Single-cell Transcriptomics of the Pancreas

Experimental & Analytical Tools to Study Organ Development and Composition

Mauro J Muraro

The work described in this thesis was performed at the Hubrecht Institute for Developmental Biology and Stem Cell Research (the Royal Netherlands Academy of Arts and Sciences, KNAW) within the framework of the research school Cancer Stem cells & Developmental biology (CS&D), which is part of the Utrecht Graduate School of Life Sciences (Utrecht University).

Cover: 'Discodip t-SNE map', by Buro Brouns graphic design. www.burobrouns.nl

Layout by Mauro J Muraro and Buro Brouns
Printed by Ridderprint. www.ridderprint.nl

ISBN: 978-94-6299-846-9

Single-cell Transcriptomics of the Pancreas

Experimental and Analytical Tools to Study Organ Development and Composition

Single-cell Transcriptomics van de Alvleesklier
Experimentele en Analystische Methodes om Orgaan Ontwikkeling
en Compositie te Bestuderen
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op

donderdag 18 januari 2018
des middags te 4.15 uur

door

Mauro Javier Gurruchaga Muraro

geboren op 15 juli 1985 te Porto Alegre, Brazilië

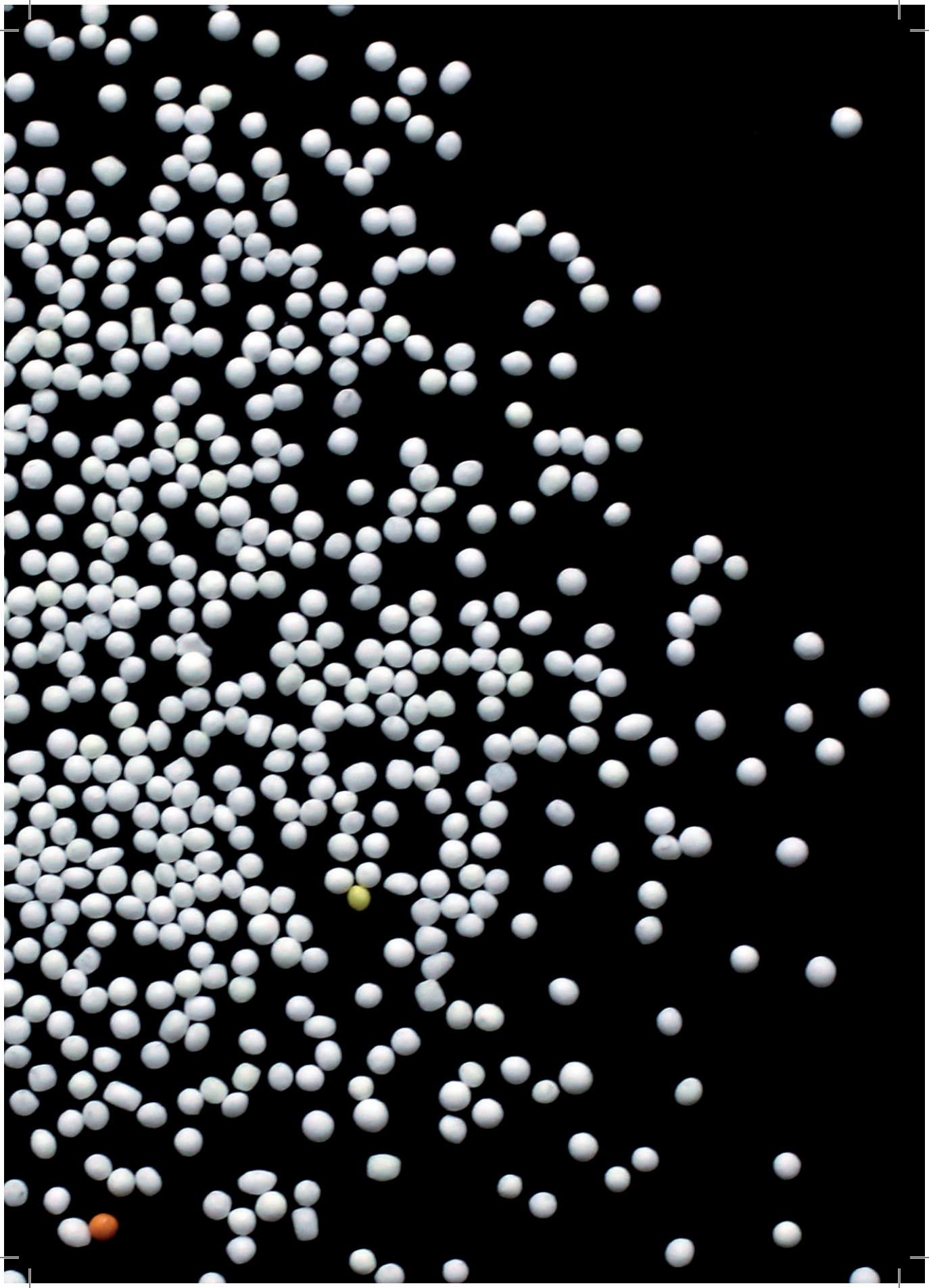
Promoter: Prof. dr. ir. A. van Oudenaarden

Table of contents

Outline and scope of the thesis

Broadly speaking, this thesis combines two subjects: Single-cell transcriptomics and pancreas biology. Single-cell mRNA transcriptomics is a relatively new field that has gone through a fast-paced development since it was first established in 2009. Pancreas homeostasis and development are two fields that have been studied for a much longer time, but that still contain a lot of open questions. In this outline and the following introduction, I will first cover the content of the 6 chapters in this thesis, after which I will place the research questions addressed in the context of both single-cell transcriptomics and pancreas biology.

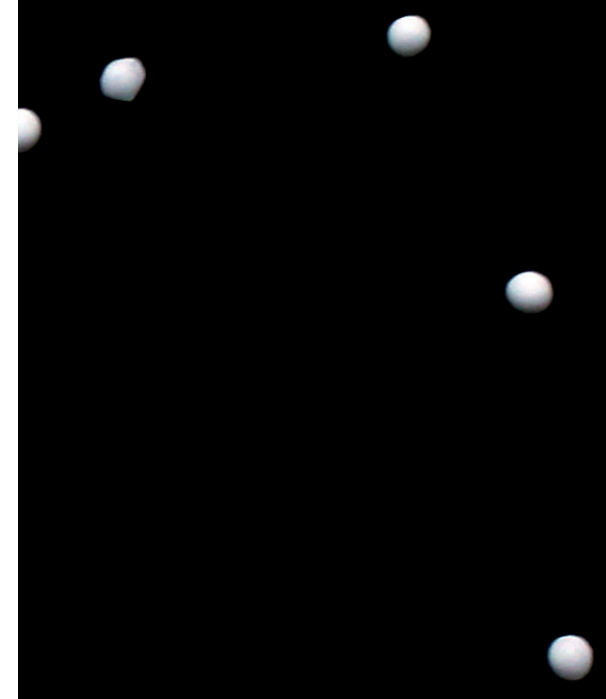
In **Chapter 1** I first describe the recent experimental and analytical progress that led to the single-cell sequencing technology used in this thesis. Next, some of the more recent work on pancreatic development and homeostasis is described, as well as the open questions in the field.
Chapter 2 describes StemID, an algorithm that predicts stem cell identity from single-cell data. This is achieved by first clustering cells based on similarity and then infering lineage relasionships between these cell clusters. This is done by projecting cells onto “highways” between the clusters and combining this with cluster entropy to form a StemID score.
In **Chapter 3**, we sequenced thousands of single cells from the adult human pancreas. Using StemID, we found clear clusters of cells corresponding to the major pancreatic cell types. We then found and validated novel alpha- and beta cell specific genes. One of these was a cell surface marker that we used to purify both cell types from a mixture of pancreatic cells. We also found novel subpopulations of beta and acinar cells.
In **Chapter 4** we applied StemID to analyze data from thousands of single cells from the developing mouse pancreas. By sequencing cells from 5 different time points spanning the second wave of pancreas development, we find and describe all mouse embryonic pancreas cell types as well as the genes important for development of alpha and beta cells from pancreatic endocrine progenitor cells.
Chapter 5 describes GateID, an algorithm that combines single-cell transcriptomics data with FACS information on each sequenced cell to predict novel FACS gates that can be used to purify cell types of choice, without the use of fluorescent reporter genes or antibodies.
Finally, **Chapter 6** contains a summary and discussion of the work presented in this thesis.



1

An Introduction to Single-Cell Sequencing and Pancreas Biology

Mauro J. Muraro



Why study single cells?

The cell is the basic structural unit of life, and each organism found on earth consists of at least one to trillions of cells. An average human, for example, is estimated to have roughly 30 trillion cells and at least 200 different cell types (Bianconi et al., 2014; James W. Valentine Allen G. Collins, 1994). In multicellular organisms, cells form tissues and organs that together perform the essential bodily functions. Ever since cell theory was first postulated by Schleiden, Schwann and Virchow (Turner, 1890) lots of effort has been put into characterizing cell types across organisms and tissues, as misbehavior of even a single cell can lead to a deadly disease such as cancer (Greaves & Maley, 2012). In other cases, like Parkinson's Disease or Diabetes, the death or malfunction of one cell type (dopaminergic neurons and beta cells respectively) can lead to a life-altering disease (Marsden, 1990; Vetere, Choudhary, Burns, & Wagner, 2014). Therefore, to understand organ function and disease, it is important to have a clear idea of the cell types of which it consists. While histology has proven to be an effective way of identifying cell types based on morphology or by the presence a small number of marker genes, it does not inform us on what is happening inside this cell in terms of its full gene expression repertoire. As the genome of all cells in an organism is virtually the same, it is the transcriptome -which reflects which subset of genes is expressed- that is most informative for probing cell state.

The most popular current method of doing this is Next Generation Sequencing, which allows us to measure the complete transcriptome of a sample in question (Shendure & Ji, 2008). To date, most whole-transcriptome studies on disease and tissue have been done on bulk material (thousands to millions of pooled cells) from a given organ or cell type. This obscures not only the contribution of each individual cell type to organ function but also makes it impossible to study heterogeneity within one cell type. This is important to keep in mind, as even cells from the same cell type can differ sufficiently in gene expression profile that it leads to phenotypic changes between them (Eldar & Elowitz, 2010; Munsky, Neuert, & van Oudenaarden, 2012). This is already apparent during early development, when a blastocyst is made up of only a few cells that exhibit stochastic heterogeneity that is responsible for developmental progression (Ohnishi et al., 2013). Even in adult tissues, heterogeneity within the same cell type is considered to influence organ function. For example, beta cells of the adult pancreas show functional heterogeneity in terms of insulin secretion, expression levels and calcium response (Bonner-Weir & Aguayo-Mazzucato, 2016; Gutierrez, Gromada, & Sussel, 2017a). To detect these differences btween cells of the same type, single-cell information is essential. Another important application of single-cell sequencing is the detection of rare cell types such as circulating tumor cells (Ramsköld et al., 2012). In the case of the pancreas, this is an important application as the pancreas is a quiescent organ with very little cell division. It is unknown weather new cells come from slow cell division of already existing cells, from transdifferentiation between cell types or from adult tissue stem cells that become activated when needed. In short, to fully grasp organ function and development, it is important to analyze the complete transcriptional state of its cells at single-cell resolution. With the advent of single-cell

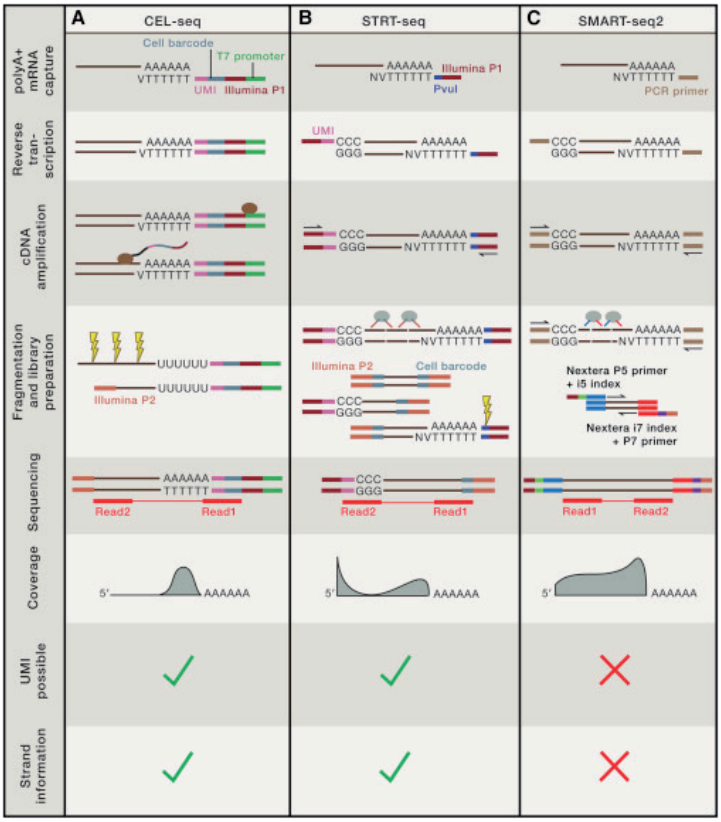


Figure 1: The three main single-cell sequencing techniques
Adapted from Grün and van Oudenaarden (2015)

transcriptomics, this is finally feasible.

Single-cell transcriptomics: experimental methods

Single-cell measurements of gene expression are considerably older than single-cell mRNA sequencing. The first successful attempts were done by Southern blot using single neurons and hematopoietic cells (Brady, Barbara, & Iscove, 1990; Eberwine et al., 1992). These studies were followed by single-cell PCR techniques that allowed the study of pre-selected sets of genes in a relatively small number of cells (Guo et al., 2010). These techniques lacked quantitative power, since levels of expression are measured as an expression ratio between known genes. Technologies based on fluorescence such as FACS or single-molecule RNA FISH then allowed mRNA measurements over many cells in a quantitative manner, but were also limited by the number of pre-selected genes that could be tested at once (Buganim et al., 2012; Klemm et al., 2014; Raj, van den Bogaard, Rifkin, van Oudenaarden, & Tyagi, 2008). In the meantime, Next Generation sequencing technologies were developed that made whole-transcriptome measurements possible, but only on samples consisting of many thousands of pooled cells. When the first single-cell transcriptome sequencing technique was established in 2009 by the Surani lab, these two fields (few genes & single-cells versus many genes & pooled cells) finally

merged to allow whole-transcriptome measurements over multiple single cells (Tang et al., 2009). Since then, other types of single-cell sequencing have been developed, such as single-cell sequencing of DNA (Gawad, Koh, & Quake, 2016), the measurement of epigenetic marks like (hydroxy) methylcytosine (Mooijman, Dey, Boisset, Crosetto, & van Oudenaarden, 2016; Smallwood et al., 2014), chromatin accessibility (Buenrostro et al., 2015; Jin et al., 2015) and small RNAs (Faridani et al., 2016). As the subject of this thesis is single-cell mRNA sequencing, we will now focus on the development of this technology in particular.

The first single-cell transcriptomics protocol by the Surani lab employed manual cell picking using a glass capillary needle to successfully amplify mRNA from single cells in a mouse blastomere. The use of manual cell picking, as well as many experimental steps that had to be performed on each single cell separately made these experiments cumbersome and expensive to perform. The next generation of sequencing methods started with the development of STRT by the Linnarson lab, where single-cell mRNA was tagged with a primer containing a poly-A tail and a cell-specific barcode that enabled multiplexing of 96 cells into one sequencing library (Islam et al., 2012). Not long after this, two more landmark methods were published, namely Smart-seq and CEL-Seq (Hashimshony, Wagner, Sher, & Yanai, 2012; Ramsköld et al., 2012). While Smart is similar to STRT since it is uses a Template Switching Oligo (TSO) to elongate barcode the cDNA on the 5'end, CEL-Seq primes the opposite, 3' part of the mRNA molecule and then produces a double stranded cDNA molecule that is subsequently used for In Vitro Transcription (IVT) (see Figure 1). Many more techniques followed these initial three studies, most of which fall into either the TSO or IVT bracket and are therefore derivatives of either STRT/SMART or CEL-Seq. The addition of Unique Molecular Identifiers (UMIs) to cellular barcodes meant that individual mRNA molecules present in each cell could be estimated quantitatively in both STRT and CEL-Seq based techniques (Kivioja et al., 2011). For a more information on the rapid development of single-cell sequencing technologies, see (Svensson, Roser, & Teichmann, 2017). These initial, technical proof-of-principle studies as well as contemporary papers were mostly done on cultured cells (Deng, Ramsköld, Reinius, & Sandberg, 2014; Tang et al., 2010). What followed was a second wave of papers analyzing primary tissues and showing that single-cell transcriptomics could be used to characterize cell types in the mouse spleen (Jaitin et al., 2014) , brain (Zeisel et al., 2015), lung (Treutlein et al., 2014), retina (Macosko et al., 2015), small intestine (Grün et al., 2015) and human pancreas (Li et al., 2016). Most of these studies, however, were done on either manually processed and/or limited to low numbers of cells. To fully capture all the relevant cell types present in a tissue, higher numbers of cells were necessary (Grün & van Oudenaarden, 2015; Shapiro, Biezuner, & Linnarsson, 2013). This led to the development of several different automated single-cell transcriptomics platforms such as Fluidigms C1 (Islam et al., 2014), that uses microfluidic chips that could process 96 cells (now also improved to 800-cell chips). The C1 was used to sequence mouse and human pancreatic cells, which yielded useful data on pancreatic cells, but also revealed some technical issues with the C1 platform, such as a low success rate of cells (only approximately 50% survives the procedure) (Xin, Kim, Okamoto, et al., 2016)(Xin, Okamoto, et al., 2016) and a high doublet rate

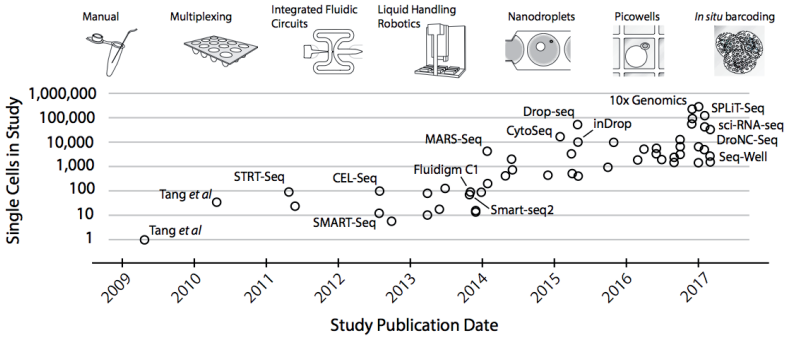


Figure 2: Scaling single-cell transcriptomics
Adapted from Svensson et al. (2017)

(27% of the cells showed expression patterns belonging to more than one cell type) (Xin, Kim, Ni, et al., 2016). Two methods have been described that combine FACS sorting with the CEL-Seq (2) protocol to allow robotic processing of thousands of cells: MARS-seq (Jaitin et al., 2014) and SORT-Seq (described in chapter 3 of this thesis) that allow a single researcher to process thousands of cells in one day while keeping the doublet rate low. The latest, most high-throughput methods are based on either Nano liter droplet emulsions and can process up to tens of thousands of cells routinely (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017) or employ the combinatorial labeling of fixed cells in situ with multiple rounds of barcoding, allowing manual processing of up to hundreds of thousands of cells (Cao et al., 2017; Rosenberg et al., 2017). Figure 2 provides a clear overview of the growing throughput with which single-cell transcriptomics techniques can process cells.

Single-cell transcriptomics: analysis

As single-cell data is becoming increasingly available and complex (more cells & better transcriptome coverage), dedicated algorithms are necessary to filter, normalize and cluster data. While some of the existing algorithms for normalization and gene expression analysis in bulk sequencing experiments can be used (Anders & Huber, 2010), new methods were needed to properly deal with the particular type of technical variability that comes with single-cell sequencing experiments (Brennecke et al., 2013; Grün, Kester, & van Oudenaarden, 2014). Quickly after the first studies on multiple cell types from the same tissue were published, different visualization methods were compared to find the method that best represents differences between cell types in single-cell datasets, resulting in t-distributed stochastic neighbor embedding (t-SNE) as the method of choice (Grün & van Oudenaarden, 2015; Maaten & Hinton, 2008). Since then, a variety of computational methods have been devised for clustering and identification of cell types present in single-cell data, such as RaceID, BackSpin, Phenograph and Scenic (Aibar et al., 2017; Grün et al., 2015; Shekhar et al., 2016; Zeisel et al., 2015). Other computational challenges of single-cell sequencing data include the removal of batch effects(Butler & Satija, 2017; Haghverdi, Lun, Morgan, & Marioni, 2017), in silico lineage reconstruction of trajectories between cell types (Grün et al., 2016; Setty et al., 2016; Trapnell et al., 2014) and limited coverage of lowly expressed

genes, such as transcription factors (Heimberg, Bhatnagar, El-Samad, & Thomson, 2016; van Dijk et al., 2017). At the moment, new computational methods for single-cell transcriptomics are appearing at a fast rate, and there is no consensus on which algorithm works best. For a reviews on the computational challenges concerning single-cell sequencing, see (Grün & van Oudenaarden, 2015; Rostom, Svensson, Teichmann, & Kar, 2017; Stegle, Teichmann, & Marioni, 2015; Wagner, Regev, & Yosef, 2016). In general, a sensible approach is to test several different algorithms and explore which one works best on the data at hand, as each technique and dataset can pose unique challenges.

Pancreas development and homeostasis

The pancreas is an organ that serves two distinct but important bodily functions: food digestion and maintenance of glucose levels in the blood. It lies in the abdomen behind the stomach and in humans is about 15 cm long. Its anatomical structure and development is relatively well conserved across mammals, birds, reptiles and other animals (SLACK 1995). It is an important organ in the context of human medicine, since its (dys) function is pivotal in two diseases: pancreatic cancer, a type of cancer associated with very poor prognosis and survival rate (Bardeesy & DePinho, 2002) and Diabetes Mellitus, a disease which currently affects 9% of the population world wide (WHO 2014). The exocrine compartment of the pancreas performs its digestive function. It consists of acinar cells that are grouped into acini that secrete a collection of enzymes like proteases and amylases (figure 3). These enzymes travel trough a network of ducts that eventually drains into the duodenum (SLACK 1995). The endocrine compartment, formed by the Islets of Langerhans, maintains glucose levels in the blood. Islets consist of 5 different cell types, each of which produces one hormone: the alpha cells (produce glucagon), beta cells (insulin), delta cells (somatostatin), PP cells (pancreatic polypeptide) and epsilon cells (ghrelin). To understand how closely related these distinct cell types are, it is important to know how the pancreas develops. We will next cover pancreatic development from organogenesis to differentiation of the cell types found in the mature pancreas.

Pancreatic development

Pancreatic development in the mouse (the subject of chapter 4) starts at embryonic day 9.5 (E9.5) with a thickening of the dorsal foregut endoderm that then protrudes into the surrounding mesenchyme. In rodents, pancreas formation is separated into several waves of development. The first wave, called the first transition, takes place between E9.5 and E12.5 and comprises morphogenic changes such as the formation of dorsal and ventral bud, the formation of a stratified epithelium and the emergence of multiple micro lumens that start growing into a pancreatic tree (figure 4). For a detailed review on pancreatic development, see (Pan & Wright, 2011). The secondary transition starts after E12.5 and encompasses the growth of the pancreatic tree by the division of multipotent pancreatic cells (MPC) that are found at the tip domain of the recently formed epithelium and that will eventually differentiate into acinar cells. These MPC's form new branches in the pancreatic tree and leave a layer of bipotent endocrine-ductal progenitors in their wake that will become

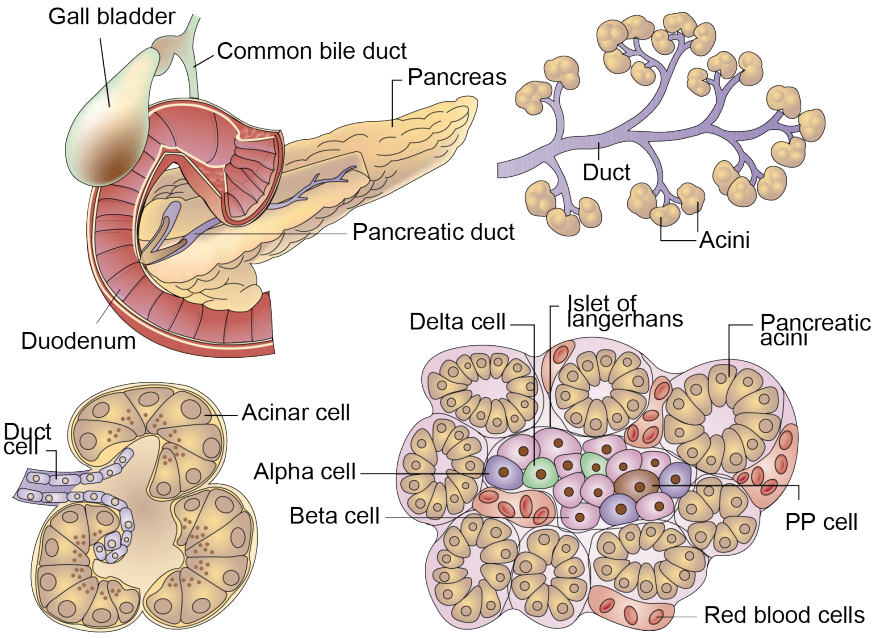


Figure 3: Pancreas anatomy
adapted from Bardeesy and DePinho (2002)

either a ductal cell or an endocrine progenitor. Endocrine progenitor cells activate Neurogenin3 and delaminate from the ductal epithelium to become differentiated endocrine cells, which will later group together to form the functional unit of the endocrine pancreas: the Islet of Langerhans. This happens through a still poorly understood process that includes an epithelial-to-mesenchymal transition (EMT) (Gouzi, Kim, Katsumoto, Johansson, & Grapin-Botton, 2011; Rukstalis & Habener, 2007). It is still unclear weather these bipotent progenitors specify an endocrine progenitor cell by symmetrical (either two duct or two endocrine progenitor cells are formed) or asymmetrical division (one duct and one endocrine progenitor), where one of the two daughter cells activates Neurog3 and delaminates from the epithelium. After the secondary transition, the acinar tissue grows and the organ increases in size. Much is still unknown about how cell type differentiation occurs in the developing pancreas. In chapter 4, we try to investigate this process by analyzing the transcriptome of thousands of single cells coming from developing mouse pancreas between E12.5 and E18.5 in order to characterize the gene expression signatures responsible for endocrine cell type differentiation.

Rationale behind this thesis

There are several open questions when it comes to pancreatic development and function that can only be truly answered by looking at the pancreas at single-cell resolution: First of all, there is a lack of markers with which to obtain pure populations of cells of each of the pancreatic cell types. Attempts have been made to purify

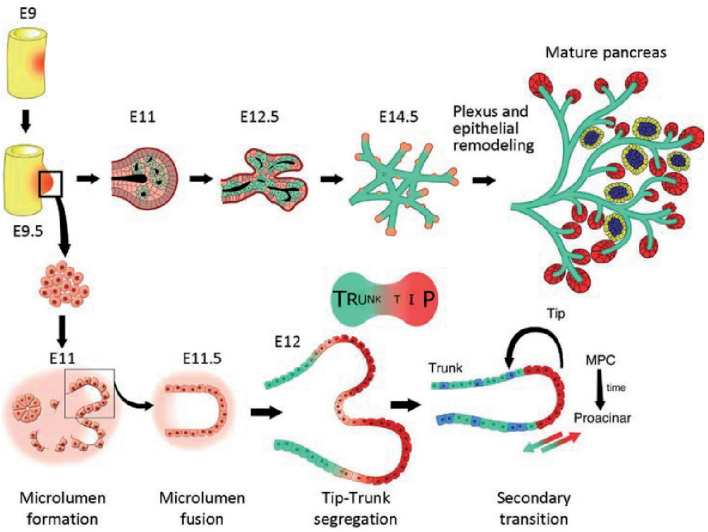


Figure 4. Pancreas development
Adapted from Pan and Wright (2011)

some of them (especially beta cells) by laser capture micro dissection (Marselli et al., 2010), their zinc content (Nica, Ongen, & Irminger, 2013) and staining with cell surface markers (Dorrell et al., 2011). This however, yielded mixed results, such as the presence of delta cell marker genes amongst the genes differentially expressed between alpha and beta cells (Table 1 in Dorell et al., 2011). The inability to obtain pure populations of one pancreatic cell type has meant that especially the rare cell types in the pancreas, such as the delta, PP and epsilon cells have been understudied, and their role in glucose homeostasis is still not clear. Using single-cell sequencing, it is possible to purify cells in silico, yielding transcriptome wide information on all the individual cell types of the pancreas without any bias stemming from isolation procedure.

Secondly, there is evidence for the existence of heterogeneity within cell types in the pancreas. Beta cells, for example, have been reported to exhibit various levels of heterogeneity (Avrahami, Klochendler, Dor, & Glaser, 2017; Bonner-Weir & Aguayo-Mazzucato, 2016; Gutierrez, Gromada, & Sussel, 2017b; Roscioni, Migliorini, Gegg, & Lickert, 2016). Using bulk transcriptome studies will average out any heterogeneity within cell types, and by looking at each cell type in single-cell resolution will help to find any subpopulations within the different pancreatic cell types (Shapiro et al., 2013).

Another important open question in pancreatic biology is the turnover of adult cells. While the origin of endocrine cells in the islets of Langerhans is known in developmental biology (Neurogenin 3 positive cells in the ductal lining) (Pan & Wright, 2011) in adults this is still unknown how the endocrine cell population is maintained. This is a particularly important question in the context of beta cells and diabetes, as this is the cell type that produces insulin and which is involved in both types of diabetes. Type 1 Diabetes is an autoimmune disease where the beta cells are destroyed by the bodies own immune system, while Type 2 diabetes is caused when the body becomes insufficiently responsive to insulin, leading to stress and

eventual death in the beta cell population (Salsali & Nathan, n.d.; Vetere et al., 2014). Evidence exists for three sources of new beta cells: slow beta cell turnover, trans differentiation from other endocrine cell types and for progenitor cells (Afelik & Rovira, 2017). The debate is still ongoing, and understanding how new beta cells arise would be very valuable for treatment of diabetes. Analyzing large numbers of cells together masks the contribution of rare cell types such as stem cells, so single-cell information could shine some light on the question of the source of new endocrine cells in the adult pancreas.
Similarly, the changes between developing endocrine cells during pancreas organogenesis are hard to describe accurately on a whole-tissue level, since every cell type covers a different developmental trajectory from its birth in the form of an MPC or an endocrine progenitor cell. While some of the regulators of endocrine cell fates are known (figure 4), this list is limited to a handful of genes. It will be informative to obtain cell type-specific information at the single-cell level for all pancreatic cell types.

Context of this thesis
At the start of the projects described here (late 2013), single-cell sequencing was still in its infancy, where most publications reported the transcriptomes of a few dozen to hundred cultured cells. Before any thorough analysis of pancreatic cells could be done, experimental and computational efforts were needed to:

1. Develop algorithms that can filter, normalize and cluster single-cell transcriptomics data so that each sequenced cell can be assigned to a cell type and to infer lineages between the different cell types in a dataset.
2. Automate single-cell mRNA sequencing in order to study many thousands of cells and develop computational methods that could store and use the resulting single-cell transcriptome and FACS data.

Chapters 2,3 and 5 describe efforts on the technical and analytical side of single-cell mRNA sequencing. Chapters 3 and 4 use these techniques to describe the cell types of the adult human pancreas and the gene expression changes during pancreas organogenesis in the mouse embryo.

References

Afelik, S., & Rovira, M. (2017). Pancreatic β-cell regeneration: Facultative or dedicated progenitors? Molecular and Cellular Endocrinology, 445, 85–94. http://doi.org/10.1016/j.mce.2016.11.008

Aibar, S., Bravo González-Blas, C., Moerman, T., Wouters, J., Huynh-Thu, V. A., Imrichová, H., … Aerts, S. (2017). SCENIC: Single-Cell Regulatory Network Inference And Clustering. bioRxiv, 1–41. http://doi.org/10.1101/144501

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology, 11(10), R106. http://doi.org/10.1186/gb-2010-11-10-r106

Avrahami, D., Klochendler, A., Dor, Y., & Glaser, B. (2017). Beta cell heterogeneity: an evolving concept. Diabetologia, 60(8), 1363–1369. http://doi.org/10.1007/s00125-017-4326-z

Bardeesy, N., & DePinho, R. a. (2002). Pancreatic cancer biology and genetics. Nature Reviews. Cancer, 2(12), 897–909. http://doi.org/10.1038/nrc949

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., … Canaider, S. (2014). An estimation of the number of cells in the human body. Annals of Human Biology, 40(6), 463–71. http://doi.org/10.3109/03014460.2013.807878

Bonner-Weir, S., & Aguayo-Mazzucato, C. (2016). Physiology: Pancreatic β-cell heterogeneity revisited. Nature, 535(7612), 365–366. http://doi.org/10.1038/nature18907

Brady, G., Barbara, M., & Iscove, N. N. (1990). Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. Methods Mol. Cell. Biol., 2(1), 17–25.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., … Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods, 10(11), 1093–5. http://doi.org/10.1038/nmeth.2645

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., … Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature, 523(7561), 486–490. http://doi.org/10.1038/nature14590

Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., … Jaenisch, R. (2012). Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. Cell, 150(6), 1209–1222. http://doi.org/10.1016/j.cell.2012.08.023

Butler, A., & Satija, R. (2017). Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. bioRxiv. Retrieved from http://biorxiv.org/content/early/2017/07/18/164889.abstract

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., … Shendure, J. (2017). Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. bioRxiv. Retrieved from http://biorxiv.org/content/early/2017/02/02/104844.abstract

Deng, Q., Ramsköld, D., Reinius, B., & Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science (New York, N.Y.), 343(6167), 193–6. http://doi.org/10.1126/science.1245316

Dorrell, C., Schug, J., Lin, C. F., Canaday, P. S., Fox, a. J., Smirnova, O., … Grompe, M. (2011). Transcriptomes of the major human pancreatic cell types. Diabetologia, 54(11), 2832–2844. http://doi.org/10.1007/s00125-011-2283-5

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., … Coleman, P. (1992). Analysis of gene expression in single live neurons. Proceedings of the National Academy of Sciences of the United States of America, 89(7), 3010–3014. http://doi.org/10.1073/pnas.89.7.3010

Eldar, A., & Elowitz, M. B. (2010). Functional roles for noise in genetic circuits. Nature, 467(7312), 167–173. http://doi.org/10.1038/nature09326

Faridani, O. R., Abdullayev, I., Hagemann-Jensen, M., Schell, J. P., Lanner, F., & Sandberg, R. (2016). Single-cell sequencing of the small-RNA transcriptome. Nature Biotechnology. http://doi.org/10.1038/nbt.3701

Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. Nature Reviews Genetics, 17(3), 175–188. http://doi.org/10.1038/nrg.2015.16

Gouzi, M., Kim, Y. H., Katsumoto, K., Johansson, K., & Grapin-Botton, A. (2011). Neurogenin3 initiates stepwise delamination of differentiating endocrine cells during pancreas development. Developmental Dynamics, 240(3), 589–604. http://doi.org/10.1002/dvdy.22544

Greaves, M., & Maley, C. C. (2012). Clonal evolution in cancer. Nature, 481(7381), 306–313. http://doi.org/10.1038/nature10762

Grün, D., Kester, L., & van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nature Methods, 11(6), 637–40. http://doi.org/10.1038/nmeth.2930

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., … van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. http://doi.org/10.1038/nature14966

Grün, D., Muraro, M. J., Boisset, J.-C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., … van Oudenaarden, A. (2016). De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. Cell Stem Cell, 1–12. http://doi.org/10.1016/j.stem.2016.05.010

Grün, D., & van Oudenaarden, A. (2015). Design and Analysis of Single-Cell Sequencing Experiments. Cell, 163(4), 799–810. http://doi.org/10.1016/j.cell.2015.10.039

Guo, G., Huss, M., Tong, G. Q., Wang, C., Li Sun, L., Clarke, N. D., & Robson, P. (2010). Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. Developmental Cell, 18(4), 675–685. http://doi.org/10.1016/j.devcel.2010.02.012

Gutierrez, G. D., Gromada, J., & Sussel, L. (2017a). Heterogeneity of the pancreatic beta cell. Frontiers in Genetics, 8(MAR), 1–9. http://doi.org/10.3389/fgene.2017.00022

Gutierrez, G. D., Gromada, J., & Sussel, L. (2017b). Heterogeneity of the Pancreatic Beta Cell. Frontiers in Genetics, 8. http://doi.org/10.3389/fgene.2017.00022

Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2017). Correcting batch effects in single-cell RNA sequencing data by matching mutual nearest neighbours. bioRxiv, 1–18. http://doi.org/10.1101/165118

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Reports, 2(3), 666–673. http://doi.org/10.1016/j.celrep.2012.08.003

Heimberg, G., Bhatnagar, R., El-Samad, H., & Thomson, M. (2016). Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. Cell Systems, 2(4), 239–250. http://doi.org/10.1016/j.cels.2016.04.001

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., & Linnarsson, S. (2012). Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. Nature Protocols, 7(5), 813–28. http://doi.org/10.1038/nprot.2012.022

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., … Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods, 11(1), 163–166. http://doi.org/10.1038/nmeth.2772

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., … Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science (New York, N.Y.), 343(6172), 776–9. http://doi.org/10.1126/science.1247651

James W. Valentine Allen G. Collins, C. P. M. (1994). Morphological Complexity Increase in Metazoans. Paleobiology, 20(2), 131–142. http://doi.org/10.2307/2401015

Jin, W., Tang, Q., Wan, M., Cui, K., Zhang, Y., Ren, G., … Zhao, K. (2015). Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. Nature. http://doi.org/10.1038/nature15740

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. Nature Methods, 9(1), 72–74. http://doi.org/10.1038/nmeth.1778

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., … Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell, 161(5), 1187–201. http://doi.org/10.1016/j.cell.2015.04.044

Klemm, S., Semrau, S., Wiebrands, K., Mooijman, D., Faddah, D. A., Jaenisch, R., & van Oudenaarden, A. (2014). Transcriptional profiling of cells sorted by RNA abundance. Nature Methods, 11(5), 549–551. http://doi.org/10.1038/nmeth.2910

Li, J., Klughammer, J., Farlik, M., Penz, T., Spittler, A., Barbieux, C., … Kubicek, S. (2016). Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types, 17(2), 178–187.

Maaten, L. Van Der, & Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605. http://doi.org/10.1007/s10479-011-0841-3

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell, 161(5), 1202–14. http://doi.org/10.1016/j.cell.2015.05.002

Marsden, C. D. (1990). Parkinson's disease. Lancet (London, England), 335(8695), 948–52. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1691427

Marselli, L., Thorne, J., Dahiya, S., Sgroi, D. C., Sharma, A., Bonner-Weir, S., … Weir, G. C. (2010). Gene Expression Profiles of Beta-Cell Enriched Tissue Obtained by Laser Capture Microdissection from Subjects with Type 2 Diabetes. PLoS ONE, 5(7), e11499. http://doi.org/10.1371/journal.pone.0011499

Mooijman, D., Dey, S. S., Boisset, J.-C., Crosetto, N., & van Oudenaarden, A. (2016). Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. Nature Biotechnology, 34(8), 852–856. http://doi.org/10.1038/nbt.3598

Munsky, B., Neuert, G., & van Oudenaarden, A. (2012). Using Gene Expression Noise to Understand Gene Regulation. Science, 336(6078), 183–187. http://doi.org/10.1126/science.1216379

Nica, A. C., Ongen, H., & Irminger, J. (2013). Cell-type , allelic and genetic signatures in the human pancreatic beta cell transcriptome Cell-type , allelic and genetic signatures in the human pancreatic beta cell transcriptome, 1554–1562. http://doi.org/10.1101/gr.150706.112

Ohnishi, Y., Huber, W., Tsumura, A., Kang, M., Xenopoulos, P., Kurimoto, K., … Hiiragi, T. (2013). Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. Nature Cell Biology, 16(1), 27–37. http://doi.org/10.1038/ncb2881

Pan, F. C., & Wright, C. (2011). Pancreas organogenesis: From bud to plexus to gland. Developmental Dynamics, 240(3), 530–565. http://doi.org/10.1002/dvdy.22584

Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A., & Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nature Methods, 5(10), 877–879. http://doi.org/10.1038/nmeth.1253

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., … Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature Biotechnology, 30(8), 777–782. http://doi.org/10.1038/nbt.2282

Roscioni, S. S., Migliorini, A., Gegg, M., & Lickert, H. (2016). Impact of islet architecture on β-cell heterogeneity, plasticity and function. Nature Reviews. Endocrinology, 12(12), 695–709. http://doi.org/10.1038/nrendo.2016.147

Rosenberg, A. B., Roco, C., Muscat, R. A., Kuchina, A., Mukherjee, S., Chen, W., … Seelig, G. (2017). Scaling single cell transcriptomics through split pool barcoding. bioRxiv, 105163. http://doi.org/10.1101/105163

Rostom, R., Svensson, V., Teichmann, S. A., & Kar, G. (2017). Computational approaches for interpreting scRNA-seq data. FEBS Letters, 1–13. http://doi.org/10.1002/1873-3468.12684

Rukstalis, J. M., & Habener, J. F. (2007). Snail2, a mediator of epithelial-mesenchymal transitions, expressed in progenitor cells of the developing endocrine pancreas. Gene Expression Patterns, 7(4), 471–479. http://doi.org/10.1016/j.modgep.2006.11.001

Salsali, A., & Nathan, M. (n.d.). A review of types 1 and 2 diabetes mellitus and their treatment with insulin. American Journal of Therapeutics, 13(4), 349–61. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16858171

Setty, M., Tadmor, M. D., Reich-zeliger, S., Angel, O., Salame, T. M., Kathail, P., … Pe, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. Nature Biotechnology, (May), 1–14. http://doi.org/10.1038/nbt.3569

Shapiro, E., Biezuner, T., & Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews. Genetics, 14(9), 618–630. http://doi.org/10.1038/nrg3542

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., … Sanes, J. R. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. Cell, 166(5), 1308–1323.e30. http://doi.org/10.1016/j.cell.2016.07.054

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechnology, 26(10), 1135–1145. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18846087

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., … Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nature Methods, 11(8), 817–820. http://doi.org/10.1038/nmeth.3035

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. Nature Reviews Genetics, 16(3), 133–145. http://doi.org/10.1038/nrg3833

Svensson, V., Roser, V.-T., & Teichmann, S. A. (2017). Moore ' s Law in Single Cell Transcriptomics. arXiv:1704.01379v1 [Q-bio.GN].

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., … Surani, M. A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell, 6(5), 468–78. http://doi.org/10.1016/j.stem.2010.03.015

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., … Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods, 6(5), 377–382. http://doi.org/10.1038/nmeth.1315

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., … Rinn, J. L. (2014). letters The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology, 32(4), 381–386. http://doi.org/10.1038/nbt.2859

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., … Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature, 509(7500), 371–5. http://doi.org/10.1038/nature13173

Turner, W. (1890). The Cell Theory, Past and Present. The Anatomy of Plants, 24, 253–287.

van Dijk, D., Nainys, J., Sharma, R., Kathail, P., Carr, A. J., Moon, K. R., … Pe'er, D. (2017). MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. bioRxiv. Retrieved from http://biorxiv.org/content/early/2017/02/25/111591.abstract

Vetere, A., Choudhary, A., Burns, S. M., & Wagner, B. K. (2014). Targeting the pancreatic β-cell to treat diabetes. Nature Reviews. Drug Discovery, 13(4), 278–89. http://doi.org/10.1038/nrd4231

Wagner, A., Regev, A., & Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nature Biotechnology, 34(11), 1145–1160. http://doi.org/10.1038/nbt.3711

Xin, Y., Kim, J., Ni, M., Wei, Y., Okamoto, H., Lee, J., … Gromada, J. (2016). Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. Proceedings of the National Academy of Sciences of the United States of America, 113(12), 3293–8. http://doi.org/10.1073/pnas.1602306113

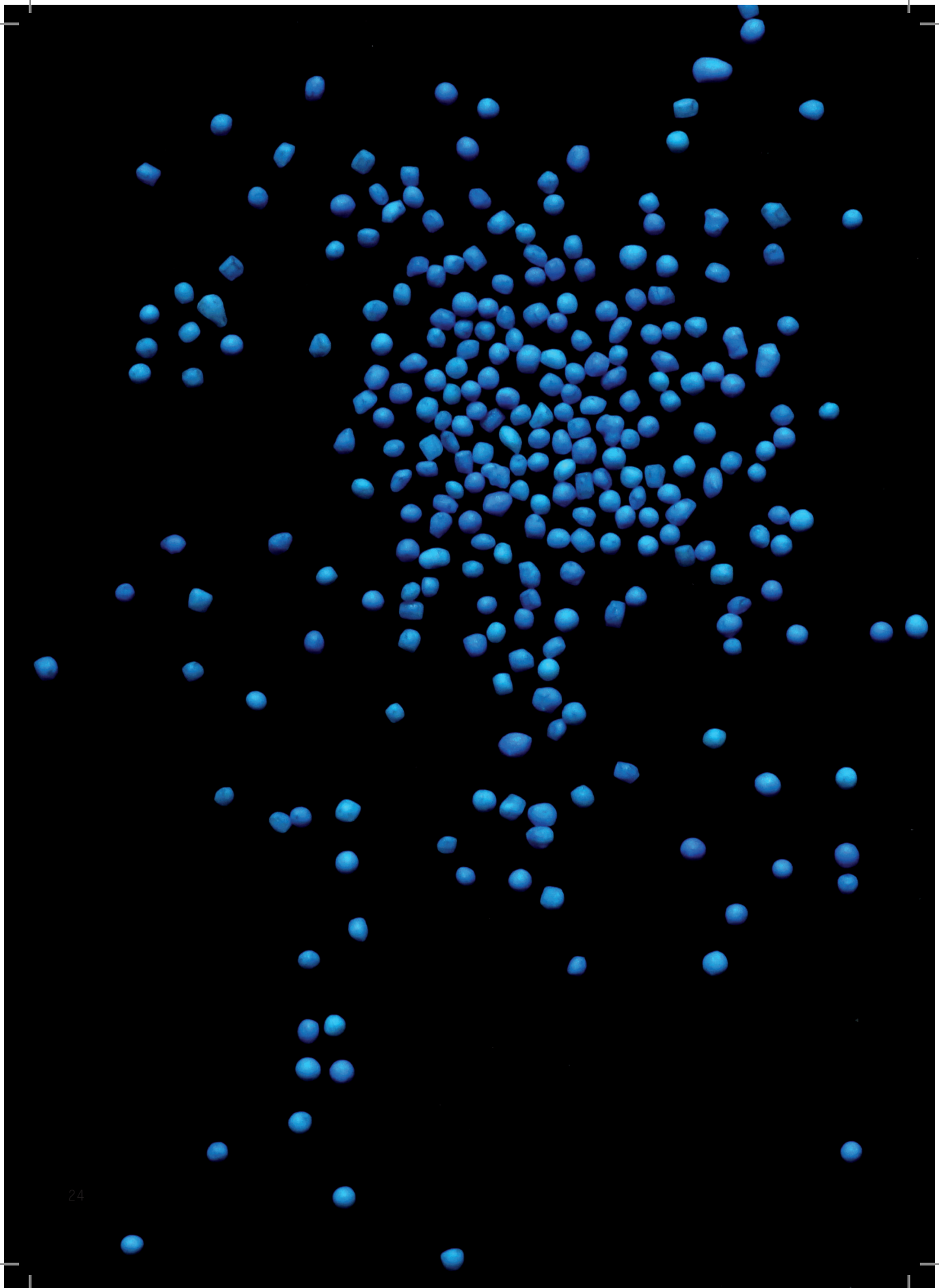
Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., … Gromada, J. (2016). RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. Cell Metabolism, 24(4), 608–615. http://doi.org/10.1016/j.cmet.2016.08.018

Xin, Y., Okamoto, H., Kim, J., Ni, M., Adler, C., Cavino, K., … Lin, C. (2016). Single-Cell RNAseq Reveals That Pancreatic b-Cells From Very Old Male Mice Have a Young Gene Signature, (August), 1–8. http://doi.org/10.1210/en.2016-1235

Zeisel, A., Munoz-Manchado, A. B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., … Linnarsson, S. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science, 347(6226), 1138–1142. http://doi.org/10.1126/science.aaa1934

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. Nature Communications, 8, 14049. http://doi.org/10.1038/ncomms14049

1



2

De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data

Dominic Grün, Mauro J. Muraro, Jean-Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike van den Born, Johan van Es, Erik Jansen, Hans Clevers, Eelco J. P. de Koning, and Alexander van Oudenaarden

Adapted from Cell Stem Cell. 2016. 19(2):266-277

ABSTRACT
Adult mitotic tissues like the intestine, skin, and blood undergo constant turnover throughout the life of an organism. Knowing the identity of the stem cell is crucial to understand tissue homeostasis and its aberrations upon disease. Here we present a computational method for the derivation of a lineage tree from single cell transcriptome data. By exploiting the tree topology and the transcriptome composition we establish StemID, an algorithm for identifying stem cells among all detectable cell types within a population. We demonstrate that StemID recovers two known adult stem cell populations, the Lgr5+ cells in the small intestine and the hematopoietic stem cells in the bone marrow. We apply StemID to predict candidate multipotent cell populations in the human pancreas, a tissue with largely uncharacterized turnover dynamics. We hope that StemID will accelerate the search for novel stem cells by providing concrete markers for biological follow-up and validation.

