1- based (Clontech), olio-(dT)-primed reverse transcription; template switching for second-strand priming; and amplification of cDNA on the C1 System. Qualitative and quantitative analysis of all single-cell cDNA products was performed on a 96 capillary fragment analyzer (Advanced Analytical). Only cell singlets, as determined by imaging, with adequate cDNA yield and quality were processed for subsequent sequencing. Fragmentation and tagmentation of cDNA was done with Nextera XT reagent (Illumina) using dual indices to prepare single-cell multiplexed libraries.

Bulk-cell RNA-seq

Bulk cells were pelleted and RNA purified using the PicoPure RNA isolation kit (Life Technologies). All RNA-seq libraries from bulk-sample RNA were generated with the same SMARTer v1 chemistry (Clontech) as for the C1 single-cell data largely following the manufacturer's instructions. Unlike the C1 workflow, after first-strand DNA synthesis, cDNA was purified using Agencourt AMPure beads (Beckman Coulter). cDNA was subsequently amplified through 12 PCR cycles. The cDNA yield and fragment size were measured on a 2100 Bioanalyzer (Agilent). For sequencing library preparation, amplified cDNA was sheared using a Covaris LE220 system to obtain fragments of ~200 bp. The fragmented cDNA was prepared for sequencing using the NEBNext DNA library prep kit for Illumina sequencing (New England Biolabs).

Sequencing, read mapping, and quality control

All sequencing was performed on a NextSeq500 (Illumina) using the 75-cycle high-output chip. RNA-seq reads were subjected to quality control using custom scripts developed at the computational sciences group at The Jackson Laboratory. Briefly, reads with >30% of bases with quality scores less than 30 were removed from the analysis, and samples with >50% of the low-quality reads were removed from the cohort. Trimmed reads were mapped to human transcriptome (GRCh37, Ensembl v70) using Bowtie 2 (Langmead and Salzberg 2012), and expression levels of all genes were estimated using RSEM (Li and Dewey 2011). Transcript per million (TPM) values as defined by RSEM were added a value of one prior to $\log_2$ transformation to avoid zeros. GRCh37 was selected for mapping to facilitate integration and comparative analyses with existing islet data sets (e.g., Parker et al. 2013; Fadista et al. 2014; van de Bunt et al. 2015) and ENCODE and NIH Roadmap data by members of the islet biology, diabetes, and functional genomics communities. The observation of expected "positive control" genes for each cell type strongly suggested that mapping to

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

GRCh37 instead of GRCh38 did not mask or alter any important conclusions that could be drawn from the data.

Single-cell sample processing and quality filtering

We used 26,616 protein-coding genes and lincRNAs from the GRCh37, Ensembl v70 build in our study. Genes with expression five or more TPMs in a sample were defined as expressed. Seventy-two single-cell samples that expressed fewer than 3500 genes according to these criteria were removed from downstream analysis.

Islet cell type classification

GMM of islet marker genes was performed on a per gene basis using the R-package *mclust_5.2* (Scrucca et al. 2016). Each single-cell sample was classified as a specific pancreatic cell type if and only if a single gene from the selected marker gene list—*INS* (beta), *GCG* (alpha), *SST* (delta), *PPY* (PP/gamma), *KRT19* (ductal), *PRSS1* (acinar), and *COL1A1* (stellate)—was expressed in the sample and none of the other marker genes were expressed. Cells expressing no marker genes were labeled as "none," and those expressing >1 marker gene were labeled as "multiple." Fluidigm released a white paper report detailing the potential for single cells to "*z*-stack" in up to 30% of capture nests on the medium (10–17 µm) Fluidigm C1 chip (http://info.fluidigm.com/rs/673-MRG-416/images/C1-Med-96-IFC-Redesign_wp_101-3328B1_FINAL.pdf). DAPI staining identified *z*-stacked islet cell doublets in 10% and 30% of capture nests from two additional C1 single-cell captures. Because the proportion of "multiple" labeled cells approximately equaled that of *z*-stacked doublets, we discarded these cells (n = 340) from subsequent analyses.

Unsupervised dimensionality reduction and hierarchical clustering

Barnes-Hut variant of t-SNE (van der Maaten 2014) was implemented using the R-package *Rtsne_0.10* (https://github.com/jkrijthe/Rtsne). ND single-cell transcriptomes were reduced to two dimensions in an unsupervised manner using genes with $\log_2$-CPM values greater than 10.5 in at least one sample. Similar analyses were repeated using only the T2D single-cell data and the combined single-cell data. Hierarchical clustering of cell transcriptomes was performed using Euclidean distance, Ward.D2 linkage, and the same gene selection criteria. Resultant "fan" dendrogram images were produced using the R-packages *dendextend_1.1.8* (Galili 2015) and *ape_3.5* (Paradis et al. 2004). Bulk islet transcriptomes were clustered using the same criteria.