INTRODUCTION
The identification of a stem cell in a tissue is a major challenge of pivotal importance. Being able to detect the stem cell population allows for powerful approaches to study cell differentiation dynamics by, for example, lineage tracing (Barker et al., 2007; Busch et al., 2015). Additionally, it provides a first step towards ex vivo propagation of primary stem cells in organoid cultures (Lancaster et al., 2013; Sato et al., 2009) important for applications in regenerative medicine. Moreover, stem cell populations relevant for disease progression such as cancer stem cells offer promising targets for therapeutic intervention. Stem cells are typically rare, which makes their discovery by traditional population-based assays very difficult. For example, it took decades of dedicated research to define the population of hematopoietic stem cells (HSCs) (Eaves, 2015) yet it remains an open question how much heterogeneity exists within this subpopulation of bone marrow cells (Wilson et al., 2015). Similarly, the discovery of intestinal stem cells (van der Flier and Clevers, 2009) took years of work and also heterogeneity within this compartment remains under debate (Buczacki et al., 2013).
The recent availability of single cell mRNA sequencing methods allows profiling of healthy and diseased tissues with single cell resolution (Grün et al., 2015; Jaitin et al., 2014; Macosko et al., 2015; Patel et al., 2014; Paul et al., 2015; Treutlein et al., 2014; Zeisel et al., 2015). The transcriptome of a cell can be interpreted as a fingerprint revealing its identity. However, biological gene expression noise (Eldar and Elowitz, 2010; Raj and van Oudenaarden, 2008) and technical noise due to amplification of minute amounts of mRNA from a single cell (Brennecke et al., 2013; Grün et al., 2014) affect the read-out and make it a challenge to discriminate cell types based on their transcriptome. By sequencing large numbers of randomly sampled single cells from a tissue it is now possible to compile a near complete inventory of cell types.
These inventories can now be screened for cell types of particular interest such as stem cells. An obvious strategy for the identification of the stem cell is the derivation of a lineage tree from single cell sequencing data. However, transcriptomes of randomly sampled cells only represent a snapshot of the system and temporal

differentiation dynamics cannot be directly derived. However, if the system of interest comprises all differentiation stages, such as the intestinal epithelium or the bone marrow, attempts can be made to infer a lineage tree by assembling single cell transcriptomes in a pseudo-temporal order. Existing approaches assume a continuous temporal change of transcript levels to assemble differentiation trajectories (Bendall et al., 2014; Haghverdi et al., 2015; Trapnell et al., 2014) but resolving the correct tree topology remains a challenge.

Here, we present a method to identify rare and abundant cell types of a system and use these cell type classifications to guide the inference of a lineage tree. We investigate general properties characterizing the position of a cell type within the lineage tree and identify the number of branches and the transcriptome uniformity of a cell type as features correlating with the degree of pluripotency. We show that our approach successfully recovers the identity of the stem cell in the intestine and in the bone marrow, two systems with a well-described stem cell population. We then use our method to predict multipotent cell populations in the adult human pancreas.

RESULTS

Robust identification of mouse intestinal cell types by RaceID2

To develop a robust approach for the inference of differentiation trajectories we used a previously published dataset from a lineage tracing experiment comprising progeny of Lgr5-positive mouse intestinal stem cells (Grün et al., 2015). This system is ideal for testing the inference of differentiation dynamics, since the lineage tree is already well characterized (Figure 1A). The continuously self-renewing intestinal epithelium is arranged in crypts and villi with a small number of Lgr5+ stem cells, also known as crypt base columnar cells (CBC), residing near the crypt bottom. These CBC cells give rise to rapidly proliferating transit amplifying (TA) cells, which migrate upward along the crypt-villus axis and develop into the terminally differentiated cell types (Barker, 2014; van der Flier and Clevers, 2009). While absorptive enterocytes constitute the most abundant cell type, the secretory lineage comprises rare cells such as mucus producing goblet, hormone secreting enteroendocrine, and Paneth cells. Labeled cells were collected five days after label induction using an Lgr5-CreERT2 construct and a Rosa26-YFP reporter with a loxP flanked transcriptional roadblock (Figure 1B).

We first improved the robustness of the initial clustering step of the RaceID algorithm by replacing the k-means clustering with k-medoids clustering (Figure S1). Second, we noticed that the previously used gap statistic (Tibshirani et al., 2001) was not ideal for determining the cluster number. Although increasing the number of clusters in many cases leads to a growing gap statistic, the decrease of the within-cluster dispersion (Tibshirani et al., 2001) saturates quickly. A further increase of the cluster number therefore reduces the cluster reproducibility. In RaceID2 we thus determine the cluster number by identifying the saturation point of the within-cluster dispersion. Together, these two changes lead to a more robust initial clustering of RaceID2 (Experimental Procedures and Figure S1).

For the intestinal lineage tracing data (see Experimental Procedures), RaceID2 recovered a larger group of Lgr5+ stem cells (cluster 2) and early progeny (cluster

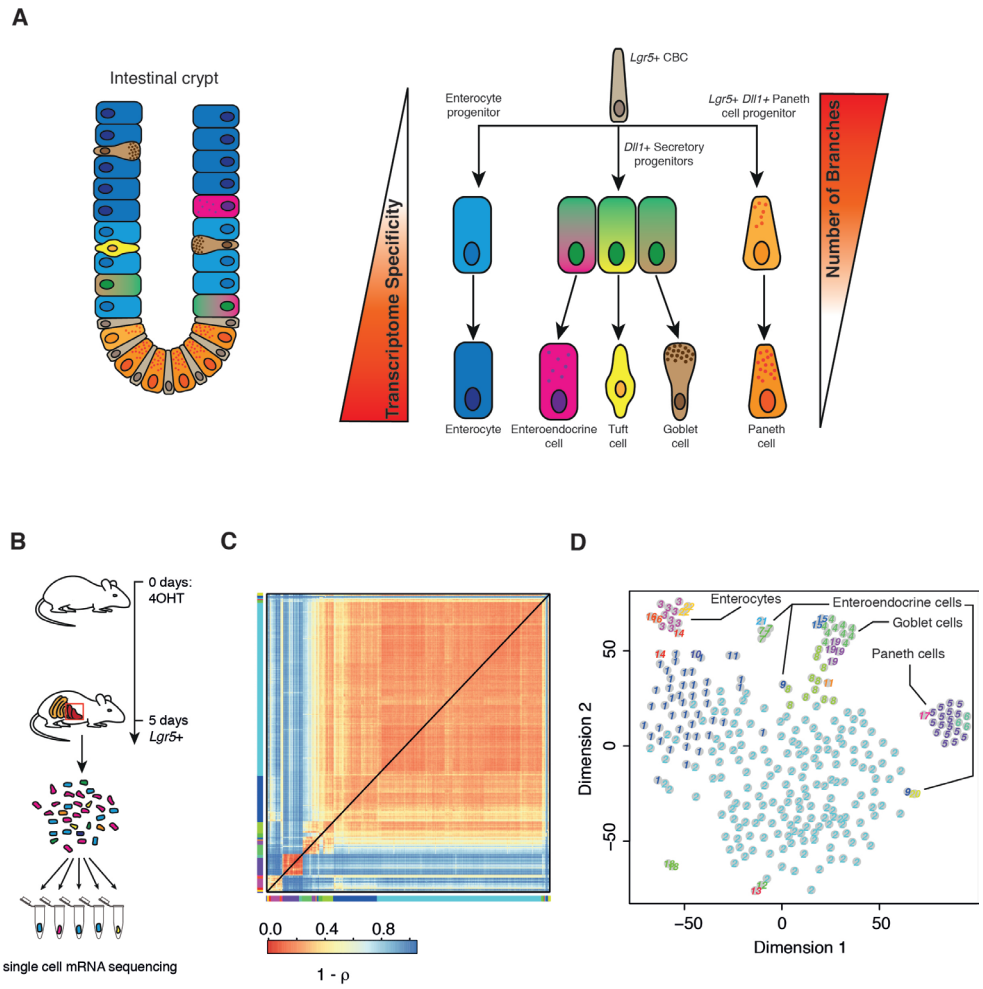


Figure 1. RaceID2 recovers intestinal cell types.
(A) The intestinal epithelium is a well-characterized differentiation system. Lgr5-positive stem cells give rise to secretory and absorptive precursors by WNT and NOTCH signaling, which further differentiate into mature intestinal cell types. (B) Summary of the lineage tracing experiment performed to sequence single 5 days old progeny of Lgr5-positive cells. (C) Heatmap of cell-to-cell transcriptome distances measured by 1 – Pearson's correlation coefficient (ρ). RaceID2 cluster are color coded along the boundaries. (D) t-SNE map representation of transcriptome similarities between individual cells. Clusters identified in (C) are highlighted with different numbers and colors and corresponding intestinal cell types identified based on known marker genes are indicated. (See also Figure S1).

1, 8), as well as the major mature cell types, i. e. enterocytes (cluster 3), goblet (cluster 4, 19), Paneth (cluster 5, 6), and enteroendocrine cells (cluster 7) (Figure 1C,D). These cell types could be unambiguously assigned based on the cluster-specific up-regulation of marker genes inferred by RaceID2 (Table S1).

Inference of the lineage tree with guided topology

One of the major challenges for the inference of differentiation pathways in a system with multiple cell lineages is the determination of branching points. To overcome this problem we predefined the topology of the lineage tree by allowing differentiation trajectories linking each pair of clusters. A putative differentiation trajectory links the medoids of two clusters and the ensemble of all inter-cluster links defines the possible topology of the lineage tree. To minimize the impact of technical noise and at the same time the computational burden, we first reduce dimensionality of the input space requiring maximal conservation of all point-to-point distances. In a second step we assign each cell to its most likely position on a single inter-cluster link. In order to find this position, the vector connecting the medoid of a cluster to one of its cells is projected onto the links between the medoid of this and all remaining clusters and the cell is assigned to the link with the longest projection after normalizing the length of each link to one. The projection also defines the most likely position of the cell on the link (Figure 2A), reflecting its differentiation state (see Experimental Procedures). If this strategy is applied to the intestinal data, only a subset of links is populated (Figure 2B). To determine links that are more highly populated than expected by chance and are therefore candidates for actual differentiation trajectories, we computed an enrichment p-value based on a comparison to a background with randomized cell positions (Figure 2B and Figure S2A). Furthermore, we reasoned that the coverage of a link by cells indicates how likely this link represents an actual differentiation trajectory and not only biased perturbations driving the transcriptome of a given cluster preferentially towards the transcriptome of another cluster without leading to actual differentiation events. We defined a link score as one minus the maximum difference between the positions of each pair of neighboring cells on link after normalizing the length of each link to one (Figure S2B). If this score is close to one, the link is densely covered with cells with only small gaps in between. If the link score is close to zero, the cell density is only concentrated near the cluster centers connected by this link. A detailed description of the algorithm is given in the Experimental Procedures. The computationally inferred intestinal lineage tree is consistent with the known lineage tree (Figure 1A). Secretory cell types (clusters 4, 5, 6, 7) populate individual branches emanating from the central Lgr5+ cluster and absorptive enterocytes (cluster 3) differentiate from the same group via a more abundant group of TA cells (cluster 1).

We compared the inferred lineage tree to the tree predicted by Monocle (Trapnell et al., 2014), a recent method for the derivation of branched lineage trees that does not rely on a predefined tree topology and found that Monocle could not resolve the different branches of secretory cells (Figure S2).

High connectivity and high transcriptome entropy reveals the identity of the stem cell

Next we attempted to predict the stem cell identity from the lineage tree. Our working definition of a stem cell for this purpose purely relies on multipotency. More precisely, we try to identify from the lineage tree the cell population with the highest degree of multipotency. We noticed that different cell types showed a variable number of populated links to other clusters. The link score is reflected by the thickness of the line in our graphical representation (Figure 2B). We also show links with low link score, since they are informative about the associated cell state. For example, a cell type with many low scoring links can fluctuate towards a diversity of fate biases, while cell types with only few links are much more canalized. These two scenarios reflect a more promiscuous transcriptome, such as expected for a stem cells, versus a more confined transcriptome as expected for a mature cell type. In our data, cluster 2, which contains cells positive for Lgr5 and other established stem cell marker (Ascl2, Clca4) (Figure 2C), was the most highly connected cluster. Another putative property of stem cells is the tendency to exhibit a more uniform composition of the transcriptome in comparison to differentiated cells. Mature cell types frequently express a small number of genes at very high levels crucial for cell-type specific functions. The transcriptome of Paneth cells, for instance, is dominated by high numbers of lysozymes and other host defense genes. The uniformity of the transcriptome is reflected by Shannon's entropy (Shannon, 1948), and this concept has previously been applied to study cellular differentiation (Anavy et al., 2014; Banerji et al., 2013; Piras et al., 2014) (see Experimental Procedures). We anticipate that the transcriptome of a multipotent cell type is not only more uniform in each individual cell. In addition, multiple state biases could coexist within this population that can give rise to diverse mature cell types upon external stimuli, or stochastically, leading to high entropy (Banerji et al., 2013; Ridden et al., 2015). For the intestinal lineage tracing data, both Paneth and goblet cells had clearly reduced entropy compared to Lgr5-positive cells while the entropy of enterocytes and enteroendocrine cells was comparable to stem cells (Figure 2D). We found that for all analyzed datasets (see below) the number of links discriminates better between multipotent and differentiated cells if rescaled by the entropy. Therefore, the simplest score that performs well in discriminating multipotent cells from the remaining cell types was a product of the entropy (after subtracting the minimal entropy observed in the system) and the number of links (see Experimental Procedures). This score exhibits a clear maximum for cluster 2 comprising the Lgr5+ stem cells (Figure 2D). We named our algorithm for the lineage tree inference and the derivation of this score StemID.

StemID recovers intestinal stem cells in a complex dataset with non-random cell-type frequencies

Next, we wanted to test if StemID could identify Lgr5+ cells in a larger and more complex data set comprising intestinal cells of various independent experiments conducted in our lab. In this dataset we combined three weeks and eight weeks Lgr5 lineage tracing data. A subset of those was enriched in secretory cells by fluorescence activated cell sorting (FACS) on CD24 (van Es et al., 2012) (Figure

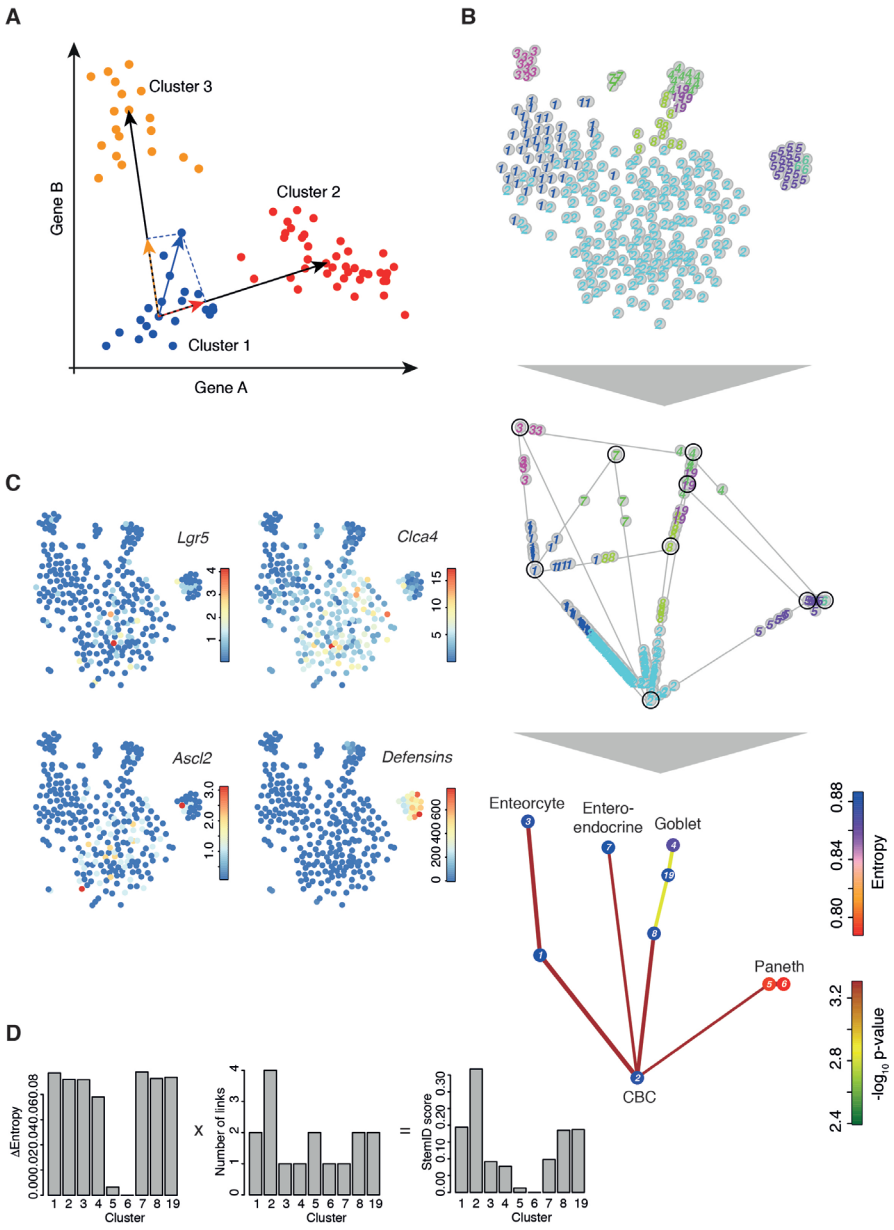


Figure 2. Lineage tree inference for intestinal stem cell progeny.
(A) Schematic representation of the method to infer differentiation trajectories (B) Outline of the method visualized in the t-SNE space. All RaceID2 clusters with >2 cells (upper panel) are connected by links and for each cell the link with the maximum projection is determined as shown in (A). Only populated links are shown (middle panel). Cluster centers are circled in black. Significant links are inferred by comparing to background distribution with randomized cell positions. Only significant links are shown (P<0.01) and color indicates -$\log_{10}$p-value. The color of the vertices indicates the entropy. The thickness indicates how densely a link is covered with cells. (C) Transcript counts (color legend) of intestinal stem cell markers Lgr5, Clca4, Ascl2 in the t-SNE map. Accumulated transcript counts across all Defensin genes, which are markers of Paneth cells are shown at the bottom right. (D) Barplot of StemID scores for all clusters. The transcriptome entropy of each cell type was computed after averaging transcript counts across all cells in a cluster (left). The lowest entropy across all cell types was subtracted for each cell type, since absolute differences were only small. This Dentropy was multiplied by the number of significant links for each cluster (middle), yielding the StemID score (right). (See also Figure S2).

S3). For both time points we also sorted non-traced CD24+ control cells (see Experimental Procedures and Figure S3). RaceID2 revealed the known intestinal cell types within this dataset based on cluster-specific expression of known cell-type marker genes and subdivided these into stages of differentiation or maturation (Figure 3A,B, Figure S3A). A full list of differentially expressed genes for each cluster is given in Table S2. For example, intestinal stem cells in cluster 7, marked by high expression of Lgr5 and Clca4 (Figure 3B), were connected directly to all secretory branches, while TA cells (cluster 5) primarily give rise to enterocytes (cluster 10) (Figure 3C and Figure S3C, D). Interestingly, we observed two distinct differentiation trajectories for Paneth cells (clusters 13,14), one via a Dll1-positive common precursor of Paneth and goblet cells (cluster 1), and another one directly connecting stem cells (cluster 7) or TA cells in cluster 5, marked by up-regulation of the cell cycle gene Pcna, directly to the mature Paneth cell clusters. Both the Dll1-dependent (van Es et al., 2012) and the direct route (Farin et al., 2014; Sawada et al., 1991), which was observed after ablation of Paneth cells, have been described. The recovery of alternative differentiation pathways demonstrates the power of our guided lineage inference. We were not able to recover this finding with a minimum spanning tree based alternative approach (Figure S3E).

We then computed the StemID score and found that the Lgr5+/Clca4+ cells (cluster 7) exhibit the highest score (Figure 3D). The second highest score was observed for cluster 21, which represents a common progenitor to Paneth and goblet cells. The TA cells in cluster 5, which our lineage inference identifies as progenitors with enterocyte fate bias, acquire the third highest StemID score.

Noticeably, Paneth cells in cluster 13 and mature goblet cells in cluster 2 show the same connectivity like the stem and progenitor cells in cluster 7, 5, and 21, but rescaling by the entropy helps correctly assign a mature state to these cells (Figure S3F). In conclusion, StemID could identify intestinal stem cells and distinguish progenitor populations from more mature intestinal cell types.

StemID recovers hematopoietic stem cells within a non-random sample of bone marrow cells

To test the performance of StemID in a different biological system we applied the algorithm to single cell sequencing data of mouse bone marrow cells. These cells were selected based on physical interactions between doublets or larger groups of cells, and are thus not sampled randomly from all cell types in the bone marrow. This dataset was complemented with $Kit^{+}Sca\text{-}1^{+}Lin^{-}CD48^{-}CD150^{+}$ hematopoietic stem cells (HSCs) (Kiel et al., 2005) sorted from the bone marrow (see Experimental Procedures and Figure S5B). Cell types identified by RaceID2 were dominated by the myeloid lineage and comprised HSCs, erythroblasts, megakaryocytes, two groups of granulocytes (neutrophils and eosinophils), macrophages, a small group of B lymphocytes and several clusters representing progenitor stages of the myeloid lineage (Figure 4A,B and S6A). A full list of differentially expressed genes for each cluster is shown in Table S3. Cluster 1 comprises almost exclusively sorted HSCs (Figure S4B). The inferred lineage tree (Figure 4C and S6C,D) indicates that HSCs differentiate into multipotent progenitor cells (cluster 5), but are also directly linked to mature lineages. HSCs and multipotent progenitors are both linked to megakaryocytes (cluster 4), to eosinophils (clusters 10 and 29), to macrophages

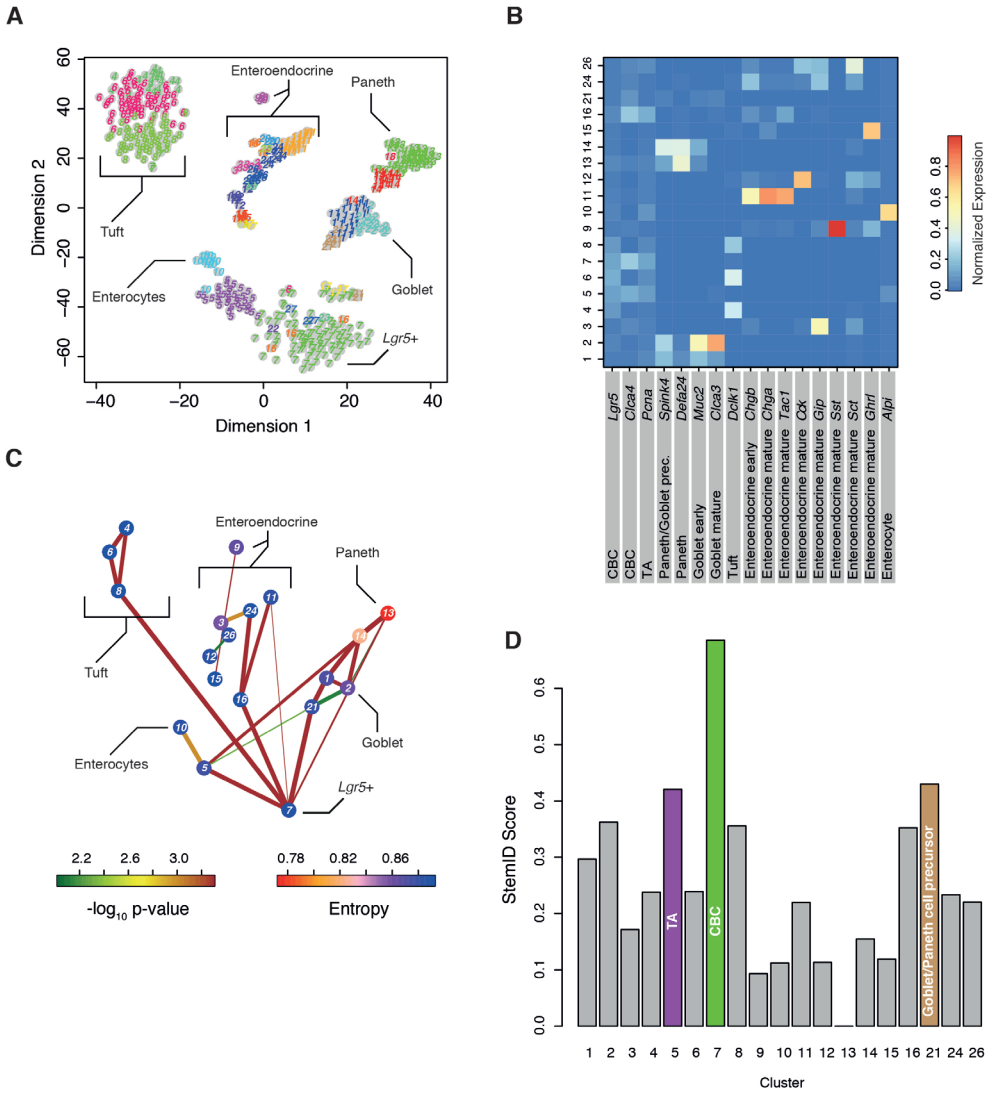


Figure 3. StemID identifies stem cells in complex non-random mixtures of intestinal cells.
(A) t-SNE map of transcriptome similarities of intestinal cells from a variety of single cell mRNA sequencing experiments (see main text and Figure S3). RaceID2 clusters are highlighted with different numbers and colors. Cell types identified based on marker gene expression are shown. (B) Heatmap showing the average expression of known cell type markers across all clusters with >5 cells. For each gene the sum of expression values over all clusters is normalized to one. (C) Inferred intestinal lineage tree. Only significant links are shown (P<0.01). The color of the link indicates the -$\log_{10}$p-value. The color of the vertices indicates the entropy. The thickness indicates the link score reflecting how densely a link is covered with cells (see Experimental procedure). (D) Barplot of StemID scores for intestinal clusters. For (B-D) only clusters with >5 cells were analyzed. (See also Figure S3, S6 and S7).

(cluster 28) and to two branches covering a spectrum of progenitor and mature states of the neutrophil (clusters 11, 3, 2, 14, 12, 22) and erythroid lineage (clusters 9, 8, 7, 6, 13), respectively. The B lymphocytes are only directly linked to the HSCs, suggesting that cluster 5 represents a myeloid progenitor population and no lymphoid progenitors were present in our sample. The inferred lineage tree is therefore consistent with the existence of a common myeloid progenitor population giving rise to erythrocytes, megakaryocytes, granulocytes and macrophages (Orkin and Zon, 2008). StemID determines the highest score for cluster 1 and therefore correctly recovers HSCs among all cell types in the mixture (Figure 4D and S6). The second highest score discriminates the mutipotent myeloid progenitors (cluster 5) from the remaining cell types and the third highest score was assigned to the earliest progenitor of the erythroblast lineage. Therefore, the level of multipotency also correlates with the StemID score bone marrow derived cells.

The high connectivity of cluster 1 provides evidence for an early fate bias already in HSCs. Moreover, the high entropy of HSCs reflects a more uniform transcriptome in individual cells of this population: the entropy distribution across all cells in this cluster is shifted in comparison to all other groups (Figure 5A). In general, the inter-cluster variability substantially exceeds the intra-cluster variability. The narrow entropy distribution of cluster 1 also rules out a strong dependence on the cell cycle. However, we also observed that 54 out of the 276 HSCs (20%) show distinct fate biases, revealed by low expression of lineage-specific marker genes (Figure 5B), a finding that is consistent with a recent report based on lineage tracing (Perié et al., 2015). Since the sensitivity of single cell sequencing is limited, this number is almost certainly an underestimation. We note, that most HSCs (112 out of 276) are assigned to the link with the multipotent progenitor (cluster 5). We cannot address if the observed fate bias persists during differentiation or if stochastic switching between distinct cell fates occurs during differentiation. Our observation is also consistent with a recent single cell transcriptome analysis showing an unexpected heterogeneity of myeloid progenitor cell populations and suggesting the existence of an early cell fate bias (Paul et al., 2015). We observe very similar sets of marker genes as found in this study, but our lineage inference permits an analysis of temporal dynamics of gene expression. As an example, we extracted all cells from the neutrophil branch (clusters 1, 11, 3, 2, 12) in pseudo-temporal order derived from the projection coordinates and clustered temporal expression profiles by using self-organizing maps (see Experimental Procedures). A z-score of gene expression values along this trajectory reveals that the RaceID2 clusters represent sets of cells with common modules of co-expressed genes and that gene expression within these modules changes smoothly over time (Figure 5C). While ribosomal protein encoding genes and other components of the translational machinery slowly decline during differentiation, other genes are transiently switched on in progenitor populations (e. g. Elane) or immature neutrophils (e. g. Ngp), or only up-regulated in mature cells (e. g. Retnlg).

Finally, we note that the identification of the HSC population by StemID is robust to changing the contribution of this population to the mixed sample. For example, when only ten HSCs are randomly selected and all others are discarded from the dataset, StemID still assigns the highest score to the small HSC cluster (data not shown).

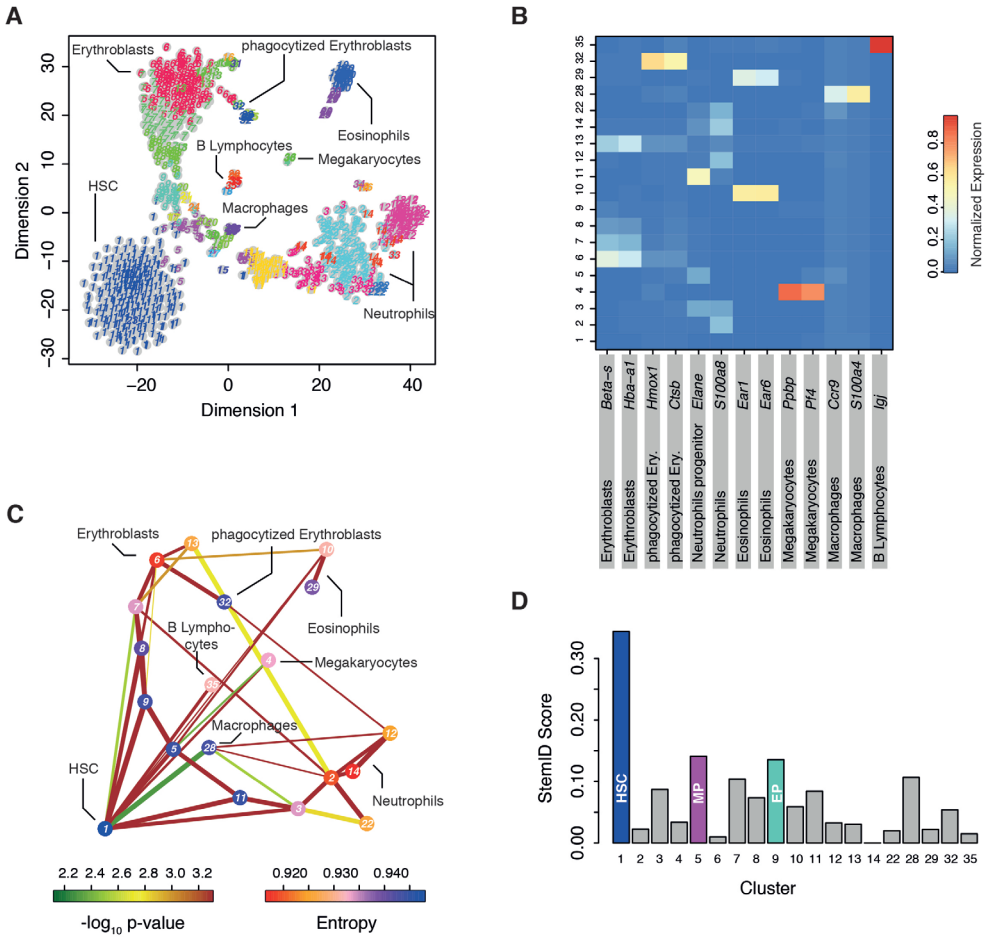


Figure 4. StemID identifies hematopoietic stem cells in non-random mixtures of bone marrow cells.

(A) t-SNE map of transcriptome similarities of hematopoietic cells sampled from physically interacting doublets or multiplets. (see main text and Figure S4). RaceID2 clusters are highlighted with different numbers and colors. Cell types identified based on marker gene expression are shown. (B) Heatmap showing the average expression of known cell type markers across all clusters with >5 cells. For each gene the sum of expression values over all clusters is normalized to one. (C) Inferred hematopoietic lineage tree. Only significant links are shown (P<0.01). The color of the link indicates the -$\log_{10}$p-value. The color of the vertices indicates the entropy. The thickness indicates the link score reflecting how densely a link is covered with cells (see Experimental procedure). (D) Barplot of StemID scores for hematopoietic clusters. HSC: Hematopoietic stem cell. MP: Myeloid progenitor. EP: Erythroblast progenitor. (See also Figure S4, S6 and S7).

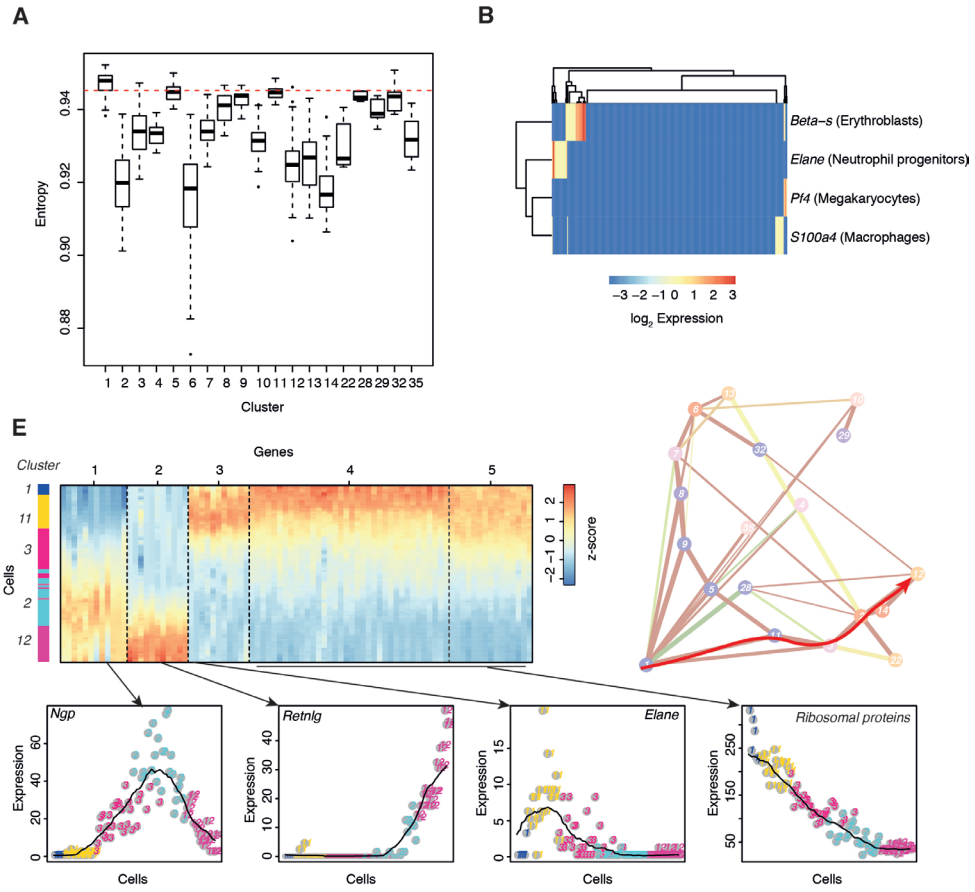


Figure 5. The multipotency of HSCs is reflected by high transcriptome entropy.
(A) Boxplot of the transcriptome entropy for all RaceID2 derived bone marrow cell types with >5 cells. Boundaries of the box represent the 25%- and 75%-quantile, the thick line corresponds to the median, and whiskers extend to th 5%- and 95%- quantile. The broken red line indicated the 25%-quantile for HSCs (cluster 1). (B) Two dimensional clustering of lineage markers in all HSCs (cluster 1). The heatmap shows logarithmic expression. (C) Self-organizing map (SOM) of z-score transformed pseudo-temporal expression profiles along the neutrophil differentiation trajectory (clusters 1, 11, 3, 2, 12) indicated by the red arrow superimposed on the lineage tree (see Experimental Procedures). The pseudo-temporal order was inferred from the projection coordinates of all cells. The color-coding on the right indicates the cluster of origin. The SOM identified five different modules of co-regulated genes. Examples are shown in the bottom panel. The clusters of origin are indicated as colors and numbers. The black line represents a moving average (window size 25). For (B-E) only clusters with >5 cells were analyzed.

In summary, StemID could successfully identify the stem cell type in a complex mixture of cells isolated from the bone marrow. The inferred lineage tree recovered known trajectories, but suggests an early cell fate bias present already in the HSC.

StemID predicts multipotent ductal cell populations among human adult pancreatic cells

After having demonstrated that StemID can robustly identify the stem cell population in two distinct biological systems, we applied the algorithm to predict multipotent cell populations in a less characterized system: the human pancreas. The pancreas consists of acinar cells, which produce the digestive enzymes, of ductal cells secreting bicarbonate to neutralize stomach acidity and of hormone producing endocrine cells that regulate hormone metabolism (Jennings et al., 2015). It is unclear, which multipotent cells maintain pancreatic homeostasis and can give rise to different mature cell types during regeneration upon injury. Although early studies have suggested that in humans these cell populations could reside within the exocrine compartment or that dedifferentiation of exocrine cells could give rise to endocrine cells (Bonner-Weir et al., 2000; Puri et al., 2015) the identity of multipotent cell populations is still unclear (Jiang and Morahan, 2014). We sequenced pancreatic cells from human donors (see Experimental Procedures) and application of RaceID2 revealed all major cell types, including different subpopulations of acinar and ductal cells, hormone producing a-, b-, d- and PP-cells, and stellate cells (Figure 6A,B and Figure S5A,B). A full list of differentially expressed genes for each cluster is shown in Table S4. In particular, we discovered novel subpopulations of ductal cells. In one of these groups (cluster 14), the cell surface glycoprotein CEACAM6 was significantly up-regulated ($P<0.01$, see Experimental Procedures), while components of the ferritin protein (FTH1, FTL), which is the major intracellular iron storage protein, were significantly up-regulated ($P<0.01$, see Experimental Procedures) in the other group (cluster 4) (Figure 6C).

The inferred lineage tree assigns a central position to the ductal cells (Figure 6D and S7C-E). Distinct sub-types of ductal cells appear to give rise to different endocrine sub-types and acinar cells. While differentiation trajectories link cluster 4 to acinar, PP-, and b-cells, cluster 14 is linked to a- and d-cells. Consistently, cluster 4 and 14 acquire the highest StemID score indicating the highest level of multipotency among the cell types detected in this system (Figure 6E and S7F). The following ranks of the StemID score were occupied by other ductal sub-types and precursor cells that give rise to two sub-states of acinar cells. Interestingly, cluster 4 also directly connects to stellate cells. Upon injury, these cells can switch to an activated state and migrate to the injured location in order to participate in tissue repair (Omary et al., 2007).

To collect further evidence that cluster 4 is an endocrine progenitor cell we plotted the expression of the cluster 4 marker FTH1 and the b-cell marker insulin (INS) in single cells residing on the differentiation trajectory connecting these two cell types. Cells were ordered by their projection coordinate. The genes exhibited smooth, anti-correlated gradients suggestive of a continuous transition between these two cell types (Figure 6F). To independently validate this observation we performed antibody staining against Insulin and FTL in human pancreatic tissue sections. We were able to detect individual cells co-expressing Insulin and FTL within ductal

structures, confirming the existence of cluster 4 cells (Figure 7A). Co-staining of Glucagon revealed that these cells specifically produce Insulin and not Glucagon (Figure 7B) as suggested by our analysis (Figure 6C). Our results suggest that the ferritin positive sub-population of ductal cells might differentiate into mature b-cells.

DISCUSSION

In this study we present an approach to identify stem cells using single cell transcriptomics data. Since the physiological state of a cell is an approximate reflection of its transcriptome, it is a reasonable assumption that cell types can be discriminated based on their transcriptome. However, determining the stem cell identity among all the rare cell types discovered also requires the derivation of a lineage tree.To address this task we combined cell type identification by RaceID2 with a tree reconstruction by guided topology. We first introduce an improved version of our previous RaceID algorithm (Grün et al., 2015) with a more robust initial clustering step: The replacement of k-means by k-medoids leads to increased robustness of clustering for all datasets analyzed in the paper. For the complex intestinal dataset (Figure 3), the fraction of clusters with Jaccard's similarity >0.7 is 40% for k-means versus 73% for k-medoids. The corresponding fractions are 58% versus 83% for the bone marrow data and 40% versus 90% for the pancreas data. To infer differentiation trajectories we assign every cell onto a specific link between its cluster of origin and another cluster based on the longest projection of the vector connecting the cluster center with the cell position onto these links. This adequately reflects how much a cell has moved from the most representative cell state in the same cluster (the medoid) towards another cell identity (or vice versa). If significantly more cells reside on a link than expected by chance, this provides strong evidence that cells of the cluster of origin exhibit a pronounced transcriptome bias towards another cell fate. In addition, if a continuum of cell states covers a given link, as evidenced by a high link score, this link represents a strong candidate for an actual differentiation trajectory. Significant links with reduced link scores, on the other hand, indicate plasticity of the connected cell types in a sense that the transcriptome of a cell type can to some extent fluctuate towards another fate.

The quality of our lineage inference is supported by the recovery of known differentiation trajectories in the intestinal epithelium and the bone marrow. Remarkably, we recovered a rare alternative differentiation pathway where Lgr5+ cells differentiate directly into Paneth cells without intermediate Dll1+ progenitors (Farin et al., 2014; Sawada et al., 1991). We could also show for the intestinal and the bone marrow data that StemID infers a lineage tree with substantially higher resolution in comparison to previously published methods (Haghverdi et al., 2015; Trapnell et al., 2014) (Figure S6).

The derived lineage tree for the bone marrow suggested, that, in contrast to the classical view of dichotomous differentiation via a hierarchy of increasingly restricted progenitor populations (Giebel and Punzel, 2008), a cell fate bias already exists at stages as early as the HSC stage (Figure 5B). This observation is consistent with a recent single cell transcriptome analysis revealing heterogeneity of the common myeloid progenitor cell population indicating early fate bias (Paul et al., 2015). Moreover, direct generation of progenitors restricted to the myeloid fate from mouse HSCs has been described in the past (Yamamoto et al., 2013), and the existence of unipotent cells within the human HSC and classically defined multipotent progenitor

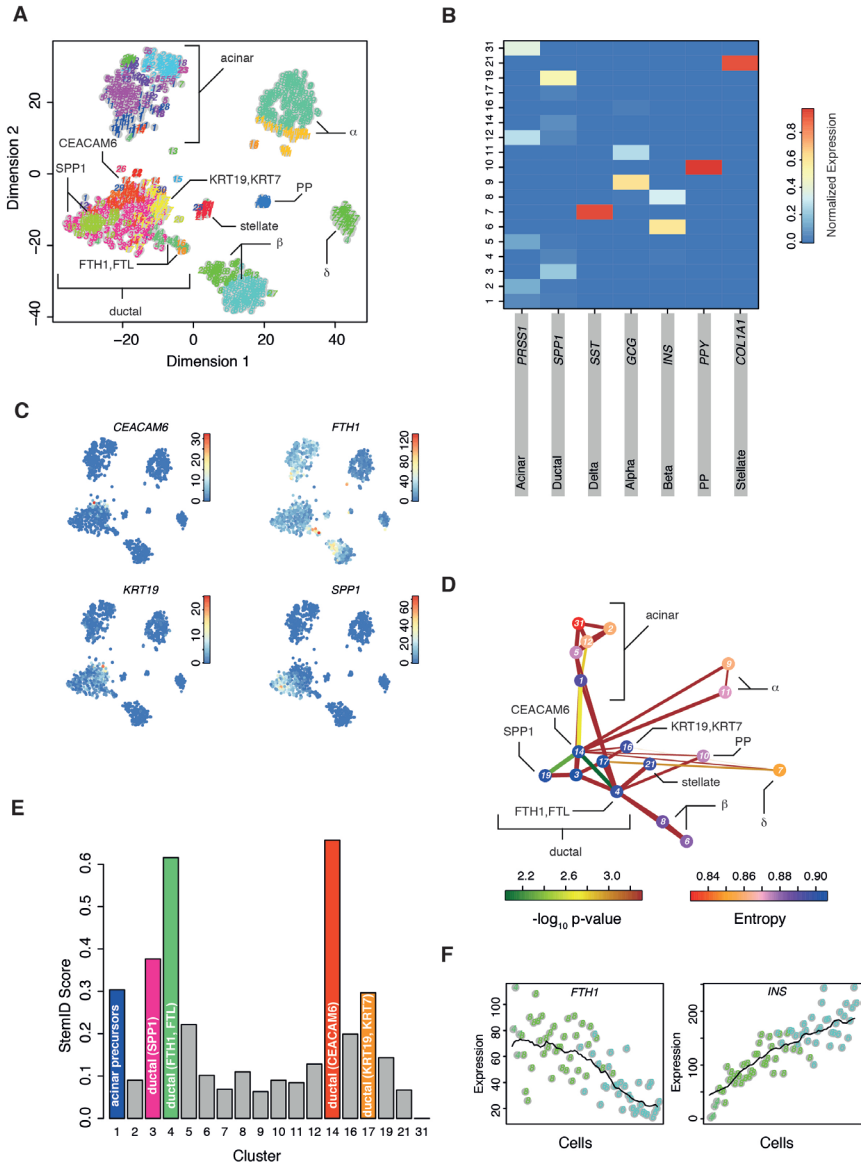


Figure 6. StemID predicts human pancreatic pluripotent cells.
(A) t-SNE map of transcriptome similarities of human pancreatic cells. RaceID2 clusters are highlighted with different numbers and colors. Cell types identified based on marker gene expression are shown. For ductal cells marker genes of sub-populations are shown. (B) Transcript counts (color legend) of ductal sub-type markers CEACAM6, FTH1, KRT19 and SPP1 are highlighted in the t-SNE map. (C) Inferred pancreatic lineage tree. Only significant links are shown (P<0.01) and color indicates -$\log_{10}$p-value. The color of the vertices indicates the entropy. The thickness indicates the link score reflecting how densely a link is covered with cells. (D) Barplot of StemID scores for pancreatic clusters. (E) Pseudo-temporal expression profiles for INS and FTH1. The transcript count is plotted for cells on the link, connecting cluster 4, 8, and 6. Cells are ordered by the projection coordinate. For (B-D) only clusters with >5 cells were analyzed. (See also Figure S5).

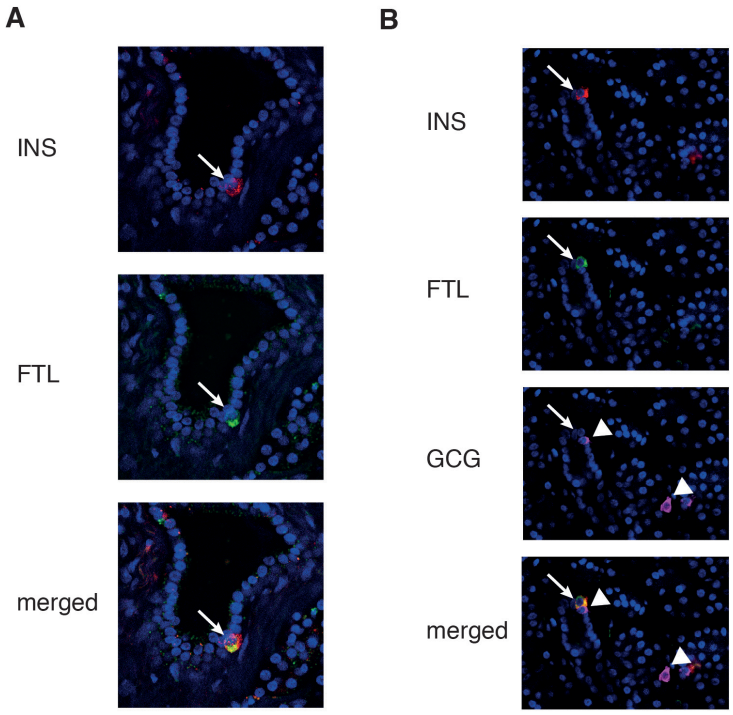


Figure 7. Validation of putative endocrine precursor cells in ductal sub-populations by antibody staining.
(A, B) Antibody staining for INS and FTH1 in human pancreatic tissue. (A) Shown is a single cell positive for INS and FTH1 residing in the lining of the duct (arrow). (B) Antibody staining for INS, FTH1 and GCG in human pancreatic tissue. Shown is a single cell positive for INS and FTH1 residing in the lining of the duct (arrow) next to a GCG expressing cell (arrowhead). Another GCG expressing cell is found nearby (arrowhead). Both GCG expressing cells are FTL negative.

populations was shown recently (Paul et al., 2015; Perié et al., 2015).
For both model systems, the StemID score, which quantifies very general properties of a cell type, i. e. the number of links and the entropy of the transcriptome, ranks RaceID2 predicted cell types by their level of pluripotency. Lgr5+ CBC cells and sorted HSCs acquire the highest score among all cell types of the intestine and the bone marrow, respectively, demonstrating the performance of our algorithm. We could further demonstrate the performance of StemID on two previously published datasets (Figure S7) for cells from developing lung epithelium (Treutlein et al., 2014) and differentiating human radial glial cells (Pollen et al., 2015).
Potential problems for the StemID algorithm arise in the absence of intermediate progenitors or the occurrence of unrelated cell types. In the absence of intermediate progenitors, StemID infers a link to a more multipotent population. For example, B lymphocytes in the bone marrow dataset are directly linked to HSCs. It is known that a spectrum of progenitors will reside on this trajectory and as we have observed for the other lineages, an early fate bias towards lymphocytes could exist in HSCs.

In the absence of intermediate progenitors, a link to a more multipotent population reflects all information on the lineage relationship that can be extracted from the data. If the stem cell itself is missing from the sample, StemID will identify the cell type with the highest level of multipotency. The presence of unrelated cell types in the mixture could lead to false positive links. However, since the feature space is high-dimensional, it is likely that none of the links between an unrelated cell type and the remaining lineage tree will be significantly populated. We also argue that links of mature cell types to related progenitor or stem cell populations were identified with high specificity (oftentimes only a single link in line with previous findings was detected). This makes the occurrence of significant links between unrelated cell types unlikely.

Finally, we used StemID to screen human adult pancreatic cells for multipotent cell populations. It is unclear, which adult pancreatic cell types can give rise to the different mature pancreatic lineages during normal tissue turnover or regeneration. Although initial evidence suggested that multipotent cells within the ductal compartment could differentiate into endocrine cells both in human and mice (Jiang and Morahan, 2014), subsequent lineage tracing experiments produced contradictory results. While mouse lineage tracing of carbonic anhydrase II (Ca2) positive ductal cells revealed that these cells give rise to b-cells upon injury (Bonner-Weir et al., 2008), lineage tracing of Sox9-, Muc1- or Hnf1b-positive cells could not confirm this finding (Furuyama et al., 2011; Kopinke and Murtaugh, 2010; Kopp et al., 2011; Solar et al., 2009). Using StemID we were able to predict distinct sub-populations of ductal cells with varying differentiation potential. While ductal cells marked by high levels of CEACAM6 are predicted to differentiate into a-, d- and PP-cells, another sub-population expressing high levels of the ferritin complex primarily appears to give rise to b-cells and acinar cells. We note that the latter sub-population does not express any of the markers used in previous lineage tracing experiments, but we caution that expression of these genes might be too low to be reliably detected by single cell mRNA sequencing. We further remark that b-cell differentiation in the adult pancreas might not be conserved between human and mouse.

We provide well-documented R source code for RaceID2 and the StemID algorithm at https://github.com/dgrun/StemID. We hope that StemID will be useful for a better understanding of differentiation dynamics in a variety of systems.

EXPERIMENTAL PROCEDURES

Lineage tracing experiments

For lineage tracing experiments we injected 0.4 mg tamoxifen into 3-month old Lgr5-CreERT2 C57Bl6/J mice bred to a Rosa26LSL-YFP reporter mice.

Isolation of crypts from mouse small intestine

Crypts were isolated from mice as described previously (Sato et al., 2009). See Supplemental Experimental Procedures for more details.

Human islet isolation, dispersion and sorting

Pancreatic cadaveric tissue was procured from a multiorgan donor program and only used if the pancreas could not be used for clinical pancreas or islet transplantation, according to national laws, and if research consent was present. Human islet isolations were performed in the islet isolation facility of the Leiden University Medical Center according to a modified protocol originally described by Ricordi et al. (Ricordi et al., 1988). See Supplemental Experimental Procedures for details on culturing and cell sorting.

Immunofluorescence

Pancreatic tissue samples were fixed overnight in 4% formaldehyde (Klinipath), stored in 70% ethanol, and subsequently embedded in paraffin. After deparaffinization and rehydration in xylene and ethanol respectively, antigen retrieval was performed in citric buffer for 20 minutes. Sections were blocked with 2% normal donkey serum and 1% lamb serum in PBS. Primary antibodies were rabbit anti-Ftl (ab69090), mouse anti-Glucagon (ab10988) and guinea pig anti-Insulin (ab7842). Alexa Fluor conjugated secondary antibodies against rabbit, mouse and guinea pig IgG (Life Technologies A11008, A10037 and A21450) were used at a dilution of 1:200. Nuclear counterstaining was done by embedding with DAPI vectashield (Vector Laboratories #H-1500). Imaging was performed on a Leica SP8 confocal microscope using hybrid detectors.

Preparation of mouse hematopoietic cells

We used C57Bl/6 female or male mice, from 23 to 52 weeks, bred in our facility. Experimental procedures were approved by the Dier Experimenten Commissie (DEC) of the KNAW, and performed according to the guidelines. Bone marrow was isolated from femur and tibia by flushing Hank's Balanced Salt Solution (HBSS, Invitrogen) without calcium or magnesium, supplemented with 1% heat-inactivated Fetal Calf Serum (FCS, Sigma). See Supplemental Experimental Procedures for details on single cell isolation.

CEL-seq library preparation

The protocol was carried out as described previously (Grün et al., 2015). See Supplemental Experimental Procedures for a detailed description.

Quantification of transcript abundance

Read mapping and quantification was done as described previously (Grün et al., 2015). See Supplemental Experimental Procedures for a detailed description.

RaceID2 and StemID
A brief overview is given in the Results section. The algorithm and follow-up analyses are described in full detail in the Supplemental Experimental Procedures.

ACCESSION NUMBERS
The accession numbers for the RNA-seq datasets reported in this paper are GSE76408, GSE76983 and GSE81076.

SUPPLEMENTAL INFORMATION
Supplemental Information includes Supplemental Experimental Procedures, seven figures and four tables.

AUTHOR CONTRIBUTIONS
D.G. and A.v.O. conceived the study; D. G. developed the algorithm and performed all computational analyses; single cell sequencing of pancreatic cells and antibody staining was performed by M. J. M with the help of G. D.; single cell sequencing of intestinal cells was performed by K. W. with the help of A. L., J. v. E. and M. v. d. B.; single cell sequencing of bone marrow cells was performed by J.-C. B.; E. J. helped with antibody stainings; D. G. wrote the manuscript and all authors read and edited the manuscript; A. v. O. supervised the project and D. G., M. J. M., K. W., and J.-C. B.; E. J. P. d. K. supervised G. D. and E. J.; H. C. supervised M. v. d. B. and J. v. E.

ACKNOWLEDGEMENTS
This work was supported by an European Research Council Advanced grant (ERC-AdG 294325-GeneNoiseControl) and a Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) Vici award.

REFERENCES

Anavy, L., Levin, M., Khair, S., Nakanishi, N., Fernandez-Valverde, S.L., Degnan, B.M., and Yanai, I. (2014). BLIND ordering of large-scale transcriptomic developmental timecourses. Development 141, 1161–1166.

Banerji, C.R.S., Miranda-Saavedra, D., Severini, S., Widschwendter, M., Enver, T., Zhou, J.X., and Teschendorff, A.E. (2013). Cellular network entropy as the energy potential in Waddington's differentiation landscape. Sci. Rep. 3, 3039.

Barker, N. (2014). Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. Nat. Rev. Mol. Cell Biol. 15, 19–33.

Barker, N., van Es, J.H., Kuipers, J., Kujala, P., van den Born, M., Cozijnsen, M., Haegebarth, A., Korving, J., Begthel, H., Peters, P.J., et al. (2007). Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature 449, 1003–1007.

Bendall, S.C., Davis, K.L., Amir, E.-A.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell 157, 714–725.

Bonner-Weir, S., Taneja, M., Weir, G.C., Tatarkiewicz, K., Song, K.H., Sharma, A., and O'Neil, J.J. (2000). In vitro cultivation of human islets from expanded ductal tissue. Proc. Natl. Acad. Sci. U. S. A. 97, 7999–8004.

Bonner-Weir, S., Inada, A., Yatoh, S., Li, W.-C., Aye, T., Toschi, E., and Sharma, A. (2008). Transdifferentiation of pancreatic ductal cells to endocrine beta-cells. Biochem. Soc. Trans. 36, 353–356.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 10, 1093–1095.

Buczacki, S.J.A., Zecchini, H.I., Nicholson, A.M., Russell, R., Vermeulen, L., Kemp, R., and Winton, D.J. (2013). Intestinal label-retaining cells are secretory precursors expressing Lgr5. Nature 495, 65–69.

Busch, K., Klapproth, K., Barile, M., Flossdorf, M., Holland-Letz, T., Schlenner, S.M., Reth, M., Höfer, T., and Rodewald, H.-R. (2015). Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. Nature 518, 542–546.

Eaves, C.J. (2015). Hematopoietic stem cells: concepts, definitions and the new reality. Blood 125, 2605–2613.

Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. Nature 467, 167–173.

van Es, J.H., Sato, T., van de Wetering, M., Lyubimova, A., Nee, A.N.Y., Gregorieff, A., Sasaki, N., Zeinstra, L., van den Born, M., Korving, J., et al. (2012). Dll1+ secretory progenitor cells revert to stem cells upon crypt damage. Nat. Cell Biol. 14, 1099–1104.

Farin, H.F., Karthaus, W.R., Kujala, P., Rakhshandehroo, M., Schwank, G., Vries, R.G.J., Kalkhoven, E., Nieuwenhuis, E.E.S., and Clevers, H. (2014). Paneth cell extrusion and release of antimicrobial products is directly controlled by immune cell-derived IFN-γ. J. Exp. Med. 211, 1393–1405.

van der Flier, L.G., and Clevers, H. (2009). Stem cells, self-renewal, and differentiation in the intestinal epithelium. Annu. Rev. Physiol. 71, 241–260.

Furuyama, K., Kawaguchi, Y., Akiyama, H., Horiguchi, M., Kodama, S., Kuhara, T., Hosokawa, S., Elbahrawy, A., Soeda, T., Koizumi, M., et al. (2011). Continuous cell supply from a Sox9-expressing progenitor zone in adult liver, exocrine pancreas and intestine. Nat. Genet. 43, 34–41.

Giebel, B., and Punzel, M. (2008). Lineage development of hematopoietic stem and progenitor cells. Biol. Chem. 389, 813–824.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods 11, 637–640.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 525, 251–255.

Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics 31, 2989–2998.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science 343, 776–779.

Jennings, R.E., Berry, A.A., Strutt, J.P., Gerrard, D.T., and Hanley, N.A. (2015). Human pancreas development. Development 142, 3126–3137.

Jiang, F.-X., and Morahan, G. (2014). Pancreatic Stem Cells Remain Unresolved. Stem Cells Dev. 23, 2803–2812.

Kiel, M.J., Yilmaz, O.H., Iwashita, T., Yilmaz, O.H., Terhorst, C., and Morrison, S.J. (2005). SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. Cell 121, 1109–1121.

Kopinke, D., and Murtaugh, L.C. (2010). Exocrine-to-endocrine differentiation is detectable only prior to birth in the uninjured mouse pancreas. BMC Dev. Biol. 10, 38.

Kopp, J.L., Dubois, C.L., Schaffer, A.E., Hao, E., Shih, H.P., Seymour, P.A., Ma, J., and Sander, M. (2011). Sox9+ ductal cells are multipotent progenitors throughout development but do not produce new endocrine cells in the normal or injured adult pancreas. Development 138, 653–665.

Lancaster, M.A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature 501, 373–379.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214.

Omary, M.B., Lugea, A., Lowe, A.W., and Pandol, S.J. (2007). The pancreatic stellate cell: a star on the rise in pancreatic diseases. J. Clin. Invest. 117, 50–59.

Orkin, S.H., and Zon, L.I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. Cell 132, 631–644.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B. V, Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396–1401.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell 163, 1663–1677.

Perié, L., Duffy, K.R., Kok, L., de Boer, R.J., and Schumacher, T.N. (2015). The Branching Point in Erythro-Myeloid Differentiation. Cell 163, 1655–1662.

Piras, V., Tomita, M., and Selvarajoo, K. (2014). Transcriptome-wide variability in single embryonic development cells. Sci. Rep. 4, 7137.

Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A., et al. (2015). Molecular Identity of Human Outer Radial Glia during Cortical Development. Cell 163, 55–67.

Puri, S., Folias, A.E., and Hebrok, M. (2015). Plasticity and Dedifferentiation within the Pancreas: Development, Homeostasis, and Disease. Cell Stem Cell 16, 18–31.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 135, 216–226.

Ricordi, C., Lacy, P.E., Finke, E.H., Olack, B.J., and Scharp, D.W. (1988). Automated method for isolation of human pancreatic islets. Diabetes 37, 413–420.

Ridden, S.J., Chang, H.H., Zygalakis, K.C., and MacArthur, B.D. (2015). Entropy, Ergodicity, and Stem Cell Multipotency. Phys. Rev. Lett. 115, 208103.

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

Sawada, M., Takahashi, K., Sawada, S., and Midorikawa, O. (1991). Selective killing of Paneth cells by intravenous administration of dithizone in rats. Int. J. Exp. Pathol. 72, 407–421.

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423, 623–656.

Solar, M., Cardalda, C., Houbracken, I., Martín, M., Maestro, M.A., De Medts, N., Xu, X., Grau, V., Heimberg, H., Bouwens, L., et al. (2009). Pancreatic exocrine duct cells give rise to insulin-producing beta cells during embryogenesis but not after birth. Dev. Cell 17, 849–860.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Statistical Methodol. 63, 411–423.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386.

Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371–375.

Wilson, N.K., Kent, D.G., Buettner, F., Shehata, M., Macaulay, I.C., Calero-Nieto, F.J., Sánchez Castillo, M., Oedekoven, C.A., Diamanti, E., Schulte, R., et al. (2015). Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. Cell Stem Cell 16, 712–724.
Yamamoto, R., Morita, Y., Ooehara, J., Hamanaka, S., Onodera, M., Rudolph, K.L., Ema, H., and Nakauchi, H. (2013). Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. Cell 154, 1112–1126.

Zeisel, A., Manchado, A.B.M., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–1142.turbed haematopoiesis from stem cells in vivo. Nature 518, 542–546.
Eaves, C.J. (2015). Hematopoietic stem cells: concepts, definitions and the new reality. Blood 125, 2605–2613.

Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. Nature 467, 167–173.

van Es, J.H., Sato, T., van de Wetering, M., Lyubimova, A., Nee, A.N.Y., Gregorieff, A., Sasaki, N., Zeinstra, L., van den Born, M., Korving, J., et al. (2012). Dll1+ secretory progenitor cells revert to stem cells upon crypt damage. Nat. Cell Biol. 14, 1099–1104.

Farin, H.F., Karthaus, W.R., Kujala, P., Rakhshandehroo, M., Schwank, G., Vries, R.G.J., Kalkhoven, E., Nieuwenhuis, E.E.S., and Clevers, H. (2014). Paneth cell extrusion and release of antimicrobial products is directly controlled by immune cell-derived IFN-γ. J. Exp. Med. 211, 1393–1405.

van der Flier, L.G., and Clevers, H. (2009). Stem cells, self-renewal, and differentiation in the intestinal epithelium. Annu. Rev. Physiol. 71, 241–260.

Furuyama, K., Kawaguchi, Y., Akiyama, H., Horiguchi, M., Kodama, S., Kuhara, T., Hosokawa, S., Elbahrawy, A., Soeda, T., Koizumi, M., et al. (2011). Continuous cell supply from a Sox9-expressing progenitor zone in adult liver, exocrine pancreas and intestine. Nat. Genet. 43, 34–41.

Giebel, B., and Punzel, M. (2008). Lineage development of hematopoietic stem and progenitor cells. Biol. Chem. 389, 813–824.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods 11, 637–640.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 525, 251–255.

Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics 31, 2989–2998.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science 343, 776–779.

Jennings, R.E., Berry, A.A., Strutt, J.P., Gerrard, D.T., and Hanley, N.A. (2015). Human pancreas development. Development 142, 3126–3137.

Jiang, F.-X., and Morahan, G. (2014). Pancreatic Stem Cells Remain Unresolved. Stem Cells Dev. 23, 2803–2812.

Kiel, M.J., Yilmaz, O.H., Iwashita, T., Yilmaz, O.H., Terhorst, C., and Morrison, S.J. (2005). SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. Cell 121, 1109–1121.

Kopinke, D., and Murtaugh, L.C. (2010). Exocrine-to-endocrine differentiation is detectable only prior to birth in the uninjured mouse pancreas. BMC Dev. Biol. 10, 38.

Kopp, J.L., Dubois, C.L., Schaffer, A.E., Hao, E., Shih, H.P., Seymour, P.A., Ma, J., and Sander, M. (2011). Sox9+ ductal cells are multipotent progenitors throughout development but do not produce new endocrine cells in the normal or injured adult pancreas. Development 138, 653–665.

Lancaster, M.A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature 501, 373–379.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214.

Omary, M.B., Lugea, A., Lowe, A.W., and Pandol, S.J. (2007). The pancreatic stellate cell: a star on the rise in pancreatic diseases. J. Clin. Invest. 117, 50–59.

Orkin, S.H., and Zon, L.I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. Cell 132, 631–644.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B. V, Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396–1401.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell 163, 1663–1677.

Perié, L., Duffy, K.R., Kok, L., de Boer, R.J., and Schumacher, T.N. (2015). The Branching Point in Erythro-Myeloid Differentiation. Cell 163, 1655–1662.

Piras, V., Tomita, M., and Selvarajoo, K. (2014). Transcriptome-wide variability in single embryonic development cells. Sci. Rep. 4, 7137.

Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A., et al. (2015). Molecular Identity of Human Outer Radial Glia during Cortical Development. Cell 163, 55–67.

Puri, S., Folias, A.E., and Hebrok, M. (2015). Plasticity and Dedifferentiation within the Pancreas: Development, Homeostasis, and Disease. Cell Stem Cell 16, 18–31.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 135, 216–226.

Ricordi, C., Lacy, P.E., Finke, E.H., Olack, B.J., and Scharp, D.W. (1988). Automated method for isolation of human pancreatic islets. Diabetes 37, 413–420.

Ridden, S.J., Chang, H.H., Zygalakis, K.C., and MacArthur, B.D. (2015). Entropy, Ergodicity, and Stem Cell Multipotency. Phys. Rev. Lett. 115, 208103.

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

Sawada, M., Takahashi, K., Sawada, S., and Midorikawa, O. (1991). Selective killing of Paneth cells by intravenous administration of dithizone in rats. Int. J. Exp. Pathol. 72, 407–421.

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423, 623–656.

Solar, M., Cardalda, C., Houbracken, I., Martín, M., Maestro, M.A., De Medts, N., Xu, X., Grau, V., Heimberg, H., Bouwens, L., et al. (2009). Pancreatic exocrine duct cells give rise to insulin-producing beta cells during embryogenesis but not after birth. Dev. Cell 17, 849–860.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Statistical Methodol. 63, 411–423.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386.

Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371–375.

Wilson, N.K., Kent, D.G., Buettner, F., Shehata, M., Macaulay, I.C., Calero-Nieto, F.J., Sánchez Castillo, M., Oedekoven, C.A., Diamanti, E., Schulte, R., et al. (2015). Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. Cell Stem Cell 16, 712–724.

Yamamoto, R., Morita, Y., Ooehara, J., Hamanaka, S., Onodera, M., Rudolph, K.L., Ema, H., and Nakauchi, H. (2013). Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. Cell 154, 1112–1126.

Zeisel, A., Manchado, A.B.M., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–1142

2

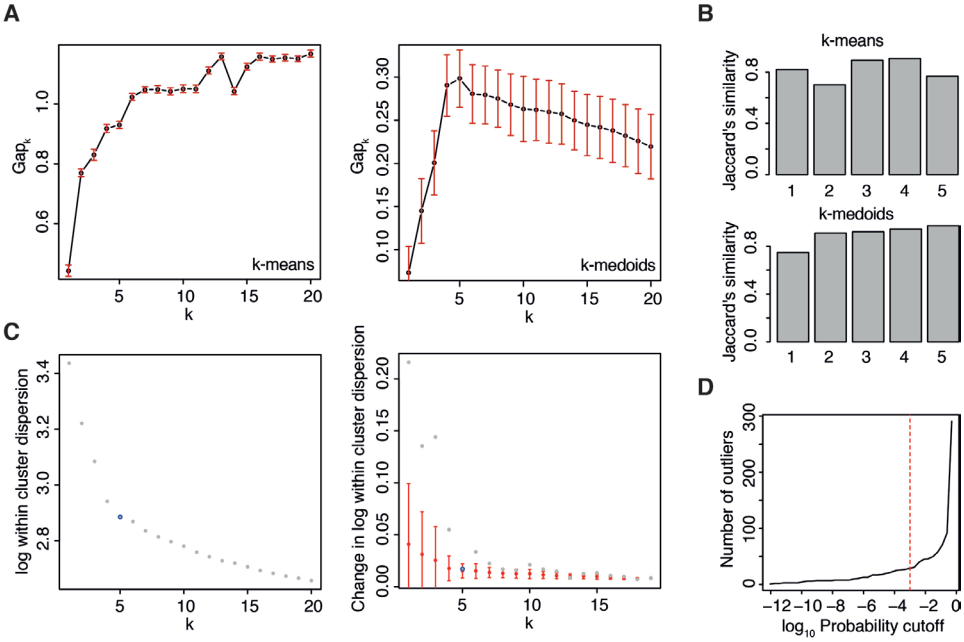

A
Gap$_k$
k-means
k-medoids
k
B
k-means
k-medoids
Jaccard's similarity
C
log within cluster dispersion
Change in log within cluster dispersion
D
Number of outliers
log$_{10}$ Probability cutoff

Figure S1. RaceID2 improves robustness of clustering. (Related to Figure 1)
(A) Gap statistitic (Tibshirani et al., 2001) computed with k-means clustering of the similarity matrix as in RaceID (left) and with k-medoids clustering using 1- pearson's correlation directly as clustering distance metric as in RaceID2 (right). (B) Jaccard's similarity computed by bootstrapping for k-means (upper panel) and k-medoids (lower panel) clustering with 5 clusters. K-medoids clustering shows higher reproducibility. (C) Criterion for the selection of the cluster number used for k-medoids clustering. If the change of the within-cluster dispersion (Tibshirani et al., 2001) upon increasing the cluster number ($k_{i+1} = k_i + 1$) is within the error of the average change upon further increase (k_{i+2}, …, k_{max}), k_i is chosen as input. The average change across cluster numbers k_{i+2}, …, k_{max} and its error is computed from a linear regression. The within-cluster dispersion as a function of k is shown on the left. The right panel shows the change of the within-cluster dispersion as a function of k and the average dispersion for higher values of k with error bars (red). In both panels the selected cluster number is circled in blue. (D) Outliers identification by RaceID2 is the same as in RaceID. Shown is the number of outliers as a function of the p-value cutoff. The red line indicates the cutoff chosen for this work ($P<10^{-3}$).

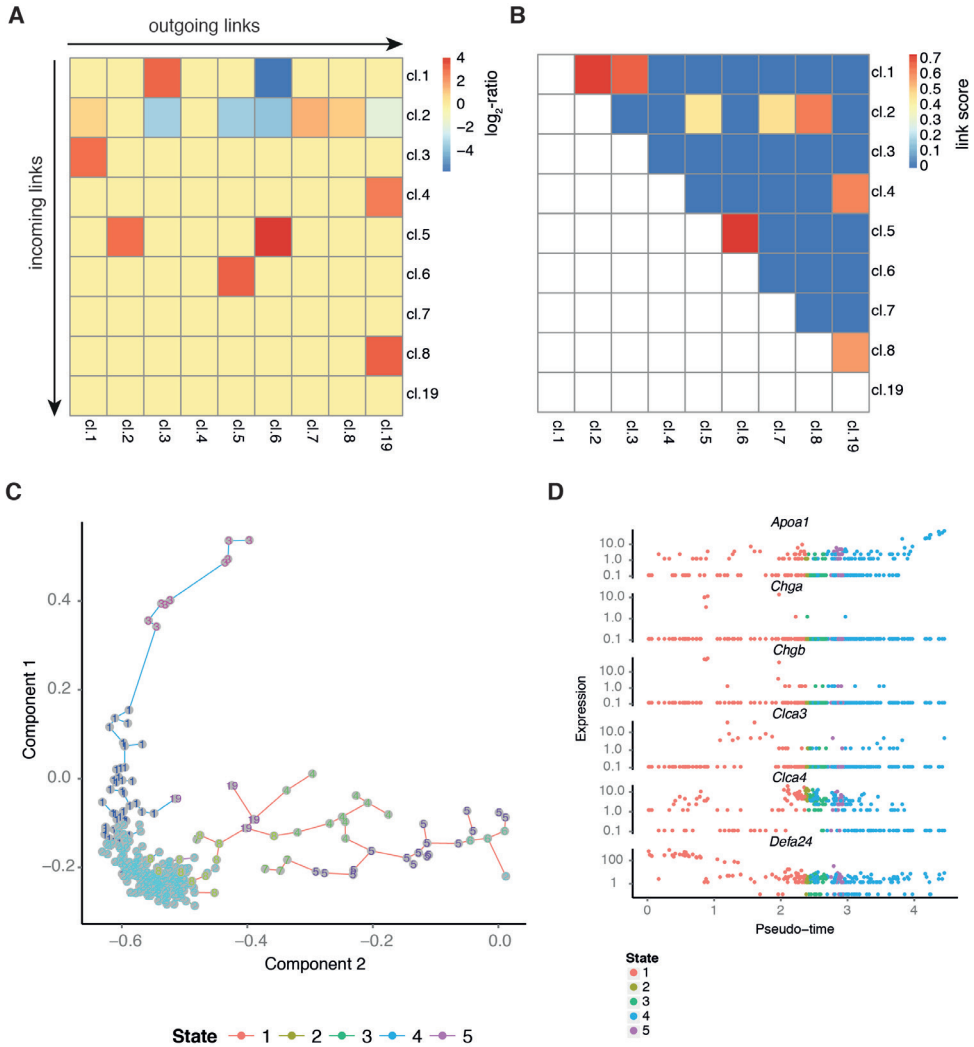

A
outgoing links
incoming links
log2-ratio
cl.1
cl.2
cl.3
cl.4
cl.5
cl.6
cl.7
cl.8
cl.19
B
link score
C
Component 1
Component 2
State 1 2 3 4 5
D
Apoa1
Chga
Chgb
Clca3
Clca4
Defa24
Expression
Pseudo−time
State
E
Clca3
Chga
Apoa1

Figure S2. Lineage inference by StemID and comparison to an alternative method for the derivation of differentiation trajectories does not resolve secretory intestinal cells. (Related to Figure 2) (A) The heatmap shows the log_2-ratio of the cell number assigned to each link between RaceID2 clusters and the expected number computed by a background model with randomized cell positions. Only significantly enriched or depleted links are highlighted. A log_2-ratio of zero is assigned to all other links. (B) The heatmap shows the link score for each pair of clusters, reflecting how densely a link between clusters is populated with cells (see Experimental Procedures). Values close to one indicate dense coverage, while values close to zero indicate that cells are concentrated near the centers of the clusters connected by the link. A higher value reflects a higher likelihood that the link represents an actual differentiation trajectory. (C-E) The Monocle (Trapnell et al., 2014) algorithm was run on the single cell transcriptomes of the 5 days Lgr5 lineage tracing data. (A) Minimum spanning tree computed by Monocle. Since 5 different cell types were observed in the data, Monocle was run with num_paths = 4. RaceID2 clusters were highlighted by numbers and colors used in Figure 1. (B) Expression of lineage markers (Apoe1: enterocytes; Chga: mature enteroendocrine cells; Chgb: early and mature enteroendocrine cells; Clca3: Goblet cells; Clca4: crypt bottom columnar cells; Defa24: Paneth cells) in cells assembled in pseudo-temporal order computed by Monocle. (C) Transcript counts (color legend) of mature lineage markers highlighted in the t-SNE map. RaceID2 clusters reliably discriminate different cell types (see Figure 2C). Monocle assigns stem, goblet, Paneth and enteroendocrine cells to one state and the inferred pseudo-temporal order does not reflect the published one shown in Figure 1A and inferred by StemID.

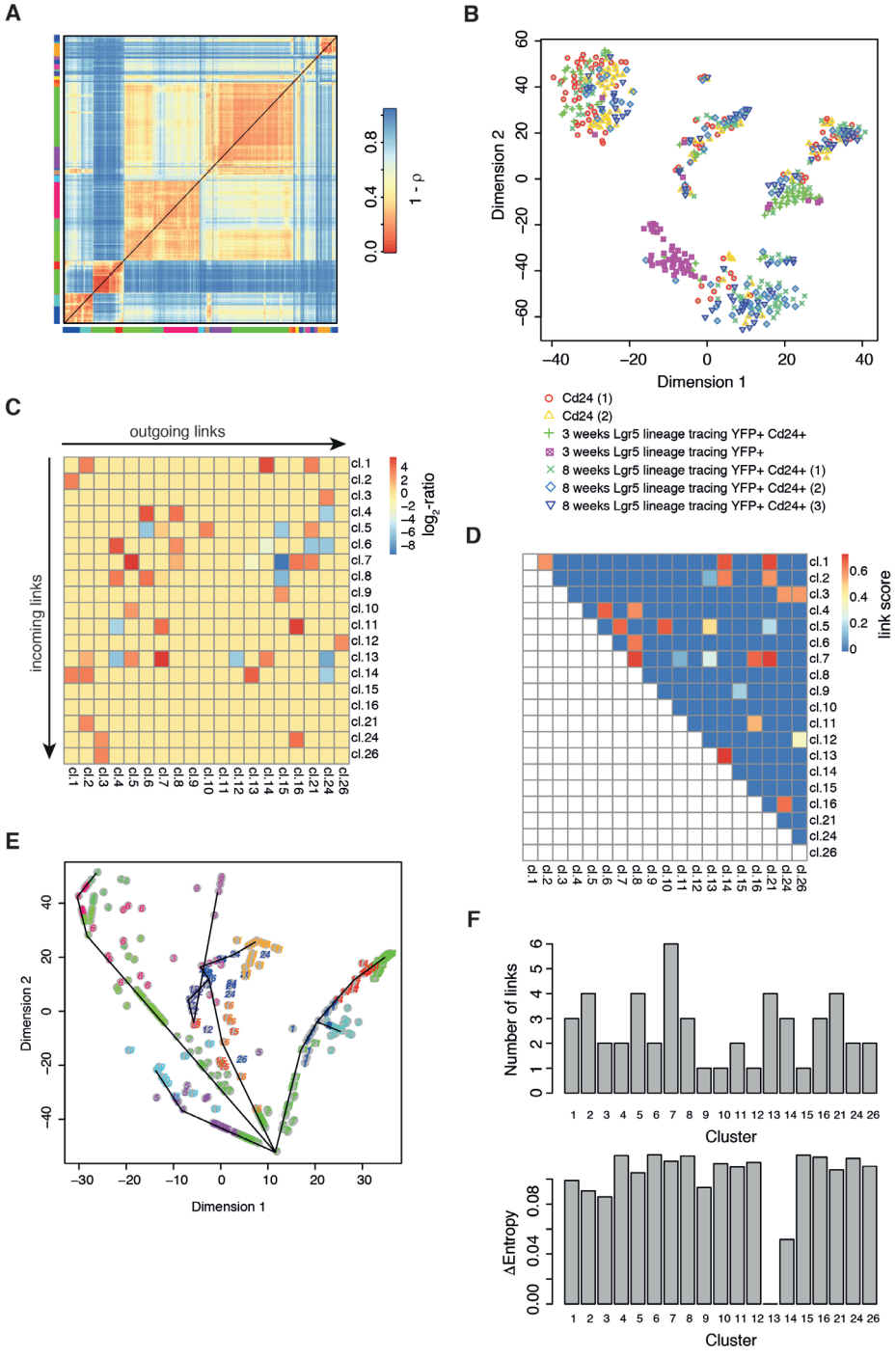

A
1 - ρ
B
Dimension 1
Dimension 2
Cd24 (1)
Cd24 (2)
3 weeks Lgr5 lineage tracing YFP+ Cd24+
3 weeks Lgr5 lineage tracing YFP+
8 weeks Lgr5 lineage tracing YFP+ Cd24+ (1)
8 weeks Lgr5 lineage tracing YFP+ Cd24+ (2)
8 weeks Lgr5 lineage tracing YFP+ Cd24+ (3)
C
outgoing links
incoming links
log2-ratio
D
link score
E
Dimension 1
Dimension 2
F
Number of links
Cluster
ΔEntropy
Cluster

Figure S3. StemID identifies stem cells in a complex intestinal dataset. (Related to Figure 3)
We ran RaceID2 and StemID on a dataset combining single mouse intestinal cell transcriptome data from a variety of experiments conducted in our lab, comprising Cd24-positive secretory cells, 3 weeks old progeny of Lgr5-positive cells and a sub-population of those positive for Cd24, and 8 weeks old Cd24-positive progeny of Lgr5-positive cells. (A) Heatmap of cell-to-cell transcriptome distances measured by 1 – Pearson's correlation (r) coefficient. RaceID2 cluster are color coded along the boundaries. (B) t-SNE map representation of transcriptome similarities between individual cells. Different experiments are highlighted with different colors and symbols. (C) The heatmap shows the $\log_2$-ratio of the cell number assigned to each link between RaceID2 clusters and the expected number computed by a background model with randomized cell positions. Only significantly enriched or depleted links are highlighted. A $\log_2$-ratio of zero is assigned to all other links. (D) The heatmap shows the link score for each pair of clusters, reflecting how densely a link between clusters is populated with cells (see Experimental Procedures). Values close to one indicate dense coverage, while values close to zero indicate that cells are concentrated near the centers of the clusters connected by the link. A higher value reflects a higher likelihood that the link represents an actual differentiation trajectory. (E) t-SNE map showing the projections of all cells as computed in a high dimensional space (see Experimental Procedures) in the embedded two-dimensional space. The black solid line indicates a minimum spanning tree connecting the cluster centers, which was computed based on the distances between cluster centers. The minimum spanning tree recovers the main differentiation trajectories, but does not identify a number of alternative trajectories revealed by the projection-based approach. (F) Barplot of the number of links (upper panel) and the Dentropy (lower panel). A comparison to the StemID score (Figure 3C) shows that neither of these quantities alone could rank the cell types by pluripotency with the same specificity as the StemID score.

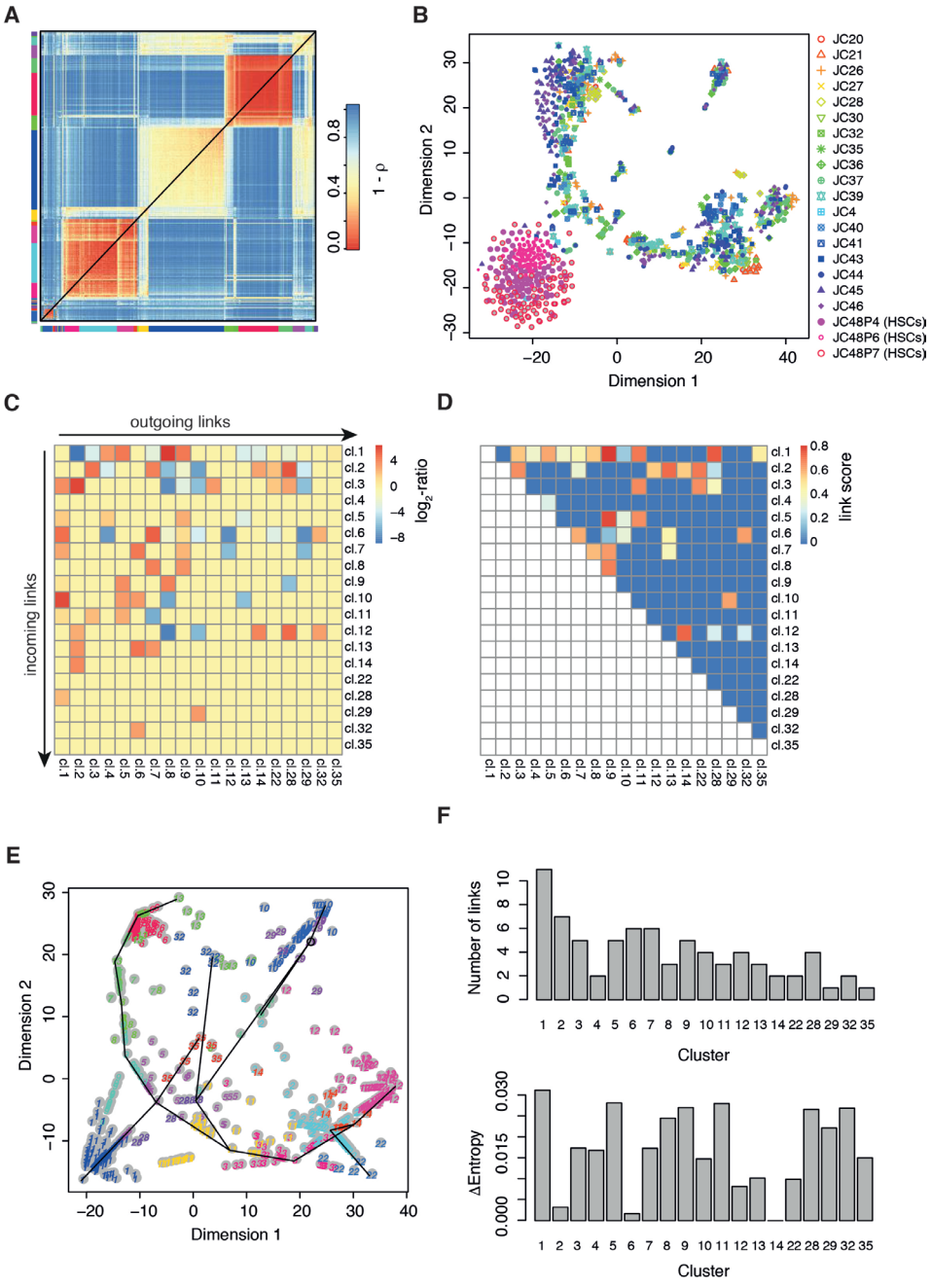

A
1 - ρ
0.0
0.4
0.8
B
Dimension 1
Dimension 2
JC20
JC21
JC26
JC27
JC28
JC30
JC32
JC35
JC36
JC37
JC39
JC4
JC40
JC41
JC43
JC44
JC45
JC46
JC48P4 (HSCs)
JC48P6 (HSCs)
JC48P7 (HSCs)
C
outgoing links
incoming links
log2-ratio
D
link score
E
Dimension 1
Dimension 2
F
Number of links
Cluster
ΔEntropy
Cluster

Figure S4. StemID identifies hematopoietic stem cells in single cells sequenced from the bone marrow. (Related to Figure 4)
We ran RaceID2 and StemID on a single cell sequencing dataset comprising mouse bone marrow cells manually isolated from interacting doublets or multiplets of cells and Kit^+ $Sca\text{-}1^+$ Lin^- $CD48^-$ $CD150^+$ hematopoietic stem cells (HSCs). (A) Heatmap of cell-to-cell transcriptome distances measured by 1 – Pearson's correlation (r) coefficient. RaceID2 cluster are color coded along the boundaries. (B) t-SNE map representation of transcriptome similarities between individual cells. Different experiments are highlighted with different colors and symbols. (C) The heatmap shows the $\log_2$-ratio of the cell number assigned to each link between RaceID2 clusters and the expected number computed by a background model with randomized cell positions. Only significantly enriched or depleted links are highlighted. A $\log_2$-ratio of zero is assigned to all other links. (D) The heatmap shows the link score for each pair of clusters, reflecting how densely a link between clusters is populated with cells (see Experimental Procedures). Values close to one indicate dense coverage, while values close to zero indicate that cells are concentrated near the centers of the clusters connected by the link. A higher value reflects a higher likelihood that the link represents an actual differentiation trajectory. (E) t-SNE map showing the projections of all cells as computed in a high dimensional space (see Experimental Procedures) in the embedded two-dimensional space. The black solid line indicates a minimum spanning tree connecting the cluster centers, which was computed based on the distances between cluster centers. The minimum spanning tree recovers the main differentiation trajectories, but does not identify a number of alternative trajectories revealed by the projection-based approach. (F) Barplot of the number of links (upper panel) and the Dentropy (lower panel). A comparison to the StemID score (Figure 4C) shows that neither of these quantities alone could rank the cell types by pluripotency with the same specificity as the StemID score.

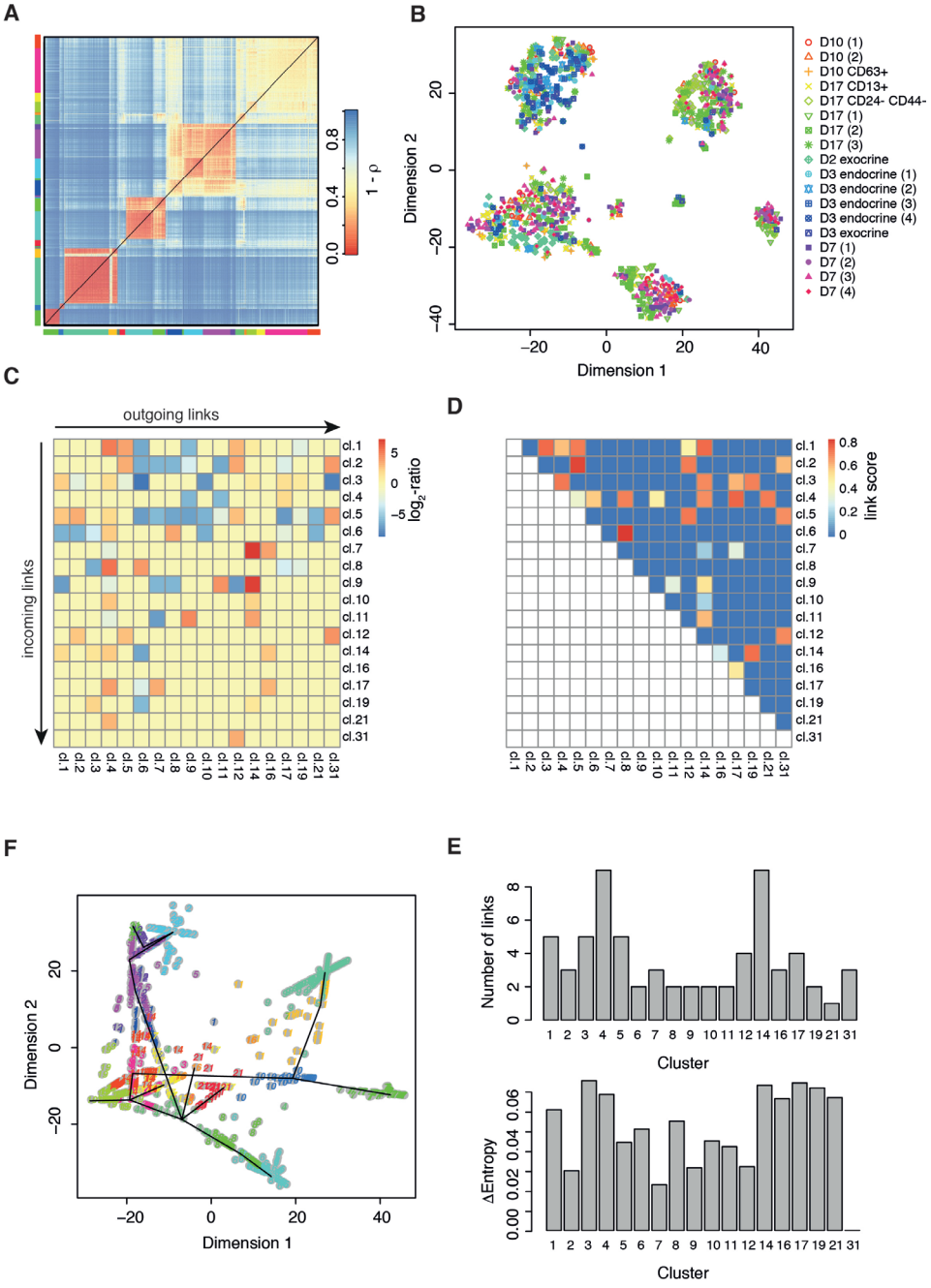

A
1 - ρ
0.0
0.4
0.8
B
Dimension 1
Dimension 2
D10 (1)
D10 (2)
D10 CD63+
D17 CD13+
D17 CD24- CD44-
D17 (1)
D17 (2)
D17 (3)
D2 exocrine
D3 endocrine (1)
D3 endocrine (2)
D3 endocrine (3)
D3 endocrine (4)
D3 exocrine
D7 (1)
D7 (2)
D7 (3)
D7 (4)
C
outgoing links
incoming links
log2-ratio
D
link score
F
Dimension 1
Dimension 2
E
Number of links
Cluster
ΔEntropy
Cluster

Figure S5. StemID predicts pluripotent cells in random mixtures of human pancreatic cells. (Related to Figure 6)

We ran RaceID2 and StemID on a single cell sequencing dataset comprising single human pancreatic cells isolated form five different donor (D2, D3, D7, D10, D17). Different enrichment strategies were applied to collect random mixture, endocrine and exocrine cells, or subsets of those. (A) Heatmap of cell-to-cell transcriptome distances measured by 1 – Pearson's correlation coefficient (r). RaceID2 cluster are color coded along the boundaries. (B) t-SNE map representation of transcriptome similarities between individual cells. Different experiments are highlighted with different colors and symbols. (C) The heatmap shows the $\log_2$-ratio of the cell number assigned to each link between RaceID2 clusters and the expected number computed by a background model with randomized cell positions. Only significantly enriched or depleted links are highlighted. A $\log_2$-ratio of zero is assigned to all other links. (D) The heatmap shows the link score for each pair of clusters, reflecting how densely a link between clusters is populated with cells (see Experimental Procedures). Values close to one indicate dense coverage, while values close to zero indicate that cells are concentrated near the centers of the clusters connected by the link. A higher value reflects a higher likelihood that the link represents an actual differentiation trajectory. (E) t-SNE map showing the projections of all cells as computed in a high dimensional space (see Experimental Procedures) in the embedded two-dimensional space. The black solid line indicates a minimum spanning tree connecting the cluster centers, which was computed based on the distances between cluster centers. The minimum spanning tree recovers the main differentiation trajectories, but does not identify a number of alternative trajectories revealed by the projection-based approach. (F) Barplot of the number of links (upper panel) and the Dentropy (lower panel). A comparison to the StemID score (Figure 6C) shows that neither of these quantities alone could rank the cell types by pluripotency with the same specificity as the StemID score.

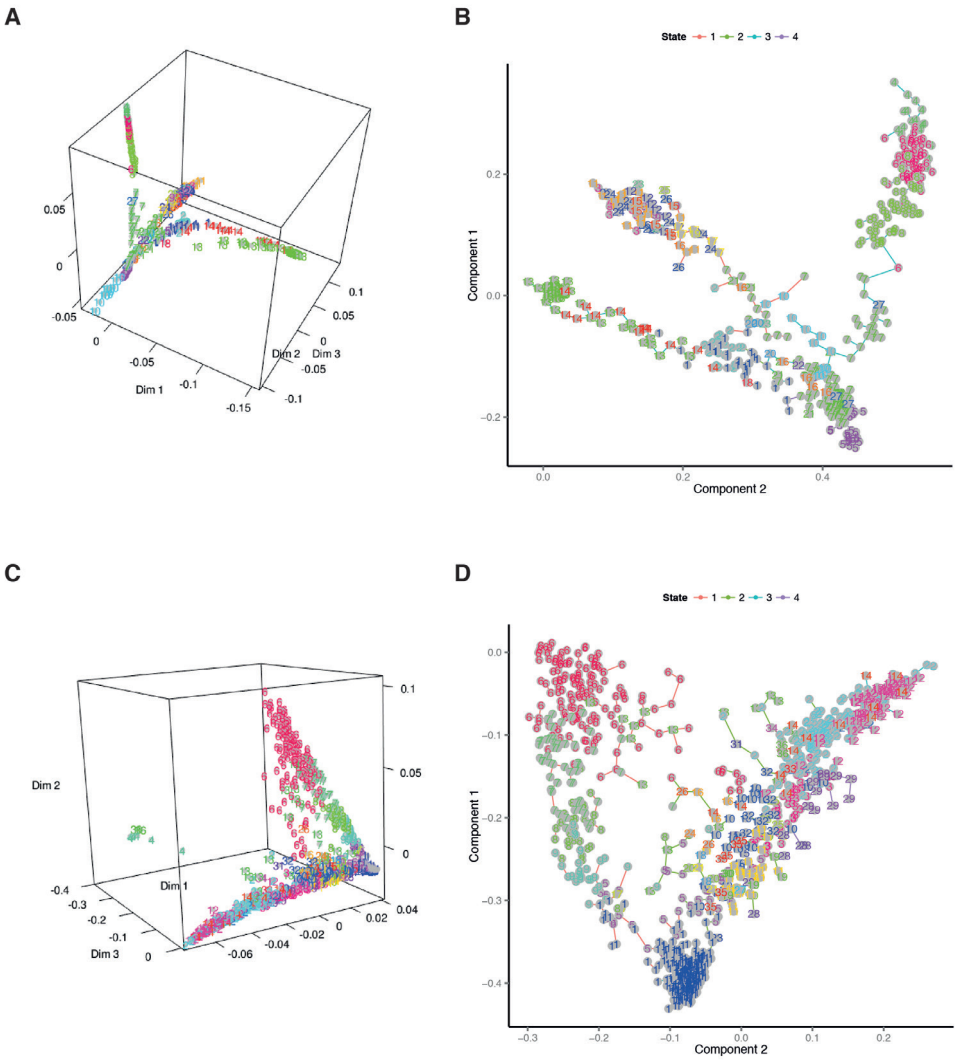

A
0.05
0
-0.05
Dim 1
0
-0.05
-0.1
-0.15
Dim 2
0.1
0.05
0
Dim 3
-0.05
-0.1
B
State 1 2 3 4
Component 1
0.2
0.0
−0.2
0.0
0.2
0.4
Component 2
C
Dim 2
0.1
0.05
0
Dim 1
-0.4
-0.3
-0.2
-0.1
Dim 3
0
-0.06
-0.04
-0.02
0
0.02
0.04
D
State 1 2 3 4
Component 1
0.0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
Component 2

Figure S6. StemID provides novel information in comparison to published methods. (Related to Figure 3 and 4)
For the complex intestinal data set (Fig. 3) and the bone marrow data (Fig. 4) we derived a lineage tree with two previously published methods. On the one hand we used Monocle (Trapnell et al., 2014), which constructs a minimum spanning tree connecting all cells based on transcriptome similarity, and on the other hand we applied a recent method based on diffusion maps (Haghverdi et al., 2015). Results of Monocle and diffusion maps are shown in (A) and (B) for the intestinal data and in (C) and (D) for the bone marrow data. For the intestinal data (A, B) both methods reveal major branches (Paneth/goblet cells, tuft cells, enterocytes, compare to Figure 3 for colors and cluster labels). However, the small clusters of different enteroendocrine cells could not be assembled onto a branched tree by any method. Moreover, none of the methods reveals that Paneth and goblet cells have a common precursor, but rather place mature Clca3 expressing goblet cells on the same branch with mature Paneth cells. Monocle does not recover the relation between TA cells and mature enterocytes. Crucially, none of these methods provides a cell type inference and a prediction of the stem cell identity. For both methods, it is not apparent from the topology that cluster 7 represents the stem cell identity.
For the bone marrow data (C, D) both methods recover the major branches of neutrophils and erythroblasts, but intermingle the low frequency cell types with myeloid precursors.