Supervised differential gene expression analysis

Differential expression analyses were performed using the Bioconductor package *edgeR_3.14.1* (Robinson et al. 2010). Gender was used as a blocking factor to account for variability between male and female patient islets. In each comparison, protein-coding genes and lincRNAs with 20 or fewer counts in at least 20% of either cell type population being compared or at least a minimum of three cells were used. Differentially expressed genes with FDR < 5% were regarded as significant results. Endocrine cell signature genes were identified by first performing the above differential expression analysis procedure between each endocrine cell type (e.g., beta vs. alpha, beta vs. delta, and beta vs. PP/gamma). Afterward, the intersection of these results was performed to identify genes exclusively enriched in the cell type. Exocrine cell signature genes were identified after pairwise comparisons between each respective exocrine cell type (e.g., acinar vs. stellate, acinar vs. ductal). Comparisons between T2D and ND single-cell transcriptomes were performed for the same cell types (e.g., T2D beta cells vs. ND aeta cells) to identify cell-type–specific differences in gene expression between T2D and ND states.

ANOVA and post-hoc analyses

For each collection of diabetes-associated and eQTL genes examined, one-way analysis of variance (ANOVA) was used to identify statistically significant differences (FDR > 5%) in islet cell-type gene expression. Following this, we performed a post hoc analysis using a THSD test to determine genes with cell-type–specific expression patterns (fold change > 4). Genes were classified as "pan-islet" if they had an average $\log_2$(CPM) expression greater than four in all islet cell types. Genes that were not enriched in a cell type or pan-islet were classified as "lowly expressed" (average $\log_2$(CPM) < 2 in all cell types), and the remaining genes were classified as "less than fourfold change." This same methodology was used to characterize expression of the previously reported alpha- and beta-specific genes from Dorrell et al. (2011b), Bramswig et al. (2013), Nica et al. (2013), and Blodgett et al. (2015). Similar methods were used to characterize expression patterns of genes nearby diabetes-related GWAS SNPs (downloaded from the GWAS Catalog, https://www.ebi.ac.uk/gwas/, and available in Supplemental Table S12). Protein-coding RNAs and lincRNAs that overlapped within one megabase upstream of and downstream from the diabetes-associated SNPs were analyzed.

Gene set enrichment analysis

The Bioconductor package *gage_2.22.0*, (Luo et al. 2009) was used with default parameters to identify enrichment (FDR < 5%) of human Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in each of the ND islet cell transcriptomes. Enriched pathways were determined by comparing each cell-type transcriptome against the other aggregate islet cell-type transcriptomes (e.g., beta vs. alpha, delta, and PP/gamma).

RNA in situ hybridization

RNA transcripts were visualized in OCT-embedded pancreatic islet sections from two ND donors (P3 and P4) using QuantiGene ViewRNA ISH cell assay kit (catalog no. QVC0001, Affymetrix). Human *INS* ViewRNA type 6 probe (Catalog no. VA6-13248-06), *SST* ViewRNA type 6 probe (catalog no. VA6-17244-06), *LEPR* ViewRNA type 1 probe (catalog no. VA1-15221-06), and *HADH* ViewRNA type 1 probe (catalog no. VA1-12106-06) were purchased from Affymetrix (Santa Clara). The assay was performed according to the cell-based ViewRNA assay protocol with a 15-min formaldehyde fixation and a 10-min protease treatment (dilution factor 1:4000). ViewRNA probes were detected at 550 nm (Cy3) and 650 nm (Cy5) using a Leica TSC SP8 confocal microscope at 63× magnification.

Islet cell subpopulation analyses

Dorrell et al. 2016 previously defined four beta cell subpopulations with differing expression of 59 genes. With this gene set, we attempted to validate the presence of these four subpopulations via unsupervised t-SNE and hierarchical clustering of all ND beta cell transcriptomes (n = 168). Similarly, Bader et al. (2016) characterized proliferative (*Fltp*$^+$/FVR$^+$) and mature (*Fltp*$^-$/FVR$^-$) mouse beta cells that showed differential expression of 996 transcripts. By using the Mouse Genome Informatics (MGI; http://www.informatics.jax.org) database, these 996 transcripts corresponded to 691 human orthologs that were detected in our data set. Beta

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

cell transcriptomes were clustered using these orthologs to attempt to identify mature and proliferating subpopulations. Finally, we used the functions available in *scran_1.04* (http://bioconductor.org/packages/release/bioc/html/scran.html) to computationally assign single-cell samples into cell cycle phases (G1, G2/M, or S phase) and investigate whether our data set contained proliferating islet cells.