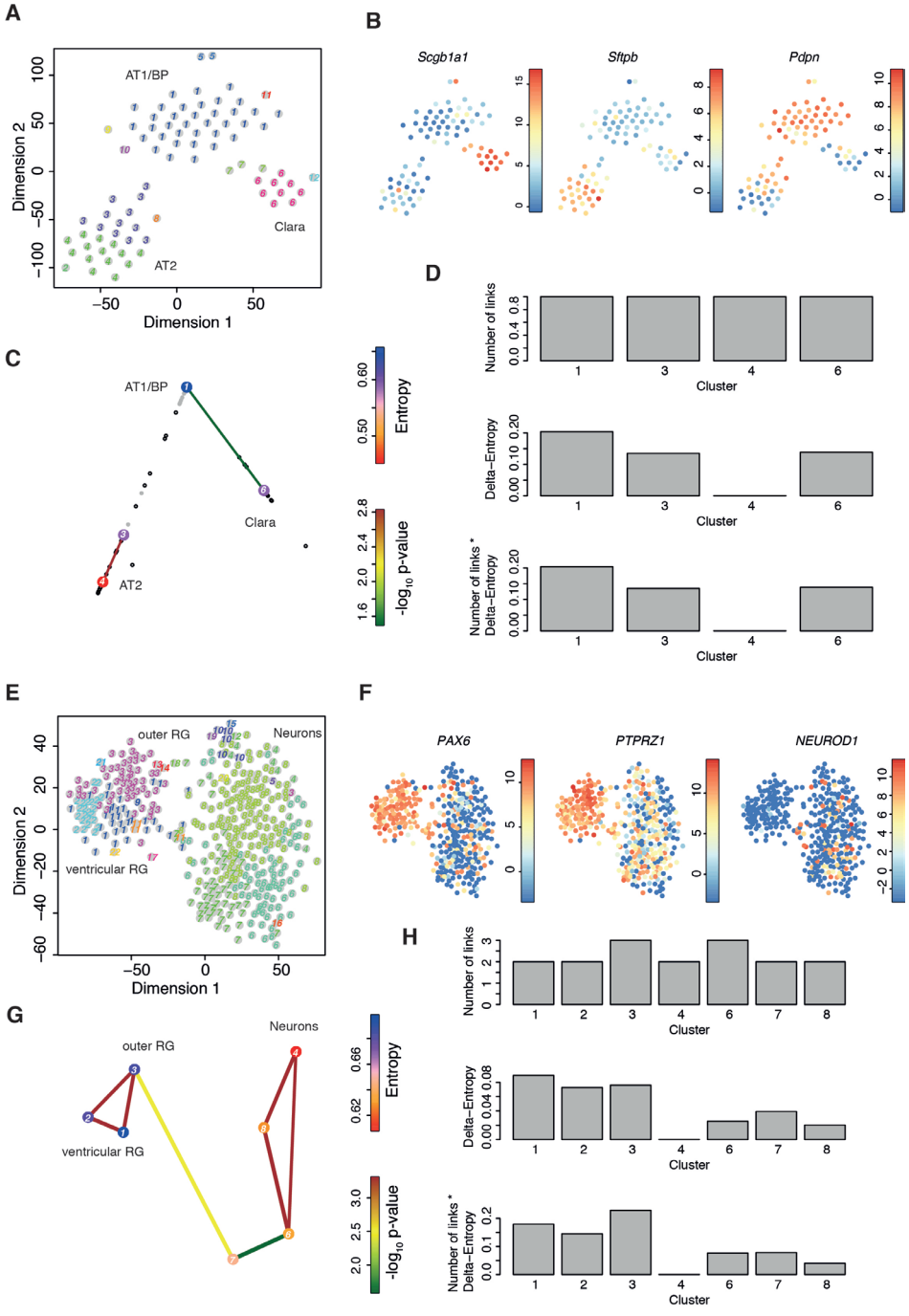
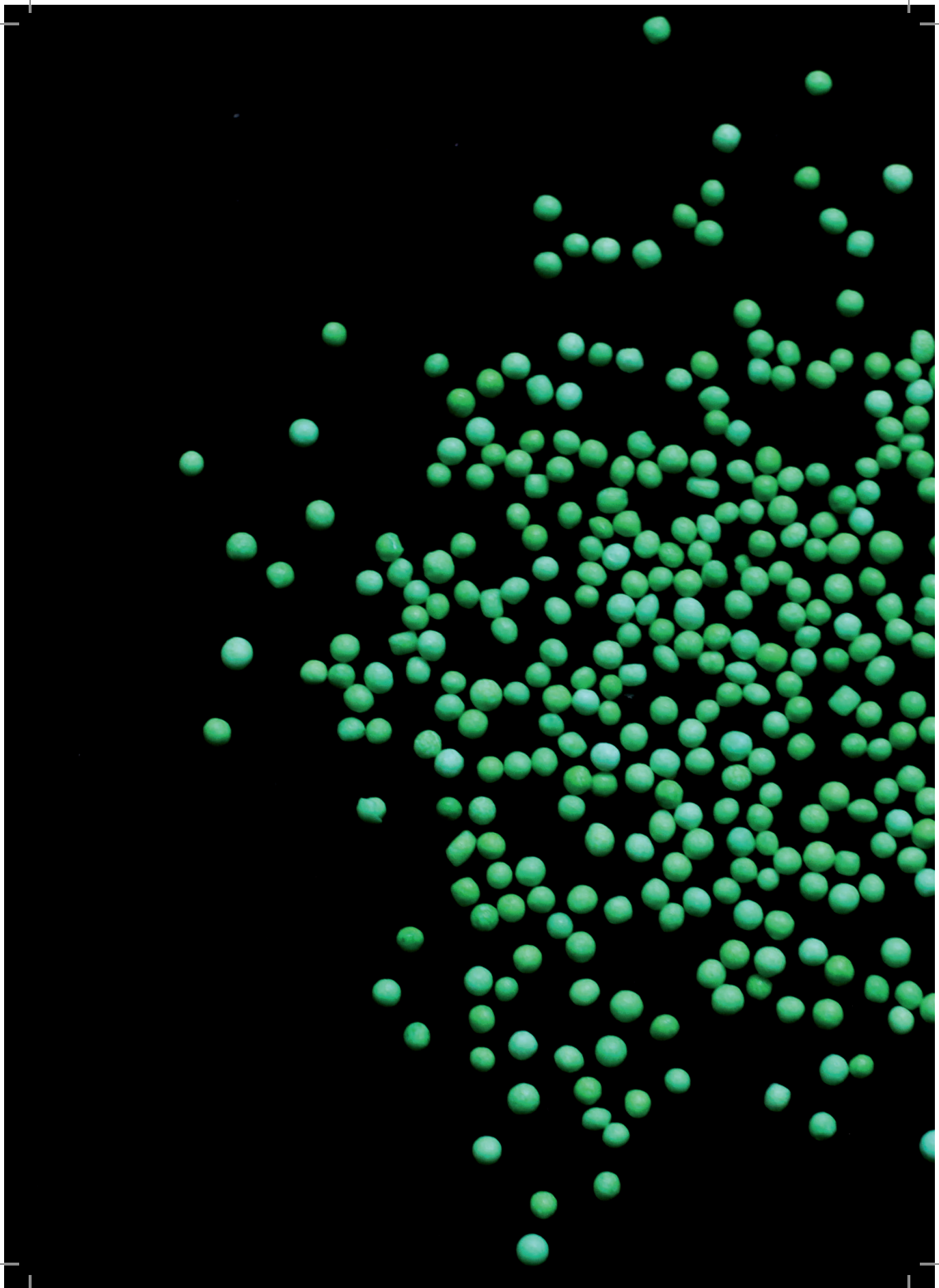

A
AT1/BP
Clara
AT2
Dimension 1
Dimension 2
B
Scgb1a1
Sftpb
Pdpn
C
AT1/BP
Clara
AT2
Entropy
-log10 p-value
D
Number of links
Delta−Entropy
Number of links * Delta−Entropy
Cluster
E
outer RG
Neurons
ventricular RG
Dimension 1
Dimension 2
F
PAX6
PTPRZ1
NEUROD1
G
outer RG
Neurons
ventricular RG
Entropy
-log10 p-value
H
Number of links
Delta−Entropy
Number of links * Delta−Entropy
Cluster

Figure S7. StemID predicts the stem cell identity for previously published data sets. (Related to Figure 3 and 4)
To test StemID on additional published datasets we searched the literature for single cell profiling of stem cell differentiation systems. We could not find suitable unique molecular identifier (UMI) based data and therefore applied StemID to read based data for the developing lung epithelium (Treutlein et al., 2014) and for developing radial glia cells (Pollen et al., 2015). Although our algorithm was not designed for read based quantification, StemID could infer correct lineage trees and correctly predict the stem cell identity in both systems. (A-D) StemID on 80 cells extracted from mouse lung epithelium at E18.5 (Treutlein et al., 2014). (A) t-SNE map showing the major populations inferred by RaceID2. Clusters are highlighted with different numbers and colors. Alveolar type 1 (AT1) and bipotential progenitors (BP) clustered together (cluster 1). Since our outlier identification is designed for UMI based quantification these subtypes remained unresolved. The other major groups correspond to Clara cells and alveolar type 2 (AT2) cells. (B) Expression of population specific markers (Treutlein et al., 2014) was highlighted in t-SNE maps on a logarithmic (log_2) scale (color legend). (C) Inferred intestinal lineage tree. Only significant links are shown ($P<0.05$). The color of the link indicates the $-log_{10}$p-value. The color of the vertices indicates the entropy. Cells are shown in the background as grey dots. A black circle indicates a significant projection component. From these cells an additional link between cluster 1 and clusters 3 and 4 can be recognized, which is marginally significant ($P\sim0.06$). (D) Barplot of StemID scores. The BP/AT1 cluster acquires the highest StemID score. With the additional marginal link the difference between cluster 1 and the other clusters would be even larger. (E-H) StemID on 393 cells from the ventricular and subventricular zone of the human cortex at gestational week 16-18 (Pollen et al., 2015). (E) t-SNE map showing the major populations inferred by RaceID2. Clusters are highlighted with different numbers and colors. Clusters 1,2 and 3 represent radial glia cells while 4,6,7,8 represent intermediate progenitors and mature neurons. (F) t-SNE map highlighting expression of radial glia markers (PAX6, PTPRZ1) and an early neuronal marker (NEUROD1) on a logarithmic (log_2) scale (color legend). Up-regulation of PTPRZ1 identifies cluster 3 as outer and cluster 1 and 2 as ventricular radial glia (RG) cells. (C) Inferred cortical lineage tree. Only significant links are shown ($P<0.05$). The color of the link indicates the $-log_{10}$p-value. The color of the vertices indicates the entropy. The thickness indicates the link score reflecting how densely a link is covered with cells (see Experimental procedure). The tree links the RG sub-types to the mature neurons (cluster 4 and 8) via a NEUROD1 expressing progenitor population (D) Barplot of StemID scores. The highest score was correctly assigned to outer RG cells, which have been shown to express self-renewal pathways (as opposed to ventricular RG cells) and differentiate into various neural and glial cell types.
For (B-D) and (F-H) only clusters with >5 cells were analyzed.

Supplemental information

supplemental tables, methods and references can be found online at: http://dx.doi.org/10.1016/j.stem.2016.05.010.



3

A Single-cell Transcriptome Atlas of the Human Pancreas

Mauro J Muraro*, Gitanjali Dharmadhikari*, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon van Gurp, Marten A. Engelse, Francoise Carlotti, Eelco J.P. de Koning, and Alexander van Oudenaarden

*Equal contribution

Adapted from Cell Systems. 2016. 3(4):385-394.e3

ABSTRACT
To understand organ function it is important to have an inventory of its cell types and of their corresponding marker genes. This is a particularly challenging task for human tissues like the pancreas, since reliable markers are limited. Hence, transcriptome-wide studies are typically done on pooled islets of Langerhans, which obscures contributions from rare cell types and of potential subpopulations. To overcome this challenge, we developed an automated platform that uses FACS, robotics, and the CEL-Seq2 protocol to obtain the transcriptomes of thousands of single pancreatic cells from deceased organ donors, allowing in silico purification of all main pancreatic cell types. We identify cell type-specific transcription factors and a subpopulation of REG3A-positive acinar cells, and show that CD24 and TM4SF4 expression can be used to sort live alpha and beta cells with high purity. This resource will be useful for developing a deeper understanding of pancreatic biology and pathophysiology of diabetes mellitus.

INTRODUCTION
Most organs consist of a variety of cell types with interdependent functions. To understand organ function and disease, genome-wide information on each cell type is crucial. Studies on pooled material detect global gene expression patterns, but represent an average dominated by the most abundant cell types. With the advent of single-cell transcriptomics it is possible to determine the transcriptome of individual cells, allowing the identification of cell types in an unbiased manner (Grün and van Oudenaarden, 2015; Kolodziejczyk et al., 2015; Trapnell, 2015; Wang and Navin, 2015). Initial single-cell transcriptomics studies were performed on cultured cells (Deng et al., 2014; Hashimshony et al., 2012; Islam et al., 2011; Klein et al., 2015; Shalek et al., 2013; Tang et al., 2010). Subsequent studies described cell types in the mouse lung (Treutlein et al., 2014), spleen (Jaitin et al., 2014), brain(Zeisel et al., 2015), retina (Macosko et al., 2015), small intestine (Grün et al., 2015) and pancreas (Xin et al., 2016). Studies on human tissue have so far been limited to fetal neurons (Johnson et al., 2015), glioblastomas (Patel et al., 2014) and a small set of human pancreatic cells (Li et al., 2016). We and others have recently used single-cell sequencing of the human pancreas to reveal subpopulations of cells that show potential as progenitors (Grün et al., 2016; Wang et al., 2016). These studies described manually processed samples and/or low numbers of cell, which limited the number of detected genes. Here, we developed a more efficient, high-throughput method to sequence primary human cells of all pancreatic cell types.
The pancreas functions as an exocrine and endocrine gland. The exocrine compartment consists of acinar cells producing digestive enzymes and ductal cells forming channels that drain into the duodenum. The endocrine compartment consists of alpha, beta, delta, PP and epsilon cells that are found in the islets of Langerhans. Insulin-producing beta cells and glucagon-producing alpha cells play a major role in glucose homeostasis and islet dysfunction is the hallmark of diabetes mellitus, a chronic metabolic disorder affecting approximately 9% of people worldwide (WHO, 2014). Functional analysis and genetic profiling is typically performed on whole islets , which masks the contribution of individual cell types to pancreas biology and disease (Bugliani et al., 2013; Cnop et al., 2014; Eizirik et al.,

2012). To study heterogeneity and classify subpopulations within known cell types single-cell resolution is essential. We developed a high throughput approach for single-cell sequencing based on the CEL-Seq2 protocol (Hashimshony et al., 2016) to create a single-cell transcriptome atlas of the human pancreas. Our method implements fluorescence activated cell sorting (FACS), which allows the user to work with low amounts of starting material. This dataset provides an unbiased view of cell types in the human pancreas at single-cell resolution, enabling comparison of gene expression patterns between cell types and detection of subpopulations within them. This resource can be mined for genes involved in pancreatic function to define novel therapeutic targets for diseases such as diabetes mellitus.

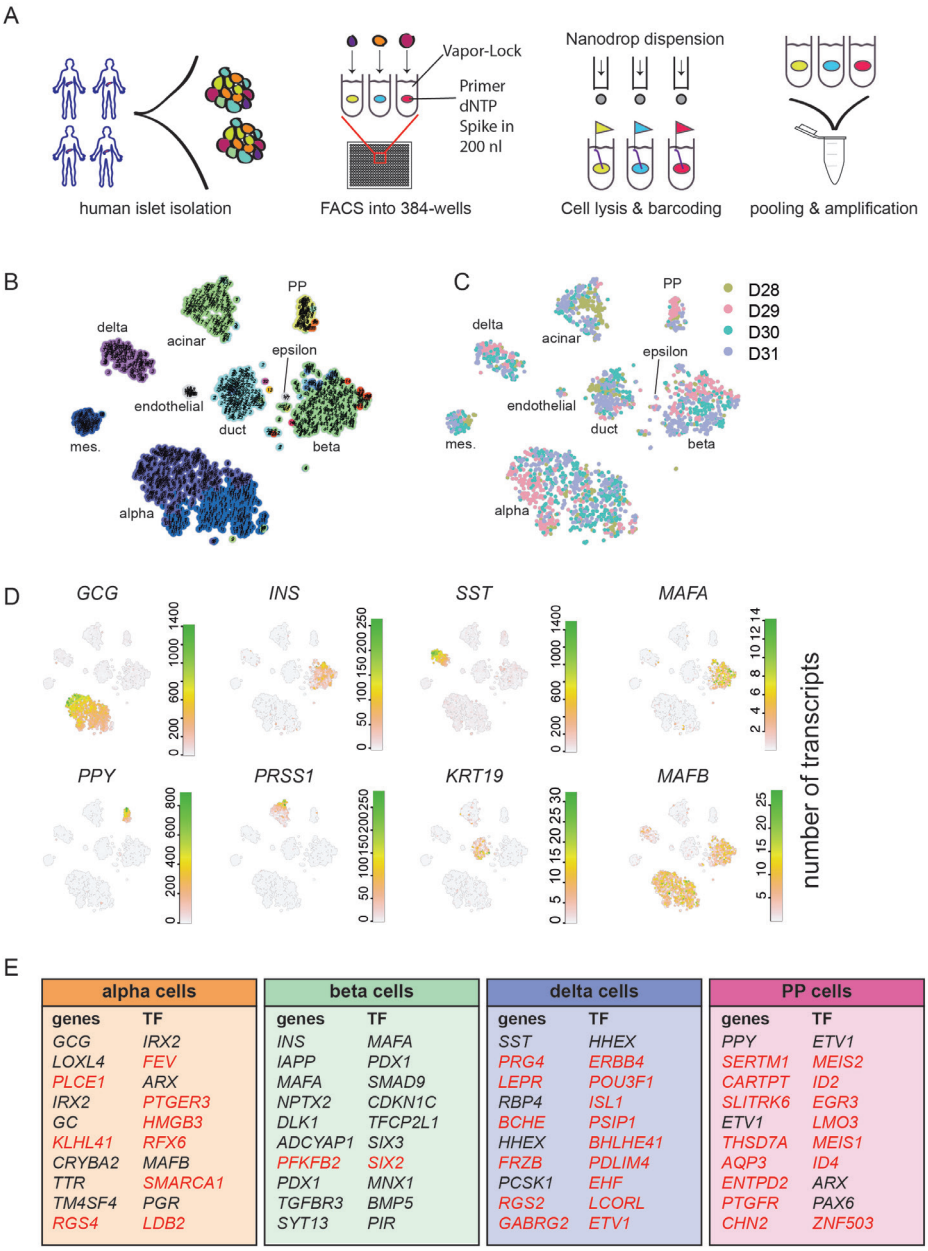
RESULTS

SORT-seq allows for deep sequencing of human pancreas cells.

To assay the transcriptomes of the various human pancreatic cell types, we obtained human pancreas material from four deceased organ donors (Figure 1A). Isolation of the Islets of Langerhans yielded 55-95% islet purity (Table S1). The non-islet cells in these preparations mainly consisted of exocrine cells. After a culture period of 3-5 days, the islets were dispersed for FACS followed by single-cell sequencing. Previously, we sorted cells from 5 different donors, which were processed manually by CEL-Seq as described (Grün et al., 2016). These yielded an average of 4262 unique transcripts and a median of 1958 detected genes per cell (Figure S1A and S1B). While useful to determine interesting progenitor cells and describe general differences between cell types, this dataset lacked the depth to fully describe the transcriptome of each pancreatic cell type. For example, comparing expression across endocrine cell types resulted in low numbers of differentially expressed genes (Figure S1C).

To more efficiently capture single-cell transcriptomes, we used FACS and robotics liquid handling to perform automated single-cell sequencing based on the CEL-Seq2 protocol (Hashimshony et al., 2016). We refer to this platform as SORT-Seq (Sorting and Robot-assisted Transcriptome Sequencing, Figure 1A). Briefly, live single cells (based on DAPI and scatter properties) are sorted into 384-well plates with 5 µl of Vapor-Lock oil containing a droplet of 100 nl of CEL-Seq primers, Spike-ins and dNTPs. For cDNA construction, cells are first lysed by heat, after which a robotic liquid handler dispenses RT and second strand mix. Cells are then pooled and the aqueous phase is extracted from the oil. The CEL-Seq2 protocol can be followed from this point onwards. Compared to the manual method, the amount of reads that could be mapped to the reference transcriptome increased from 15% to 45%. Additionally, the number of unique transcripts per cell also increased (median of 14604 compared to 4262, Figure S1D), as did the number of genes detected per cell (median of 4497 compared to 1958, Figure S1E). This resulted in more complex single-cell libraries with more differentially expressed genes between cell types (Figure S1F)

To investigate if we could detect the expected pancreatic cell types we used StemID, an approach we developed for inferring the existence of stem cell populations from single-cell transcriptomics data (Grün et al., 2016). StemID calculates all pairwise cell-to-cell distances (1 – Pearson correlation) and uses this to cluster similar cells



E

alpha cells		beta cells		delta cells		PP cells	
genes	TF	genes	TF	genes	TF	genes	TF
GCG	*IRX2*	*INS*	*MAFA*	*SST*	*HHEX*	*PPY*	*ETV1*
LOXL4	*FEV*	*IAPP*	*PDX1*	*PRG4*	*ERBB4*	*SERTM1*	*MEIS2*
PLCE1	*ARX*	*MAFA*	*SMAD9*	*LEPR*	*POU3F1*	*CARTPT*	*ID2*
IRX2	*PTGER3*	*NPTX2*	*CDKN1C*	*RBP4*	*ISL1*	*SLITRK6*	*EGR3*
GC	*HMGB3*	*DLK1*	*TFCP2L1*	*BCHE*	*PSIP1*	*ETV1*	*LMO3*
KLHL41	*RFX6*	*ADCYAP1*	*SIX3*	*HHEX*	*BHLHE41*	*THSD7A*	*MEIS1*
CRYBA2	*MAFB*	*PFKFB2*	*SIX2*	*FRZB*	*PDLIM4*	*AQP3*	*ID4*
TTR	*SMARCA1*	*PDX1*	*MNX1*	*PCSK1*	*EHF*	*ENTPD2*	*ARX*
TM4SF4	*PGR*	*TGFBR3*	*BMP5*	*RGS2*	*LCORL*	*PTGFR*	*PAX6*
RGS4	*LDB2*	*SYT13*	*PIR*	*GABRG2*	*ETV1*	*CHN2*	*ZNF503*

Figure 1. SORT-seq allows for deep sequencing of human pancreas cells.
(A) Experimental workflow for SORT-seq. Islets were isolated from human donors. Cells were dispersed and sorted into 384-wells plates with mineral oil, containing 100nl of CEL-Seq2 primers, dNTPs and Spike-ins. RT mix was then distributed by the Nanodrop II. After second strand synthesis, material was pooled and amplified prior to RNA library preparation.
(B) Visualization of k-medoid clustering and cell-to-cell distances using t-distributed stochastic neighbor embedding (t-SNE). Each dot represents a single cell. Colors and numbers indicate clusters and cell type names are indicated with their corresponding cluster(s).
(C) t-SNE map highlighting donor source. Each color represents one donor.
(D) t-SNE maps highlighting the expression of marker genes for each of the six main pancreatic cell types. Transcript counts are given in linear scale.
(E) Tables denoting the top 10 differentially expressed genes and transcription factors (TF) when comparing one cell type to all other cells in the dataset ($P<10^{-6}$). Genes whose cell type specificity was previously unknown in the human pancreas are marked in red.

into clusters that correspond to the cell types present in the tissue (Figure S1G). This resulted in well-separated cell clusters with low intra- and high inter-cluster cell-to-cell distances as visualized in t-SNE maps (Figure 1B) (Maaten and Hinton, 2008). These maps were also used to highlight expression of specific genes across all cells (Figure 1D). To test if the donor source influenced cluster formation, we plot donor contribution to the clusters in Figure 1C, showing that none of the clusters consists of cells from only one donor. When we compared all cells from each cell type of one donor to that of all others, we did not find major differences between donors. Most significantly differentially expressed genes differ by less than two fold. As expected, XIST was upregulated in all cell types of D30 (Table S2), the only female donor of the set. The donor independent clustering shows StemID groups cells based on cell type, rather than donor.

We found the clusters to highly express markers for all pancreatic cell types (Figure 1D). We found cluster specific expression of GCG (alpha cells), INS (beta), SST (delta), PPY (PP), PRSS1 (acinar), KRT19 (duct) and COL1A1 (mesenchyme) (Figure 1D and S1H). Since we did not detect clusters with either epsilon or endothelial cells we looked for expression of the markers GHRL or ESAM. Indeed, we found two clusters of cells exclusively expressing these markers, and manually annotated them as epsilon and endothelial cells (Figure 1B).

Notably, we also detected the expression of MAFA and MAFB, transcriptional regulators important for determining the identity of endocrine cell types (Nishimura et al., 2008) (Figure 1D). MAFA expression is restricted to beta cells while MAFB expression is found in both alpha and beta cells, as previously reported in mouse (Dai et al., 2012).

We next set out to generate a resource with which to compare pancreatic cell types and mine their transcriptomes for interesting genes. To this end, we compared all alpha cells (clusters expressing high GCG), beta, epsilon, delta, PP, duct, acinar, mesenchymal, and endothelial cells based on their distinct clustering from other cell types. Each group of cells was compared to all other cell groups, yielding a list of differentially expressed genes. The top 10 of each list can be found in Figures 1E and Figure S1I, and the full list in Table S3. We then selected only those genes that have been reported to function as transcription factors by using the TFcheckpoint database (Chawla et al., 2013) (Table S4). Several genes and transcription factors found here have never been reported as markers for specific cell types of the human pancreas (Figure 1E).

Apart from the classically known alpha cell transcription factors IRX2, ARX, (Dorrell et al., 2011a) and PGR (Doglioni et al., 1990), our analysis reveals transcription factors FEV, PTGER3 (Kimple et al., 2013) SMARCA1 (Rankin and Kushner, 2010) HMGB3 and RFX6 (Piccand et al., 2014) that to our knowledge have not been reported to be enriched in alpha cells and have been previously implicated in beta cell function. Some of these factors have broader expression across other endocrine cell types, such as RFX6, but are most highly expressed in alpha cells.

Classical beta cell markers like INS, MAFA and PDX1 (Kulkarni, 2004) top the beta cell list, and we detect PFKFB2 (Arden et al., 2008), a gene thought to regulate insulin secretion, and the transcription factor SIX2. To our knowledge, neither PFKFB2 nor SIX2 have been reported previously in beta cells. SIX2 is known to interact with the transcription factor TCF7L2 (Xu et al., 2014), a well-known SNP for type 2 diabetes

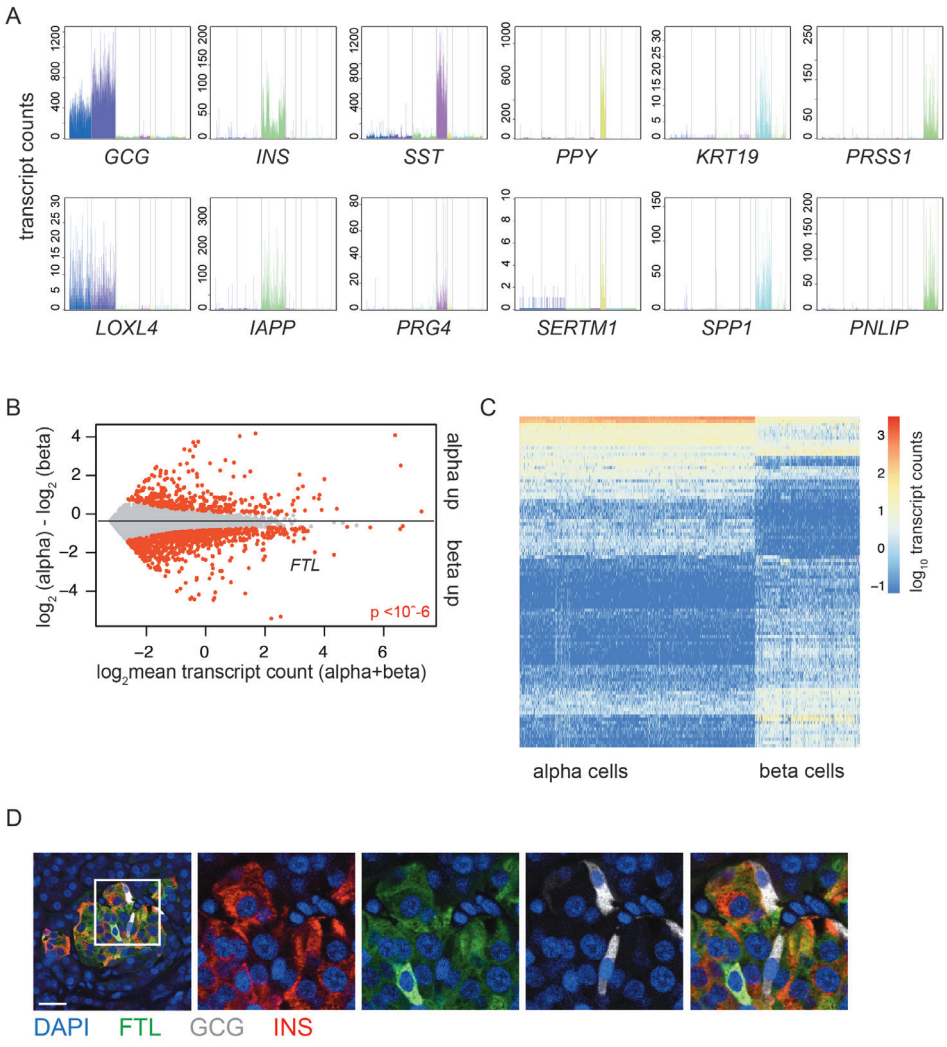


Figure 2. Cluster-restricted gene expression patterns and identification of cell-type specific genes
(A) Expression of well-known marker genes (top) and the most differentially expressed gene (bottom) in each of the six main pancreatic cell types. If the most differentially expressed gene is also a canonical marker gene, the next most differentially expressed gene is shown. Gene expression values are plotted on the Y-axis. Each bar represents a cell and cells are grouped by cluster with a specific color in the following order: alpha, beta, delta, PP, duct and acinar.
(B) Differential gene expression analysis between alpha and beta cells reveals 1376 differentially expressed genes. Grey dots indicate genes; red dots indicate significant genes ($P<10^{-6}$). Y-axis indicates log2 fold change, the X-axis the mean transcript count over both groups of cells.
(C) Heat map of the top 100 differentially expressed genes between alpha and beta cells. Rows are genes, columns are cells. Log2 expression of transcript counts for genes is plotted. Columns are ordered based on cell type (alpha left, beta right). Genes are grouped based on hierarchical clustering.
(D) Immunohistochemistry for ferritin light subunit (FTL, green) glucagon (GCG, gray) and insulin (INS, red) with counterstaining for DAPI (blue) on human pancreatic tissue sections. Scale bar is 25 μM.

(Grant et al., 2006). This makes it interesting for further investigations in the context of beta cell function.

Apart from the classical SST and HHEX expression in delta cells (Zhang et al., 2014) genes like LEPR and GHSR imply a possible role of leptin and ghrelin on delta cell function. PP cells have substantial expression of genes related to neuronal cells, which hints towards the developmental proximity of PP and neuronal cells. This has been previously described by others in the context of beta cells (Arntfield and van der Kooy, 2011; Le Roith et al., 1982)

In summary, these gene lists confirm markers and reveal gene expression patterns in the endocrine cell types that can be further investigated for their roles in cellular identity and function.

Cluster-restricted gene expression patterns and identification of cell-type specific genes

We next analyzed each cluster in detail to see if the remaining differentially expressed genes corroborated the initial identification of the six major pancreatic cell types. To investigate to what extent gene expression patterns are shared between cell types, we focused on the expression of both the top differentially expressed genes and classical marker genes (Figure 2A). Especially the expression of hormones was restricted to individual clusters, taking up one fifth of the transcriptome, while being near-zero in other clusters. For most clusters, the top differentially expressed genes were documented markers (Table S3). For example, INS and IAPP were co-expressed in beta cells, LOXL4 with GCG (alpha cells) and PNLIP with PRSS1. PRG4 was most highly expressed in delta cells after SST. Ductal markers SPP1 and KRT19 were relatively lowly expressed but limited to the ductal cluster. Further inspection of the top differentially expressed genes per cluster yielded new cell type-specific genes, such as ALDH1A1, which was enriched in alpha cells and co-expressed with GCG (Figure S2C, S2D).

Going further down the list of differentially expressed genes continued to show cell type-restricted patterns (Figure S2A). To test if we could use StemID clustering to compare different types of cells, we determined differentially expressed genes between all endocrine and exocrine cells. This yielded 2858 genes that were differentially expressed ($P < 10^{-6}$ after Benjamini-Hochberg correction). Clear separation of endocrine and exocrine was visible by plotting the top 100 differentially expressed genes (Figure S2B). This list consisted of many genes related to endocrine function, proving single-cell sequencing yields useful data on specific pancreatic cell types. This allowed us to continue exploring differences between more closely related cell types such as alpha and beta cells, which yielded a list of 1376 differentially expressed genes ($P<10^{-6}$) (Figure 2B).

Plotting these differences in expression patterns showed clear cell type-specific patterns (Figure 2C). Not surprisingly,canonical marker genes for alpha and beta cells (GCG, MAFA, IAPP, CHGB, INS, INS-IGF2, SCG2, PCSK1 and PCSK2). were in the list, as were genes found in studies that analyzed enriched populations of alpha or beta cells, such as TTR, which is specific in mouse alpha cells (Dorrell et al., 2011b), NPTX2 in beta cells (Figure S2A) (Nica et al., 2013) and GC in human alpha cells (Ackermann et al., 2016). We also identified several previously unreported cell type-specific genes for both alpha (CRYBA2, TM4SF4, ALDH1A1) and beta

cells (ID1, RBP4, SQSTM1, MT1X, FTL, FTH1) (Ackermann et al., 2016; Benner et al., 2014). Interestingly, many of these beta cell-specific genes have been linked to T2D or to oxidative and/or ER stress responses (Åkerfeldt and Laybutt, 2011; Chen et al., 2001; Orino et al., 2001; Yang et al., 2005). To validate our results, we visualized protein levels of FTL and ALDH1A1 in tissue sections of human pancreas. Ftl expression was visible in insulin producing cells and absent from GCG positive alpha cells in islets of Langerhans (Figure 2D). Aldh1a1 expression appeared to be quite similar in acinar cells and alpha cells, whereas in general higher mRNA expression was observed in alpha cells (Figure S2C and S2D). Within the Islets of Langerhans, we detected Aldh1a1 expression only in glucagon positive alpha cells, but not in other cells in the islets.

GO-term analysis reveals cell-type specific gene expression patterns relevant to diabetes and glucose metabolism.

We used EnrichR (Chen et al., 2013) to perform GO-term analysis on the full list of genes differentially expressed in each cell type compared to all other pancreatic cell types. We determined the top 15 enriched GO terms for alpha, beta, delta and PP cells (Figure S3A). Additionally, we provide the lists of GO terms for each type, along with the genes that are involved in this GO term (Table S5). Parsing the file for alpha cell-related GO terms, shows the inositol receptor ITPR1 to be involved in insulin secretion. ITPR1 has previously been associated with a diabetic phenotype in mice (Ye et al., 2011), (Figure S3C). GO-terms, like negative regulation of nervous system development are found highest in PP-cells, indicating these cells have a more neuronal nature. The serotonin transporter SLC6A4 is found predominantly in PP cells (Figure S3C) and has a well-documented role in neurons and in behavior (Murphy and Lesch, 2008). To focus on differences between cell types in more detail, we performed GO-term analysis on gene sets obtained after comparing beta cells to alpha, delta and PP cells separately (Figure S3B). In particular, delta cells show more hits in behavior and synaptic transmission. The ghrelin receptor GHSR is involved in several of these processes. This receptor is only present in delta cells (Figure S3C), indicating a role for ghrelin in delta cell function, which has been recently demonstrated in mice (Digruccio et al., 2016). These results are an example of how genes obtained in this resource can be used to do GO-term analysis. By zooming in on specific genes from interesting terms, we can generate hypotheses regarding cell-type specific processes in the human pancreas.

Outlier identification reveals heterogeneity within acinar and beta cells

We set out to analyze cellular heterogeneity by detecting outliers within specific populations of cells using the RaceID algorithm (Grün et al., 2015).. The most striking results were found in beta and acinar cells, where we found subpopulations of cells with distinct gene expression patterns. In beta cells, the most significant genes dictating this heterogeneity were SRXN1, SQSTM1 and three ferritin subunits FTH1P3, FTH1 and FTL (Figure 3A, Figure S4A). All these genes were highly expressed in cluster 2 (Figure S4A) and are implicated in response to endoplasmic reticulum and oxidative stress (Orino et al., 2001; Zhou et al., 2015; Rantanen et al., 2013). The main acinar cluster split into four clusters, of which cluster 2 showed high levels of REG3A expression (Figure 3C, 3D and S4B), while the acinar marker PRSS1 was expressed in all clusters, but highest in a group of cells in cluster 3 and 4 (Figure S4C).

To confirm the existence of subpopulations of REG3A positive acinar cells, we stained sections of human pancreas for Reg3a and Prss1. Scattered individual Reg3a/Prss1 double positive cells were observed (Figure S4D) in acinar tissue. More interestingly, we also detected large clusters of brightly Reg3a/Prss1 positive acinar cells close to Islets of Langerhans (Figure 3E).
To characterize subpopulations obtained in silico in more depth, we averaged the expression profiles of all single cells belonging to the different subpopulations. By averaging and pooling the transcriptomes from these groups of cells, we achieve transcriptome coverage more similar to bulk sequencing experiments (Table S6).
In summary, we detected subpopulations of beta cells expressing higher levels of FTH1 and validated acinar subpopulations expressing high levels of REG3A. This subtype of acinar cell merits more investigation, since the role of REG3A in pancreas biology is unclear.

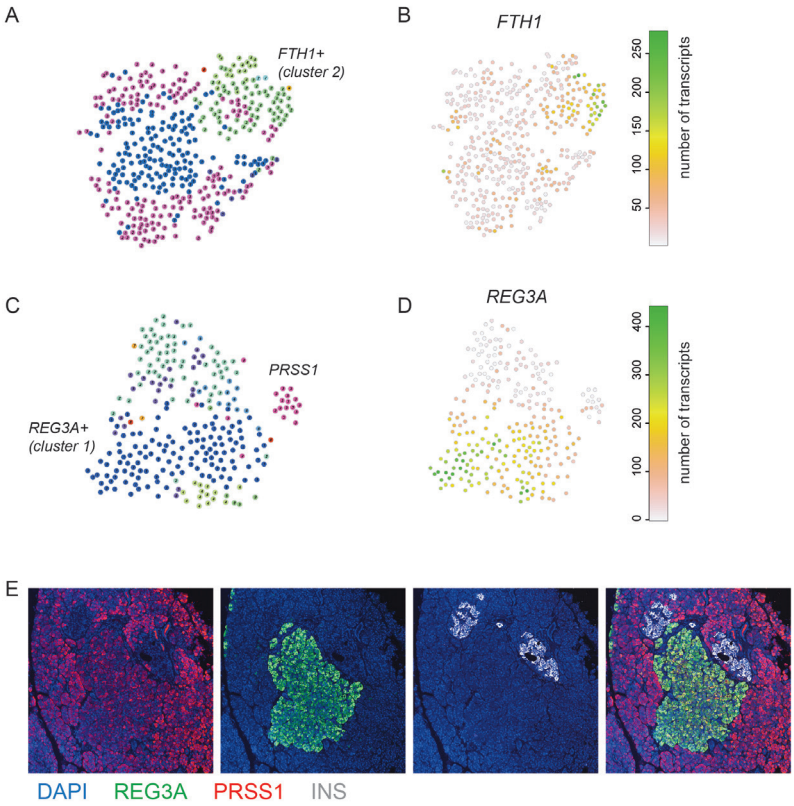


Figure 3. Outlier identification reveals heterogeneity within acinar and beta cells.
(A) t-SNE map of RaceID clusters after clustering of only beta cells.
(B) t-SNE map highlighting the expression of FTH1.
(C) t-SNE map of RaceID clusters after clustering of only acinar cells.
(D) t-SNE map highlighting the expression of REG3A.
(E) Immunohistochemistry for reg3A (green), trypsin (red) and insulin (gray) with counterstaining for DAPI (blue). Scale bar is 75 μM.

Enrichment of alpha and beta cells based on cell-surface markers

We next mined our transcriptome resource for novel cell-surface markers to enrich specific pancreatic cell types using live-cell cell sorting. As a proof of principle, we set out to deplete the exocrine fraction from islet isolations of low purity. We found cell surface markers CD24 and CD44 were restricted to acinar and ductal clusters (Figure 4A). Next, we prepared six FACS sorted libraries, two with only live cells, and four with negative selection for CD24 and CD44 (Figure 4B). This yielded compact clusters of cells that corresponding to the main pancreatic cell types (Figure S5B). Nearly all endocrine cells were derived from the negatively selected libraries (Figure S5A), demonstrating the efficiency of the predicted cell surface markers. Notably, alpha cells seemed to be preferentially enriched with this strategy (Figure S5A).

To test if we could enrich for one pancreatic cell type, we explored alpha cell cell surface markers, finding TM4SF4, a tetraspanin family member that has been linked to pancreatic development (Anderson et al., 2011) and is specifically expressed in alpha cells, with lower expression in PP cells (Figure 4C). To verify the membrane-localized expression of TM4SF4 in alpha cells we performed imaging flow cytometry analysis on fixed cells that were co-stained with either glucagon or insulin and Tm4sf4 antibodies. We found Tm4sf4 to be localized at the membrane of alpha cells, but not of beta cells (Figure 4D). To test if this antibody can be used to enrich for alpha cells we processed 8 libraries from an endocrine-rich islet extraction (Table S1): four libraries were composed of live cells, two were CD24- / TM4SF4+ and two were CD24- / TM4SF4-. We found the main endocrine pancreatic cell types after clustering (Figure S5C). Libraries sorted for Tm4sf4 consisted of >85% alpha cells. When selecting against Tm4sf4 and CD24, alpha cells were depleted from the resulting population and enrichment for beta cells became possible with similar purity (Figure 4F).

In conclusion, this shows that our resource can be used to mine for genes with a specific subcellular location in a pancreatic cell type of choice. Table S7 provides a list of cell-type enriched cell-surface markers in each of the main pancreatic cell types.

DISCUSSION

Scarcity of material, lack of reliable cell-surface markers and analysis of pooled populations of cells often hamper analysis of human organ cell type composition. Most importantly methods relying on pooled cells average gene expression profiles over thousands of cells, masking any heterogeneity to be found within one cell type and potentially missing interesting intermediate cell types. To overcome these challenges we have sequenced single cells from donor pancreata from four different donors using SORT-Seq, a FACS-compatible, automated version of the CEL-Seq2 protocol.. We readily detected several clusters corresponding to the canonical pancreatic cell types, allowing us to purify cell-types in silico for further analysis. Due to consideration for transplantation, the islets obtained for this study were cultured for 3-5 days prior to dispersion to single cells and FACS. It should be noted that culture conditions might affect the varied pancreatic cell types differently (progenitor cells are more likely to be affected than terminally differentiated cell types). However, shorter culture times for human islets are difficult to achieve, and

we could not detect any major biases between donors, irrespective of their culture times. It is also important to keep in mind that since the efficiency of single-cell sequencing (especially when using manual TRIzol-based methods) is of the order of 10% (Grün et al., 2014), lowly expressed genes are detected only sporadically. However, sequencing many cells enabled us to detect transcription factors, rare cell types, and to detect heterogeneity within canonical pancreatic cell types such as acinar and beta cells. To further test the predictive power of this resource, we describe a panel of cell-surface markers that were specifically expressed in exocrine or alpha cells. Using these markers we were able to enrich for endocrine cells, alpha and beta cells.
In conclusion, we present this dataset as a resource that can be used to study pancreas composition and function with single-cell resolution. We envision broad applicability of this single-cell transcriptome atlas of the human pancreas to improve our understanding of pancreas biology and diabetes research.

Author Contributions
M.J.M., G.D., E.J.P.d.K and A.v.O. conceived the project. M.J.M. and G.D. carried out experiments. M.A.E. supervised the human islet isolation procedure.
D.G. helped with StemID. N.G., T.D., E.J., L.vG. and F.C. aided with experiments. M.J.M, G.D. and A.v.O analyzed the data. M.J.M., G.D., E.J.P.d.K and A.v.O wrote the manuscript.

Acknowledgments
We thank Tamar Hashimshony and Itai Yanai for sharing CEL-Seq2. We thank USF for sequencing, Anko de Graaf for help with microscopy, and Reinier van der Linden for FACS. Many thanks to Nicola Crosetto for ideas on automation of CEL-Seq. This work was supported by a ERC Advanced grant (ERC-AdG 294325-GeneNoiseControl), NWO VICI awards and grants from Stichting DON, the Dutch Diabetes Research Foundation and the JDRF.

Methods
Experimental model details
Human cadaveric donor pancreata were procured through a multiorgan donor program. Pancreatic tissue was only used if the pancreas could not be used for clinical pancreas or islet transplantation, only if research consent was given and according to national laws. In total, 4 human donor pancreata were procured (3 male, 1 female). See Supplemental table 1 for details on donor age, sex and BMI.

Human islet isolation, dispersion and sorting
Human islet isolations from pancreatic tissue were performed in the islet isolation facility of the Leiden University Medical Center according to a modified protocol originally described by Ricordi et al. (Ricordi et al., 1988). Islets were cultured in CMRL 1066 medium (5.5 mM glucose) (Mediatech) supplemented with 10% human serum, 20 µg/ml ciprofloxacin, 50 µg/ml gentamycin, 2 mM L-glutamin, 0.25 µg/ml fungizone, 10 mM HEPES and 1.2 mg/ml nicotinamide for 3-6 days. Islets were maintained in culture at 37°C in a 5% CO_2 humidified atmosphere. Medium was refreshed the day after isolation and every 2-3 days thereafter until cell sorting. The

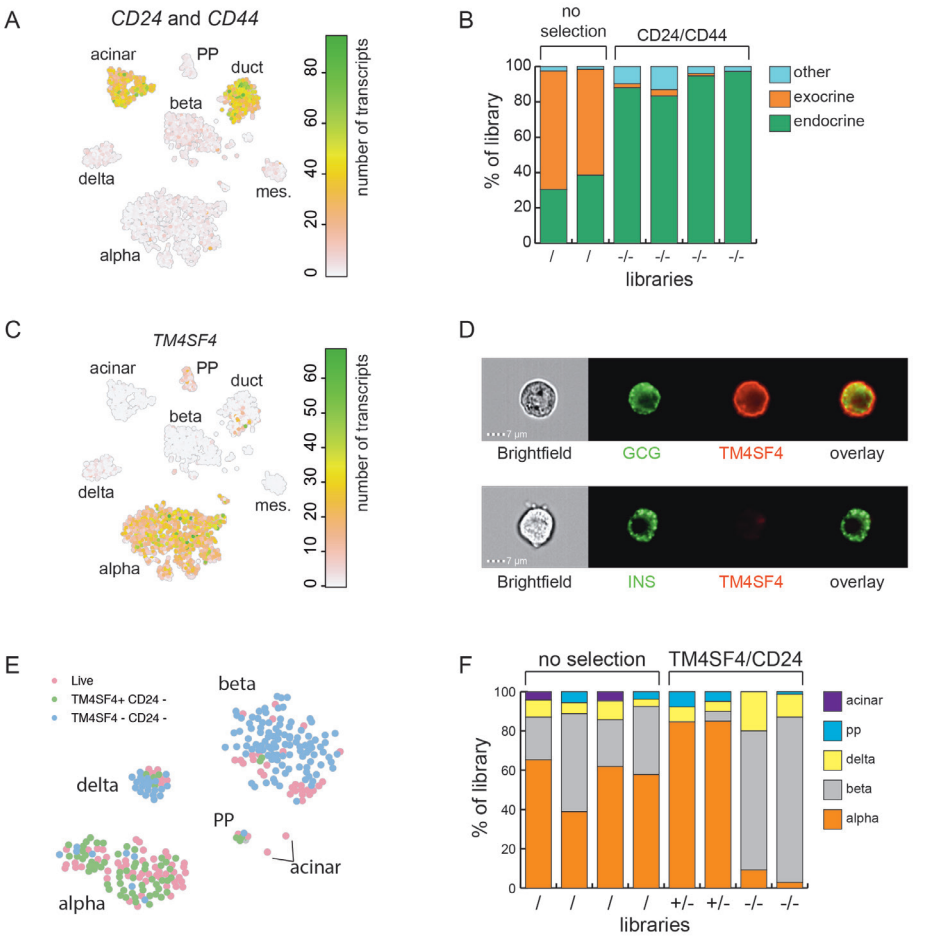


Figure 4. Enrichment of alpha and beta cells based on cell-surface markers
(A) t-SNE map highlighting the combined expression of CD24 and CD44.
(B) Results of FACS enrichment based on selection against CD24 / CD44. Two libraries were selected for live staining (/), four against CD24 and CD44 expression (-/-). Y-axis indicates the portion of the library consisting of a particular cell type. Colors indicate cell types.
(C) t-SNE map highlighting the expression of TM4SF4.
(D) Imagestream analysis of dispersed, fixed single-cells from human pancreas. Left panel shows a bright field image of the cell, then immunostaining against glucagon (green) and Tm4sf4 (red). Lower panel shows Insulin in green.
(E) t-SNE map highlighting libraries from a Tm4sf4 / Cd24 sort. Cells that were not stained are in pink. Cells sorted for Tm4sf4+ / Cd24- are in green, Tm4sf4- / Cd24-in blue.
(F) Results of FACS enrichment based on selection for Tm4sf4/ Cd24. Four libraries were selected for live staining (/). Two libraries were Tm4sf4+ / Cd24- (+/-) and two were Tm4sf4- / Cd24- (-/-). Y-axis indicates the portion of the library consisting of a particular cell type. Colors indicate cell types.

islets were cultured for 3.-5 days after islet isolation. Culture time depended on the decision time needed for considering islets for transplantation and FACS.
For cell sorting cultured Islets were briefly washed in cold PBS. The islet pellet was then suspended in 1 ml of Accutase (Stemcell technologies) per 5000 islet equivalents and incubated at 37 degrees with gentle intermittent shaking for 8-10 minutes until the islets were dispersed into single cells. The digestion process was stopped using an excess volume of cold RPMI medium containing 10% FCS. The dispersed tissue was washed briefly with cold PBS followed by filtering through a sieve to get rid of any debris and undigested material. DAPI was added to access the viability of the cells.. The tissue was stored on ice until sorting using a FACS Aria II or FACSJazz (BD biosciences). Live single cells (based on DAPI exclusion and forward/side scatter properties) were sorted into 384-well hard shell plates (Biorad) with 5 µl of vapor-lock (Qiagen) containing 100-200 nl of RT primers, dNTPs and synthetic mRNA Spike-Ins and immediately spun down and frozen to -80˚C. For cells sorted on cell surface markers; filtered, dispersed cells were incubated with FITC-CD24 (BD, 560992), PE-CD44 (Cell signaling, 8724S) and/or APC-TM4SF4 (BD, FAB7998A) antibodies for 30 min post dispersion on ice, followed by brief washing and sorting as above.

Single-cell mRNA sequencing of single cells
For SORT-seq, cells were lysed by 5 minutes at 65˚C, after which RT and second strand mixes were dispersed with the Nanodrop II liquid handling platform (GC biotech). Aqueous phase was separated from the oil phase after pooling all cells in one library, followed by IVT transcription. The CEL-Seq2 protocol was used for library prep. Primers consisted of a 24 bp polyT stretch, a 4bp random molecular barcode (UMI), a cell-specific 8bp barcode, the 5' Illumina TruSeq small RNA kit adapter and a T7 promoter. mRNA of each cell was then reverse transcribed, converted to double-stranded cDNA, pooled and in vitro transcribed for linear following the CEL-Seq 2 protocol (Hashimshony et al, 2016). Illumina sequencing libraries were then prepared with the TruSeq small RNA primers (Illumina) and sequenced paired-end at 75 bp read length the Illumina NextSeq.

Immunofluorescence and Imaging Flow cytometry
Pancreatic tissue samples were fixed overnight in 4% formaldehyde (Klinipath), stored in 70% ethanol, and subsequently embedded in paraffin. Sections were deparaffinized in xylene and rehydrated in a series of ethanol, followed by heat assisted antigen retrieval in citric buffer (pH 6.0). Sections were blocked by incubating with 2% normal donkey serum and 1% lamb serum in PBS. Primary antibodies included rabbit anti-Ftl (ab69090), mouse anti-Glucagon (ab10988) and guinea pig anti-Insulin (ab7842), mouse anti-trypsin-1 (sc-137077), rabbit anti-Reg3a (ab134309) and rabbit anti-Aldh1a1 (ab23375). Sections were incubated in with primary antibody in PBS/1% lamb serum at 4˚C overnight. Alexa Fluor 488-, 568- and 647- conjugated secondary antibodies against rabbit, mouse and guinea pig IgG as appropiate (Life Technologies A11008, A10037 and A21450) were diluted 1:200 and incubated at room temperature for 1 hour. Nuclear counterstaining was done with DAPI and by additionally embedding with DAPI vectashield (Vector Laboratories #H-1500). Imaging was done on a Leica SP8 confocal microscope using hybrid detectors.

TM4SF4 staining on alpha versus the beta cells was performed on fixed, stained single cells from dispersed human islets. Dispersed Islet cells were fixed with 4%PFA and washed using 2% FCS/PBS, followed by permeabilization using Perm/Wash buffer from BD Cytofix/Cytoperm Fixation/Permeabilization Kit (Cat. 554717) 15 minutes at room temperature. The samples were incubated with antibodies diluted in Perm/Wash buffer rabbit anti glucagon (1:200) or guinea pig anti insulin (1:200) or anti TM4SF4-APC (1:50) for 30 minutes at room temperature. Alexa Fluor 488- conjugated secondary antibodies (directly or in biotin-streptavidin system) against rabbit, and guinea pig as appropriate (Life Technologies A11008) were diluted 1:200 and incubated at room temperature for 30 minutes. These samples were imaged using Amnis® Imagestream^X Mark II Imaging Flow cytometer (EMD Millipore, WA USA) with 488 nm and 642 nm lasers respectively. Analysis was done using the IDEAS® software.

Data analysis

Paired-end reads from illumina sequencing were aligned to the human transcriptome with BWA (Li and Durbin, 2009). Read 1 was used for assigning reads to correct cells and libraries, while read 2 was mapped to gene models. Reads that mapped equally well to multiple locations were discarded. Read counts were first corrected for UMI barcode by removing duplicate reads that had identical combinations of library, cellular, and molecular barcodes and were mapped to the same gene. Transcript counts were then adjusted to the expected number of molecules based on counts, 256 possible UMI's and poissonian counting statistics.

Samples were normalized by downsampling to a minimum number of 6000 transcripts. StemID was used to cluster cells and to perform outlier analysis. Differentially expressed genes between two subgroups of cells were identified similar to a previously published method (Anders and Huber, 2010). First, a negative binomial distribution was calculated reflecting the gene expression variability within each subgroup based on the background model for the expected transcript count variability computed by StemID (Grün et al., 2016). Using these distributions a p-value for the observed difference in transcript counts between the two subgroups is computed as described in Anders and Huber, 2010. These p-values were corrected for multiple testing by the Benjamini-Hochberg method.

Data Resources

The single-cell sequencing data described in this study was uploaded to GEO under accession number GSE85241

References

Ackermann, A.M., Wang, Z., Schug, J., Naji, A., and Kaestner, K.H. (2016). Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Mol. Metab. 5, 233–244.

Åkerfeldt, M.C., and Laybutt, D.R. (2011). Inhibition of Id1 augments insulin secretion and protects against high-fat diet-induced glucose intolerance. Diabetes 60, 2506–2514.

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

Anderson, K.R., Singer, R. a, Balderes, D. a, Hernandez-Lagunas, L., Johnson, C.W., Artinger, K.B., and Sussel, L. (2011). The L6 domain tetraspanin Tm4sf4 regulates endocrine pancreas differentiation and directed cell migration. Development 138, 3213–3224.

Arden, C., Hampson, L.J., Huang, G.C., Shaw, J.A.M., Aldibbiat, A., Holliman, G., Manas, D., Khan, S., Lange, A.J., and Agius, L. (2008). A role for PFK-2/FBPase-2, as distinct from fructose 2,6-bisphosphate, in regulation of insulin secretion in pancreatic beta-cells. Biochem. J. 411, 41–51.

Arntfield, M.E., and van der Kooy, D. (2011). β-Cell evolution: How the pancreas borrowed from the brain: The shared toolbox of genes expressed by neural and pancreatic endocrine cells may reflect their evolutionary relationship. Bioessays 33, 582–587.

Benner, C., van der Meulen, T., Cacéres, E., Tigyi, K., Donaldson, C.J., and Huising, M.O. (2014). The transcriptional landscape of mouse beta cells compared to human beta cells reveals notable species differences in long non-coding RNA and protein-coding gene expression. BMC Genomics 15, 620.

Bugliani, M., Liechti, R., Cheon, H., Suleiman, M., Marselli, L., Kirkpatrick, C., Filipponi, F., Boggi, U., Xenarios, I., Syed, F., et al. (2013). Microarray analysis of isolated human islet transcriptome in type 2 diabetes and the role of the ubiquitin-proteasome system in pancreatic beta cell dysfunction. Mol. Cell. Endocrinol. 367, 1–10.

Chawla, K., Tripathi, S., Thommesen, L., Laegreid, A., and Kuiper, M. (2013). TFcheckpoint: a curated compendium of specific DNA-binding RNA polymerase II transcription factors. Bioinformatics 29, 2519–2520.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14, 128.

Chen, H., Carlson, E.C., Pellet, L., Moritz, J.T., and Epstein, P.N. (2001). Overexpression of Metallothionein in Pancreatic ??-Cells Reduces Streptozotocin-Induced DNA Damage and Diabetes. Diabetes 50, 2040–2046.

Cnop, M., Abdulkarim, B., Bottu, G., Cunha, D. a., Igoillo-Esteve, M., Masini, M., Turatsinze, J.V., Griebel, T., Villate, O., Santin, I., et al. (2014). RNA sequencing identifies dysregulation of the human pancreatic islet transcriptome by the saturated fatty acid palmitate. Diabetes 63, 1978–1993.

Dai, C., Brissova, M., Hang, Y., Thompson, C., Poffenberger, G., Shostak, a., Chen, Z., Stein, R., and Powers, a. C. (2012). Islet-enriched gene expression and glucose-induced insulin secretion in human and mouse islets. Diabetologia 55, 707–718.

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343, 193–196.

Digruccio, M.R., Mawla, A.M., Donaldson, C.J., Noguchi, G.M., Vaughan, J., Cowing-zitron, C., Meulen, T. Van Der, and Huising, M.O. (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.

Doglioni, C., Gambacorta, M., Zamboni, G., Coggi, G., and Viale, G. (1990). Immunocytochemical localization of progesterone receptors in endocrine cells of the human pancreas. Am. J. Pathol. 137, 999–1005.

Dorrell, C., Schug, J., Lin, C.F., Canaday, P.S., Fox, a. J., Smirnova, O., Bonnah, R., Streeter, P.R., Stoeckert, C.J., Kaestner, K.H., et al. (2011a). Transcriptomes of the major human pancreatic cell types. Diabetologia 54, 2832–2844.

Dorrell, C., Grompe, M.T., Pan, F.C., Zhong, Y., Canaday, P.S., Shultz, L.D., Greiner, D.L., Wright, C. V, Streeter, P.R., and Grompe, M. (2011b). Isolation of mouse pancreatic alpha, beta, duct and acinar populations with cell surface markers. Mol. Cell. Endocrinol. 339, 144–150.

Eizirik, D.L., Sammeth, M., Bouckenooghe, T., Bottu, G., Sisino, G., Igoillo-Esteve, M., Ortis, F., Santin, I., Colli, M.L., Barthson, J., et al. (2012). The human pancreatic islet transcriptome: expression of candidate genes for type 1 diabetes and the impact of pro-inflammatory cytokines. PLoS Genet. 8, e1002552.
Grant, S.F.A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., et al. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat. Genet. 38, 320–323.

Grün, D., and van Oudenaarden, A. (2015). Design and Analysis of Single-Cell Sequencing Experiments. Cell 163, 799–810.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods 11, 637–640.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature.

Grün, D., Muraro, M.J., Boisset, J.-C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., Clevers, H., et al. (2016). De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. Cell Stem Cell 1–12.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Rep. 2, 666–673.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., et al. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 17, 77.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 21, 1160–1167.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 343, 776–779.

Johnson, M.B., Wang, P.P., Atabay, K.D., Murphy, E. a, Doan, R.N., Hecht, J.L., and Walsh, C. a (2015). Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. Nat. Neurosci. 18.

Kimple, M.E., Keller, M.P., Rabaglia, M.R., Pasker, R.L., Neuman, J.C., Truchan, N.A., Brar, H.K., and Attie, A.D. (2013). Prostaglandin E2 receptor, EP3, is induced in diabetic islets and negatively regulates glucose- and hormone-stimulated insulin secretion. Diabetes 62, 1904–1912.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The technology and biology of single-cell RNA sequencing. Mol. Cell 58, 610–620.

Kulkarni, R.N. (2004). The islet beta-cell. Int. J. Biochem. Cell Biol. 36, 365–371.

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

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

Maaten, L. Van Der, and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214.

Murphy, D.L., and Lesch, K.-P. (2008). Targeting the murine serotonin transporter: insights into human neurobiology. Nat. Rev. Neurosci. 9, 85–96.

Nica, A.C., Ongen, H., and Irminger, J. (2013). Cell-type , allelic and genetic signatures in the human pancreatic beta cell transcriptome Cell-type , allelic and genetic signatures in the human pancreatic beta cell transcriptome. 1554–1562.

Nishimura, W., Kondo, T., Salameh, T., Khattabi, I.E., Dodge, R., Bonner-Weir, S., and Sharma, A. (2008). A switch from MafB to MafA expression accompanies differentiation to pancreatic beta-cells. October 28, 4439–4448.

Orino, K., Lehman, L., Tsuji, Y., Ayaki, H., Torti, S. V, and Torti, F.M. (2001). Ferritin and the response to oxidative stress. Biochem. J. 357, 241–247.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B. V, Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396–1401.

Piccand, J., Strasser, P., Hodson, D.J., Meunier, A., Ye, T., Keime, C., Birling, M.-C., Rutter, G.A., and Gradwohl, G. (2014). Rfx6 maintains the functional identity of adult pancreatic β cells. Cell Rep. 9, 2219–2232.

Rankin, M.M., and Kushner, J.A. (2010). Aging induces a distinct gene expression program in mouse islets. Islets 2, 345–352.

Rantanen, K., Pursiheimo, J.-P., Hogel, H., Miikkulainen, P., Sundstrom, J., and Jaakkola, P.M. (2013). p62/SQSTM1 regulates cellular oxygen sensing by attenuating PHD3 activity through aggregate sequestration and enhanced degradation. J. Cell Sci. 126, 1144–1154.

Ricordi, C., Lacy, P.E., Finke, E.H., Olack, B.J., and Scharp, D.W. (1988). Automated method for isolation of human pancreatic islets. Diabetes 37, 413–420.

Le Roith, D., Shiloach, J., and Roth, J. (1982). Is there an earlier phylogenetic precursor that is common to both the nervous and endocrine systems? Peptides 3, 211–215.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498, 236–240.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., and Surani, M.A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell 6, 468–478.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. 25, 1491–1498.

Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M. a, and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371–375.

Wang, Y., and Navin, N.E. (2015). Advances and applications of single-cell sequencing technologies. Mol. Cell 58, 598–609.

Wang, Y.J., Schug, J., Won, K., Liu, C., Naji, A., Avrahami, D., Golson, M.L., and Kaestner, K.H. (2016). Single cell transcriptomics of the human endocrine pancreas. 1–49.

WHO (2014). Global status report on noncommunicable diseases 2014. World Health 176.

Xin, Y., Kim, J., Ni, M., Wei, Y., Okamoto, H., Lee, J., Adler, C., Cavino, K., Murphy, A.J., Yancopoulos, G.D., 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.

Xu, J., Liu, H., Park, J.-S., Lan, Y., and Jiang, R. (2014). Osr1 acts downstream of and interacts synergistically with Six2 to maintain nephron progenitor cells during kidney organogenesis. Development 141, 1442–1452.

Yang, Q., Graham, T.E., Mody, N., Preitner, F., Peroni, O.D., Zabolotny, J.M., Kotani, K., Quadro, L., and Kahn, B.B. (2005). Serum retinol binding protein 4 contributes to insulin resistance in obesity and type 2 diabetes. Nature 436, 356–362.

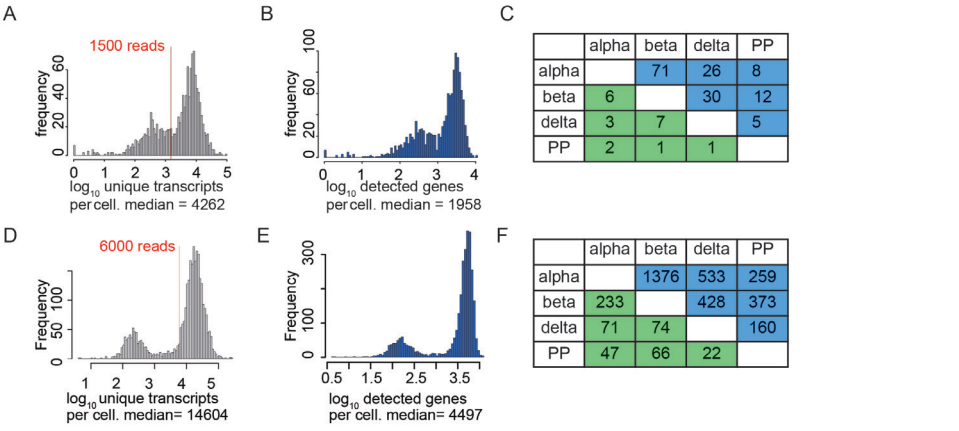
Ye, R., Ni, M., Wang, M., Luo, S., Zhu, G., Chow, R.H., and Lee, A.S. (2011). Inositol 1,4,5-trisphosphate receptor 1 mutation perturbs glucose homeostasis and enhances susceptibility to diet-induced diabetes. J. Endocrinol. 210, 209–217.

Zeisel, A., Munoz-Manchado, A.B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (80-.). 347, 1138–1142.

Zhang, J., McKenna, L.B., Bogue, C.W., and Kaestner, K.H. (2014). The diabetes gene Hhex maintains δ-cell differentiation and islet function. Genes Dev. 28, 829–834.

Zhou, Y., Duan, S., Zhou, Y., Yu, S., Wu, J., Wu, X., Zhao, J., and Zhao, Y. (2015). Sulfiredoxin-1 Attenuates Oxidative Stress via Nrf2/ARE Pathway and 2-Cys Prdxs After Oxygen-Glucose Deprivation in Astrocytes. J. Mol. Neurosci. 55, 941–950

3



	alpha	beta	delta	PP
alpha		71	26	8
beta	6		30	12
delta	3	7		5
PP	2	1	1	

	alpha	beta	delta	PP
alpha		1376	533	259
beta	233		428	373
delta	71	74		160
PP	47	66	22	

I

epsilon cells		duct cells		acinar cells		mesenchymal cells		endothelial cells	
genes	TF	genes	TF	genes	TF	genes	TF	genes	TF
GHRL	*VTN*	*SPP1*	*ONECUT2*	*PNLIP*	*GATA4*	*COL1A1*	*WNT5A*	*FLT1*	*SOX18*
ANXA13	*EBF1*	*CFTR*	*LITAF*	*REG1B*	*MECOM*	*COL1A2*	*SNAI2*	*KDR*	*RGCC*
PHGR1	*BMP7*	*AQP1*	*SOX4*	*PRSS1*	*NR5A2*	*COL3A1*	*NOTCH3*	*CD93*	*SMAD6*
ACSL1	*CDKN2A*	*ALDH1A3*	*DAB2*	*ALB*	*ZFP36L1*	*COL6A3*	*FBN1*	*ESAM*	*ERG*
FRZB	*PROX1*	*KRT19*	*CREB5*	*PRSS3P2*	*CSDA*	*FN1*	*HEYL*	*SOX18*	*PRDM1*
SPTSSB	*ARX*	*CRP*	*HLA-DQB1*	*CPA2*	*CEBPD*	*SFRP2*	*PRRX1*	*PECAM1*	*TCF4*
ASGR1	*ZKSCAN1*	*DEFB1*	*WWTR1*	*CTRB2*	*CREB3L1*	*COL5A1*	*UACA*	*ESM1*	*NOTCH4*
HEPACAM2		*CEACAM6*	*PPARGC1A*	*CEL*	*XBP1*	*SPARC*	*AEBP1*	*PASK*	*SNAI1*
VTN		*MMP7*	*PKHD1*	*PLA2G1B*	*LGR4*	*COL15A1*	*TBX3*	*SLCO2A1*	*NKX2-3*
SERPINA1		*TSPAN8*	*NFIB*	*CELA3A*	*NUPR1*	*SERPINE1*	*FOXF2*	*PLVAP*	*ETS1*

Figure S1. SORT-Seq allows for deep sequencing of human pancreas cells, Related to Figure 1.
(A) Histogram of the total detected transcripts per cell for cells of the five donors processed by manual CEL-Seq. On the X-axis are the $\log_{10}$ detected unique transcripts per cell. Y-axis is the frequency. The minimum number of unique transcripts per cell used as cutoff for downsampling and analysis is indicated in red (1500).
(B) Histogram of genes detected per cell for cells of the first five donors processed by manual CEL-Seq. X-axis are the genes detected per cell. Y-axis is the frequency.
(C) Table indicating the differentially expressed genes (blue) and transcription factors (green) when comparing across the different endocrine cell types from data prepared by manual CEL-Seq.
(D) Histogram of the total detected transcripts per cell for cells of the four donors (SORT-Seq) used in this study. On the X-axis are the $\log_{10}$ detected unique transcripts per cell. Y-axis is the frequency. . The minimum number of unique transcripts per cell used as cutoff for downsampling and analysis is indicated in red (6000).
(E) Histogram of genes detected per cell for cells of the four donors (SORT-Seq) used in this study. X-axis are the genes detected per cell. Y-axis is the frequency. On average 1891 genes were detected.
(F) Table indicating the differentially expressed genes (blue) and transcription factors (green) when comparing across the different endocrine cell types from data prepared by SORT-Seq.
(G) Heat map showing distances between cellular transcriptomes obtained by sequencing. Clustering was performed by StemID (Grün et al, 2016). Distances are calculated as 1 – Pearson correlation and used as input for k-medoid clustering. Each line represents a cell and cells are grouped by cluster. Black lines indicate clusters, as do color bars and numbers on the axes, which match the colors and numbers in Figure 1B.
(H) t-SNE maps highlighting cell type-specific expression of pancreatic marker genes. Transcript counts are given in linear scale. Green indicates high expression.
(I) Tables denoting the top 10 differentially expressed genes and transcription factors (TF) when comparing one of the pancreatic cell types to all other cells in the dataset ($P<10^{-6}$). Continuation of Figure 1E.

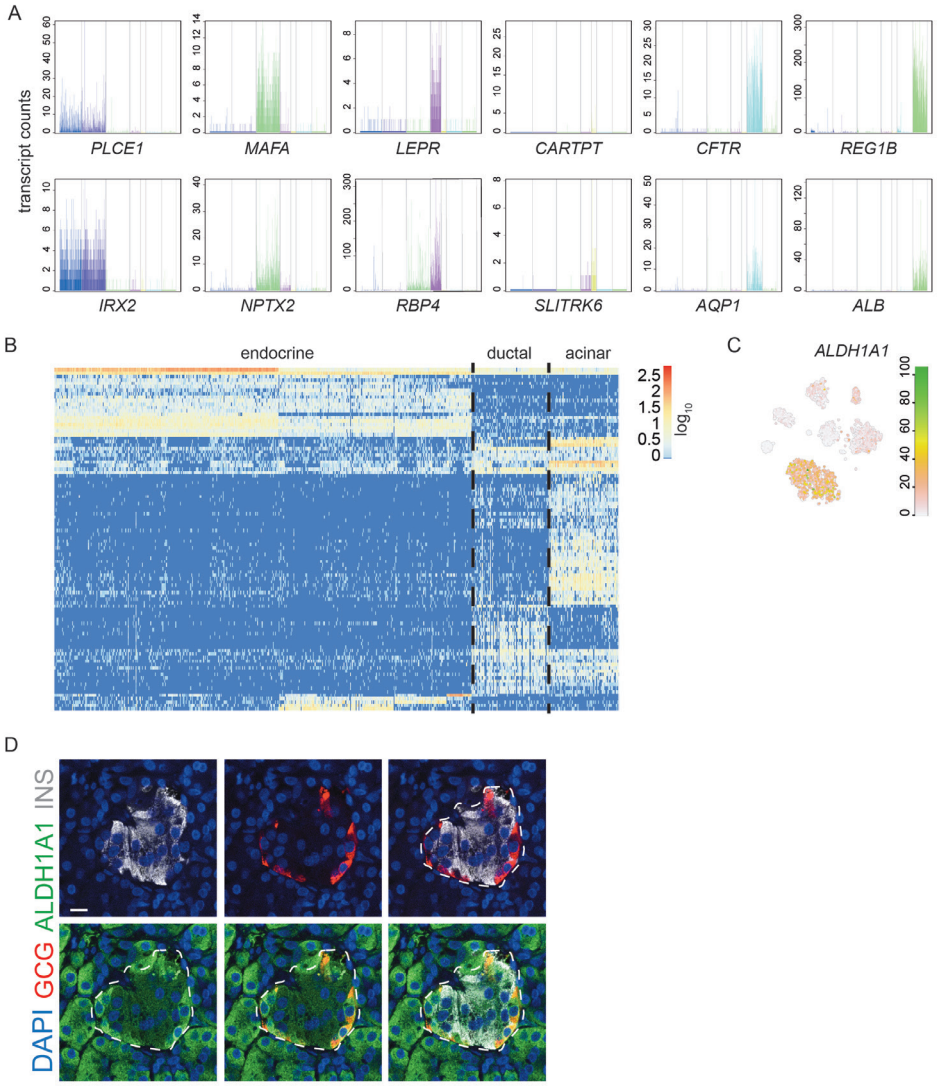

A
transcript counts
PLCE1
MAFA
LEPR
CARTPT
CFTR
REG1B
IRX2
NPTX2
RBP4
SLITRK6
AQP1
ALB
B
endocrine
ductal
acinar
log10
C
ALDH1A1
D
DAPI GCG ALDH1A1 INS

Figure S2. Cluster-restricted gene expression patterns and identification of new cell-type specific genes, Related to Figure 2.
(A) Expression of second (top) and third (bottom) most differentially expressed genes in each of six of the main pancreatic cell types. Down-sampled gene expression values are plotted on the Y-axis. Each bar represents a cell and cells are grouped by cluster with a specific color in the following order: alpha, beta, delta, PP, duct and acinar cells. If the most differentially expressed gene was also a canonical marker gene, the third and fourth most differentially expressed genes are shown.
(B) Heat map of the top 100 differentially expressed genes between endocrine and exocrine cell types. Rows are genes, columns are cells. Dashed lines indicate separation between acinar, ductal and endocrine cells. Log_2 expression of transcript counts for genes is plotted where red is high expression. Genes are grouped based on hierarchical clustering.
(C) t-SNE map highlighting the expression ALDH1A1. Transcript counts are given in linear scale. Green indicates high expression.
(D) Immunohistochemistry for ALDH1A1 (green) glucagon (red) and insulin (gray) with counterstaining for DAPI (blue) on human pancreatic tissue sections. Co-staining for INS and GCG identifies an Islet of Langerhans (marked by white dashed line). Co-staining of ALDH1A1 with GCG and INS shows overlap in the alpha cells, but not the beta cells inside the islet of Langerhans. Surrounding acinar cells express ALDH1A1 as well. Scale bar is 25 μM.

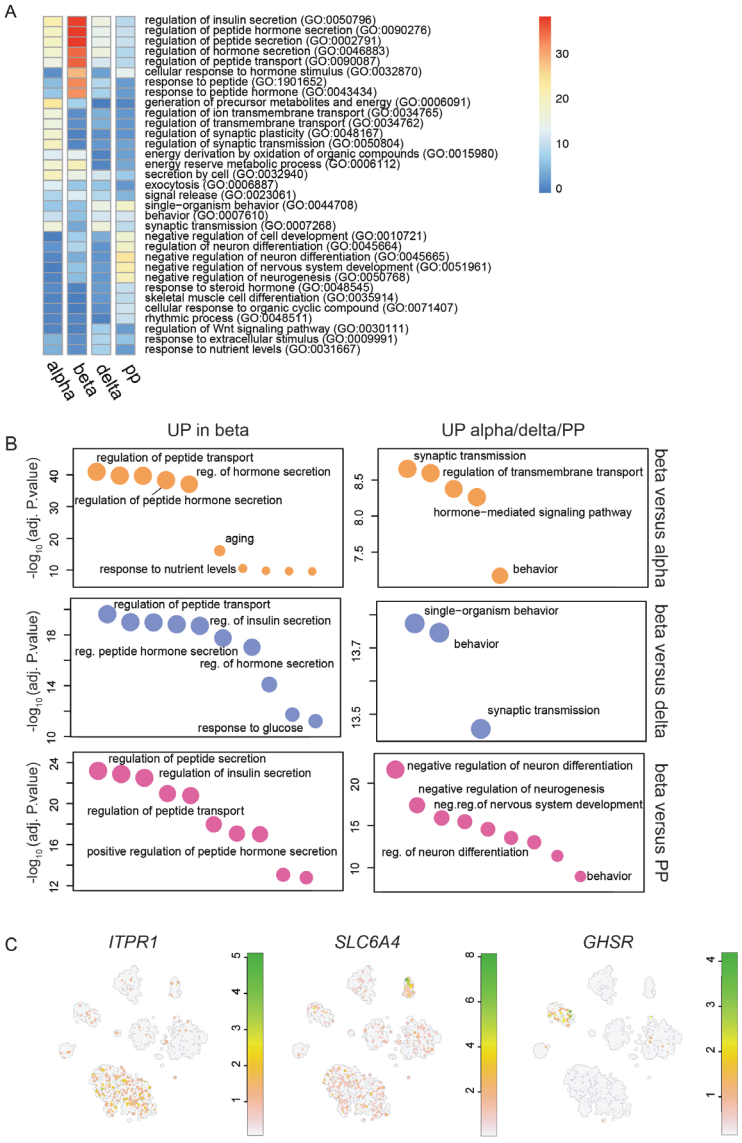


Figure S3. GO-term analysis reveals cell-type specific gene expression patterns relevant to endocrine biology and glucose metabolism.
(A) Heatmap showing the combined list of top 15 enriched GO terms for genes differentially expressed in endocrine cell types. Color indicates 1/p-value value so that red indicates a high score.(B) Plot showing top 10 enriched GO terms for genes differentially expressed between beta cells compared to the three other endocrine cell types. The left column shows GO terms for genes with higher expression in beta cells, the right column shows GO terms of genes with higher expression in each of the other endocrine cell types. Terms are ordered on p-value on the x-axis, with the most significant on the left. Names of relevant terms are highlighted. (C) t-SNE map genes found upon GO-term analysis with alpha, beta or delta specific expression. Green indicates high expression.

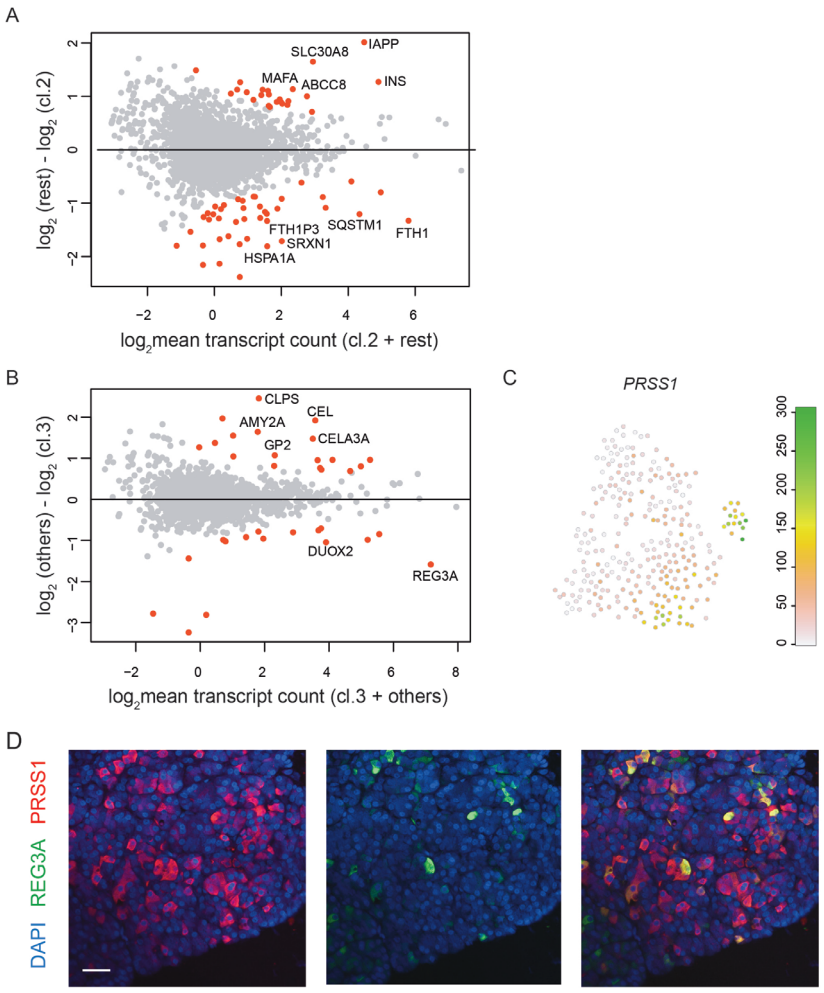


Figure S4. Outlier identification shows heterogeneity within acinar and beta cells, Related to Figure 3.

(A) Differential gene expression analysis of beta subclusters (FTH1-high cluster 2 versus the rest of the cells). Grey dots indicate genes, red dots indicate significant genes (P<10^-6).

(B) Differential gene expression analysis between the acinar subclusters (REG3A-high cluster 1 versus the rest of the cells). Grey dots indicate genes, red dots indicate significant genes (P<10^-6).

(B) t-SNE map highlighting the expression PRSS1 across all acinar cells. Transcript counts are given in linear scale. Green indicates high expression.

(D) Immunohistochemistry showing protein expression for REG3A (green), and PRSS1 (red) with counterstaining for DAPI (blue). Scale bar is 25 µM.

Figure S5. FACS Enrichment of endocrine and alpha cells based on novel cell-surface markers, Related to Figure 4

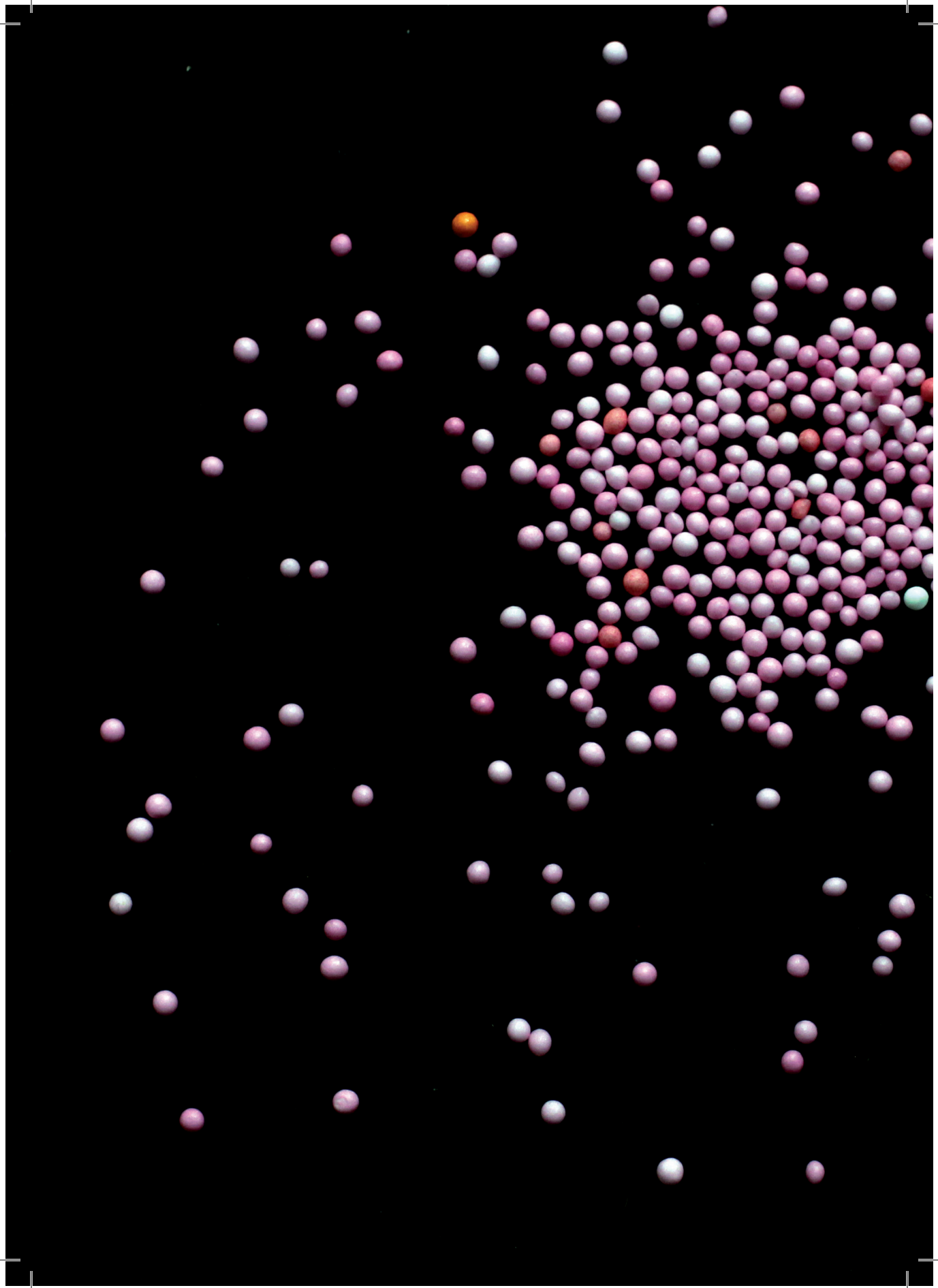
(A) t-SNE map highlighting the cells coming from the different FACS gating strategies. Each strategy is one color. Names of cell types are indicated next to their corresponding clusters. Cells sorted on only live (DAPI) marker are pink. Cells sorted against CD24 and CD44 expression are green.

(B) t-SNE maps highlighting the expression of the main pancreatic marker genes in libraries obtained by sorting for live or CD24/CD44 negative cells. Green indicates high expression.

(C) t-SNE map highlighting the expression of the main pancreatic marker genes in libraries obtained by sorting for live or CD24-/TM4SF4+/- cells. Green means high expression.

Supplemental information

Supplemental Information includes five figures, seven tables, and one data file and can be found with this article online at http://dx.doi.org/10.1016/j.cels. 2016.09.002.



4

The Dynamics of Pancreas Development Resolved by Single-Cell Transcriptomics

Léon van Gurp[1,*], Mauro J. Muraro[1,*], Tim Dielen[1], Lina Seneby[1], Gitanjali Dharmadhikari[1], Gerard Gradwohl[2], Alexander van Oudenaarden[1], Eelco J. P. de Koning[1,3]

*Equal contribution

[1] Hubrecht Institute
KNAW and University Medical Center Utrecht
Uppsalalaan 8, 3584CT Utrecht, the Netherlands

[2] Institut de Génétique et de Biologie Moléculaire et Cellulaire
Université de Strasbourg, Strasbourg, France

[3] Department of Internal Medicine
Leiden University Medical Center, Leiden, the Netherlands

Abstract
During pancreatic development, endocrine cells appear when Neurog3 positive cells delaminate from the ductal lining of the pancreatic epithelium. Little is known about how these endocrine progenitors dynamically develop into adult islet cells. Here, we characterize the temporal, lineage-specific developmental programs involved by performing single-cell mRNA sequencing of thousands of cells from embryonic day 12.5 to 18.5. We found clusters of cells corresponding to all major pancreatic cell types and were able to deconstruct the gene expression patterns involved in pancreatic cell type specification by ordering of all single cells in pseudotime from Neurog3+ progenitors to alpha or beta cell fate. Endocrine progenitors could be subdivided in Neurog3 expressing and Neurog3 negative cells that were more committed towards endocrine lineages. We found dozens of genes that show progenitor-specific expression patterns during developmental time, many of which have no currently known function in endocrine development. Finally, we validated three of these novel genes (Megf11, Nhlh1 and Chgb) that marked specific subpopulations of endocrine progenitors. This resource allows dynamic profiling of embryonic pancreas development at single-cell level and reveals previously unknown gene signatures for different endocrine progenitor states as well as for alpha and beta cell differentiation.

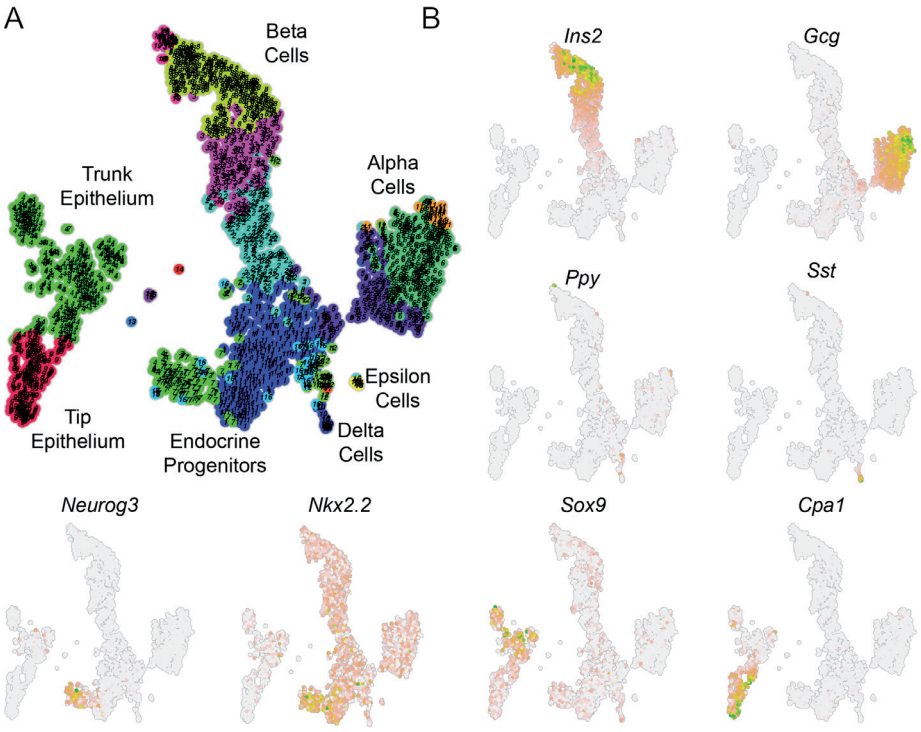
Introduction
Mouse pancreatic development consists of two main differentiation phases: the primary and secondary transition. The secondary transitional phase, characterized by segregation of the pancreatic epithelium into ductal tip and trunk domains, takes place between embryonic age (E)12.5 and E15.5 (1, 2). In the trunk region, lateral inhibition determines which cells will differentiate into mature ductal cells, and which towards an endocrine cell fate (3-5) . The cells with an endocrine fate are marked by Neurogenin-3 (*Neurog3*), a key transcription factor during endocrine differentiation (6). After expression of *Neurog3*, endocrine progenitors delaminate from the ductal lining to form the Islets of Langerhans in the mesenchyme surrounding the pancreatic epithelium (7, 8). These endocrine progenitors, which stop proliferation after commitment to the endocrine lineage (9), make endocrine fate choices based on the expression of a number of key transcription factors. An early fate choice is determined by the balance in expression between mutually inhibitory genes *Pax4* or *Arx*, which push cells towards an alpha, beta or delta cell phenotype (10-12). Later fate choices towards a beta cell phenotype involve *Pax6*, *Neurod1*, *Nkx2.2*, *Nkx6.1*, *Mafa* and *Mafb* (13-16). Expression of these markers is well described, but the precise dynamic interplay between these transcription factors and other genes involved in these transitions remain poorly understood. Transcriptome-wide information for each of the developing cell types would be informative to clarify the cell fate choices involved in pancreas development.
Single cell mRNA sequencing is a relatively new approach that provides transcriptome-wide gene expression information from individual cells (17-20). The strongest advantages over traditional bulk sequencing are the possibilities to investigate gene expression patterns for each cell type individually, to probe heterogeneity within cells of the same type and to identify rare cell types within

a population. With traditional bulk sequencing, thousands to millions of cells of different types are analyzed together, where the most abundant cell type dominates the results (21). Others and we have recently analyzed gene expression in mature human and murine pancreatic tissue at the single cell level (22-28). These studies show how single cells from the endocrine and exocrine pancreas can be clustered based on gene expression profiles into alpha, beta, gamma, delta, epsilon, ductal and acinar clusters. Heterogeneity within these clusters was validated at the tissue level to find rare subpopulations of cells (22, 23, 25). A new challenge lies in the unbiased identification of progenitor cells from single-cell datasets, which can be done in an unbiased manner using cellular entropy (29) in combination with connectivity between clusters of cells (30).
Here, we used SORT-seq (22) to sequence individual cells from multiple time points during the secondary transition of the pancreas. All cells were combined into a single dataset, and then grouped into clusters by StemID (30) that correspond to all cell types of the embryonic pancreas. This dataset was used to follow tissue maturation through time, characterize progenitor clusters, identify the distinct endocrine cell types of the pancreas that arise from these progenitors and reveal dynamic regulation of genes involved in these fate choices.

Results

Unbiased transcriptome profiling of single cells from different developmental timepoints identifies all known pancreatic cell types

Transcriptomic profiles of embryonic pancreatic cells from MIP-GFP mice of E12.5, E13.5, E14.5, E15.5 and E18.5 were pooled into a single dataset containing 4921 cells and 19696 genes (after removal of non-pancreatic cell types, see methods). Both GFP positive and negative cells were sorted independently (Figure S1A). After filtering for lowly expressed genes and cells with low total numbers of unique transcripts, a dataset of 2636 cells and 5899 genes remained. Cells expressed an average of 10820 unique transcripts per cell, divided over an average 2951 genes per cell (Figure S1B-C). Of these cells, 1850 were GFP positive and 786 were GFP negative. By overlaying single-cell FACS information with transcriptome information, we could detect differences between GFP positive and negative cells on FACS parameters FSC (GFP-neg $31.2 \pm 0.3 \times 10^3$ vs GFP-pos $34.7 \pm 0.2 \times 10^3$, $p < 0.01$) and SSC (GFP-neg $18.2 \pm 0.3 \times 10^3$ vs GFP-pos $26.4 \pm 0.4 \times 10^3$, $p < 0.01$), indicating that the GFP positive cells were on average both larger and have a higher internal complex than GFP negative cells (Figure S1D).
Clustering of all cells revealed 19 unique clusters (Figure 1A). We found cluster-specific gene expression signatures that contained markers of all major pancreatic cell types (Figure S1F and supplementary data 1) By overlaying these gene expression patterns with a t-SNE map of all clusters we could clearly map all major pancreatic cell types to specific (groups of) clusters (Figure 1A and 1B). We found alpha cells (marked by *Gcg*), beta cells (*Ins2*), delta cells (*Sst)*, epsilon cells (*Ghrl*, Figure S1E), endocrine progenitors (*Neurog3)*, and epithelial trunk and tip cells (Sox9 and *Cpa1*, resp). Located between the Neurog3+ progenitor cluster and all hormone producing endocrine cell types, we found a cluster (cluster 7) expressing several transcription factors crucial for endocrine cell type differentiation in the


A
Beta Cells
Trunk Epithelium
Alpha Cells
Epsilon Cells
Tip Epithelium
Endocrine Progenitors
Delta Cells
B
Ins2
Gcg
Ppy
Sst
Neurog3
Nkx2.2
Sox9
Cpa1

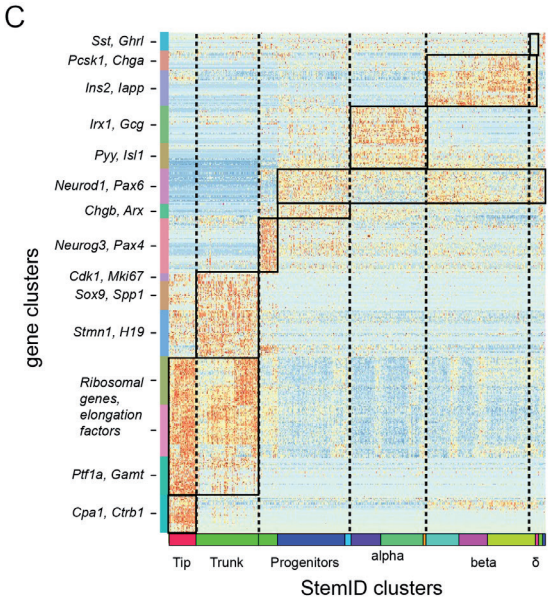

C
Sst, Ghrl
Pcsk1, Chga
Ins2, Iapp
Irx1, Gcg
Pyy, Isl1
Neurod1, Pax6
Chgb, Arx
Neurog3, Pax4
Cdk1, Mki67
Sox9, Spp1
Stmn1, H19
Ribosomal genes, elongation factors
Ptf1a, Gamt
Cpa1, Ctrb1
gene clusters
Tip
Trunk
Progenitors
alpha
beta
δ
StemID clusters

Figure 1. Unbiased transcriptome profiling of single cells from different developmental time-points identifies all known pancreatic cell types

A) t-SNE map of a combined dataset of embryonic pancreatic cells from embryonic ages E12.5 to E18.5. Cells in the dataset grouped in 19 clusters, which are indicated by colour and number. The main clusters of this dataset could be identified as being tip- or trunk epithelium, endocrine progenitors, and alpha- beta- and delta cell progenitors. B) Specific expression of marker genes *Ins2*, *Gcg*, *Ppy*, *Sst*, *Neurog3*, *Nkx2.2*, *Sox9* and *Cpa1*. C) Heatmap of gene clusters of the most differentially expressed genes on the Y-axis. These were picked from the top differentially expressed genes in each StemID cluster and are ordered according to those on the X-axis. Representative genes for each gene are labeled. Numbers and colors on the X-axis indicate StemID clusters.

pancreas, such as *Nkx2.2* (Figure 1B). Since it does not yet show expression of more differentiated endocrine cell markers, we assume this is a population of cells that are still in progenitor stage, and will refer to this cluster as progenitor 2. Besides the mesenchymal and blood cell types already filtered before clustering (see methods), we found small populations of sympathetic neurons (marked by *Hand2*) and mast cells (*Cma1*, Figure S1E). We next set out to visualize the different gene expression signatures across different cell types in our dataset. To achieve this, we selected only the differentially expressed genes in the StemID clusters corresponding to alpha, beta, delta, progenitor, tip and trunk cells. These were then clustered into 15 different groups and plotted in a heatmap (Figure 1C). This map shows groups of genes with expression that is specific to Tip/Trunk, progenitor populations and endocrine cell types. Interestingly, some groups of genes show expression that spans several cell types as identified by StemID. For example, we found a group of genes marking cells in either trunk or progenitor clusters that are dividing (*Mki67 and Cdk1*). The clusters that span endocrine cell types show genes like *Neurod1* and *Pax6,* verifying that the StemID clusters correspond to the known cell types in the developing pancreas.

In short, by unbiased clustering of single-cell transcriptomes from E12.5 to E18.5, we could detect all major pancreatic cell types in our data, allowing us to describe their genome-wide expression patterns of developing embryonic cell types for the first time. We supply cell cluster specific gene expression patterns in supplemental data 1.

Alpha and beta cell progenitors show a more differentiated transcriptome at later embryonic ages

To see the effect of individual embryonic ages on the composition of the dataset, we generated a t-SNE map that visualizes which cell is derived from which embryonic age (figure 2A). In some clusters, cells from the different embryonic ages are mixed, such as in the cluster with Neurog3 expressing cells (cluster 7, red box). Other clusters show a temporal gradient, with cells from early embryonic ages on one side and cells from later embryonic ages on the other side. This is indicated by the blue box, containing the clusters that express *Ins2* (combined clusters 2, 3, 8 and 19).

To explore why cells organize based on embryonic age, we drew a t-SNE map that visualizes the entropy per cell (Figure 2B) (37). Cells with higher entropy have a more uniform transcriptome, and these cells are generally considered to be more undifferentiated. The entropy of the *Neurog3* expressing cells in cluster 7 (red box) is significantly higher than the *Ins2* expressing cells in clusters 2, 3, 8

and 19 (blue box) (*Ins2* cells 8.542 ± 0.005 vs *Neurog3* cells 8.245 ± 0.010, $p < 0.01$). More importantly, while the entropy in cluster 7 is uniform between cells, there is an entropy gradient in clusters 2, 3, 8 and 19, correlating with the embryonic age of the cells. This indicates that, in contrast to *Neurog3* expressing cells, *Ins2* expressing cells have a less uniform transcriptome at later embryonic ages. This is clear from the levels of expression of *Ins2*, which reach one sixth of all total detected transcripts in some E15.5 and E18.5 cells (Figure 2D). We next set out to quantify the contribution of each time point to each cell type. Figure 2C shows that both progenitor 1 and progenitor 2 cluster cells consist of mostly cells from the first 4 time points, while cells from the alpha and beta clusters are progressively populated by cells from later time points. Interestingly, we found that the timepoint also influences the cells obtained by using the MIP-GFP reporter. Upon clustering of each time point individually, two things became clear. First, that different cell types cluster more apart at later embryonic ages. Secondly, that enriching for GFP positive cells at E12.5 mostly yielded alpha cells and only a few other endocrine cells (Figure S2A). The reverse is true at E18.5, and this reinforces the idea that alpha cells are the first endocrine cell type to appear in the developing pancreas .