Data access

Raw sequence data from this study have been submitted to the databases of NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under accession number SRP075970 and BioProject (http://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA323853. Processed data sets from this study have been submitted to Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE86473. UCSC Genome Browser tracks of aggregate ND islet single-cell-type transcriptomes are available at http://genome.ucsc.edu/ and may be accessed with the user name "lawlorn" and session name "Islet_Single_Cell_Type_Transcriptomes." The source code used to produce the figures and tables in this paper is available in the Supplemental_Methods_Source_Code folder as suggested by Hoffman (2016).

Acknowledgments

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs, through the Peer Reviewed Medical Research Program under award no. W81XWH-16-1-0130 and by the laboratory startup funds to M.L.S. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense. We thank Vishal Kumar Sarsani for uploading data to the Sequence Read Archive (SRA), Jane Cha for aiding in figure design and artwork, and Anna Lisa Lucido for help with manuscript preparation. We thank Palucka laboratory members Jan Martinek and Tina Wu for help on in situ RNA analyses and members of the Stitzel, Ucar, and Parker laboratories, Travis Hinson, Karolina Palucka, and FUSION Consortium members for helpful discussions on the study.

References

Abdallah BM, Ditzel N, Laborda J, Karsenty G, Kassem M. 2015. DLK1 regulates whole-body glucose metabolism: a negative feedback regulation of the osteocalcin-insulin loop. *Diabetes* **64:** 3069–3080.

Alkhatatbeh MJ, Enjeti AK, Acharya S, Thorne RF, Lincz LF. 2013. The origin of circulating CD36 in type 2 diabetes. *Nutr Diabetes* **3:** e59.

Andersson SA, Olsson AH, Esguerra JL, Heimann E, Ladenvall C, Edlund A, Salehi A, Taneera J, Degerman E, Groop L, et al. 2012. Reduced insulin secretion correlates with decreased expression of exocytotic genes in pancreatic islets from patients with type 2 diabetes. *Mol Cell Endocrinol* **364:** 36–45.

Arda HE, Li L, Tsai J, Torre EA, Rosli Y, Peiris H, Spitale RC, Dai C, Gu X, Qu K, et al. 2016. Age-dependent pancreatic gene regulation reveals mechanisms governing human β cell function. *Cell Metab* **23:** 909–920.

Bader E, Migliorini A, Gegg M, Moruzzi N, Gerdes J, Roscioni SS, Bakhti M, Brandl E, Irmler M, Beckers J, et al. 2016. Identification of proliferative and mature β-cells in the islets of Langerhans. *Nature* **535:** 430–434.

Baetens D, Malaisse-Lagae F, Perrelet A, Orci L. 1979. Endocrine pancreas: Three-dimensional reconstruction shows two types of islets of Langerhans. *Science* **206:** 1323–1325.

Bakay M, Pandey R, Hakonarson H. 2013. Genes involved in type 1 diabetes: an update. *Genes* **4:** 499–521.

Blodgett DM, Nowosielska A, Afik S, Pechhold S, Cura AJ, Kennedy NJ, Kim S, Kucukural A, Davis RJ, Kent SC, et al. 2015. Novel observations from next-generation RNA sequencing of highly purified human adult and fetal islet cell subsets. *Diabetes* **64:** 3172–3181.

Bramswig NC, Everett LJ, Schug J, Dorrell C, Liu C, Luo Y, Streeter PR, Naji A, Grompe M, Kaestner KH. 2013. Epigenomic plasticity enables human pancreatic α to β cell reprogramming. *J Clin Invest* **123:** 1275–1284.

Brissova M, Fowler MJ, Nicholson WE, Chu A, Hirshberg B, Harlan DM, Powers AC. 2005. Assessment of human pancreatic islet architecture and composition by laser scanning confocal microscopy. *J Histochem Cytochem* **53:** 1087–1097.

Butler AE, Janson J, Bonner-Weir S, Ritzel R, Rizza RA, Butler PC. 2003. β-Cell deficit and increased β-cell apoptosis in humans with type 2 diabetes. *Diabetes* **52:** 102–110.

Butler AE, Dhawan S, Hoang J, Cory M, Zeng K, Fritsch H, Meier JJ, Rizza RA, Butler PC. 2016. β-Cell deficit in obese type 2 diabetes, a minor role of β-cell dedifferentiation and degranulation. *J Clin Endocrinol Metab* **101:** 523–532.