The entropy change at different embryonic ages is clearly exemplified by how individual genes are strongly differentially expressed during development. To illustrate this, we divided the endocrine compartment of our dataset into three fragments: *Neurog3* expressing endocrine progenitors (cluster 7), developing alpha cells (clusters 5, 6 and 11), and developing beta cells (clusters 2, 3, 8 and 19). Within these compartments, we identified key differentially expressed genes per compartment and calculated their average expression for each embryonic age (Figure S2B). In *Neurog3* expressing cells from cluster 7 many genes do not show a very distinct expression profile, except for a few genes that are downregulated at later embryonic ages like *Megf11*, *Cck* and *Arx*. In alpha cell progenitors (clusters 5, 6 and 11) some specific genes are strongly upregulated during embryonic development (*Tm4sf4*, *Gpx3* and *Scg2*), while other genes are strongly downregulated (*Slc38a5*, *Fev* and *Cck*). Finally, in beta cell progenitors *Ins1*, *Ins2* and *Chga* show markedly increased expression while *Gip* is strongly downregulated at later embryonic ages. To summarize, we found that endocrine clusters show a marked decrease in entropy and increasingly contain cells from later time points, reflecting a more specialized transcriptome dominated by hormonal transcripts such as Insulin or Glucagon. Furthermore, time point contributions in progenitor clusters and early alpha and beta cells are more mixed than for the more differentiated clusters, suggesting that progenitor cells extracted at the same time point are not necessarily developmentally synchronized.

Pseudo-temporal ordering of cells shows the genetic differentiation dynamics in the alpha and beta cell lineages

We next asked if we could order all cells of a specific lineage by their develelopmental stage, rather than the embryonic age at which they where extracted. Using the StemID algorithm (30), a lineage tree was generated that projects all cells onto paths between the medoids of each cluster and shows all significant connections between clusters (figure 3A). Using this lineage tree, we ordered cells into

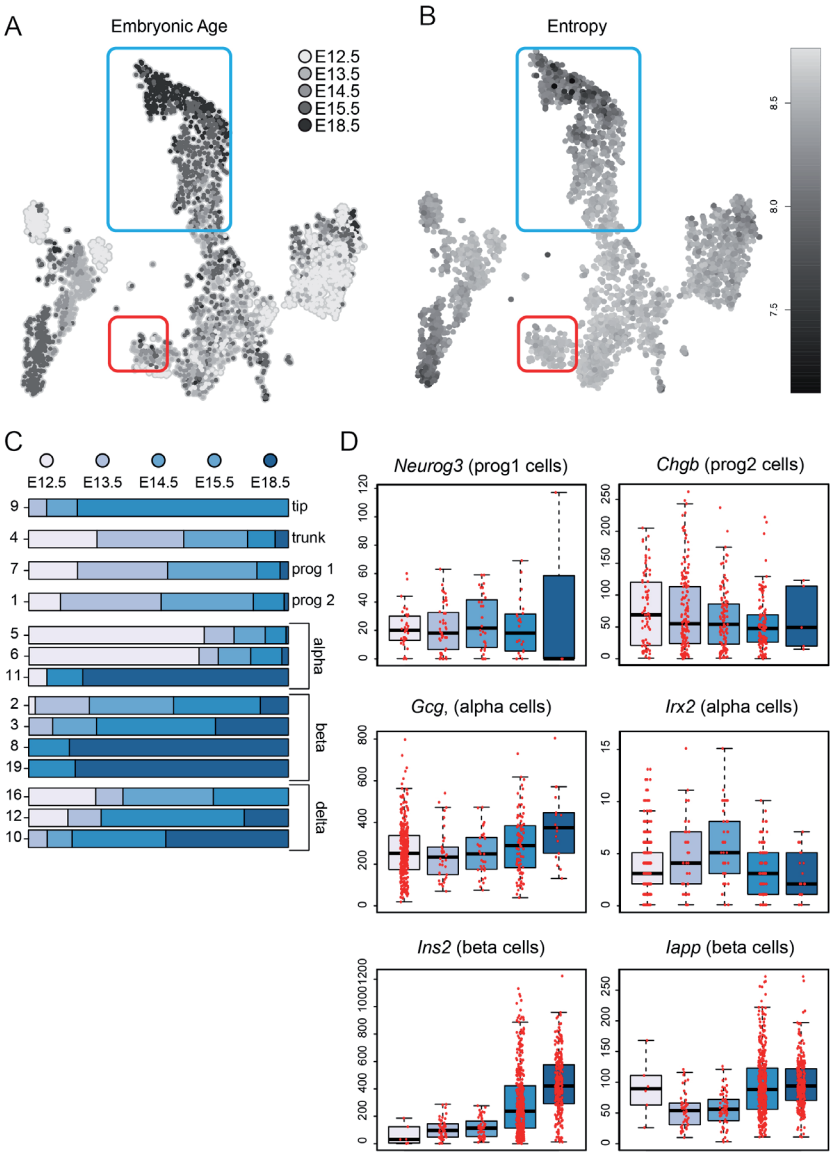


Figure 2. Alpha and beta cell progenitors show a more differentiated transcriptome at later embryonic ages

A) t-SNE map that shows from which embryonic age (E12.5, E13.5, E14.5, E15.5 or E18.5) a cell was collected. The red box indicates cluster 7, which contains Neurog3 expressing endocrine progenitor cells. The blue box indicates clusters 2, 3, 8 and 19, which contain Ins2 expressing beta cell progenitor cells . B) t-SNE map that shows cellular entropy. A lower entropy value indicates a less uniform transcriptome. The red and blue boxes indicate Neurog3 positive endocrine progenitor cells (cluster 7) and Ins2 expressing beta cell progenitor cells (clusters 2, 3, 8 and 19. C) Contribution of each time point to cell types. Clusters are grouped per cell type. D) Boxplots showing expression of 2 marker genes for three main cell types (progenitors, alpha and beta cells), ordered by time point of extraction. Each dot represents one cell.

developmental pseudo timelines that represent alpha and beta cell development. For both, development begins in the epithelial tip cluster 9, then progresses through the epithelial trunk cluster 4 to the *Neurog3* expressing endocrine progenitor cluster 7 and the Chgb expressing endocrine progenitor cluster 1. Based on this algorithm cluster 1 appears to be at a late progenitor intersection for alpha and beta cell lineages. From there, the pathways diverge towards clusters 5, 6 and 11 for developing alpha cells and clusters 2, 3, 8 and 19 for developing beta cells.
To explore which genes are responsible for development from endocrine progenitors to the most differentiated alpha and beta cells in our dataset, we generated self-organizing maps (SOM) for both alpha- and beta cell trajectories. Genes were assigned to 1 of 25 SOM clusters based on their expression profiles through pseudo time (Figure 3B-C, Supplemental Data 2 and 3). For both alpha and beta branches, SOM clusters can be identified that contain genes that are most strongly expressed in a specific part of the developmental trajectory. For example, in the alpha branch, SOM cluster 25 contains genes with very high expression in the Neurog3 progenitor cluster, SOM cluster 21 genes that peak the second progenitor cluster while SOM clusters 5 to 8 contain genes that are very highly expressed in the tip of the alpha branch. Equally, for the beta branch, SOM clusters 1 to 3 contain genes that are highly expressed in tip of the beta branch and could be designated late beta cell progenitors. We next asked how specific genes were dynamically expressed in pseudo-time. For this, we expression maps for 16 key genes in alpha and beta cell development along the clusters organized in developmental pseudotime (Figure 3D and S3A). In the alpha cell branch, progenitor markers such as *Arx*, *Pax4*, and *Chgb* peak in cluster 1 and are gradually downregulated during maturation. The early marker *Slc38a5* peaks in early alpha cells (cluster 5) while adult markers like *Gcg*, *Ttr* and *Irx2* peak in the more differentiated alpha cells (cluster 6). In the beta cell branch progenitor markers such as *Neurod1*, *Mafb*, *Pax4* and *Arx* peak in the second progenitor cluster (cluster 1). The early marker Dlk1 peaks in early beta cells (cluster 3), while adult markers such as *Ins1*, *Ins2*, *Iapp* and *Nkx6.1* peak in the later beta cell clusters 8 and 19. Beyond the genes known for having a function in pancreatic development, we find many genes with no known role in pancreas development that show a distinct, branch-specific pattern, like Hopx (Supplemental data 3), a gene that has been implicated in pancreatic cancer, but never has been reported to have a cell type specific role during development (Waraya et al., 2012).

In conclusion, we show that the genes found by clustering genes along developmental pseudotime contain important pancreas development-related candidates like *Pax4*, *Arx* and *Neurog3*. For each SOM cluster(s) we find many more genes that behave in a specific fashion along development to alpha or beta cells, providing many novel targets for understanding pancreatic development.

Nhlh1, Megf11 and Chgb are heterogeneously expressed in Neurog3 positive endocrine progenitor cells
Since we identified two distinct clusters with common endocrine progenitor cells (clusters 1 and 7) that show specific expression of genes along developmental pseudotime, we set out to determine the global transcriptomic differences between

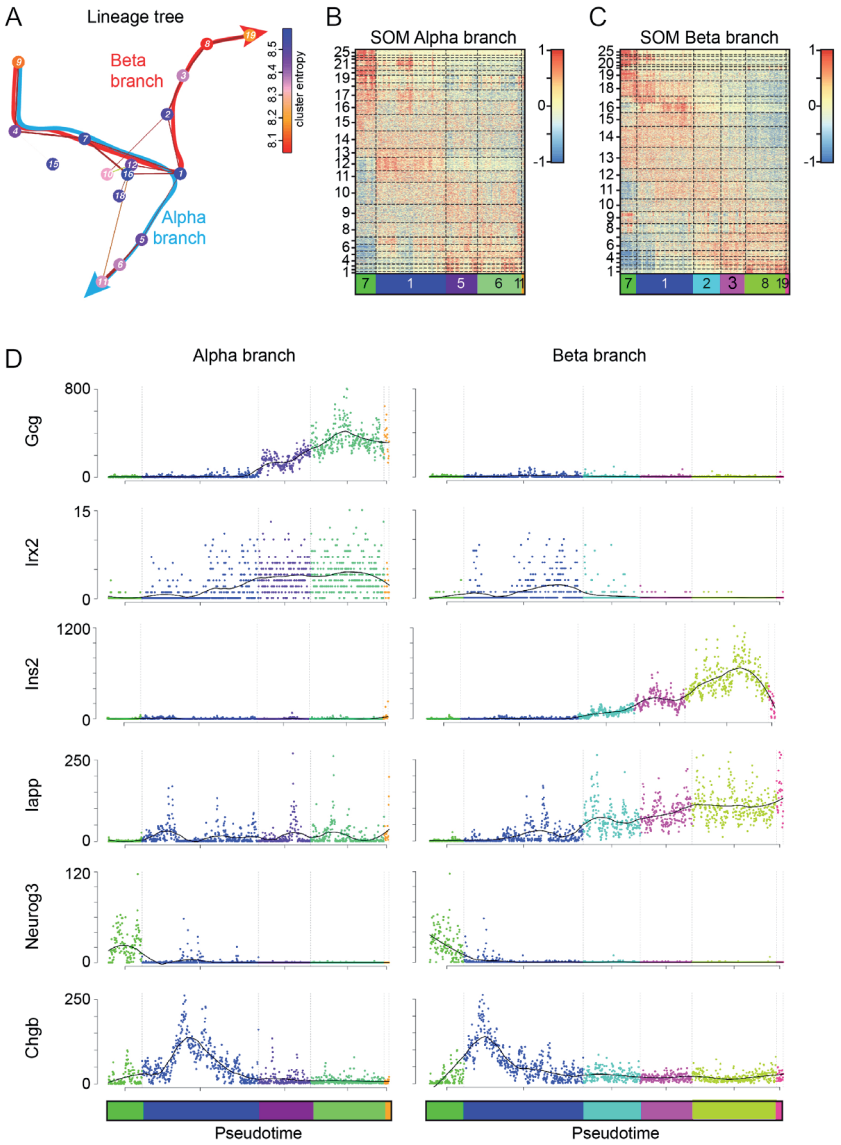


Figure 3: Pseudo-temporal ordering of cells shows differentiation dynamics in the alpha and beta cell lineages

A) Using stemID, a lineage tree was calculated that shows the average entropy (red to blue) for every cluster and the connectivity (thickness of line) between clusters. From this, two branches were designated, for alpha and beta cell development. These branches follow connected clusters from high entropy to low entropy. B-C) Self-organizing maps (SOM) were generated for both alpha and beta branches as depicted in A, starting from Neurog3+ progenitor cells to the most mature alpha or beta cells. Cells were ordered in pseudotime (x-axis, clusters 7, 1, 5, 6, 11 for alpha cells (B) and clusters 7,1,2,3,8,19 for beta cells(C)). Genes were divided into 25 clusters based on expression profiles (y-axis). Expression of genes is indicated by color; from low (blue) to high (red). D) Pseudo time plots for alpha and beta branches as indicated in B-C. All cells from Neurog3+ to the most differentiated alpha and beta cells are on ordered left-to-right on the X-axis. Expression of 7 marker genes is shown on the Y-axis.

these two groups of endocrine progenitor cells. Differential gene expression was calculated between these clusters (Figure 4A, supplemental data 4). *Neurog3* and *Pax4* were upregulated in cluster 7 while key marker genes for different endocrine cell types like *Ins2*, *Gcg*, *Ppy* and *Sst*, *Irx2*, *Ttr* and *Iapp* were upregulated in cluster 1. This indicates that cluster 1 represents a progenitor cluster that is already more specified towards endocrine cell fates than the Neurog3+ cells in cluster 7. Interestingly, some of the cluster 1-specific genes were mesenchymal markers like Vimentin, indicating that this cluster might correspond to cells that are delaminating from the epithelial lining. The most distinctly upregulated gene in cluster 1 was *Chgb*. This gene also clearly peaks in cluster 1 cells along the trajectory towards alpha and beta cells compared to clusters both before and after it (Figure 3D). Several of the novel marker genes (e.g. *Megf11* and *Nhlh1*) that are co-expressed with Neurog3 in cluster 7 (Figure S3C) also appear in the list of differentially expressed genes between the two progenitor clusters, prompting us to verify their expression *In Situ.* We then set out to perform immunohistochemistry for these markers. First, we stained E15.5 embryonic pancreata from Neurog3-YFP mice for Megf11, a protein identified to be involved in mosaic patterning of neurons in the retina (39) and found that the protein was expressed in a subpopulation of Neurog3 expressing cells (figure 4b). The protein was always found in the apical domain of the cells, and this correlated with low expression of E-cadherin (*Cdh1*). Nhlh1, a basic helix-loop-helix transcription factor like Neurog3 implicated in neuronal development (40, 41), was expressed in all Neurog3 expressing cells but also in some Neurog3 negative cells that were closely located to Neurog3 positive cells (figure 4C). Chgb expression was present in Neurog3 positive/Cdh1^{high} positive cells and cluster of Neurog3 negative/Cdh1^{low} positive cells, showing that this marker indeed is heterogeneously expressed in endocrine progenitor cells of the developing pancreas. In short, we found clear differences between two different clusters of endocrine progenitor cells, of which cluster 1 shows signs of a more developmentally advanced state. We verified this by showing *in situ* that some of the markers described here are heterogeneously expressed on protein level in the developing pancreas.

Discussion

In this manuscript, we describe a developmental roadmap for endocrine cells in the developing pancreas, based on single cell transcriptome profiling of cells from the beginning of the secondary transition of pancreas development (E12.5) until just prior to birth (E18.5). We were able to enrich for all endocrine cell types in the developing pancreas by using pancreas from MIP-GFP mice. While in the adult pancreas of this mouse model GFP positivity is restricted to mature beta cells, insulin gene expression at the transcriptional level is present in many endocrine cell types during embryonic development (44). In our dataset there is a temporal effect: GFP positive cells that express high levels of *Gcg* were sorted on E12.5, after which the percentage of *Gcg* cells that express GFP diminishes and *Ins2* cells that express GFP increases towards E18.5. During the intermediate stages (E13.5, E14.5 and E15.5), some GFP positive cells were detected that expressed high levels of *Ppy* and *Sst*. We also identified pancreatic epithelial cells, which can be separated into tip or trunk domains based on expression of either *Sox9* or *Cpa1* (1, 7). Importantly, we found populations of endocrine progenitor cells organized in two

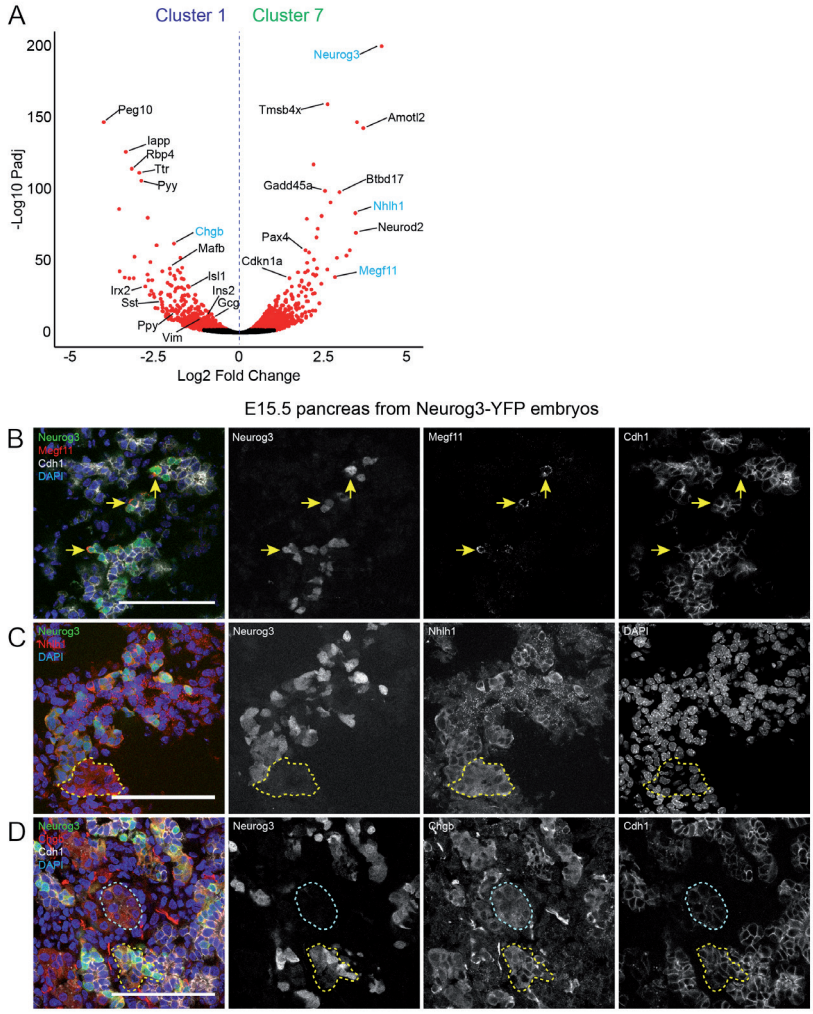


Figure 4: Nhlh1, Megf11 and Chgb are heterogeneously expressed in Neurog3 positive endocrine progenitor cells

A) Volcano plot of differential expression between endocrine progenitor clusters that do (cluster 7) and do not (cluster 1) express Neurog3. The difference in gene expression is plotted against the significance of differential expression. Each dot represents a single gene, red dots are differentially expressed with $p < 0.01$. B) Sections of E15.5 embryonic pancreas from mice expressing YFP under transcriptional control of the Neurog3 promoter (green) were stained for Megf11 (red) and Cdh1 (grey). Nuclei were stained with DAPI (blue). Scalebar 100 µm. C) Sections of E15.5 embryonic pancreas from mice expressing YFP under transcriptional control of the Neurog3 promoter (green) were stained for Nhlh1 (red). Almost all Neurog3 expressing cells also express Nhlh1. Nuclei were stained with DAPI (blue). Clusters of Nhlh1+/Neurog3- cells could be found near Neurog3+ cells, indicated by yellow dotted lines. Scalebar 100 µm. D) Sections of E15.5 embryonic pancreas from mice expressing YFP under transcriptional control of the Neurog3 promoter (green) were stained for Chgb (red) and Cdh1 (grey). Clusters of cells within the ductal lining (Cdh1 high) were positive for Chgb, and many of these cells co-expressed Neurog3, as indicated by yellow dotted lines. Outside the ductal lining, clusters of Chgb+/Neurog3- clusters were identified (Cdh1 low), as indicated by blue dotted lines. Scalebar 100 µm.

clusters that could be separated based on the expression of either *Neurog3* (45) or *Chgb* (46). The dataset initially also contained few cells with non-pancreatic cell types, like mesenchymal cells (expressing high levels of *Col1a1*), endothelial cells (expressing high levels of *Cd93*), erythrocytes (expressing high levels of *Hbb-y*) and immune cells (expressing high levels of *Apoe*), which were discarded from the dataset prior to analysis.

We found cells at an early embryonic age have higher entropy than cells from later embryonic ages, indicating that cells from later ages are generally more adult-like in their transcriptomic signature. We confirmed this by showing that expression of specific adult marker genes such as *Spp1*, *Ins2* and *Gcg* strongly increases at later embryonic ages. Interestingly, *Neurog3* expressing cells have similar entropy independent of the embryonic age they come from, indicating that cells that go through *Neurog3* driven endocrine differentiation are developmentally similar regardless of the embryonic age at which the cell was captured (7, 45).

To follow cells trough developmental stage rather than isolation time, we generated pseudo-temporal maps for alpha and beta cell using the StemID algorithm (30). This allowed us to position cells in developmental order for both alpha- and beta cell maturation. The maturation of cells could be followed by increasing levels of *Gcg* and *Ins2* expression over pseudo-time, and by the temporal expression of *Neurog3*, *Arx*, *Pax4*, *Nkx6.1* and *Mafa* (11, 45, 51, 52). It also allowed us to detect temporal expression patterns of genes that have no prior reported role in pancreatic development, like *Chgb,* the expression of which peaks in cells found immediately after high *Neurog3* expressing cells. We then found several other genes with differential expression between the two progenitor clusters and validated the heterogenous expression of these progenitor markers by immunohistochemistry. We found that Megf11 was expressed in few Neurog3 positive cells and was localized at the apical domain of cells, potentially being involved in the delamination of endocrine progenitors from the ductal lining (53). Nhlh1 is expressed in every Neurog3 expressing cell, and in clusters of cells in the ductal lining next to Neurog3 expressing cells, making the Neurog3 expressing cells a subset of the Nhlh1 expressing cells. As the Nhlh1 positive, Neurog3 negative cells express Cdh1 as well, this indicates that Nhlh1 is expressed prior to Neurog3 and might have an upstream function in relation to endocrine differentiation. Cells in the Chgb+ progenitor cluster show increased expression of all hormone genes and some important islet cell identity genes such as *Irx2* and *Iapp*, when compared to the Neurog3+ progenitor cluster. This means they might represent endocrine progenitors at a later developmental stage than the cells in the Neurog3+ cluster. Immunostaining for Chgb revealed that many cells in the embryonic pancreas are positive for this protein, including clusters of cells with low Cdh1 expression which may form the islets of Langerhans, and clusters of Neurog3 positive cells in the ductal lining. Thus, we suggest that high RNA expression of Chgb indicates endocrine progenitors in a later stadium of development than Neurog3 does, while protein expression seems to be more ubiquitous.

Methods

Tissue preparation

Mouse embryos that express green fluorescent protein under transcriptional control of the mouse insulin promoter (MIP-GFP mice, Jackson Laboratories #006864) were isolated at embryonic age (E)12.5, E13.5, E14.5, E15.5 and E18.5. Mouse embryos carrying a transgene expressing YFP on the Neurog3 locus (Neurog3-YFP) (31) were isolated at E15.5. From these embryos, the pancreas was isolated as described previously (32). Pancreases from MIP-GFP embryos were digested into single cells using TrypLE (Thermo Fisher #12605010) containing 10 µg/ml pulmozyme (Roche, Basel, Switzerland), and washed with PBS containing 10% FBS (Thermo Fisher #10500064). Cells were stored on ice until they were sorted using FACS. DAPI (Sigma Aldrich #D9542, 20 µg/ml) or TO-PRO3 (Thermo Fisher #T3605, 1µM) was added to cell suspensions immediately before sorting to distinguish between live and dead cells. Pancreases from Neurog3-YFP embryos were fixed in freshly prepared 4% PFA for 2 hour at 4C, then cryoprotected in 30% sucrose solution for 6 hours and frozen in tissue freezing medium (Leica #14020108926).

FACS sorting

Cells were sorted as single cells into hard-shell 384 wells PCR plates (BioRad) containing 100 or 200 nl of RT primers, dNTPs and ERCC spike-ins, and 5 µl vapor-lock (Qiagen) using a FACSJazz or FACSAria II (BD biosciences) as described previously (22). For every embryonic age, MIP-GFP negative cells as well as MIP-GFP positive cells (to enrich for endocrine cell types) were sorted into separate plates. Sorting was performed using index sorting; allowing us to couple FACS obtained data like FCS, SSC and GFP intensity to our transcriptome data. Sorted cells in plates were snap frozen on dry ice and stored at -80 °C.

Processing of sorted cells

Cells were processed using SORT-seq (22), a high-throughput, FACS based single-cell sequencing protocol based on CEL-seq2 (20). In short, cells were thawed on ice and lysed at 65 °C, after which reverse transcription (RT) and second strand reactions were done. RT reactions were performed using primers containing a polyT tail, a 4 or 6 basepair unique molecular identifier (UMI) sequence, a cell-specific barcode sequence (8 basepair), an illumina 5' adapter and a T7 promoter sequence. After RT, each mRNA molecule was thus uniquely labeled. Contents from all wells in a plate were pooled into a single library after second strand synthesis. RNA in these libraries underwent linear amplification using in vitro transcription, followed by fragmentation to lengths between 200 and 1000 bp. Then, another RT reaction using random hexamer primers containing the illumina 3' adapter was performed and libraries were amplified using PCR. Sequencing was performed on Illumina NextSeq (paired-end, 75 bp). Approximately 90'000 reads per cell were sequenced.

Data processing

Sequenced reads were mapped to a reference transcriptome based on the mouse genome release mm10 as described previously (18, 22). In short, reads with the same UMI – barcode – transcript combination were likely caused by PCR over-

amplification and were thus counted as a single read, and the number of reads per transcript per cell were used to calculate transcript abundance using poissonian counting statistics (18). Data from all plates were pooled into a single dataset, containing 7296 cells and 19788 genes. Cells expressing more than 3 transcripts *Apoe*, 3 transcripts *Cd93*, 5 transcripts *Col1a1*, or 5 transcripts *Hbb-y* were excluded from the dataset, as they represented cell populations unrelated to the endocrine or exocrine pancreas. Data were downsampled to 6000 transcripts per cell, and cells with fewer transcripts were excluded. Genes needed to be expressed with a minimum of 4 transcripts in at least 2 cells or were excluded from further analysis. The genes *Malat1*, *Lars2* and *Rn45s*, which were strongly upregulated in some libraries and are linked to cellular stress (33-35), were excluded from the dataset. K-medoids clustering and outlier detection was performed using the StemID algorithm (18, 22), and differential expression between groups of cells was calculated as described previously (36). Clusters were identified using the top upregulated genes compared to all other cells in the dataset.
Connectivity and stemness of clusters was calculated using the StemID algorithm (30). In short, a lineage tree is calculated between clusters by projecting all cells onto the nearest vector running between the medoids of each cluster of cells. Significant enrichment of cells found between two clusters was determined by comparing it to a random background model. Self-organizing maps were generated for both alpha- and beta cell development. Cells were ordered in pseudo-time based on stemID clustering, after which every gene is assigned to 1 of 25 clusters based on their expression pattern through pseudo-time. This way, 25 clusters are created in which each represents a gene set with a unique expression profile during pseudo-temporal development. All data were processed using custom R scripts (www.r-project.org).

Immunohistochemistry
Sections from E15.5 Neurog3-YFP pancreases were cut at 10 um thickness. For immunohistochemical staining, sections were incubated with primary antibodies at 37 °C for 1 hour or at 4 °C overnight. Antibodies were used to stain for E-cadherin (Cdh1, 1:1000, BD 610182), Multiple EGF-like domains 11 (Megf11, 1:50, Novus Biologicals NBP2-14226), Nescient Helix-Loop-Helix 1 (Nhlh1, 1:200, Bioss USA bs-11901r), Chromogranin-b (Chgb, 1:500, Novus Biologicals NB600-1516), green fluorescent protein (GFP, 1:1000, Aves Labs GFP-1010), alexa 488 goat anti chicken (1:500, Thermo Fisher A11039), alexa 568 donkey anti rabbit (1:200, Thermo Fisher A10042) and alexa 633 goat anti mouse (1:200, Thermo Fisher A21050). Nuclei were stained using DAPI (1ng/µl, Thermo Fisher D1306).

References

1. Zhou Q, et al. (2007) A multipotent progenitor domain guides pancreatic organogenesis. Dev Cell 13(1):103-114.

2. Pan FC, et al. (2013) Spatiotemporal patterns of multipotentiality in Ptf1a-expressing cells during pancreas organogenesis and injury-induced facultative restoration. Development 140(4):751-764.

3. Magenheim J, et al. (2011) Ngn3(+) endocrine progenitor cells control the fate and morphogenesis of pancreatic ductal epithelium. Dev Biol 359(1):26-36.

4. Shih HP, et al. (2012) A Notch-dependent molecular circuitry initiates pancreatic endocrine and ductal cell differentiation. Development 139(14):2488-2499.

5. Kim YH, et al. (2015) Cell cycle-dependent differentiation dynamics balances growth and endocrine differentiation in the pancreas. PLoS Biol 13(3):e1002111.

6. Gu G, Dubauskaite J, & Melton DA (2002) Direct evidence for the pancreatic lineage: NGN3+ cells are islet progenitors and are distinct from duct progenitors. Development 129(10):2447-2457.

7. Seymour PA, et al. (2007) SOX9 is required for maintenance of the pancreatic progenitor cell pool. Proc Natl Acad Sci U S A 104(6):1865-1870.

8. Grapin-Botton A, Seymour PA, & Gradwohl G (2015) Pairing-up SOX to kick-start beta cell genesis. Diabetologia 58(5):859-861.

9. Miyatsuka T, Kosaka Y, Kim H, & German MS (2011) Neurogenin3 inhibits proliferation in endocrine progenitors by inducing Cdkn1a. Proc Natl Acad Sci U S A 108(1):185-190.

10. Collombat P, et al. (2005) The simultaneous loss of Arx and Pax4 genes promotes a somatostatin-producing cell fate specification at the expense of the alpha- and beta-cell lineages in the mouse endocrine pancreas. Development 132(13):2969-2980.

11. Collombat P, et al. (2003) Opposing actions of Arx and Pax4 in endocrine pancreas development. Genes Dev 17(20):2591-2603.

12. Sosa-Pineda B, Chowdhury K, Torres M, Oliver G, & Gruss P (1997) The Pax4 gene is essential for differentiation of insulin-producing beta cells in the mammalian pancreas. Nature 386(6623):399-402.

13. Artner I, et al. (2010) MafA and MafB regulate genes critical to beta-cells in a unique temporal manner. Diabetes 59(10):2530-2539.

14. Mastracci TL, Anderson KR, Papizan JB, & Sussel L (2013) Regulation of Neurod1 contributes to the lineage potential of Neurogenin3+ endocrine precursor cells in the pancreas. PLoS Genet 9(2):e1003278.

15. Schaffer AE, et al. (2013) Nkx6.1 controls a gene regulatory network required for establishing and maintaining pancreatic Beta cell identity. PLoS Genet 9(1):e1003274.

16. St-Onge L, Sosa-Pineda B, Chowdhury K, Mansouri A, & Gruss P (1997) Pax6 is required for differentiation of glucagon-producing alpha-cells in mouse pancreas. Nature 387(6631):406-409.

17. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, & Teichmann SA (2015) The technology and biology of single-cell RNA sequencing. Mol Cell 58(4):610-620.

18. Grun D, et al. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 525(7568):251-255.

19. Grun D & van Oudenaarden A (2015) Design and Analysis of Single-Cell Sequencing Experiments. Cell 163(4):799-810.

20. Hashimshony T, et al. (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol 17:77.

21. Kanter I & Kalisky T (2015) Single cell transcriptomics: methods and applications. Front Oncol 5:53.

22. Muraro MJ, et al. (2016) A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst.

23. Segerstolpe A, et al. (2016) Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell Metab 24(4):593-607.

24. 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.

25. Wang YJ, et al. (2016) Single-Cell Transcriptomics of the Human Endocrine Pancreas. Diabetes 65(10):3028-3038.

26. Li J, et al. (2016) Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. EMBO Rep 17(2):178-187.

27. Xin Y, et al. (2016) RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. Cell Metab 24(4):608-615.

28. 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(2):208-222.

29. Banerji CR, et al. (2013) Cellular network entropy as the energy potential in Waddington's differentiation landscape. Sci Rep 3:3039.

30. Grun D, et al. (2016) De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. Cell Stem Cell 19(2):266-277.

31. Mellitzer G, et al. (2004) Pancreatic islet progenitor cells in neurogenin 3-yellow fluorescent protein knock-add-on mice. Mol Endocrinol 18(11):2765-2776.

32. Petzold KM & Spagnoli FM (2012) A system for ex vivo culturing of embryonic pancreas. J Vis Exp (66):e3979.

33. Yao J, et al. (2016) Long non-coding RNA MALAT1 regulates retinal neurodegeneration through CREB signaling. EMBO Mol Med 8(4):346-362.

34. Schild C, et al. (2014) Mitochondrial leucine tRNA level and PTCD1 are regulated in response to leucine starvation. Amino Acids 46(7):1775-1783.

35. Yoshikawa M & Fujii YR (2016) Human Ribosomal RNA-Derived Resident MicroRNAs as the Transmitter of Information upon the Cytoplasmic Cancer Stress. Biomed Res Int 2016:7562085.

36. Anders S & Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11(10):R106.

37. Shannon CE (1948) A Mathematical Theory of Communication. Bell Syst. Tech. J. 27:379–423, 623–656.

38. DiGruccio MR, 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(7):449-458.

39. Kay JN, Chu MW, & Sanes JR (2012) MEGF10 and MEGF11 mediate homotypic interactions required for mosaic spacing of retinal neurons. Nature 483(7390):465-469.

40. Murdoch JN, Eddleston J, Leblond-Bourget N, Stanier P, & Copp AJ (1999) Sequence and expression analysis of Nhlh1: a basic helix-loop-helix gene implicated in neurogenesis. Dev Genet 24(1-2):165-177.

41. De Smaele E, et al. (2008) An integrated approach identifies Nhlh1 and Insm1 as Sonic Hedgehog-regulated genes in developing cerebellum and medulloblastoma. Neoplasia 10(1):89-98.

42. Shih HP, Wang A, & Sander M (2013) Pancreas organogenesis: from lineage determination to morphogenesis. Annu Rev Cell Dev Biol 29:81-105.

43. Mastracci TL & Sussel L (2012) The Endocrine Pancreas: insights into development, differentiation and diabetes. Wiley Interdiscip Rev Membr Transp Signal 1(5):609-628.

44. Katsuta H, et al. (2010) Single pancreatic beta cells co-express multiple islet hormone genes in mice. Diabetologia 53(1):128-138.

45. Rukstalis JM & Habener JF (2009) Neurogenin3: a master regulator of pancreatic islet differentiation and regeneration. Islets 1(3):177-184.

46. Lukinius A, Stridsberg M, & Wilander E (2003) Cellular expression and specific intragranular localization of chromogranin A, chromogranin B, and synaptophysin during ontogeny of pancreatic islet cells: an ultrastructural study. Pancreas 27(1):38-46.

47. Millington-Ward S, et al. (2004) RNAi of COL1A1 in mesenchymal progenitor cells. Eur J Hum Genet 12(10):864-866.

48. Galvagni F, et al. (2016) CD93 and dystroglycan cooperation in human endothelial cell adhesion and migration adhesion and migration. Oncotarget 7(9):10090-10103.

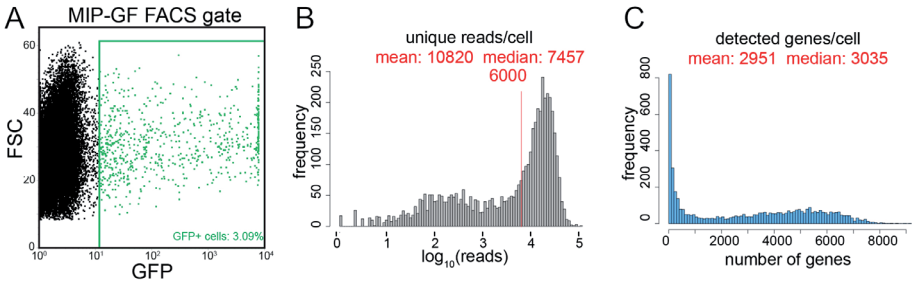
49. Finch JT, Perutz MF, Bertles JF, & Dobler J (1973) Structure of sickled erythrocytes and of sickle-cell hemoglobin fibers. Proc Natl Acad Sci U S A 70(3):718-722.

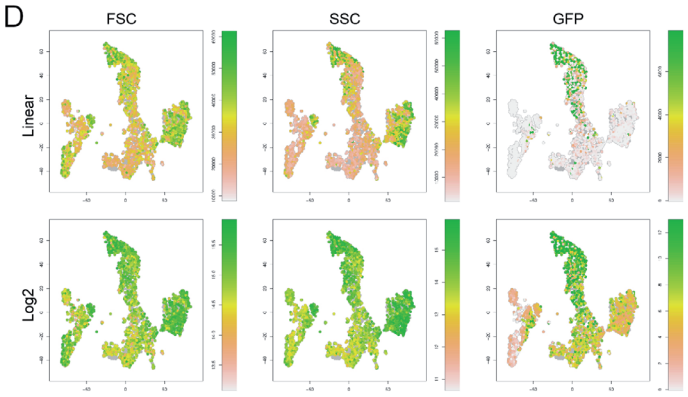
50. Dose J, Huebbe P, Nebel A, & Rimbach G (2016) APOE genotype and stress response - a mini review. Lipids Health Dis 15:121.

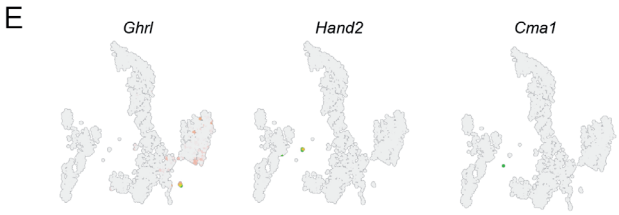
51. Sander M, et al. (2000) Homeobox gene Nkx6.1 lies downstream of Nkx2.2 in the major pathway of beta-cell formation in the pancreas. Development 127(24):5533-5540.

52. Kaneto H, et al. (2008) PDX-1 and MafA play a crucial role in pancreatic beta-cell differentiation and maintenance of mature beta-cell function. Endocr J 55(2):235-252.

53. Gouzi M, Kim YH, Katsumoto K, Johansson K, & Grapin-Botton A (2011) Neurogenin3 initiates stepwise delamination of differentiating endocrine cells during pancreas development. Dev Dyn 240(3):589-604.







F

Cluster	# Cells	Top 10 upregulated genes										ID
1	466	Chgb	Fev	Arx	Glud1	Runx1t1	Hmgn3	Dbpht2	Hap1	Gnao1	Rap1b	endocrine progenitor
2	223	Mapt	Nnat	Gng12	Atp2a2	Adra2a	Sytl4	Mlxipl	Papss2	Scgn	Iapp	beta cell progenitor
3	202	Npy	Nnat	Arhgap36	Sytl4	Ppp1r1a	Atp2a2	Iapp	Scg2	Ins2	Hadh	beta cell progenitor
4	429	Spp1	Adamts1	Cyr61	Sparc	Vim	Cxcl12	Ccnd1	Hmga2	Capn6	Wfdc2	epithelial trunk
5	209	Serpina1b	Serpina1a	Slc38a5	Cartpt	Irx1	Tmem27	Irx2	Ttr	Peg10	Gcg	alpha cell progenitor
6	294	Gcg	Cartpt	Irx1	Edn3	Irx2	Tmem27	Asb4	Serpina1b	Etv1	Ttr	alpha cell progenitor
7	135	Neurog3	Nhlh1	Neurod2	Pax4	Miat	Epb4.2	Cck	Mfap4	Btbd17	Gadd45a	endocrine progenitor
8	331	Ins1	Ins2	Prlr	G6pc2	Prss53	Slc30a8	Sdf2l1	Slc2a2	Iapp	Ero1lb	beta cell progenitor
9	196	Ctrb1	Pnliprp1	Cela2a	Amy2b	Clps	Spink3	Cel	Try10	Cela1	Cela3b	epithelial tip
10	18	Sst	Hhex	Ppy	Fam159b	Ptprz1	Arg1	Sez6l	Pyy	Cd24a	Rbp4	delta cell progenitor
11	20	Ppy	Rgs4	Gcg	Pyy	Gast	Pcsk1n	Gpx3	Etv1	Tmsb15l	Ccnd2	alpha\gamma cell progenit
12	39	Ppy	Pyy	Gm4216	Gast	Lrpprc	Peg10	Gip	Smarca1	Meis2	Dbpht2	gamma cell progenitor
13	3	Cma1	Cpa3	Tpsb2	Mcpt4	Hdc	Srgn	Gata2	Il1rl1	Ifitm1	Pla2g7	-
14	2	S100a9	Stfa1	S100a8	Ngp	BC100530	Gm5483	Camp	Stfa3	Stfa2	Lyz1	-
15	7	Stmn2	Gap43	Prph	Phox2b	Rtn1	L1cam	Phox2a	Tubb2b	Vip	Elavl3	-
16	26	Cdkn1a	Mboat4	Hhex	Arg1	Tmem184c	Runx1t1	Ghrl	Cacna2d1	Isl1	Arx	endocrine progenitor
17	1	Cartpt	Gast	Ostm1	Cbln4	Jag1	Ppy	Nfia	Plekha3	Ctxn2	Trf	-
18	9	Ghrl	Irs4	Acsl1	Foxd3	Anpep	Diras2	Mboat4	Alcam	Maged2	Enpp1	ghrelin cell progenitor
19	26	Ppy	Ins1	Enpp2	Iapp	Prss53	Ptms	Slc30a8	G6pc2	Npy	Pcsk1	beta\gamma cell progenitc

Figure S1.
A) FACS scatter plot for GFP and FSC showing MIP-GFP enrichment of endocrine cells. Green area indicates enriched cells. B) Histogram of total unique reads per cell. red line indicates the cutoff used to filter cells and for downsampling. C) Histogram showing number of detected genes per cell. D) Overlay of FACS index information with t-SNE map for FSC, SSC and GFP channels. Linear and Log scales are shown. E) t-SNE maps for genes Ghrl, Hand2 and Cma1. F) Table containing top differentially expressed genes per cluster and if possible an anntotation of the identity of that cluster.

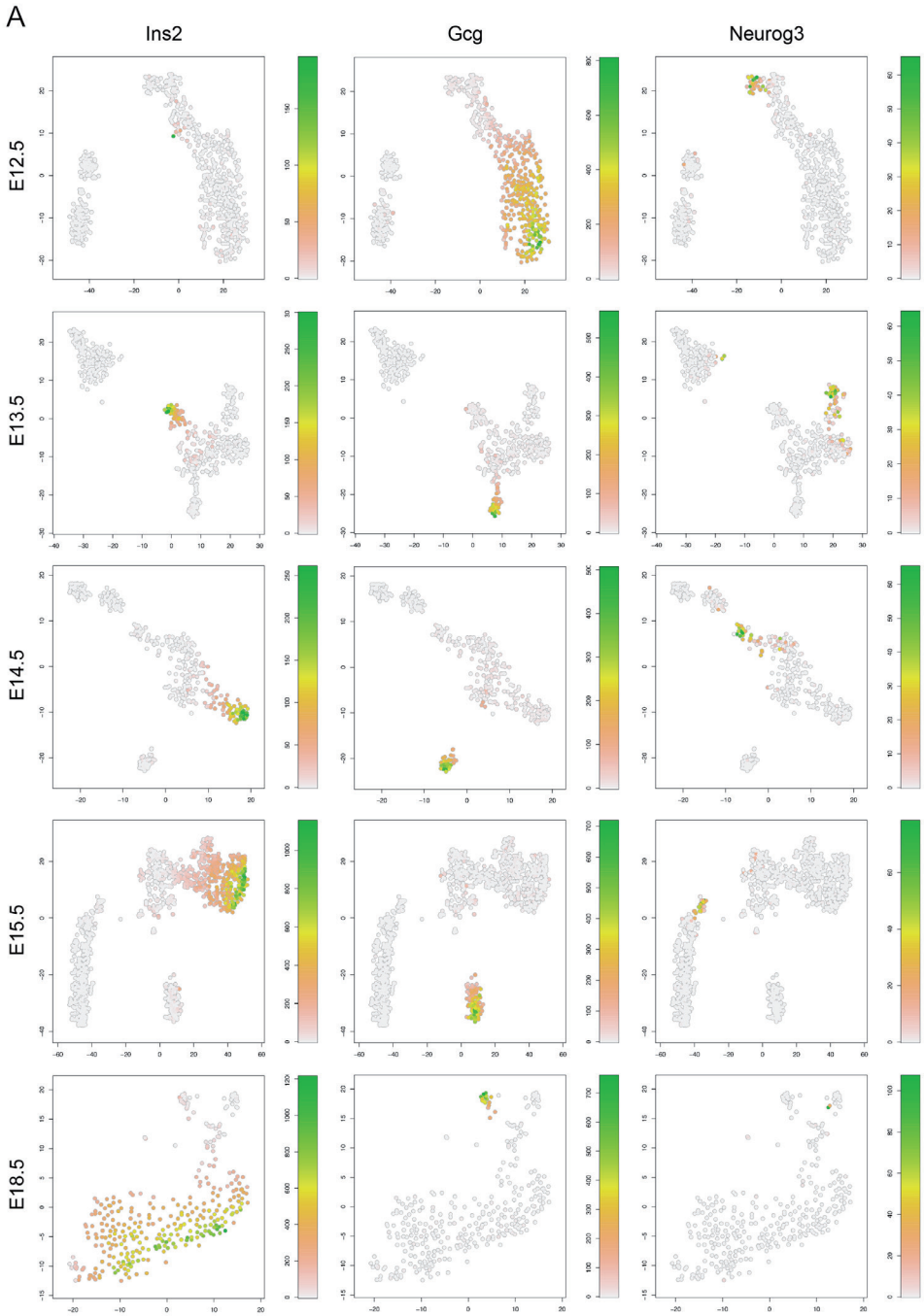


Figure S2a.
t-SNE maps for expression of three pancreatic marker genes (Ins2, Gcg, Neurog3) across the different timepoints used in this study. Color scale indicates expression.

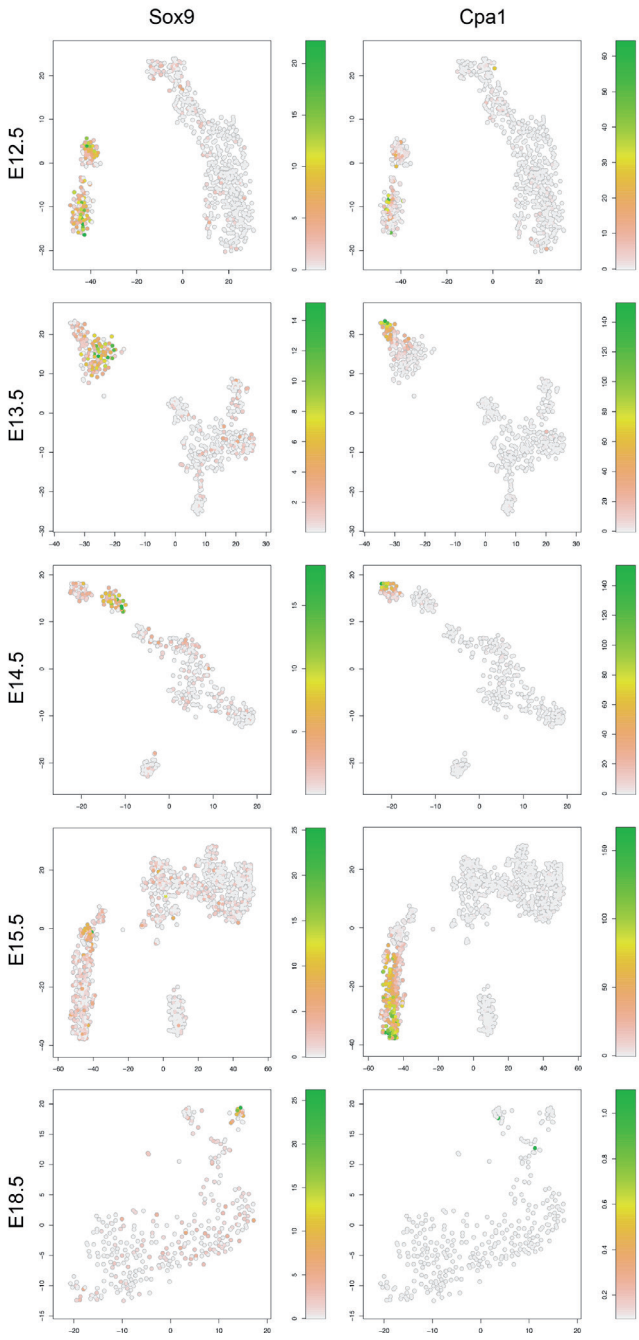


Figure S2b.
t-SNE maps for expression of two pancreatic marker genes (Sox9 and Cpa1) across the different time-points used in this study. Color scale indicates expression.