Cabrera O, Berman DM, Kenyon NS, Ricordi C, Berggren PO, Caicedo A. 2006. The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proc Natl Acad Sci* **103:** 2334–2339.

Cerf ME. 2013. β Cell dysfunction and insulin resistance. *Front Endocrinol* **4:** 37.

Chacón MR, Miranda M, Jensen CH, Fernández-Real JM, Vilarrasa N, Gutiérrez C, Näf S, Gomez JM, Vendrell J. 2008. Human serum levels of fetal antigen 1 (FA1/Dlk1) increase with obesity, are negatively associated with insulin sensitivity and modulate inflammation *in vitro*. *Int J Obes* **32:** 1122–1129.

Chen WM, Erdos MR, Jackson AU, Saxena R, Sanna S, Silver KD, Timpson NJ, Hansen T, Orrù M, Grazia Piras M, et al. 2008. Variations in the *G6PC2/ABCB11* genomic region are associated with fasting glucose levels. *J Clin Invest* **118:** 2620–2628.

Chen VP, Gao Y, Geng L, Parks RJ, Pang YP, Brimijoin S. 2015. Plasma butyrylcholinesterase regulates ghrelin to control aggression. *Proc Natl Acad Sci* **112:** 2251–2256.

Cinti F, Bouchi R, Kim-Muller JY, Ohmura Y, Sandoval PR, Masini M, Marselli L, Suleiman M, Ratner LE, Marchetti P, et al. 2016. Evidence of β-cell dedifferentiation in human type 2 diabetes. *J Clin Endocrinol Metab* **101:** 1044–1054.

Cnop M, Welsh N, Jonas JC, Jörns A, Lenzen S, Eizirik DL. 2005. Mechanisms of pancreatic β-cell death in type 1 and type 2 diabetes. *Diabetes* **54:** S97–S107.

Collombat P, Hecksher-Sørensen J, Krull J, Berger J, Riedel D, Herrera PL, Serup P, Mansouri A. 2007. Embryonic endocrine pancreas and mature β cells acquire α and PP cell phenotypes upon *Arx* misexpression. *J Clin Invest* **117:** 961–970.

Dabbs DJ. 2013. *Diagnostic immunohistochemistry*. Elsevier Health Sciences, Philadelphia, PA.

Dayeh TA, Olsson AH, Volkov P, Almgren P, Rönn T, Ling C. 2013. Identification of CpG-SNPs associated with type 2 diabetes and differential DNA methylation in human pancreatic islets. *Diabetologia* **56:** 1036–1046.

Dayeh T, Volkov P, Salö S, Hall E, Nilsson E, Olsson AH, Kirkpatrick CL, Wollheim CB, Eliasson L, Rönn T, et al. 2014. Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. *PLoS Genet* **10:** e1004160.

DiGruccio MR, Mawla AM, Donaldson CJ, Noguchi GM, Vaughan J, Cowing-Zitron C, van der Meulen T, Huising MO. 2016. Comprehensive α, β and δ cell transcriptomes reveal that ghrelin selectively activates δ cells and promotes somatostatin release from pancreatic islets. *Mol Metab* **5:** 449–458.

Donath MY, Ehses JA, Maedler K, Schumann DM, Ellingsgaard H, Eppler E, Reinecke M. 2005. Mechanisms of β-cell death in type 2 diabetes. *Diabetes* **54:** S108–S113.

Dorrell C, Abraham SL, Lanxon-Cookson KM, Canaday PS, Streeter PR, Grompe M. 2008. Isolation of major pancreatic cell types and long-term culture-initiating cells using novel human surface markers. *Stem Cell Res* **1:** 183–194.

Dorrell C, Grompe MT, Pan FC, Zhong Y, Canaday PS, Shultz LD, Greiner DL, Wright CV, Streeter PR, Grompe M. 2011a. Isolation of mouse pancreatic α, β, duct and acinar populations with cell surface markers. *Mol Cell Endocrinol* **339:** 144–150.

Dorrell C, Schug J, Lin CF, Canaday PS, Fox AJ, Smirnova O, Bonnah R, Streeter PR, Stoeckert CJ Jr, Kaestner KH, et al. 2011b. Transcriptomes of the major human pancreatic cell types. *Diabetologia* **54:** 2832–2844.

Dorrell C, Schug J, Canaday PS, Russ HA, Tarlow BD, Grompe MT, Horton T, Hebrok M, Streeter PR, Kaestner KH, et al. 2016. Human islets contain four distinct subtypes of β cells. *Nat Commun* **7:** 11756.

Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL, et al. 2010. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* **42:** 105–116.

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

Eglen RM, Reddy H, Watson N, John Challiss RA. 1994. Muscarinic acetylcholine receptor subtypes in smooth muscle. *Trends Pharmacol Sci* **15:** 114–119.

Eto K, Tsubamoto Y, Terauchi Y, Sugiyama T, Kishimoto T, Takahashi N, Yamauchi N, Kubota N, Murayama S, Aizawa T, et al. 1999. Role of NADH shuttle system in glucose-induced activation of mitochondrial metabolism and insulin secretion. *Science (New York, N.Y.)* **283:** 981–985.

Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL, Taneera J, Storm P, Osmark P, Ladenvall C, Prasad RB, et al. 2014. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc Natl Acad Sci* **111:** 13924–13929.

Fridlyand LE, Philipson LH. 2010. Glucose sensing in the pancreatic β cell: a computational systems analysis. *Theor Biol Med Model* **7:** 15.

Galili T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31:** 3718–3720.

Gautam D, Han S-J, Heard TS, Cui Y, Miller G, Bloodworth L, Wess J. 2005. Cholinergic stimulation of amylase secretion from pancreatic acinar cells studied with muscarinic acetylcholine receptor mutant mice. *J Pharmacol Exp Ther* **313:** 995–1002.

González-Barroso MM, Giurgea I, Bouillaud F, Anedda A, Bellanné-Chantelot C, Hubert L, de Keyzer Y, de Lonlay P, Ricquier D. 2008. Mutations in *UCP2* in congenital hyperinsulinism reveal a role for regulation of insulin secretion. *PLoS One* **3:** e3850.

Gunton JE, Kulkarni RN, Yim S, Okada T, Hawthorne WJ, Tseng YH, Roberson RS, Ricordi C, O'Connell PJ, Gonzalez FJ, et al. 2005. Loss of *ARNT/HIF1*β mediates altered gene expression and pancreatic-islet dysfunction in human type 2 diabetes. *Cell* **122:** 337–349.

Halban PA, Polonsky KS, Bowden DW, Hawkins MA, Ling C, Mather KJ, Powers AC, Rhodes CJ, Sussel L, Weir GC. 2014. β-Cell failure in type 2 diabetes: postulated mechanisms and prospects for prevention and treatment. *Diabetes Care* **37:** 1751–1758.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589.

Hoffman JI. 2016. Reproducibility: archive computer code with raw data. *Nature* **534:** 326.

Hoffmann A, Spengler D. 2012. Transient neonatal diabetes mellitus gene *Zac1* impairs insulin secretion in mice through Rasgrf1. *Mol Cell Biol* **32:** 2549–2560.

Hrvatin S, Deng F, O'Donnell CW, Gifford DK, Melton DA. 2014. MARIS: method for analyzing RNA following intracellular sorting. *PLoS One* **9:** e89459.

Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA. 2016. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17:** 29.

Kameswaran V, Bramswig NC, McKenna LB, Penn M, Schug J, Hand NJ, Chen Y, Choi I, Vourekas A, Won KJ, et al. 2014. Epigenetic regulation of the DLK1-MEG3 microRNA cluster in human type 2 diabetic islets. *Cell Metab* **19:** 135–145.

Kapoor RR, Heslegrave A, Hussain K. 2010. Congenital hyperinsulinism due to mutations in *HNF4A* and *HADH*. *Rev Endocr Metab Disord* **11:** 185–191.

Khandekar N, Berning BA, Sainsbury A, Lin S. 2015. The role of pancreatic polypeptide in the regulation of energy homeostasis. *Mol Cell Endocrinol* **418:** 33–41.

Kong KC, Tobin AB. 2011. The role of M_3-muscarinic receptor signaling in insulin secretion. *Commun Integr Biol* **4:** 489–491.

Krauss S, Zhang C-Y, Scorrano L, Dalgaard LT, St-Pierre J, Grey ST, Lowell BB. 2003. Superoxide-mediated activation of uncoupling protein 2 causes pancreatic β cell dysfunction. *J Clin Invest* **112:** 1831–1842.

Kulzer JR, Stitzel ML, Morken MA, Huyghe JR, Fuchsberger C, Kuusisto J, Laakso M, Boehnke M, Collins FS, Mohlke KL. 2014. A common functional regulatory variant at a type 2 diabetes locus upregulates *ARAP1* expression in the pancreatic β cell. *Am J Hum Genet* **94:** 186–197.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12:** 323.

Li J, Klughammer J, Farlik M, Penz T, Spittler A, Barbieux C, Berishvili E, Bock C, Kubicek S. 2016. Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Rep* **17:** 178–187.