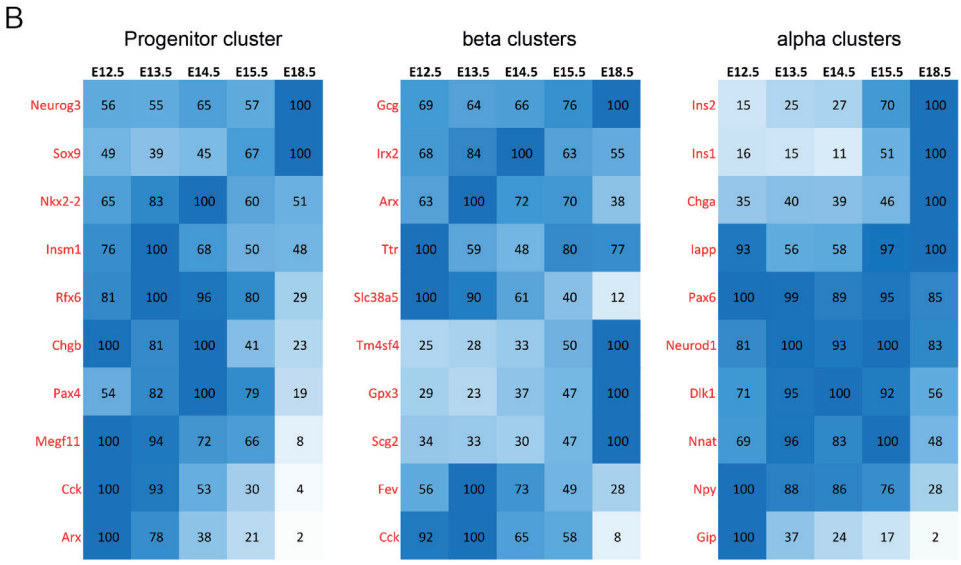


Figure S3.
Heatmaps showing expression of a set of pancreatic marker genes across the different timepoints used in this study for thee groups of clusters: progenitor, beta and alpha clusters. Color indicates expression.

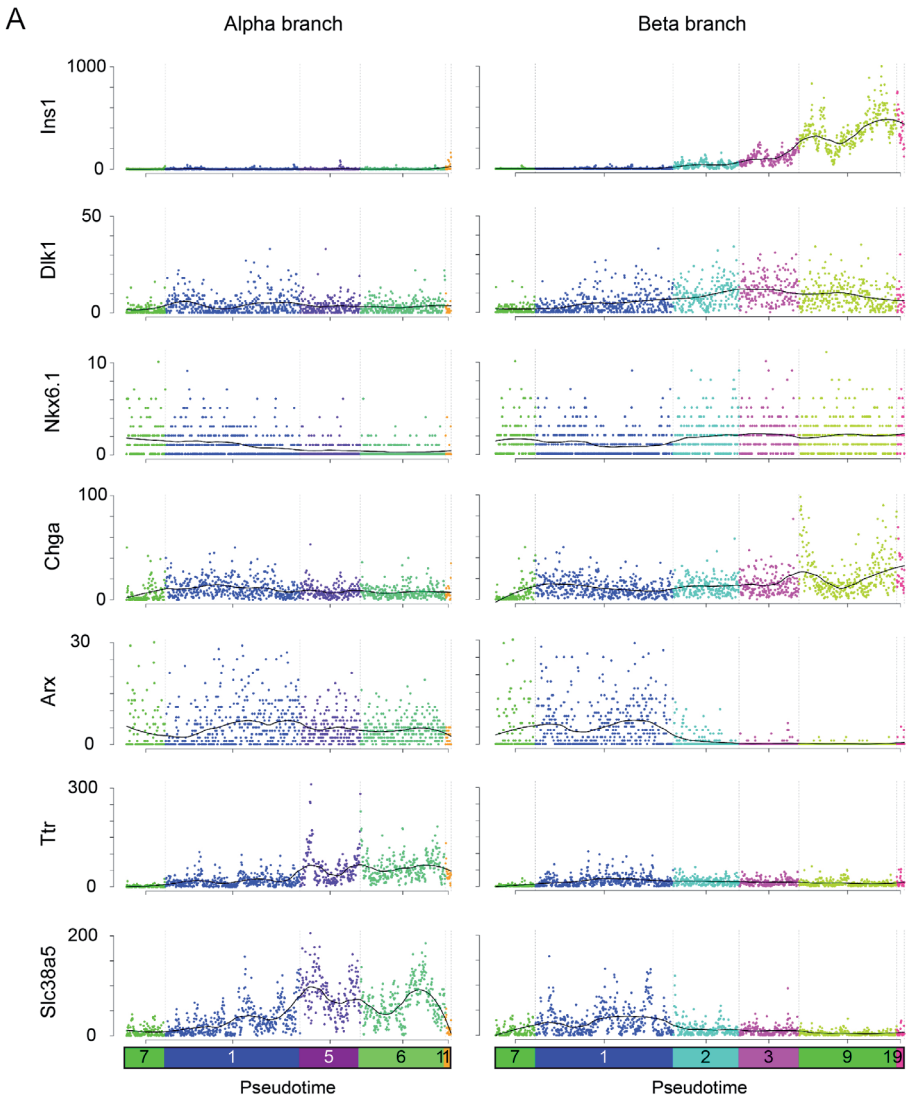


Figure S4.
Pseudo-time plots for all cells in a specific StemID branch. Left column shows clusters in the alpha branch, starting from the progenitor cluster (7) and ending in the most mature alpha cell cluster (11). Right column shows clusters in the beta branch, starting from the same progenitor cluster (7) and ending in the most mature beta cluster (16). Dots indicates cells and cells are colored according to cluster (see X-axis). Each “row” represents a different pancreatic marker gene with expression plotten on the Y-axis.

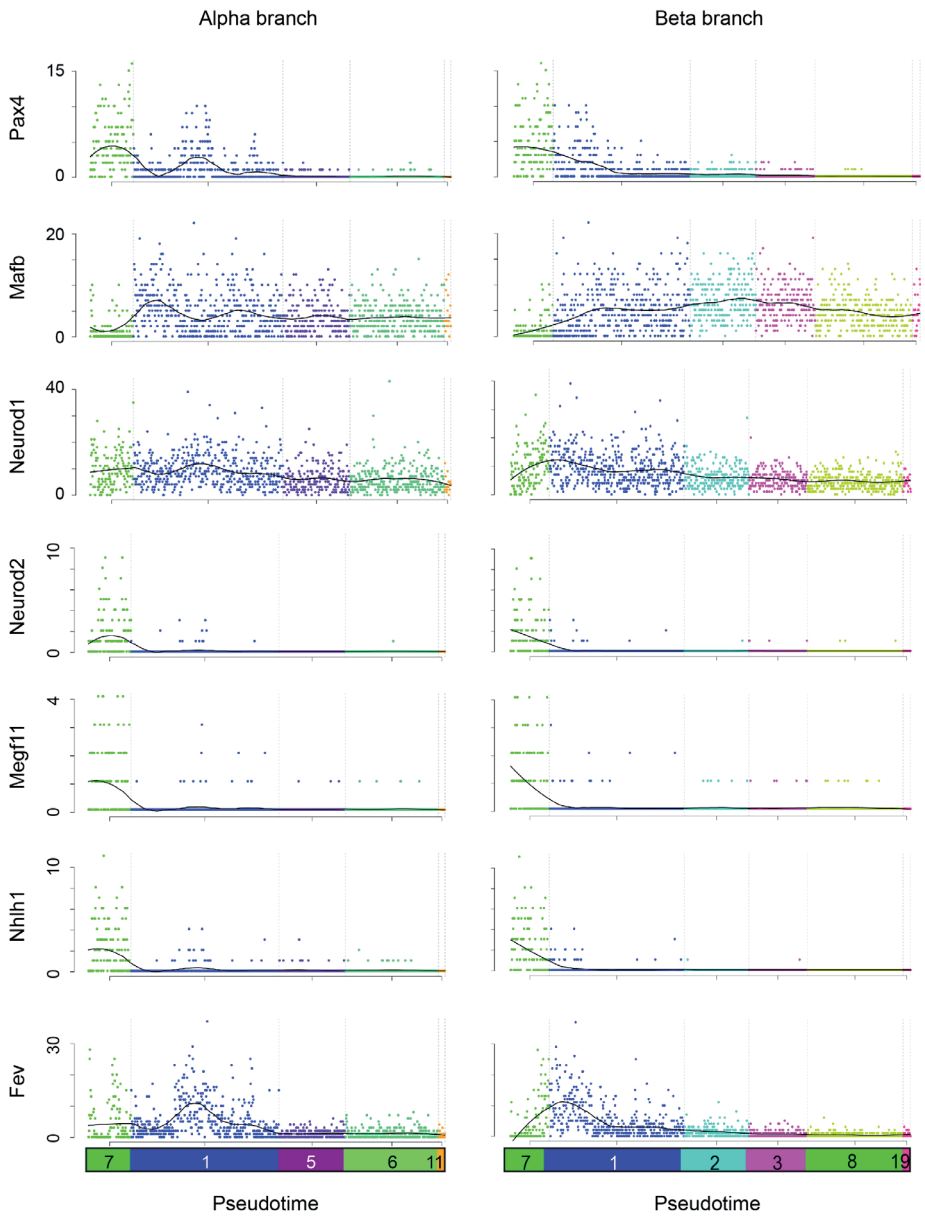
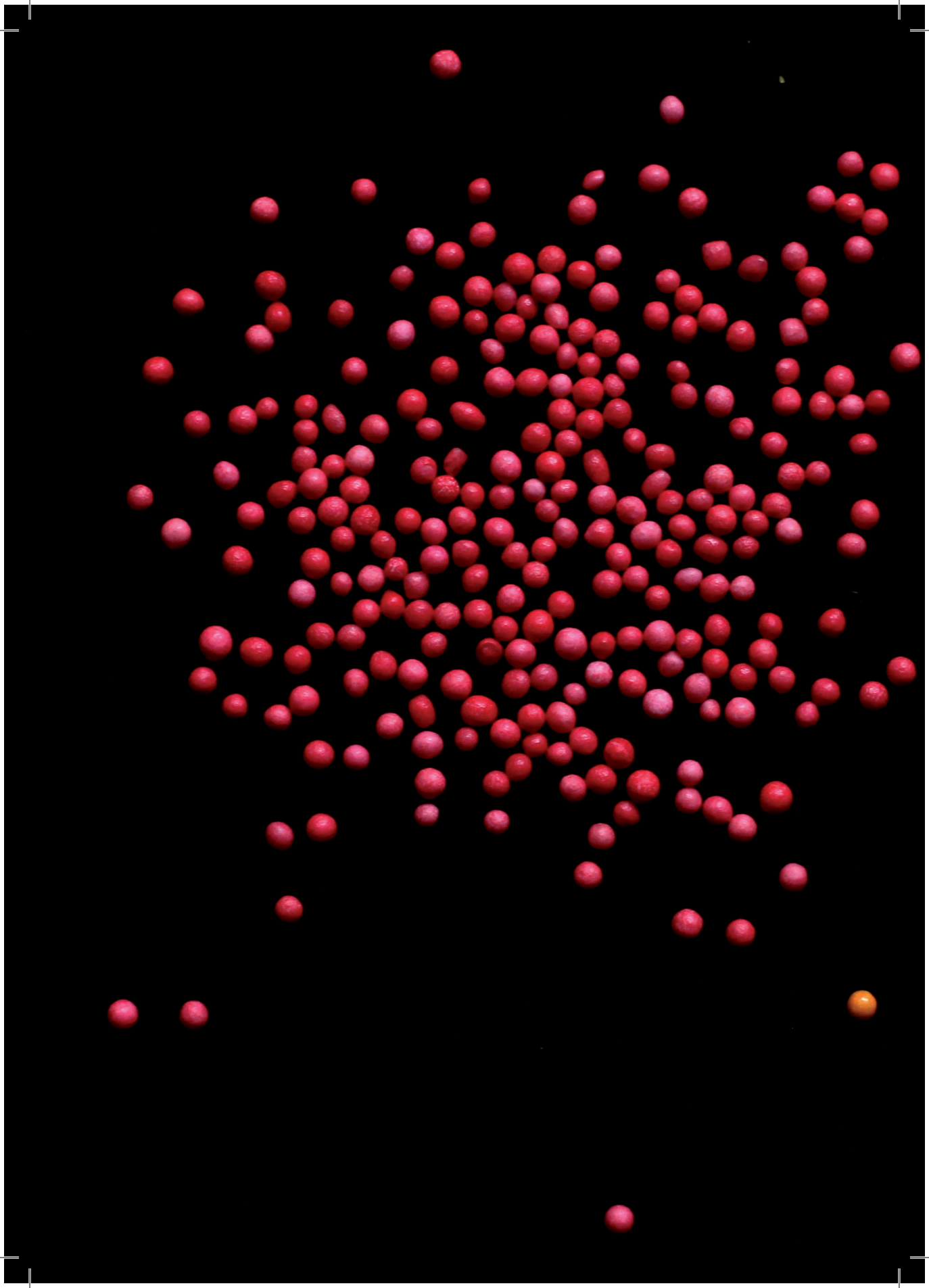


Figure S4.
Pseudo-time plots for all cells in a specific StemID branch. Left column shows clusters in the alpha branch, starting from the progenitor cluster (7) and ending in the most mature alpha cell cluster (11). Right column shows clusters in the beta branch, starting from the same progenitor cluster (7) and ending in the most mature beta cluster (16). Dots indicates cells and cells are colored according to cluster (see X-axis). Each "row" represents a different pancreatic marker gene with expression plotten on the Y-axis.



5

Cell Sorting Trained by Single-Cell Transcriptome Data Allows Cell Type Purification Without Using Fluorescent Markers

Aditya Barve[1,2,*], Chloé S Baron[1,2,*], Mauro J Muraro[1,2,*], Gitanjali Dharmadhikari[1,2], Reinier van der Linden[1,2], Eelco J. P. de Koning[1,2,3] & Alexander van Oudenaarden[1,2]

*Equal contribution

[1]Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences) Utrecht, The Netherlands.

[2]University Medical Center Utrecht Cancer Genomics Netherlands, Utrecht, The Netherlands.

[3]Leiden University Medical Center, Department of Medicine, Section of Nephrology and Section of Endocrinology, Leiden, the Netherlands.

ABSTRACT
Traditional cell type enrichment using fluorescence activated cell sorting (FACS) relies on methods that specifically label the cell type of interest. Here we propose GateID, a computational method that combines single-cell transcriptomics for unbiased cell type identification with FACS index sorting to purify cell types. We validate GateID by purifying various cell types from the zebrafish kidney marrow and the human pancreas without resorting to antibodies or transgenes.

INTRODUCTION
The ability to enrich for different cell types from heterogeneous tissues underpins much of current biological and clinical research. Methods using FACS to enrich cells use reporter transgenes or fluorescent antibodies that are specific for the cell type of interest. However, limited availability of specific antibodies or - in case of reporter constructs - the need for genetic manipulation, limit this approach. Here, we describe GateID, an optimization algorithm that combines single-cell FACS and transcriptome information with a goal to predict FACS gates for cell types that were identified in an unbiased manner by the single-cell mRNA-sequencing. It benefits from two technological breakthroughs: single-cell transcriptomics and FACS index sorting. Recent studies have demonstrated that a combining single-cell transcriptomics with index sorting can be used to improve existing sorting gates[1–3]. GateID takes this approach further by purifying cell types based on general properties such as cell size and granularity, nuclear staining, cellular proliferation, and mitochondrial activity.

RESULTS
We start with sorting single cells while recording index data followed by single-cell transcriptomics of the sorted cells using the SORT-Seq method[4] (Fig. 1, steps a-c). After assigning each cell to a specific type, we merge this information with the associated flow cytometry index sorting information. GateID takes as input this merged dataset along with the desired cell type one wishes to purify (Fig. 1, step d). At the core of GateID is an optimization algorithm that attempts to predict gates to obtain the maximum number of desired cells while minimizing the number of undesired cells. It iterates this procedure through all combinations of FACS channels and subsequently through combinations of gates to predict best gates in terms of purity and yield (see Online Methods, Fig. 1, steps e-f). To ensure that the gates change according the variability between the sample being sorted and the sample on which the gates were predicted, the gates are normalized in real-time (Fig. 1, step g). After sorting, cells are single-cell sequenced and purity is evaluated (Fig. 1, step h).

To test our method, we first decided to focus on the adult zebrafish whole kidney marrow (WKM), the primary site of production of all hematopoietic cells. Traditionally, their isolation relies on limited number of transgenic lines or manual gating subject to high variability. The efficiency of isolation is often determined by morphological analysis and/or immunohistochemistry[5,6]. We first generated a dataset of single live WKM hematopoietic cells ($DAPI^-$) using SORT-Seq and recorded FACS index data in 12 dimensions (scatter and fluorescent dimensions), resulting in 1252 cells from 3

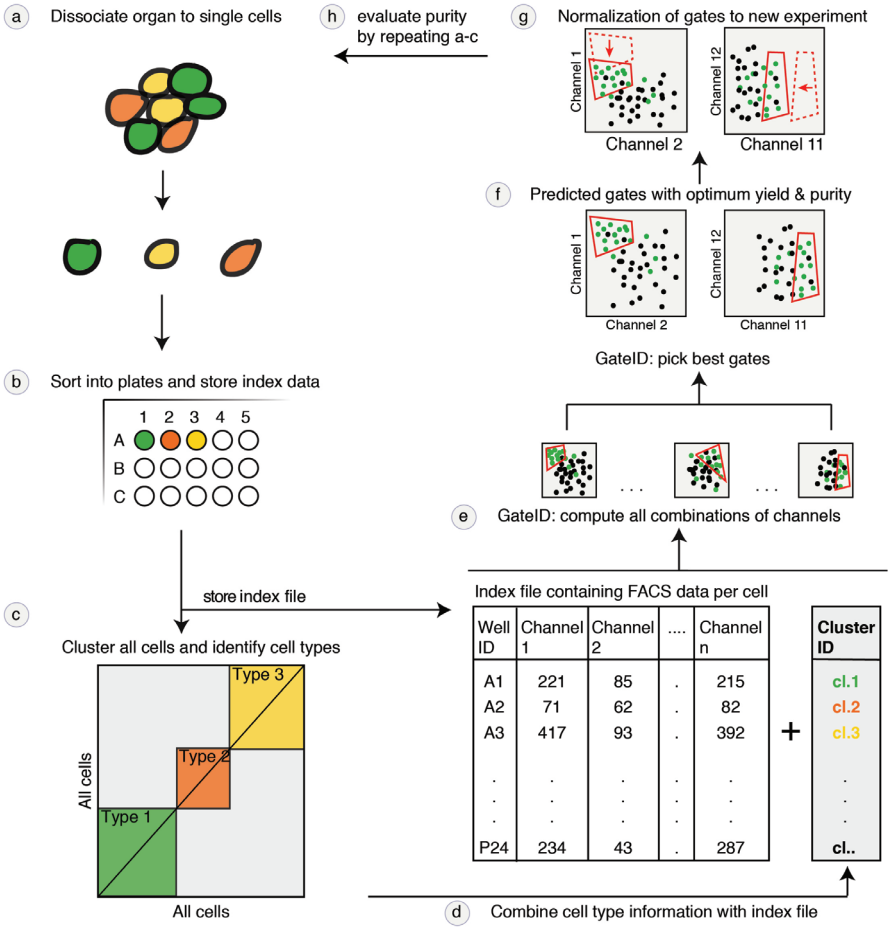


Figure 1. Workflow of GateID experiments.
(a) Dissociation of tissue into single cells just prior to FACS sorting. (b) Cells are sorted into 384 well plates and FACS single-cell index data is stored for each well. (c) Cells are clustered according to their transcriptome similarities and annotated as cell types based on the genes that label cell types expected in the tissue. (d) FACS index data is linked with cell type information. (e) GateID combines all possible channels in silico and draws gates around the desired cell type. (f) The combination of gates that gives the best combination of purity and yield is chosen. (g) After analyzing a few thousand cells, gates are normalized to for the chosen channels so the experimental gates are re-adjusted for the current sort. (h) Purity of the GateID enriched population of cells is evaluated by repeating steps a-c.

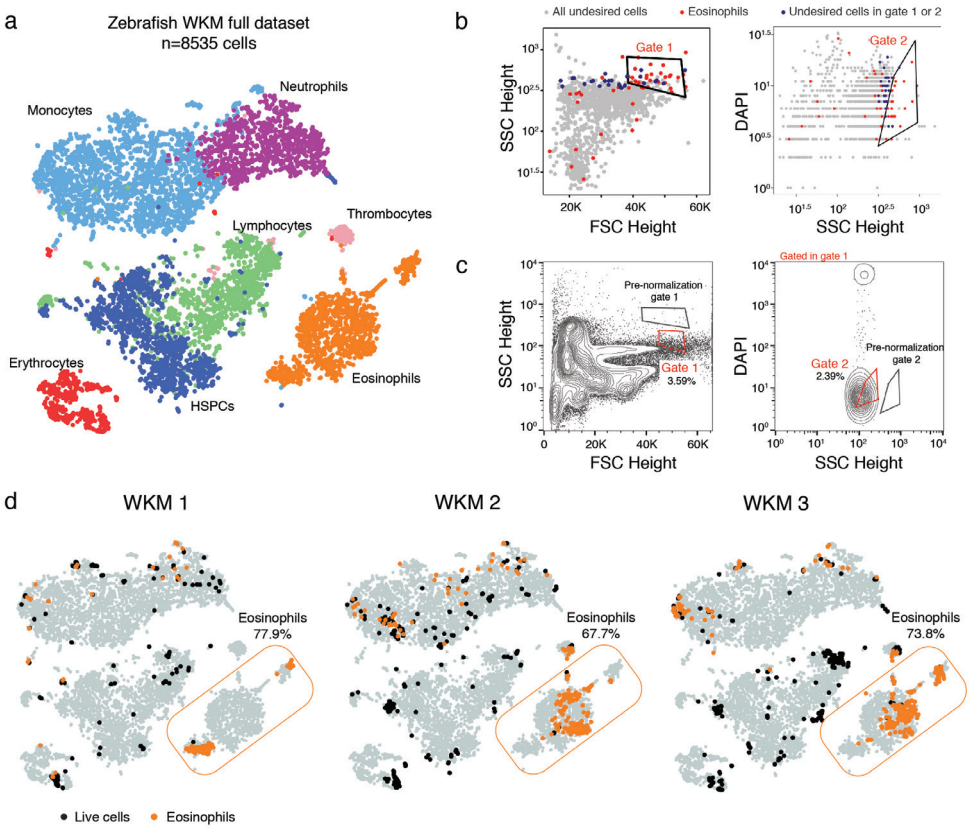


Figure 2. Zebrafish training dataset and eosinophil enrichment
(a) A t-SNE map of zebrafish WKM full dataset (training datasets and enrichment experiment datasets). Single cells are colored based on cell type identification. (b) GateID predicted gates to isolate eosinophils from unstained WKM. Gates were predicted on training dataset 1. Red points show desired cells (eosinophils) present in training dataset 1 and blue points show undesired cells falling in the other gates. A blue colored undesired cell inside a gate denotes an impure cell that will be sorted. (c) Contour plots of unstained WKM cells showing experimental sorting gates for eosinophils for WKM2 experiment (representative example of all other eosinophil enrichment experiments). Black gates represent predicted gates prior to normalization while red gates show normalized sorting gates. Sorted cells passed through gate 1 and gate 2. Percentages of events within each gate are indicated. (d) t-SNE showing eosinophil enrichments for three independent experiments. Grey points represent all cells from the WKM dataset. For each experiment, black points indicate unenriched cells and orange points indicate GateID enriched eosinophils. The eosinophil percentage in each the GateID library is indicated.

zebrafish, while excluding a large proportion of erythrocytes (Supplementary Fig. 1a). Using cell clustering and known markers, we identified 7 hematopoietic cell types[8–11] (Supplementary Fig. 1b,c, see Online Methods). We first aimed to isolate eosinophils. GateID predicted a yield of 51% and a purity of 91% to isolate eosinophils using a combination of two gates (Fig. 2b). We validated these gates by normalizing them (Fig. 2c) and sorting an enriched (GateID) and unenriched (live single cells in gate shown in Supplementary fig. 1a) cell population for comparison purposes (Fig. 2d). To ensure high confidence in our purity estimates, we clustered all zebrafish GateID experiments together resulting in 8535 cells (see Online Methods, Fig. 2a). We then evaluated the enrichment of each of our experiments based on this data set. The above-mentioned experiment to enrich eosinophils achieved an experimental purity between 67.7% and 77.9% purity (Fig. 2d) even with as low eosinophil content as 0.7% in the unenriched population (Supplementary Fig. 2a). The reason the experimental purity is lower than predicted lies in the individual variation that can be observed in both cell type composition and in FACS measurements per experiment (Supplementary Fig. 2b), which clearly shows variation in the distribution of FACS measurements. Supplementary fig. 2c shows that contaminating cells (black points) intermingle with enriched eosinophils (orange cells) and are thus difficult to eliminate. The contaminating population in all experiments consisted mainly of myeloid cells. This is not surprising, since eosinophils and myeloid cells occupy partly overlapping FACS regions[12]. Importantly, the enriched eosinophils from each experiment clustered with the unenriched population (Fig. 2d, black and orange points respectively). This shows that GateID does not bias for a subpopulation of eosinophils. Finally, to compare GateID to manual gating, we isolated eosinophils as described earlier[12] (Supplementary Fig. 2d). This manual gating yielded lower enrichment compared to GateID and revealed a strong myeloid contamination (Supplementary Fig. 3e).

We next aimed to isolate additional hematopoietic cell types from the WKM. Our initial dataset (Supplementary Fig. 1b) was obtained using $DAPI^-$ WKM cells with a limited number of cells per individual. While this dataset was sufficient to enrich gates for eosinophils, GateID was unable to predict gates with satisfying purity and yield for HSPCs, lymphocytes or myeloid cells (Supplementary fig 1d). Each curve in the figure represents the results of the best combination of gates for each cell type, highlighting the trade-off between increasing purity at the cost of yield. For brevity's sake, all yield vs. purity values for gates with an inferior trade-off and internal to these curves are not shown. Supplementary fig. 1d clearly shows that we required another approach that would allows us to purify these cell types. To overcome this challenge, we hypothesized that staining with generic cellular dyes would be an easy and reliable approach to enhance cell type separation in FACS measurement space. We chose MitoTracker, a fluorescent die that reflects mitochondrial abundance and activity, and CFSE, which binds to cytoplasmic proteins (see Online Methods). To characterize the effect of staining, we used another zebrafish WKM and split it in two parts, one of which was stained with MitoTracker and CFSE, while the other was stained only with DAPI and index sorted. After identifying the relevant cell types based on the transcriptome, we evaluated all two gate combinations for the enrichment of each cell type in $MitoTracker^+$ $CFSE^+$ (referred to as "stained"

here onwards) and DAPI$^-$ (referred to as “unstained” here onwards) samples. Fig. 3a shows the result of purity (y-axis) vs. yield (x-axis). Results on the unstained samples (dashed lines) while often reaching high purity have lower yield compared to the stained samples (solid lines). It is important to remember that these samples are limited in the number of cells (~250 for each sample) and do not represent possible contaminating cells that could be observed in a larger sample, thereby predicting higher purities than would be obtained experimentally or with a larger dataset.

Fig. 3a thus shows that MitoTracker and CFSE were indeed able to increase distances between cell populations. We therefore generated a larger dataset of 1201 stained WKM hematopoietic cells from one zebrafish (Supplementary Fig. 3a). Using this new dataset, GateID predicted a yield of 20% and a maximum purity of 90.5% to isolate HSPCs using a combination of two gates, one of them using the MitoTracker fluorescent channel (Supplementary Fig. 3b, Supplementary table 1). Experimentally, we were able to enrich HSPCs to purities above 95% (Fig. 3bc, n=2, 96.4% and 96.9% purity) wherein enriched HSPCs clustered together with the unenriched HSPC population for each experiment. Not surprisingly, Supplementary fig. 3c shows that the projection of GateID enriched HSPCs on the classical dimensions of FSC height and SSC height tallies with what is published (Fig1 in ref 12). To benchmark GateID, we compared it to a classical method of enriching HSPCs based on their low expression of cd41 (Supplementary Fig. 3ef)[14,15]. Enriched HSPCs from the cd41^{low} fraction from cd41-EGFP transgenic zebrafish yielded an inferior purity compared to GateID predicted gates (Supplementary Fig. 3fg). Surprisingly, the enriched HSPCs were contaminated by lyz$^+$ myeloid cells. Contamination by this population was minimal using GateID because of the high distance in FACS space of HSPCs and myeloid cells. This result suggested that lyz$^+$ myeloid cells reside partially in the cd41^{low} WKM fraction, an observation that would have gone undetected without the combination of single-cell FACS and transcriptome information.

Next, we used GateID to isolate lymphocytes, using four FACS dimensions including the CFSE fluorescent channel (Supplementary Fig. 4a-c, Supplementary table 1). Experimentally we obtained unbiased enrichment between 74.5% and 91.8% (Fig. 3b-e, Supplementary Fig 4c, n=4). In silico, we tested the efficiency of lymphocyte manual gating as lymphocytes are characterized by their small FSC height and SSC height properties (Supplementary fig. 4b). The manual gate yielded 60.9% purity and exhibited HSPC contamination (Supplementary Fig. 4de).

We then challenged GateID to isolate a subset of myeloid cells. lyz$^+$ and mmp$^+$ myeloid cells are strongly intermingled in a large cloud containing myelomonocytes in side scatter height vs. forward scatter height[12]. However, one of our predicted gates used the CFSE dimension (Supplementary Fig. 5a, Supplementary table 1. We succeeded in enriching this particular subset to purities between 55.2 and 86.5% (Fig. 3e-g, n=3). We find the enriched population to overlap with the one present in the live population in t-SNE space while the highest source of contamination was lyz$^+$ myeloid cells (Supplementary Fig. 5c), and note that is it quite difficult to separate

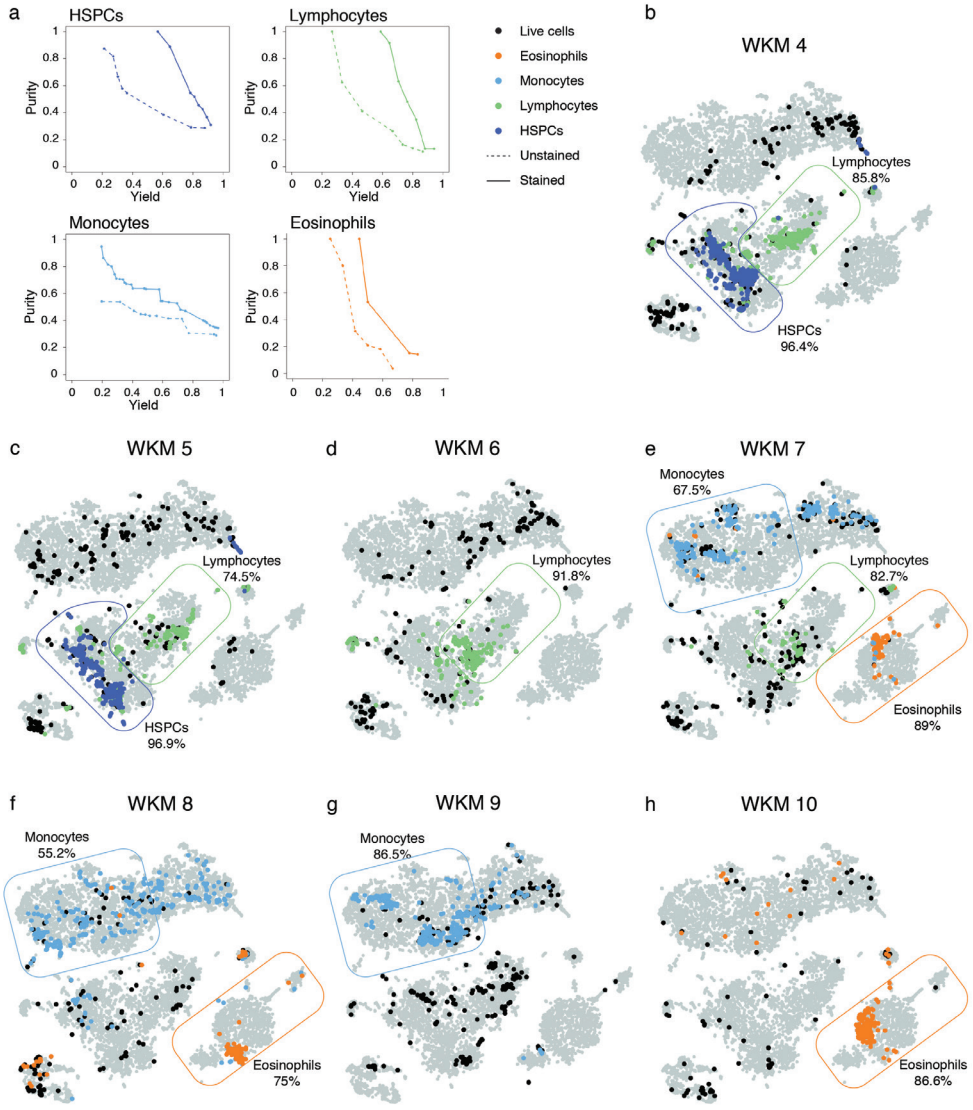


Figure 3. Further enrichment of zebrafish celltypes
(a) Curves showing trade-off between yield and purity of GateID solutions for HSPCS, lymphocytes, monocytes and eosinophils on a stained (solid line) and unstained (dashed line) dataset. (b-h) t-SNE showing cell type enrichments for seven independent experiments. The number of cell types enriched per WKM varies from one (e.g. WKM 6) to three (e.g. WKM 7). Grey points represent all cells from the WKM dataset. For each experiment, black points indicate unenriched cells and colored points indicate GateID enriched libraries. The corresponding cell type percentage in the GateID library is indicated.

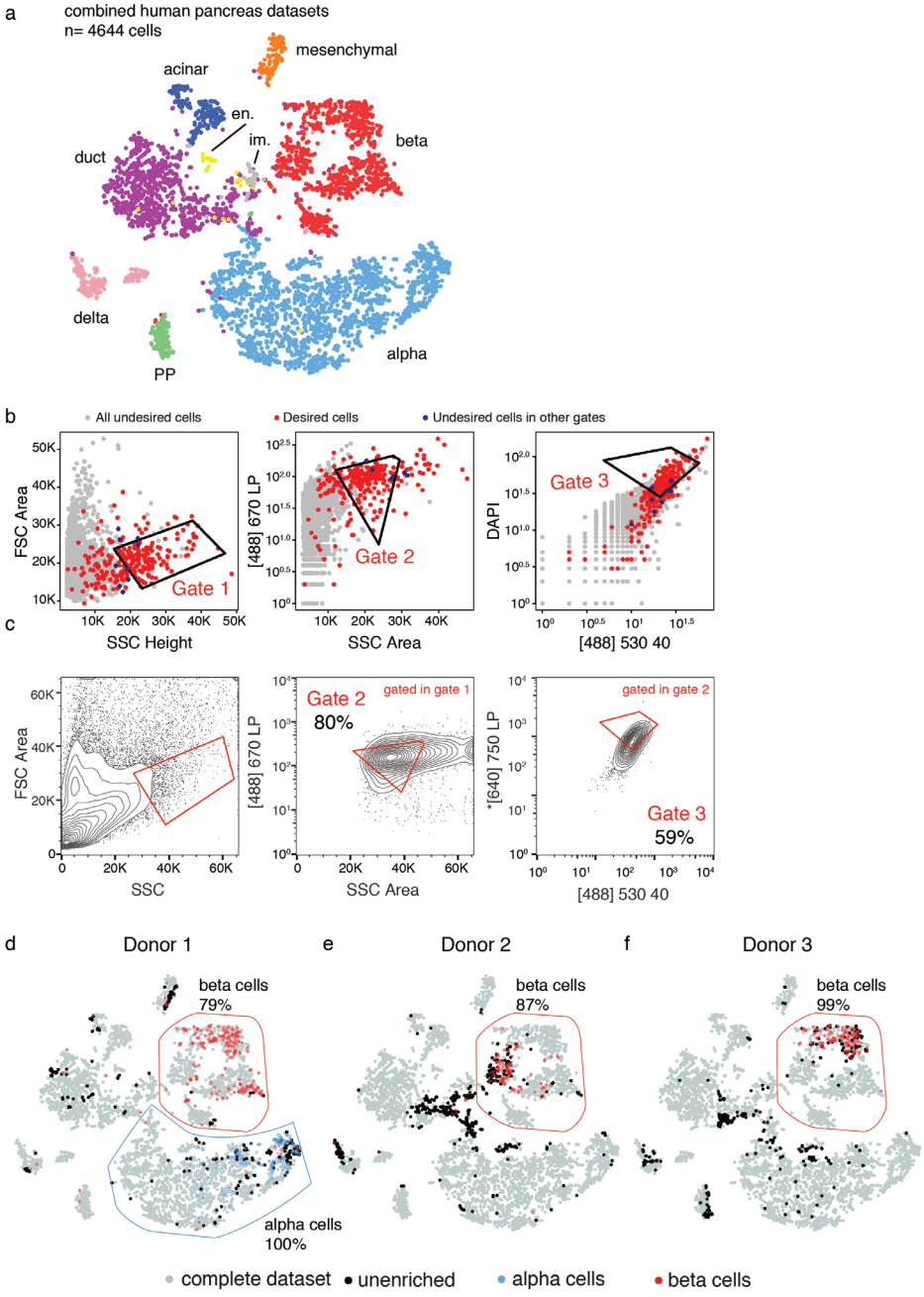


Figure 4
(a) t-SNE map of all human pancreatic data used in this study (training datasets and GateID enrichment datasets). Single cells are colored based on cell type identification. (b) GateID predicted gates to isolate beta cells from unstained human pancreas. Gates were predicted on the second training dataset. Red points show desired cells (beta cells) present in training dataset and blue points show undesired cells falling in the other gates. (c) Contour plots of unstained human pancreas cells showing experimental sorting gates based on (c) for beta cells in donor 2 (representative example both donor 2 and donor 3). Sorted cells passed through gate 1 AND gate 2 AND gate 3. Percentages of events within each gate are indicated. (d) t-SNE map showing enrichment for alpha cells (blue) and beta cells (red) from donor 1 as based on training dataset 1. Unenriched cells from the same donor are shown in black. (e-f) t-SNE maps showing enrichment for beta cells (red) from donors 2 and 3 as based on training dataset 2. Unenriched cells from the same donor are shown in black.

lyz+ cells from mmp+ cells. Finally, for sake of completeness, we repeated the eosinophil enrichment using GateID on stained WKM cells and obtained marginally higher purities (83.5% on average) when compared to unstained cells (73.13% on average) (Supplementary Fig. 6a-c, n=3). Overall, we demonstrate GateID's ability to enrich multiple zebrafish hematopoietic cells to high purity solely using generic dyes. GateID proved more robust for cell types we enrich here when compared to the tested manual gating strategies that use FACS scatter properties or fluorescent transgenic lines.

We next set out to apply GateID in a human, clinical setting. We and others previously sequenced single cells from islets of Langerhans obtained from human cadaveric material (reviewed in [16]) to describe the transcriptomes of the 6 major pancreatic cell types (alpha, beta, delta, PP, acinar and ductal cells) implicated in the pathogenesis of diabetes. Unfortunately, no reliable markers exist that can select for populations of alpha, beta and delta cells to high purity. Previous efforts[17] in obtaining enriched populations of alpha and beta cells by using antibodies are unclear, as delta cell markers are found in the enriched beta cell population, indicating a strong contamination from delta cells (table 1 in [17]). We thus set out to enrich alpha and beta cells to high purity from DAPI stained human pancreas using GateID. For the human pancreas, we used two distinct datasets and used GateID to enrich alpha and beta cells from 1 human donor and beta cells from two additional donors. All the cells obtained in these experiments are were clustered together in Fig 4a. We identified cell types based on the markers as described previously[4]. We initially used one of the donors from our previous dataset[4] (D30) as a training dataset (Supplementary Fig. 7a, called training dataset 1, n=678 cells). Upon sorting with GateID predicted gates (Supplementary Fig. 7c-d), we obtained an 100% pure alpha population and a 79% pure beta cell population (Fig.4a and Supplementary Fig. 7d). Some batch effects between different donors are clear by plotting the beta cells from different GateID sorts on the t-SNE map containing all cells, but it is important to note that GateID itself did not introduce a bias towards any specific subpopulations. This is clear from overlaying the GateID enriched cells over the cells of the same type from the unenriched fraction (Figure 4d). The contamination in the beta cell gate stemmed from several different cell types present in the dataset. Since GateID had predicted gates of 100% purity, we hypothesized that the cause of this contamination stemmed from the relatively low number of cells in the training dataset. To test this hypothesis, we built a larger training dataset of 2255 cells (Supplementary Fig. 7b, training dataset 2), where we again could identify all pancreatic cell types. GateID predicted gates of 25.83 % yield and 99% purity for beta cells (Fig. 4b-c). We used these GateID gates to enrich for beta cells and tested them on pancreatic material from two additional human donors (Fig. 4c shows a representative example of the gates drawn on FACS scatter plots from donor 3). For both donors, we obtained a high purity of beta cells (87 and 99%, Fig. 4e-f). GateID enriched beta cells did not separate from live beta cells from the unenriched fraction in t-SNE space. These results show that GateID can faithfully predict gates for both alpha and beta cells from the human pancreas, allowing us to purify these cell types to high purity for the first time.

In short, we have described a novel computational method that combines single-cell transcriptomic and FACS data to predict FACS gates that allow cell type enrichment without the aid of transgenes or antibodies. To demonstrate the effectiveness of GateID, we enriched four major hematopoietic cell types from the zebrafish WKM, a tissue for which transgenes labelling specific cell types are labor intensive to generate and antibodies are limited. Our approach is sufficiently robust to enrich for rare cell types like eosinophils, a cell type that accounts for 1~5% of zebrafish hematopoietic cells (Fig. 1e, Supplementary Fig. 7b). GateID also does better than classical methods of cell type enrichment as shown for examples concerning eosinophils, HSPCs and lymphocytes (Supplementary figs. 2d-e, 3e-g and 4d-e). Our approach also allows purification of more than one cell type from one animal as shown by purifying eosinophils, lymphocytes and monocytes from WKM7 (Fig. 3e). We showed that GateID can also enrich unlabelled human alpha and beta cells from the islets of Langerhans. This is especially important for human tissues where purification of cell types is completely restricted to the (limited) availability of antibodies. It is important to highlight the role of variability in each sample in comparison to training data and its effects on enrichment using GateID (Fig.2a,c, and Supplementary Fig. 4c). Variability can spring from the experimental protocol, variable proportion of input cell types and different statistical properties for each cell type in each sample. Such variability is often hard to predict and take into account[18]. However, GateID offers different normalization strategies that can handle sample to sample variation (see Online Methods). We envisage a broad application of GateID to make purification of any given cell type easier and to allow enrichment of cell types never isolated before.

REFERENCES

1. Paul, F. et al. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell 163, 1663–1677 (2015).

2. Schulte, R. et al. Index sorting resolves heterogeneous murine hematopoietic stem cell populations. Exp. Hematol. 43, 803–811 (2015).

3. Wilson, N. K. et al. Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. Cell Stem Cell 16, 712–24 (2015).

4. Muraro, M. J. et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst. 3, 385–394.e3 (2016).

5. Traver, D. et al. Transplantation and in vivo imaging of multilineage engraftment in zebrafish bloodless mutants. Nat. Immunol. 4, 1238–46 (2003).

6. Wittamer, V., Bertrand, J. Y., Gutschow, P. W. & Traver, D. Characterization of the mononuclear phagocyte system in zebrafish. Blood 117, 7126–7135 (2011).

7. Grün, D. et al. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. Cell Stem Cell 19, 266–277 (2016).

8. Kobayashi, I. et al. Comparative gene expression analysis of zebrafish and mammals identifies common regulators in hematopoietic stem cells. Blood 115, e1–e9 (2010).

9. Moore, F. E. et al. Single-cell transcriptional analysis of normal, aberrant, and malignant hematopoiesis in zebrafish. J. Exp. Med. 213, 979–92 (2016).

10. Macaulay, I. C. et al. Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. Cell Rep. 14, 966–977 (2016).

11. Carmona, S. J. et al. Single-cell transcriptome analysis of fish immune cells provides insight into the evolution of vertebrate immune cell types. Genome Res. 27, 451–461 (2017).

12. Balla, K. M. et al. Eosinophils in the zebrafish: prospective isolation, characterization, and eosinophilia induction by helminth determinants. Blood 116, 3944–54 (2010).

13. Fukunaga, K. Introduction to statistical pattern recognition. (Academic Press, 1990).

14. Ma, D., Zhang, J., Lin, H. -f., Italiano, J. & Handin, R. I. The identification and characterization of zebrafish hematopoietic stem cells. Blood 118, 289–297 (2011).

15. Bertrand, J. Y., Kim, A. D., Teng, S. & Traver, D. CD41+ cmyb+ precursors colonize the zebrafish pronephros by a novel migration route to initiate adult hematopoiesis. Development 135, 1853–62 (2008).

16. Carrano, A. C., Mulas, F., Zeng, C. & Sander, M. Interrogating islets in health and disease with single-cell technologies. Mol. Metab. 6, 991–1001 (2017).

17. Dorrell, C. et al. Transcriptomes of the major human pancreatic cell types. Diabetologia 54, 2832–44 (2011).

18. Dataset Shift in Machine Learning. (The MIT Press, 2008). doi:10.7551/mitpress/9780262170055.001.0001

19. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. Nat. Methods 11, 637–640 (2014).

Methods

Tissue isolation

The WKM of WT and cd41-GFP zebrafish were isolated as described previously[1]. Briefly, after a ventral midline incision the internal organs were removed. The kidney was carefully dissected and collected in PBS supplemented with FCS. To mechanically dissociate the single hematopoietic cells, the tissue was passed multiple times through a 1 ml low-bind pipet tip. The cells were filtered (70um and 40um cell strainers (VWR)) and washed. The pellet of hematopoietic cells was resuspended in PBS/FCS supplemented with DAPI (dilution 1/2000, Thermo Fisher) to assess cell viability. In case of staining, the pellet of hematopoietic cells was resuspended in PBS/FCS supplemented with both MitoTracker and CFSE (dilution 1/4000) and incubated at room temperature for 10 minutes. Cells were washed and resuspended in PBS/FCS supplemented with DAPI as described above. $DAPI^-$ single cells were sorted (BD FACSJazz™) and erythrocytes with low forward and side scatter were excluded as described in Supplementary Fig. 1a). Human pancreas isolation was done as described previously[2].

Single-Cell mRNA Sequencing of Single Cells

We used SORT-seq[2] to sequence the transcriptome from single cells and store FACS information from single cells (index files). All sorts were carried out using BD FACSJazz™. Unless mentioned otherwise, we used the following protocol for both model systems mentioned in this study. We lysed cells by incubating them at 65°C for 5 minutes, and then used Nanodrop II liquid handling platform (GC biotech) to dispense RT and second strand mixes. The aqueous phase was separated from the oil phase after pooling all cells into one library, followed by IVT transcription. The CEL-Seq2 protocol was used for library prep[3]. Primers consisted of a 24 bp polyT stretch, a 4 or 6bp random molecular barcode (UMI), a cell-specific 8bp barcode, the 5′ Illumina TruSeq small RNA kit adaptor and a T7 promoter. We used TruSeq small RNA primers (Illumina) for preparation of Illumina sequencing libraries and then paired-end sequenced them at 75 bp read length using Illumina NextSeq at approximately 45 million and 30 million reads for zebrafish kidney marrow and human pancreatic libraries respectively.

Data analysis

Zebrafish WKM and human pancreas were analysed separately as follows. For each model system we analysed, paired-end reads were aligned to the transcriptome of that model system using BWA[4]. We used Read 1 for assigning reads to correct cells and libraries, while read 2 was mapped to gene models. Only reads mapping to unique locations were kept. We corrected read counts for UMI barcodes by removing duplicate reads that had identical combinations of library, cellular, and molecular barcodes and were mapped to the same gene. Transcripts were counted using 256 UMI barcodes for the human pancreas (donor 1) and 4096 UMI barcodes for the other human donors and the zebrafish kidney. The counts were then adjusted using Poissonian counting statistics to yield the expected number of molecules as described in.

Data was normalized by median normalization to a minimum number of 1000 transcripts and genes expressing at least three transcripts in at least two cells were retained for zebrafish WKM. Pancreatic data was median normalized to 4000 transcripts and only genes expressing 5 transcripts in at least 3 cells were retained for downstream analysis. We then computed the Pearson's distance (1 - p) between cells. To cluster cells, we used a method previously published in ref. 5. Briefly, we used hierarchical clustering ('hclust' R function with 'ward.D2' method) to cluster cells. To identify the number of clusters, we used 'cutreeDynamic' along with the 'hybrid' method which allows the user to specify a 'deepSplit' parameter controlling the sensitivity of clustering. We evaluated 100 subsamples of our data by randomly selecting 90% of the genes in the dataset, specifying the 'deepSplit' parameter as an integer from 0 to 4 and evaluating the average silhouette width of the number of clusters. This procedure resulted in identifying the correct cell types for both data sets of the zebrafish WKM data and and the pancreatic data.

While evaluating the results of our enrichment experiments, we clustered all data together to ensure maximum confidence in resulting purity estimates. For zebrafish, this involved clustering both training datasets and enrichment experiments (WKM 1-10) resulting in 8535 cells in all. For the pancreas data, clustering both training data sets and data from three donors resulted in a total of 4644 cells.

Differentially expressed genes between two subgroups of cells were identified similar to a previously published method[6]. Briefly, we started by modelling the background expected transcript count variability. We then identified genes in each subgroup that were variably expressed by representing gene expression of each gene as a negative binomial distribution. We then computed Benjamini-Hochberg corrected p-values for the observed difference in transcript counts between the two subgroups as described earlier[7] and identified differentially expressed genes (adjusted p-value < 0.01). Such genes were then used to annotate specific cell types within each model system based on known published literature.

For the zebrafish WKM data, we selected the the topmost ten genes for each cell type ordered by their log fold change in expression when comparing the gene's expression in a specific cell cluster compared to other cell clusters taken together (Supplementary figure 1c). Some known marker genes, especially for HSPCs and lymphocytes do not make the top ten list. We manually added them to our list of differentially expressed genes. We then used hierarchical clustering to cluster genes in seven clusters (one for each cell type). We found that our manually added genes, namely, meis1b, myb (denoting HSPCs[8]) and pax5, cd79b (denoting lymphocytes[8]) clustered in the appropriate clusters and do not show expression elsewhere (Supplementary figure 1c).