Liu S, Trapnell C. 2016. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* **5**: 182.

Liu J, Hunter CS, Du A, Ediger B, Walp E, Murray J, Stein R, May CL. 2011. Islet-1 regulates *Arx* transcription during pancreatic islet α-cell development. *J Biol Chem* **286:** 15352–15360.

Liu J, Li J, Li W-J, Wang C-M. 2013. The role of uncoupling proteins in diabetes mellitus. *J Diabetes Res* **2013:** e585897.

Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. 2009. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10:** 161.

Lyssenko V, Nagorny CL, Erdos MR, Wierup N, Jonsson A, Spégel P, Bugliani M, Saxena R, Fex M, Pulizzi N, et al. 2009. Common variant in *MTNR1B* associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nat Genet* **41:** 82–88.

MacDonald PE, Joseph JW, Rorsman P. 2005. Glucose-sensing mechanisms in pancreatic β-cells. *Philos Trans R Soc B Lond B Biol Sci* **360:** 2211–2225.

MacDonald MJ, Longacre MJ, Stoker SW, Kendrick M, Thonpho A, Brown LJ, Hasan NM, Jitrapakdee S, Fukao T, Hanson MS, et al. 2011. Differences between human and rodent pancreatic islets: low pyruvate carboxylase, ATP citrate lyase, and pyruvate carboxylation and high glucose-stimulated acetoacetate in human pancreatic islets. *J Biol Chem* **286:** 18383–18396.

Manolio TA. 2010. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363:** 166–176.

Marobbio CMT, Giannuzzi G, Paradies E, Pierri CL, Palmieri F. 2008. α-Isopropylmalate, a leucine biosynthesis intermediate in yeast, is transported by the mitochondrial oxalacetate carrier. *J Biol Chem* **283:** 28445–28453.

Marroqui L, Dos Santos RS, Fløyel T, Grieco FA, Santin I, Op de Beeck A, Marselli L, Marchetti P, Pociot F, Eizirik DL. 2015. *TYK2*, a candidate gene for type 1 diabetes, modulates apoptosis and the innate immune response in human pancreatic β-cells. *Diabetes* **64:** 3808–3817.

Marselli L, Thorne J, Dahiya S, Sgroi DC, Sharma A, Bonner-Weir S, Marchetti P, Weir GC. 2010. Gene expression profiles of β-cell enriched tissue obtained by laser capture microdissection from subjects with type 2 diabetes. *PLoS One* **5:** e11499.

Mathison A, Liebl A, Bharucha J, Mukhopadhyay D, Lomberk G, Shah V, Urrutia R. 2010. Pancreatic stellate cell models for transcriptional studies of desmoplasia-associated genes. *Pancreatology* **10:** 505–516.

Modali SD, Parekh VI, Kebebew E, Agarwal SK. 2015. Epigenetic regulation of the lncRNA MEG3 and its target c-MET in pancreatic neuroendocrine tumors. *Mol Endocrinol* **29:** 224–237.

Mohlke KL, Boehnke M. 2015. Recent advances in understanding the genetic architecture of type 2 diabetes. *Hum Mol Genet* **24:** R85–R92.

Molina J, Rodriguez-Diaz R, Fachado A, Jacques-Silva MC, Berggren PO, Caicedo A. 2014. Control of insulin secretion by cholinergic signaling in the human pancreatic islet. *Diabetes* **63:** 2714–2726.

Muller YL, Piaggi P, Hanson RL, Kobes S, Bhutta S, Abdussamad M, Leak-Johnson T, Kretzler M, Huang K, Weil EJ, et al. 2015. A *cis*-eQTL in *PFKFB2* is associated with diabetic nephropathy, adiposity and insulin secretion in American Indians. *Hum Mol Genet* **24:** 2985–2996.

Nica AC, Ongen H, Irminger JC, Bosco D, Berney T, Antonarakis SE, Halban PA, Dermitzakis ET. 2013. Cell-type, allelic, and genetic signatures in the human pancreatic β cell transcriptome. *Genome Res* **23:** 1554–1562.

Nussey S, Whitehead S. 2001. *The endocrine pancreas*. BIOS Scientific Publishers, Oxford, UK.

Ohta Y, Kosaka Y, Kishimoto N, Wang J, Smith SB, Honig G, Kim H, Gasa RM, Neubauer N, Liou A, et al. 2011. Convergence of the insulin and serotonin programs in the pancreatic β-cell. *Diabetes* **60:** 3208–3216.

Palmieri L, Vozza A, Agrimi G, De Marco V, Runswick MJ, Palmieri F, Walker JE. 1999. Identification of the yeast mitochondrial transporter for oxaloacetate and sulfate. *J Biol Chem* **274:** 22184–22190.

Pandol SJ. 2011. The exocrine pancreas. In *Colloquium series on integrated systems physiology: from molecule to function to disease* (ed. Granger DN, Granger JP). Morgan & Claypool Life Sciences, San Rafael, CA.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20:** 289–290.

Parker SC, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, van Bueren KL, Chines PS, Narisu N; NISC Comparative Sequencing Program, et al. 2013. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci* **110:** 17921–17926.

Paulmann N, Grohmann M, Voigt JP, Bert B, Vowinckel J, Bader M, Skelin M, Jevsek M, Fink H, Rupnik M, et al. 2009. Intracellular serotonin modulates insulin secretion from pancreatic β-cells by protein serotonylation. *PLoS Biol* **7:** e1000229.

Pearson TA, Manolio TA. 2008. How to interpret a genome-wide association study. *JAMA* **299:** 1335–1344.

Pepin E, Guay C, Delghingaro-Augusto V, Joly E, Madiraju SR, Prentki M. 2010. Short-chain 3-hydroxyacyl-CoA dehydrogenase is a negative regulator of insulin secretion in response to fuel and non-fuel stimuli in INS832/13 β-cells. *J Diabetes* **2:** 157–167.

Prentki M, Nolan CJ. 2006. Islet β cell failure in type 2 diabetes. *J Clin Invest* **116:** 1802–1812.

Reichert M, Rustgi AK. 2011. Pancreatic ductal cells in development, regeneration, and neoplasia. *J Clin Invest* **121:** 4572–4578.

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140.
Schuit FC, Huypens P, Heimberg H, Pipeleers DG. 2001. Glucose sensing in pancreatic β-cells. *Diabetes* **50:** 1–11.
Schwitzgebel VM. 2014. Many faces of monogenic diabetes. *J Diabetes Investig* **5:** 121–133.
Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. mclust 5: clustering, classification and density estimation using Gaussian Finite mixture models. *The R Journal* **8:** 289–317.
Segerstolpe Å, Palasantza A, Eliasson P, Andersson EM, Andréasson AC, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, et al. 2016. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab* **24:** 593–607.
Senniappan S, Arya VB, Hussain K. 2013. The molecular mechanisms, diagnosis and management of congenital hyperinsulinism. *Indian J Endocrinol Metab* **17:** 19–30.
Sheedy FJ, Grebe A, Rayner KJ, Kalantari P, Ramkhelawon B, Carpenter SB, Becker CE, Ediriweera HN, Mullick AE, Golenbock DT, et al. 2013. CD36 coordinates NLRP3 inflammasome activation by facilitating the intracellular nucleation from soluble to particulate ligands in sterile inflammation. *Nat Immunol* **14:** 812–820.
Shiju TM, Mohan V, Balasubramanyam M, Viswanathan P. 2015. Soluble CD36 in plasma and urine: a plausible prognostic marker for diabetic nephropathy. *J Diabetes Complications* **29:** 400–406.
Simpson EH. 1951. The interpretation of interaction in contingency tables. *J R Stat Soc Series B (Methodol)* **13:** 238–241.
Sugden MC, Holness MJ. 2011. The pyruvate carboxylase-pyruvate dehydrogenase axis in islet pyruvate metabolism: going round in circles? *Islets* **3:** 302–319.
Talchai C, Xuan S, Lin HV, Sussel L, Accili D. 2012. Pancreatic β-cell dedifferentiation as mechanism of diabetic β-cell failure. *Cell* **150:** 1223–1234.
Taneera J, Lang S, Sharma A, Fadista J, Zhou Y, Ahlqvist E, Jonsson A, Lyssenko V, Vikman P, Hansson O, et al. 2012. A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets. *Cell Metab* **16:** 122–134.
Trapnell C. 2015. Defining cell types and states with single-cell genomics. *Genome Res* **25:** 1491–1498.
van de Bunt M, Manning Fox JE, Dai X, Barrett A, Grey C, Li L, Bennett AJ, Johnson PR, Rajotte RV, Gaulton KJ, et al. 2015. Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. *PLoS Genet* **11:** e1005694.
van der Maaten L. 2014. Accelerating T-SNE using tree-based algorithms. *J Mach Learn Res* **15:** 3221–3245.
Vikman J, Ma X, Hockerman GH, Rorsman P, Eliasson L. 2006. Antibody inhibition of synaptosomal protein of 25 kDa (SNAP-25) and syntaxin 1 reduces rapid exocytosis in insulin-secreting cells. *J Mol Endocrinol* **36:** 503–515.
Wallace C, Smyth DJ, Maisuria-Armer M, Walker NM, Todd JA, Clayton DG. 2010. The imprinted *DLK1-MEG3* gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nat Genet* **42:** 68–71.
Wang Z, York NW, Nichols CG, Remedi MS. 2014. Pancreatic β-cell dedifferentiation in diabetes and re-differentiation following insulin therapy. *Cell Metab* **19:** 872–882.
Wang YJ, Schug J, Won KJ, Liu C, Naji A, Avrahami D, Golson ML, Kaestner KH. 2016. Single cell transcriptomics of the human endocrine pancreas. *Diabetes* **65:** 3028–3038.
Westermark P, Andersson A, Westermark GT. 2011. Islet amyloid polypeptide, islet amyloid, and diabetes mellitus. *Physiol Rev* **91:** 795–826.
Xin Y, Kim J, Ni M, Wei Y, Okamoto H, Lee J, Adler C, Cavino K, Murphy AJ, Yancopoulos GD, et al. 2016. Use of the fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. *Proc Natl Acad Sci* **113:** 3293–3298.
Yulyaningsih E, Loh K, Lin S, Lau J, Zhang L, Shi Y, Berning BA, Enriquez R, Driessler F, Macia L, et al. 2014. Pancreatic polypeptide controls energy homeostasis via *Npy6r* signaling in the suprachiasmatic nucleus in mice. *Cell Metab* **19:** 58–72.
Zhang J, McKenna LB, Bogue CW, Kaestner KH. 2014. The diabetes gene *Hhex* maintains δ-cell differentiation and islet function. *Genes Dev* **28:** 829–834.
Zhao HL, Sui Y, Guan J, Lai FM, Gu XM, He L, Zhu X, Rowlands DK, Xu G, Tong PC, et al. 2008. Topographical associations between islet endocrine cells and duct epithelial cells in the adult human pancreas. *Clin Endocrinol* **69:** 400–406.

Received July 12, 2016; accepted in revised form November 16, 2016.

Downloaded from genome.cshlp.org on February 20, 2017 - Published by Cold Spring Harbor Laboratory Press

Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes

Nathan Lawlor, Joshy George, Mohan Bolisetty, et al.

Genome Res. published online November 18, 2016
Access the most recent version at doi:10.1101/gr.212720.116

Supplemental Material http://genome.cshlp.org/content/suppl/2017/01/17/gr.212720.116.DC1

P<P Published online November 18, 2016 in advance of the print journal.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at http://creativecommons.org/licenses/by/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

© 2017 Lawlor et al.; Published by Cold Spring Harbor Laboratory Press

Nathan Lawlor
20th Anniversary Servier-IGIS Symposium Poster Abstract

Identification of islet cell type-specific alterations in type 2 diabetes using single cell genomics

Pancreatic islets consist of at least 5 distinct endocrine cell types, including insulin secreting beta cells, which work to carefully regulate and maintain homeostatic blood glucose levels. Alterations in islet cell type proportion and/or function have been implicated in type 2 diabetes (T2D) pathophysiology, but the precise pathogenic changes elicited in each cell type remain poorly defined. Using a rapid capture and quantitative droplet-based single cell transcriptome platform (10X Genomics), we quantified and compared transcriptome profiles of 29,101 islet endocrine cells from 10 donors (5 T2D and 5 non-diabetic (ND); median 2,829 genes detected per cell; 181 genes exhibited islet cell-type-specific expression). Single cell transcriptomes from T2D and ND islets clustered into distinct endocrine cell type populations, with alpha and beta cells further segregating into two apparent sub-populations. Expression of genes related to endoplasmic reticulum stress response was higher in both minor alpha and beta cell populations. However, proportions of these populations did not significantly differ between T2D and ND donors. We identified 58 genes differentially expressed (FDR < 5%) in T2D cell types (37 in one specific cell type, 21 in multiple cell types). Interestingly, this included a set of differentially expressed genes (*PPY, S100A6, INS,* and *NPY*) in T2D beta cells recently reported for a genetic mouse model of chronic beta cell depolarization, suggesting T2D islets contain beta cells with comparably impaired identity and chronic depolarization. Together, these results provide growing insights into the precise cellular and molecular differences in T2D islets. Ongoing efforts to expand the cohort and complete functional (epi)genomic analyses of the differentially expressed genes in human islets and their constituent cell types will allow us to better decipher the (epi)genetics of islet dysfunction in T2D and to distinguish the causal vs. consequential nature of these observed differences.