Gate prediction methodology

The goal of GateID is to predict gates towards sorting a desired cell type from a mixture of multiple cell types. In other words, we want to purify a specific cell type to maximum purity while sorting a sufficient fraction of the desired cells. Recent advances in flow cytometry allow users to index sort, which is to save and associate

flow cytometry readouts pertinent to each sorted cell. After performing single-cell mRNA sequencing, one can then merge this information with the cell type annotation (Fig. 1d) for each cell. Such a merged data set forms the starting point for GateID, and we refer to it as training data.

We treat gate prediction as an optimization problem, wherein predicted gates should allow a minimal number of undesired cells while maximizing the number of desired cells. The algorithm takes as input a matrix with FACS measurements and cell type annotation for each cell. It then prompts the desired cell type and the minimum yield required by the user. Yield is defined as a percentage of desired cells (of the total number of desired cells) that are predicted to pass through the predicted gates. Because we attempt to enrich cells without using cell-specific antibodies or transgenes, our predicted flow cytometry gates are in two dimensions as opposed to univariate histograms. GateID first predicts a gate for each pair of flow cytometer channels, comprising scatter and fluorescence channels, where each gate is represented as a polygon with four vertices. The starting gate is computed by setting its vertices to represent the 2nd and 98th percentile in each of the x and y axis and functions as the starting point for the optimization algorithm. We use a two-step optimization as follows for the prediction of a gate -

1. The first step finds a gate that contains at least the user-specified minimum yield for desired cells while minimizing the number of undesired cells in the gate. Fitness of each solution is thus defined by the number of undesired cells in the gate. The highest fitness is the complete absence of undesired cells within the gate. The requirement of minimum yield is enforced by assigning the worst fitness (equivalent to the total number of undesired cells in the data set) to a solution not adhering to this constraint.
2. The second step takes as input the solution (gate) of the first step and tries to maximize the yield while disallowing an increase in undesired cells. Fitness in this step is thus defined as the number of desired cells within the gate. Best fitness is achieved when all desired cells are sorted by the gate. The requirement of maximum number of undesired cells is enforced by assigning the worst fitness of zero yield to a solution not adhering to the constraint.

By default, each step is run for 20000 iterations. While evaluating fitness at each iteration, we only allow solutions involving convex polygons thereby dismissing non-convex shapes that may result in overfitting on the training data.

Once gates for each pair of FACS channels are predicted, gate combinations can then be evaluated in logical conjunction (AND combination) such as all combinations of two gates, all combinations of three gates or a higher order. For example, experiments in this study were carried out on BD FACSJazz™, which records cytometry readouts in twelve channels, six scatter and six fluorescence channels. There are thus C(12,2) = 66 channel pairs and 66 gates. 66 gates can be further combined to yield 2145 pairwise gate combinations (C(66,2)) evaluated in an AND configuration, meaning a cell has to pass through both gates to be sorted. However, optimizing each gate separately and combining them later could be thought of as inferior to optimizing the same gates together. This is because

optimization together allows an increase in yield while reducing impurity in a coordinated fashion. While optimizing all 2145 pairwise gates is possible, the number quickly explodes thereafter to 45760 (combinations of 3 gates) and 720720 for 4 gate combinations.

This leads us to a more intuitive approach of recursive gating: once gates for each combination of FACS channels are predicted (66 gates in this study), the best gate in terms of purity is selected. This gate is paired with each other gate and re-optimised together. This process is repeated until 100% purity is reached, no overall improvement is observed in the subsequent iteration or if the number of gates exceeds a user-defined preset limit.

Even if there are differences in methods mentioned above, both approaches predict gates that are comparable in yield and purity, demonstrated by experiments enriching eosinophils from the unstained sample (Fig 2b, d), wherein the first method was used versus experiments enriching eosinophils from the stained sample, where the recursive method was used (Supplementary fig. 6). Both experiments yielded similar purities for enrichment of eosinophils.

As stated above, the objective function of the optimization procedure is to predict gates that allow a minimal number of undesired cells while maximizing the number of desired cells. This presents a discrete surface for optimization. In addition, single-cell RNA-seq along with flow cytometry results in a limited number of cells, wherein the complete variance of each cell type population may not be captured sufficiently, especially for rarer cell populations. To address these problems, we chose a derivative-free, fast and robust optimization algorithm called MA-LS-Chains, which combines an evolutionary algorithm along with a local search and is available as an R package (Rmalschains[9]). Such algorithms are known to converge faster and more reliably without being trapped in local optima (references within [9]). While theoretically any robust global optimization algorithm may suffice, a comparison with other algorithms (Supplementary Fig. 8, and see below) shows that MA-LS-Chains is both fast and optimizes to the best purity. This is not surprising in the light of the “no free lunch” theorems, which state that certain optimization algorithms may do better than others for a certain kind of problem[10].

The procedure above states in brief how gates are predicted. However, every sorted biological sample is different owing to multiple sources of variability, for example, variability is introduced during tissue isolation and subsequent sorting. An added layer of variability springs from fluctuating proportion of each cell type per isolation and variability in the statistical properties for each cell type in FACS space. For instance, the inconsistency in the proportion of each cell type can be readily observed by comparing the unenriched barplots in Supplementary figures 2c and 4c for the zebrafish, and Supplementary fig. 6d for human pancreas. Such inconsistency is further exacerbated by an overall shift in the distribution of all points demonstrated in Supplementary Fig. 2a (WKM1-3). For example, the distribution of high forward scatter height changes from a maximum of ~500 (WKM1) to ~100 (WKM2 and WKM3). Such variability requires that GateID predicted gates also change with respect to the current sort in real time (Fig 1g).

The first approach to deal with such variability is to standardize the values for the vertices of gates to the unenriched population of cells of the training data set using z-normalization. During the enrichment sort, one can analyse sufficient events (~10000 events) and use the mean and the standard deviation of the population of the current sort to normalize gates using the reverse of z-normalization procedure. Another method for gate normalization is elaborate and requires machine learning. Briefly, one first trains a machine learning classifier to classify the desired cell type based on the training data. The current sort, however, could have a different overall distribution of points, different cell type proportion therefore changing the statistical variance in different dimensions. This is referred to as data shift in machine learning approaches and is known to create a problem for classifiers[11]. Thus, data from the current sort needs to be normalized to the target distribution of the training data for each FACS channel. To do this, we use the non-linear qspline normalization[12] used to compare different microarray chips to each other. Once the new data is normalized in this fashion, we classify the cells therein as desired and undesired cells using the trained classifier. We next z-normalize predicted gates to the desired cells from our training data and renormalize them to the predicted desired cells in the new data from the current sort. Because we again use the mean and standard deviation of predicted cells to re-normalize our gates, we note that the prediction of the desired cells in the data from the current sort is perhaps of less importance than the classifier identifying a tight cluster of desired cells in the correct region of FACS space. This approach accounts for high variability in cell-type proportions from experiment to experiment, as opposed to the reverse z-normalization strategy on the complete set of points, which accounts for overall variabilities in the distribution of the whole data set.

Gate prediction for zebrafish WKM and Human pancreatic alpha and beta cells
To predict gates for eosinophils from the unstained zebrafish WKM, we used GateID to optimize gates on each of the pairs of FACS channels (66 gates) and then computed the best combination of two gates in an AND combination (Supplementary table 1). Gates were normalized using the mean normalization method for each of the eosinophil sorts from the unstained WKM. For all experiments concerning hematopoietic cell types in the stained WKM (HSPCs, lymphocytes, mmp+ myeloid cells and eosinophils), we used the recursive gating method. The recursive gating for these cell types predicted two gates, which were used for their respective enrichment. As one can observe from the eosinophil enrichment, both methods yielded experimentally similar results (Fig 2d and Fig. 3e, f, h). Gates were normalized for each sort using the machine learning based normalization and the qspline normalization method.

For alpha and beta cells from the human islets of Langerhans, we optimized gates for each of the pairs of FACS channels (66 gates) and computed the best combination of gates in AND configuration. Gates for alpha cell and beta cells predicted from the smaller training data (d30, Supplementary fig. 7a) were normalized using the mean normalization relying on the whole population of cells, as were the beta cell gates for the second donor, based on the second training dataset (Supplementary fig. 7b). To compare normalization methods, beta cells from the third donor were

normalized using both the mean normalization and the qspline normalization method. This resulted in high purity beta cell enrichment (96 and 99% resp.) for both methods (supplementary fig. 7e and 8c)

Comparison of different optimization algorithms

Different optimization algorithms may perform variably for different optimization tasks. To check if our choice of using MA-LS-Chains was indeed the best, we evaluated eight different optimization algorithms (Supplementary Fig. 8). These were controlled random search (CRS, R package: nloptr[13,14]), continuous genetic algorithm (GA, R package: GA[15]), MA-LS-Chains (R package: Rmalschains[9]), bounded Hooke-Jeeves (HJK, R package: dfoptim[16]), bounded Nelder-Mead (NMK, R package: dfoptim[16]), simulated annealing (SA, R package: GenSA[17]), DEoptim (R package: RcppDE[18]), bound optimization with quadratic approximation (BOQA, R package: nloptr[19]). We randomly chose two gates to optimize together using the stained WKM and HSPCs as the desired cells. For each optimization algorithm, we optimized those gates for maximum purity with at least a 20% yield. We repeated this process 100 times while choosing two random gates to optimize every iteration and recorded the purity of each optimization algorithm.

References

1. Stachura, D. L. & Traver, D. in Methods in Cell Biology 75–110 (2011). doi:10.1016/B978-0-12-387036-0.00004-9

2. Muraro, M. J. et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst. 3, 385–394.e3 (2016).

3. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 17, 77 (2016).

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

5. Scialdone, A. et al. Resolving early mesoderm diversification through single-cell expression profiling. Nature 535, 289–293 (2016).

6. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. Nat. Methods 11, 637–640 (2014).

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

8. Tang, Q. et al. Dissecting hematopoietic and renal cell heterogeneity in adult zebrafish at single-cell resolution using RNA sequencing. J. Exp. Med. jem.20170976 (2017). doi:10.1084/jem.20170976

9. Bergmeir, C., Molina, D. & Benítez, J. M. Memetic Algorithms with Local Search Chains in R : The Rmalschains Package. J. Stat. Softw. 75, 1–33 (2016).

10. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. IEEE Trans. Evol. Comput. 1, 67–82 (1997).

11. Dataset Shift in Machine Learning. (The MIT Press, 2008). doi:10.7551/mitpress/9780262170055.001.0001

12. Workman, C. et al. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. 3, research0048 (2002).

13. Price, W. L. Global optimization by controlled random search. J. Optim. Theory Appl. 40, 333–348 (1983).

14. Kaelo, P. & Ali, M. M. Some Variants of the Controlled Random Search Algorithm for Global Optimization. J. Optim. Theory Appl. 130, 253–264 (2006).

15. Scrucca, L. GA: A Package for Genetic Algorithms in R. J. Stat. Softw. 53, 1–37 (2013).

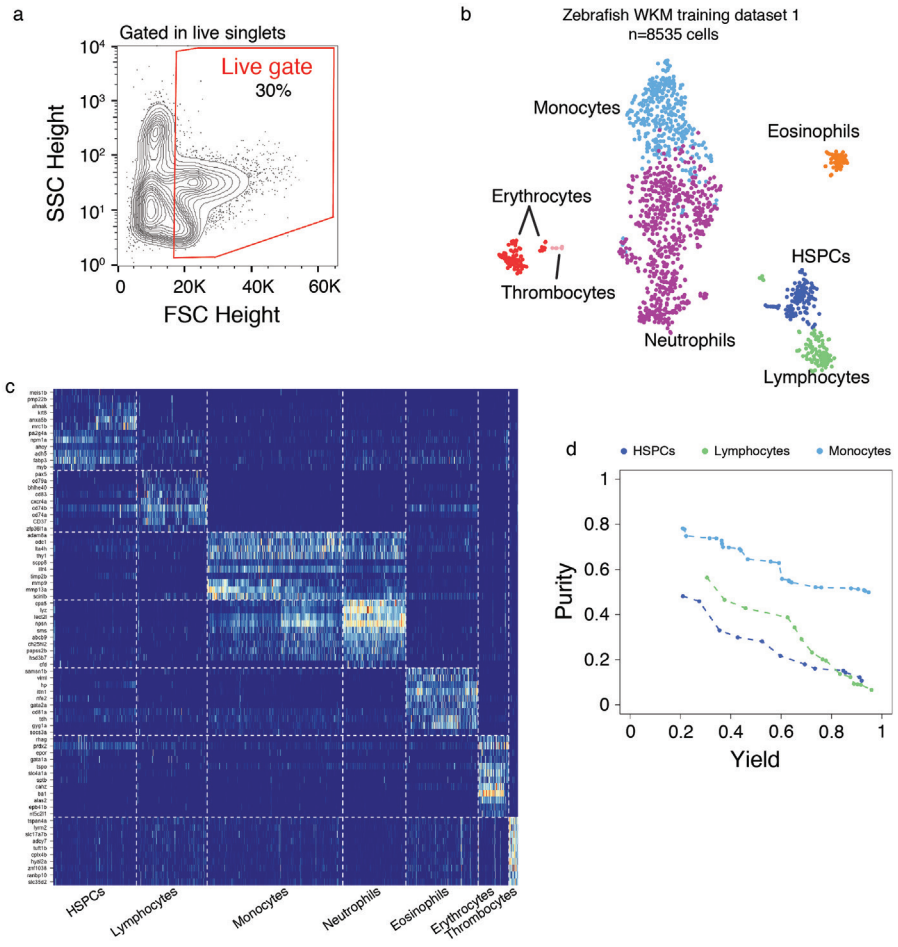
16. Kelley, C. T. Iterative Methods for Optimization. (Society for Industrial and Applied Mathematics, 1999). doi:10.1137/1.9781611970920

17. Xiang, Y., Gubian, S., Suomela, B. & Hoeng, J. Generalized Simulated Annealing for Global Optimization: The GenSA Package An Application to Non-Convex Optimization in Finance and Physics. R J. 5, (2013).

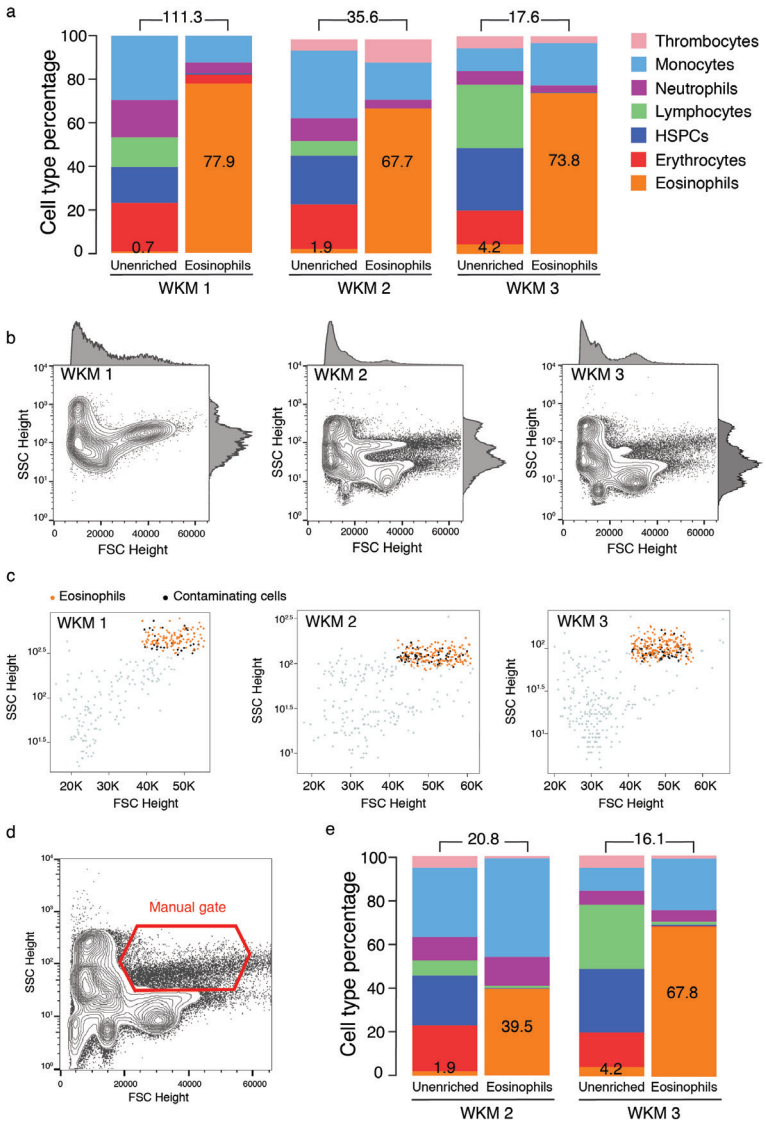
18. Mullen, K., Ardia, D., Gil, D., Windover, D. & Cline, J. DEoptim: An R package for Global Optimization by Differential Evolution. J. Stat. Softw. 40, 1–26 (2011).

19. Powell, M. J. D. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical report (2009).

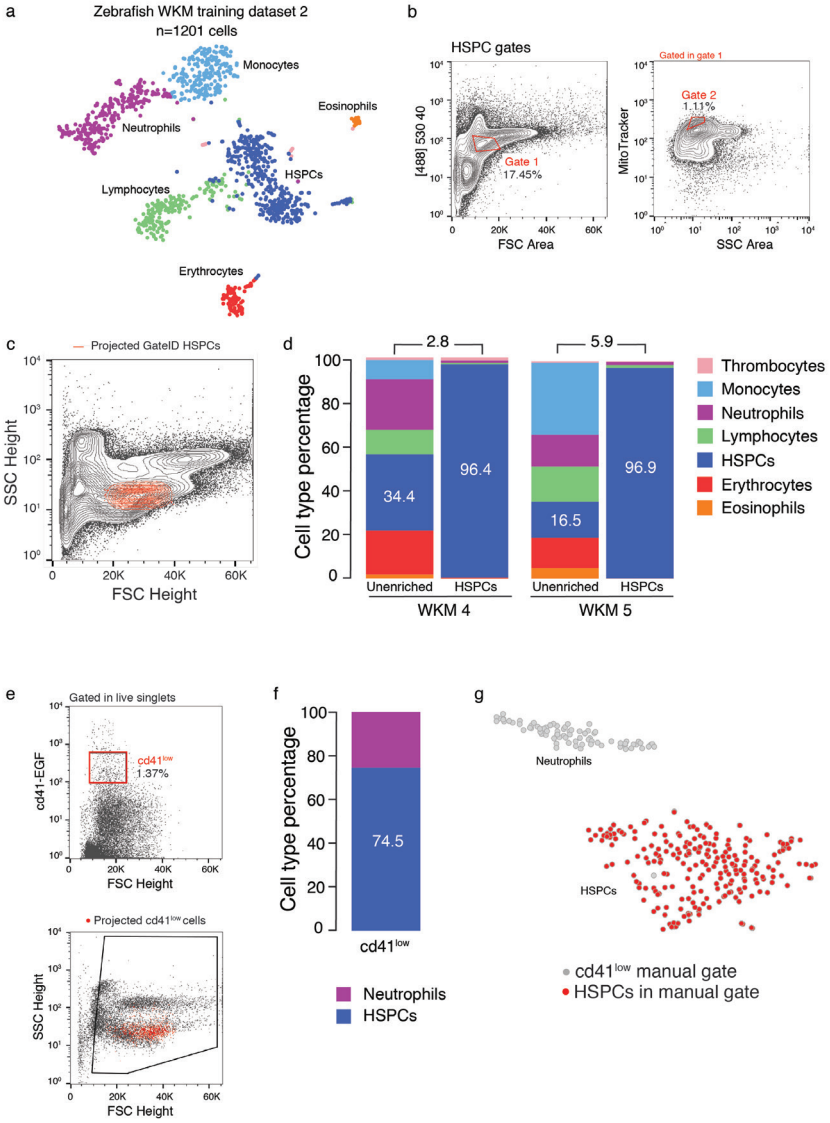
5



Supplementary Figure 1
(a) FSC Height and SSC Height contour plot of sorted live WKM cells. Events represented are live singlets from the total WKM population. Majority of erythrocytes were excluded by excluding events with low FSC-Height. (b) A t-SNE map of zebrafish WKM training dataset 1. Single cells are colored based on cell type identification. (c) Heat map showing marker genes for all hematopoietic cell types identified in the WKM full dataset. (d) Curves showing trade-off between yield and purity of GateID solutions for HSPCS, lymphocytes and monocytes for the DAPI stained sample. All gates for a given cell type with lower purity or yield are internal to these curves and are not shown.

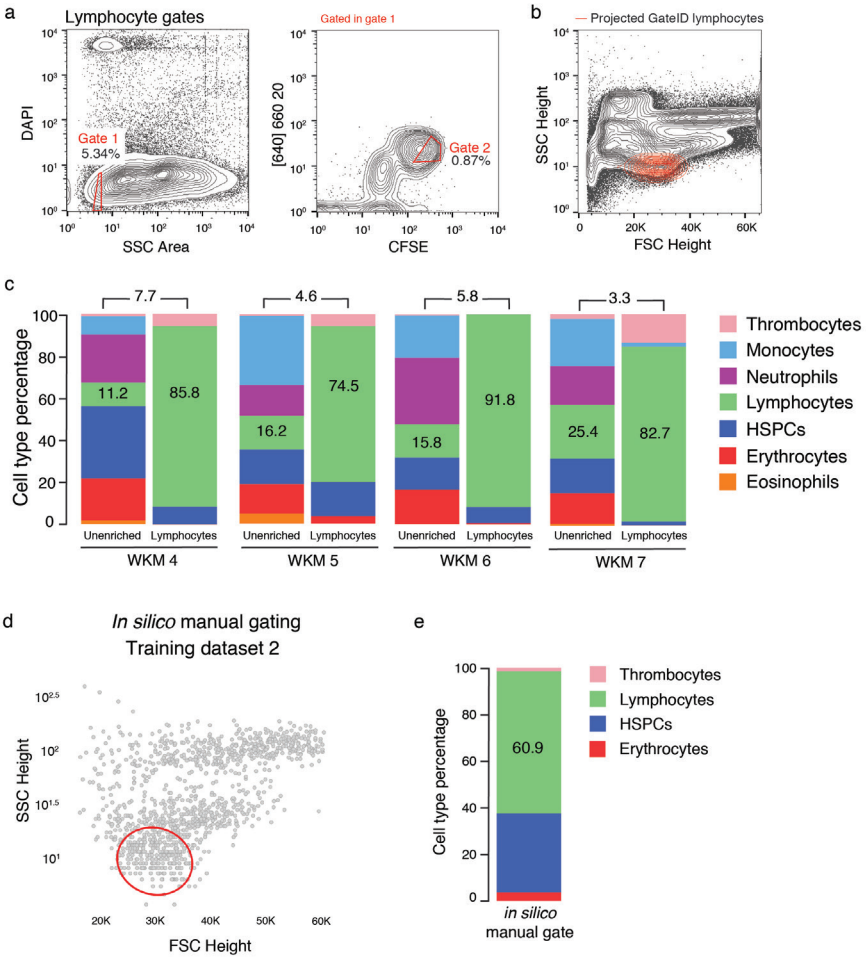


Supplementary Figure 2
(a) Barplot indicating cell type percentages for unenriched and GateID enriched libraries for all eosinophil experiments (WKM 1-3). Numbers above the bars indicate the eosinophil fold enrichment between unenriched and GateID enriched libraries. **(b)** FSC Height and SSC Height contour plots of ungated WKM cells for all eosinophils enrichment experiments (WKM 1-3). Histograms on each plot show population density is both FSC and SSC Height channels. **(c)** Plots showing sorted unenriched and GateID enriched cells for all eosinophil experiments (WKM 1-3) in FSC and SSC Height. Sorted eosinophils are highlighted in orange and sorted non-eosinophil contaminating cells are represented in black. All grey points are live cells and all colored points (red and blue) are GateID enriched cells. **(d)** FSC Height and SSC Height contour plot of all WKM cells. The eosinophil manual gate used in WKM 2 experiment is represented in red (representative for WKM 2 and WKM 3 experiments). **(e)** Barplot indicating cell type percentages for unenriched and eosinophil manual gate libraries. Numbers above the bars indicate the eosinophil fold enrichment between unenriched and manual gated libraries.

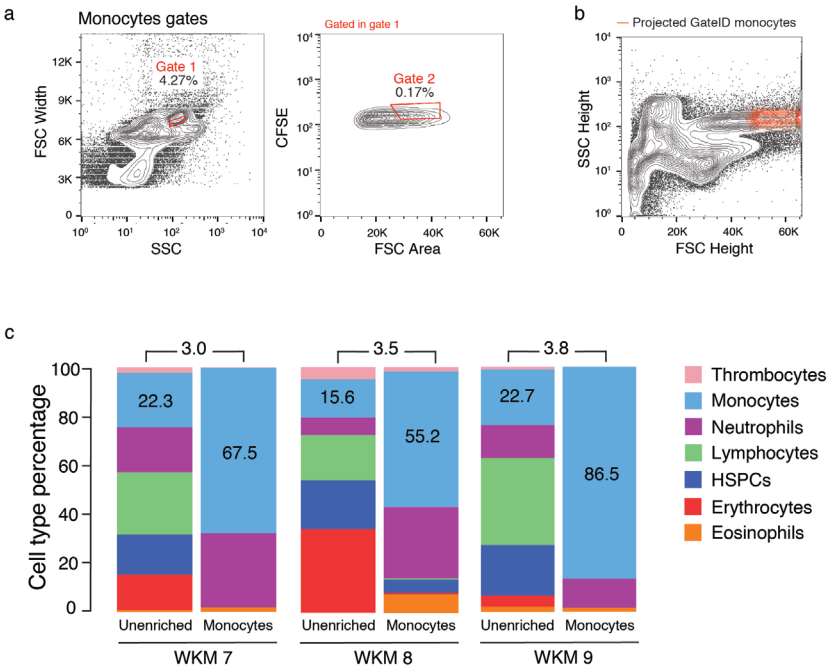


Supplementary Figure 3

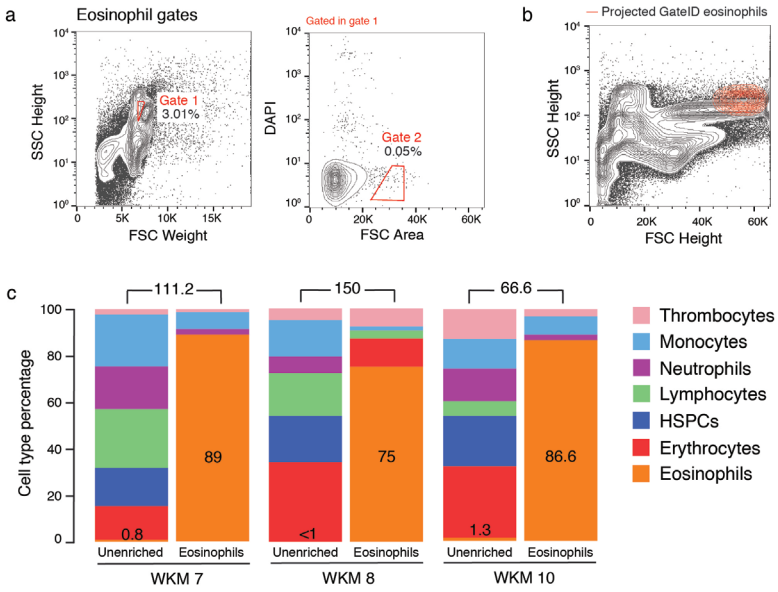
(a) A t-SNE map of zebrafish WKM training dataset 2. Single cells are colored based on cell type identification. **(b)** Contour plots of MitoTracker and CFSE stained WKM cells showing experimental sorting gates for HSPC for the WKM 4 experiment (representative example of all other HSPC experiments). Sorted cells passed through gate 1 and gate 2. Percentages of events within each gate are indicated. **(c)** Projection of the sorted GateID HSPCs (WKM 4) in FSC Height vs. SSC Height. **(d)** Barplot indicating cell type percentages for unenriched and GateID enriched libraries for all HSPCs experiments (WKM 4,5). Numbers above the bars indicate the HSPC fold enrichment between unenriched and GateID enriched libraries. **(e)** Upper panel - FSC Height vs. cd41-EGFP dot plot of live singlet WKM cells. The cd41^{low} gate is represented in red. Lower panel - projection of the cd41^{low} sorted cells in FSC Height vs. SSC Height. **(f)** Barplot indicating cell type percentages for sorted cd41^{low} cells. Percentage in the barplot indicates HSPC percentage in the sorted library. **(g)** t-SNE map showing sorted cd41^{low} cells. Non HSPCs are represented in grey and HSPCs in red.



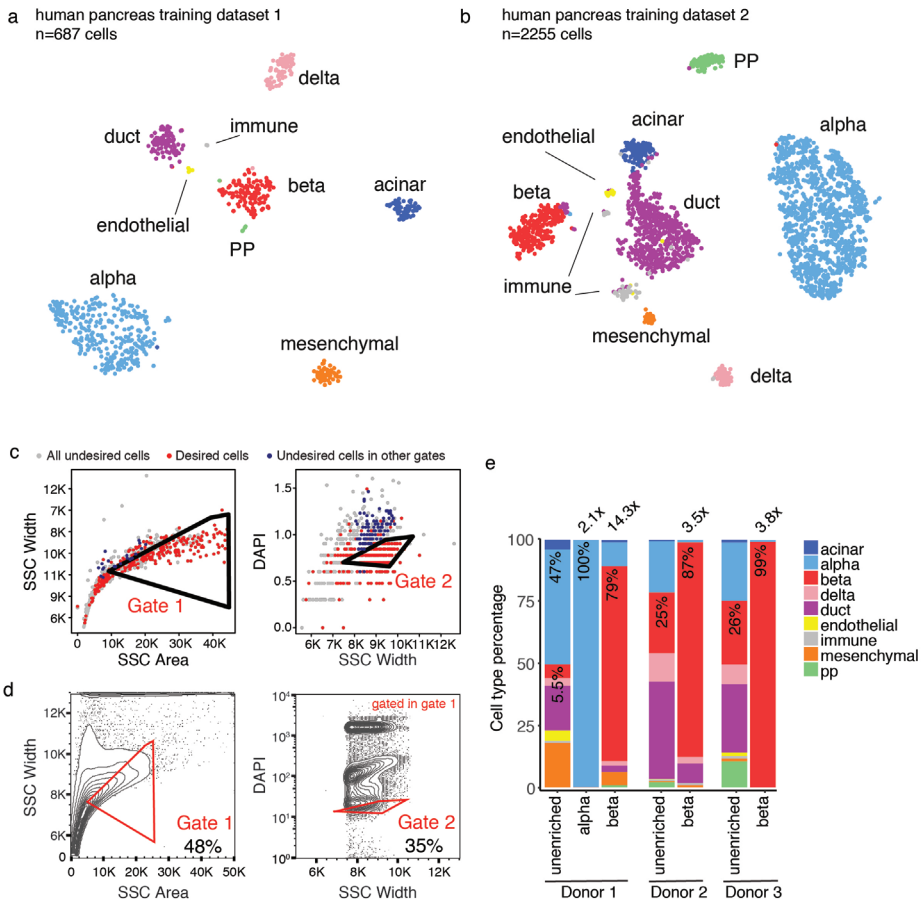
Supplementary Figure 4
(a) Contour plots of MitoTracker and CFSE stained WKM cells showing experimental sorting gates for lymphocytes for WKM 4 experiment (representative example of all other lymphocyte experiments). Sorted cells passed through gate 1 and gate 2. Percentages of events within each gate are indicated. **(b)** Projection of the sorted GateID lymphocytes (WKM 4) in FSC Height vs. SSC Height. **(c)** Barplots indicating cell type percentages for unenriched and GateID enriched libraries for all lymphocyte experiments (WKM 4-7). Numbers above the bars indicate the lymphocyte fold enrichment between unenriched and GateID enriched libraries. **(d)** Design of ***in silico*** reconstruction of the manual gate for lymphocyte enrichment. Cells from training dataset 2 are represented in grey and manual gate is drawn in red. **(e)** Barplots indicating cell type percentages for the lymphocyte ***in silico*** manual gate. Percentage in the barplot indicates lymphocyte percentage in the ***in silico*** manual gate.



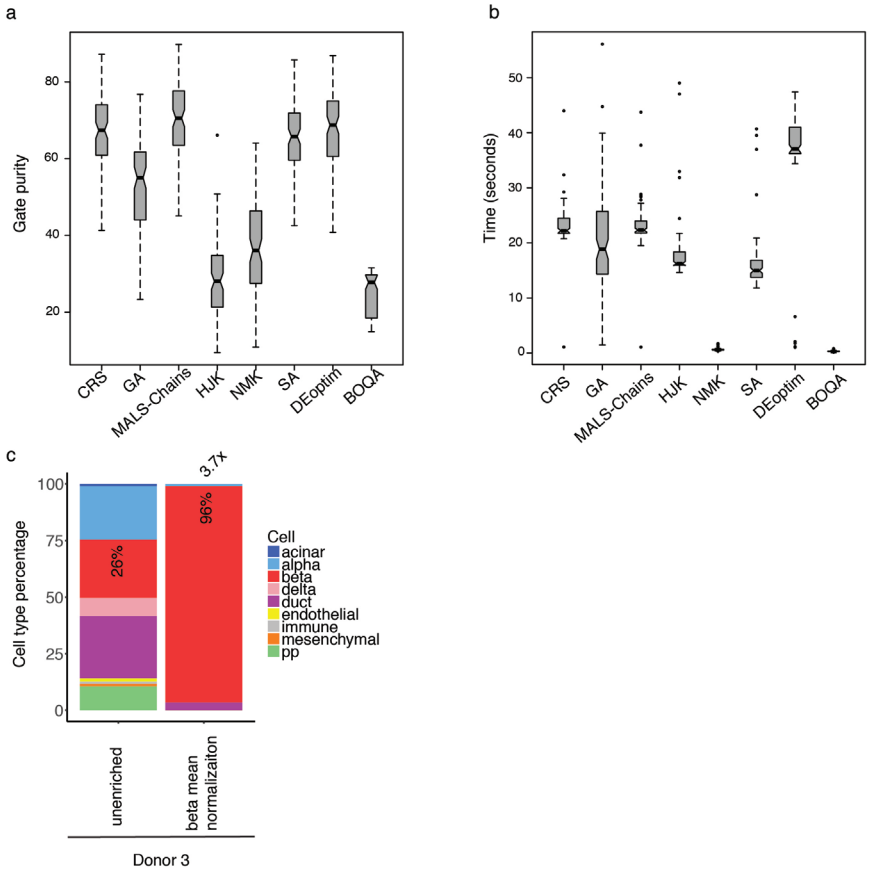
Supplementary Figure 5
(a) Contour plots of MitoTracker and CFSE stained WKM cells showing experimental sorting gates for monocytes for WKM 7 experiment (representative example of all other monocyte experiments). Sorted cells passed through gate 1 and gate 2. Percentages of events within each gate are indicated. **(b)** Projection of the sorted GateID monocytes (WKM 7) in FSC Height vs. SSC Height. **(c)** Barplot indicating cell type percentages for unenriched and GateID enriched libraries for all monocyte experiments (WKM 7-9). Numbers above the bars indicate the monocyte fold enrichment between unenriched and GateID enriched libraries.



Supplementary Figure 6
(a) Contour plots of MitoTracker and CFSE stained WKM cells showing experimental sorting gates for eosinophils for WKM 7 experiment (representative example of all other eosinophil experiments). Sorted cells passed through gate 1 and gate 2. Percentages of events within each gate are indicated. **(b)** Projection of the sorted GateID eosinophils (WKM 7) in FSC Height vs. SSC Height. **(c)** Barplot indicating cell type percentages for unenriched and GateID enriched libraries for all eosinophil experiments (WKM 7, 8 and 10). Numbers above the bars indicate the eosinophil fold enrichment between unenriched and GateID enriched libraries.



Supplementary Figure 7
(a) t-SNE map of human pancreas training dataset 1. Single cells are colored based on cell type identification. **(b)** t-SNE map of human pancreas training dataset 2. Single cells are colored based on cell type identification. **(c)** GateID predicted gates to isolate alpha cells from (DAPI^{-}) human pancreas. Gates were predicted on training dataset 1. Red points show desired cells (alpha cells) present in training dataset and the blue points show undesired cells falling in the other gate. **(d)** Contour plots of unstained human pancreas cells showing experimental sorting gates for alpha cells. Sorted cells passed through gate 1 AND gate 2. Percentages of events within each gate are indicated. **(e)** Barplots indicating cell type proportions in each sequenced library (384 cells) for ungated and GateID alpha or beta cell enriched libraries. All experiments were clustered together to call cell types. Percentages in the barplot indicate alpha or beta cell percentages in that library. Numbers above the bars indicate the alpha and beta cell fold enrichment between unenriched and GateID enriched libraries.

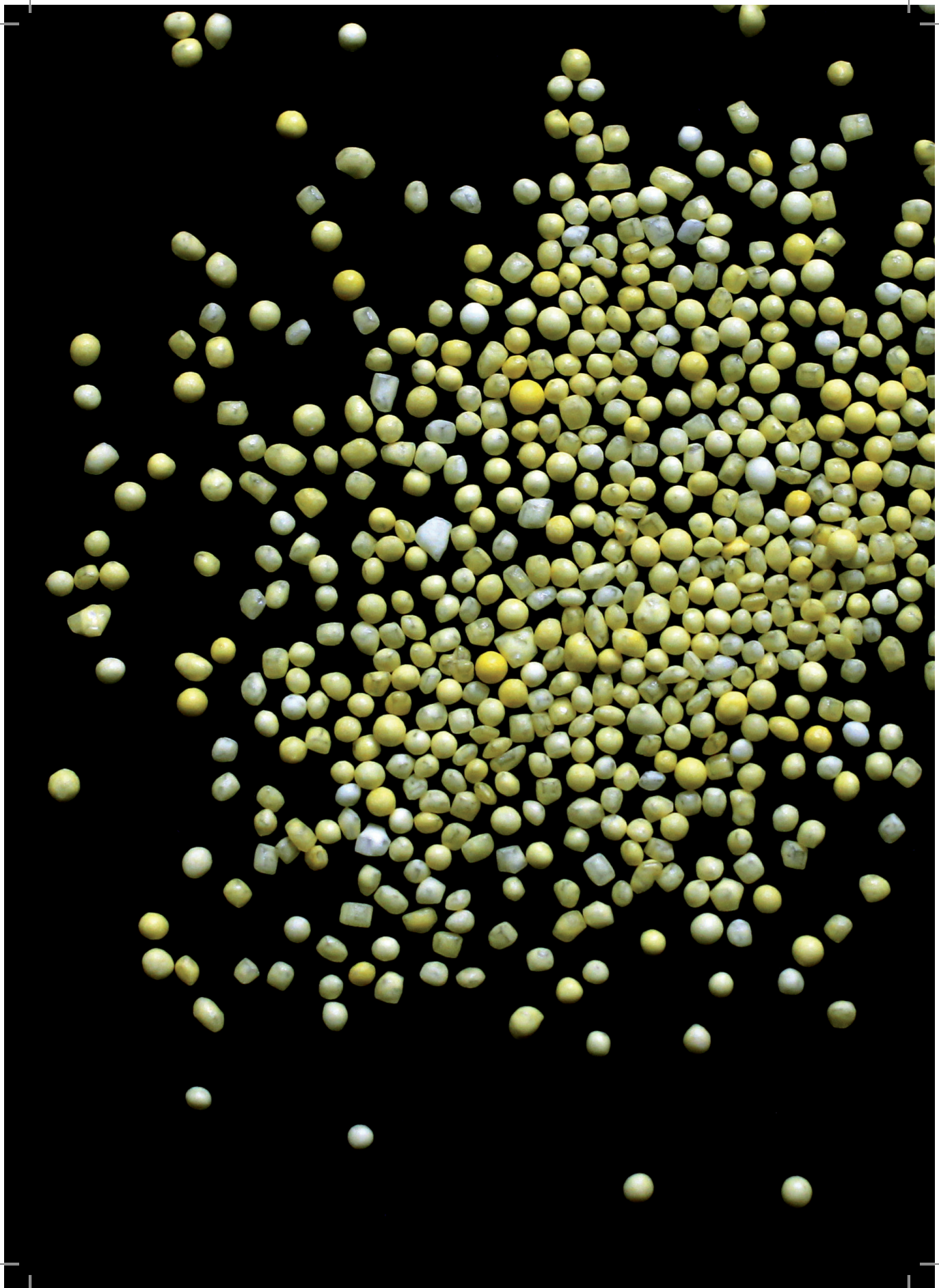


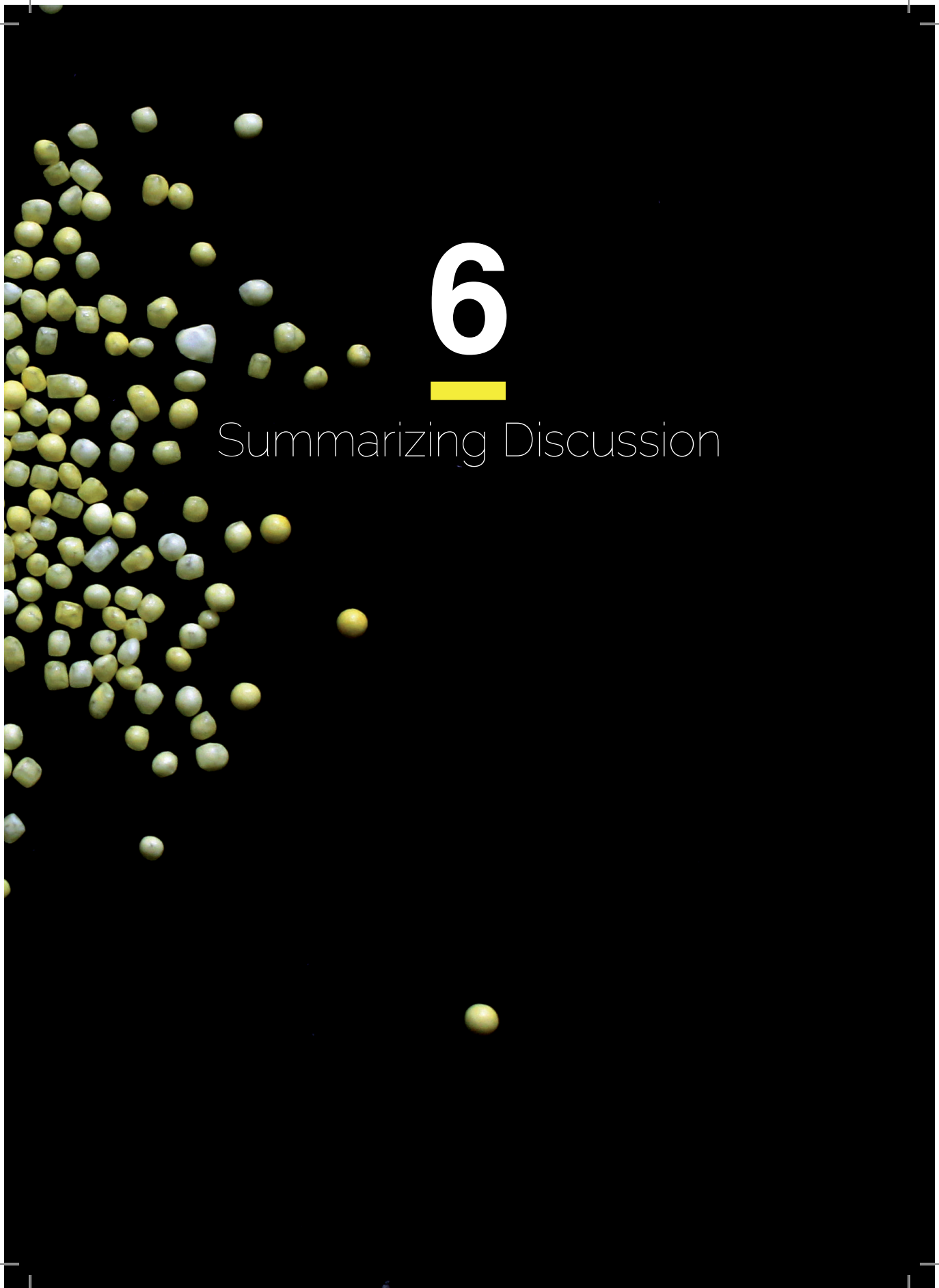
Supplementary Figure 8
(a) Purity estimate for 100 samples of gate optimization for a pair of gates using different optimization algorithms. The figure shows that MA-LS-Chains shows the best purity in comparison to 8 different optimization algorithms used here. **(b)** Time (in seconds) 100 samples of gate optimization for a pair of gates using different optimization algorithms. NMK and BOQA algorithms are fast but at the cost of substandard solution for the gate prediction problem. **(c)** Barplots indicating cell type proportions in each sequenced library (384 cells) for unenriched and GateID beta cell enriched libraries. All experiments were clustered together to call cell types. Percentages in the barplot indicate beta cell percentages in that library. Numbers above the bars indicate the beta cell fold enrichment between unenriched and GateID enriched libraries.

WKM #	Cell type	Training dataset	Fraction in training(%)	Predicted yield (%)
WKM 1	Eosinophils	1 (unstained)	3.8	51
WKM 2	Eosinophils	1 (unstained)	3.8	51
WKM 3	Eosinophils	1 (unstained)	3.8	51
WKM4	HSPCs	2 (stained)	13.4	20
WKM5	HSPCs	2 (stained)	13.4	20
WKM4	Lymphocytes	2 (stained)	7.9	20
WKM5	Lymphocytes	2 (stained)	7.9	20
WKM6	Lymphocytes	2 (stained)	7.9	20
WKM7	Lymphocytes	2 (stained)	7.9	20
WKM 7	Mmp+ myeloid cells	2 (stained)	20.1	20.5
WKM 8	Mmp+ myeloid cells	2 (stained)	20.1	20.5
WKM 9	Mmp+ myeloid cells	2 (stained)	20.1	24.43
WKM 7	Eosinophils	2 (stained)	2.09	23
WKM 8	Eosinophils	2 (stained)	2.09	23
WKM 10	Eosinophils	2 (stained)	2.09	23

Supplemental table 1 (continuation on next page)
Information on the different zebrafish sorts (each called WKM +#), the different cell types purified with them, the method (stained or not) and their expeced and obtained purity and yield

Predicted purity (%)	Sorted fraction in unenriched (%)	Experimental yield (%)	Experimental purity (%)
81	0.7	74.6	77.9
81	1.9	60.5	67.7
81	4.2	82.3	73.8
90.5	34.4	8.2	96.4
90.5	16.5	16.8	96.9
97.5	11.2	21.4	85.8
97.5	16.2	5.8	74.5
97.5	25.4	1.5	82.7
97.5	15.8	21.3	91.8
98	22.3	18.6	67.5
98	15.6	14.7	55.2
98	22.7	9.1	86.5
100	0.8	6.83	89
100	NA	2.5	75
100	1.3	17.6	86.6





6

Summarizing Discussion

Discussion, outlook and summary

We have covered several subjects in this thesis that can be classified into two main research efforts: In the first, we develop experimental and analytical techniques to automate single-cell mRNA sequencing and to subsequently analyze the data generated with this technique. The technical part of automating single-cell transcriptomics is described in the Figure 1 of chapter 3, while the first and last chapter are dedicated to algorithms that deal with single-cell data. The second research effort is the application of these techniques to pancreatic biology in an attempt to address some of the open questions in the field.

In **chapter 2** we present StemID, an algorithm that uses single-cell mRNA sequencing data to order cells into clusters and then predicts lineage relationships between these clusters. The clustering happens by comparing cells in cell-to-cell distance space (as calculated from their transcriptome). StemID then projects all cells onto vectors that run between the medoids (the most representative cell for that cluster) of all clusters. By comparing how many cells populate these projections or "pathways" between clusters to a background model of randomly assigned projections, StemID then assigns a connectivity score to each cluster. Stem cell-like clusters are assumed to have a high connectivity score, as the cells in them can differentiate towards several cell types and will therefore project cells onto more than one pathway. A fully differentiated cell type is assumed to project cells onto one pathway (usually its direct progenitor). This score is combined with the entropy of each cluster, which is a measure for the uniformity of the average transcriptome in that cluster. A multipotent cell type that has not differentiated yet is assumed to have a homogenous transcriptome (many genes expressed at equal levels), while a fully differentiated cell will have high entropy (high numbers of transcripts from a small number of genes). Good examples of this are adult pancreatic endocrine cells, which have transcriptomes that are dominated by transcripts of one hormone (sometimes one fifth of the transcriptome; see figure 2A in chapter 3). Both entropy and connectivity scores are then combined into a StemID score, which indicates the likelihood that a given cluster is a stem cell cluster. StemID was trained on data from the mouse small intestine and bone marrow immune cells, where it correctly identified LGR5+ stem cells and hematopoietic stem cells as the most pluripotent cells in the dataset. StemID was then tested on human pancreas cells, where we found a rare subpopulation of pancreatic duct cells that express both FTL and INS (beta cell markers). Upon validating the existence of these cells in human tissue sections, we indeed find rare Ftl+ Insulin+ ductal cells. How often these cells appear across more donors and what their function is are two questions that remain open. It would be interesting to find out if these cells are shared across many individuals and even across species. If they can be found in the mouse, making a knockout model for this gene could be highly interesting. Ftl proteins are essential in sequestering iron molecules from the cytoplasm and are implicated in the cellular response to oxidative stress (Orino et al., 2001). Since both metal metabolism and oxidative stress response are important for beta cell health, these proteins might be implicated in the pathenogenesis of diabetes (Bonfils et al., 2015).

In **chapter 3** we used StemID to examine the adult human pancreas. We found clusters corresponding to all main pancreatic cell types, which allowed us to provide a resource that can be used to mine the transcriptome of each cell type, including that of rare cell types like delta and pp cells. While slightly more mundane an effort than looking for stem cells, it is a useful one, since transcriptomes of these cell types had never been described in detail due to the difficulties in obtaining pure populations of these cell types. This is exemplified by the number of groups that was simultaneously working on similar projects: since the first small-scale study on single-cell transcriptomics of the human pancreas (Li et al., 2016), at least 6 similar studies have followed, some of which also analyzed data from diabetic human donors. These studies have been reviewed in (Carrano, Mulas, Zeng, & Sander, 2017). Since these studies each were done by different labs and with different single-cell mRNA sequencing techniques, this offers a unique possibility to find lab & technique-specific false positives. This is important, since others from our group have found that some subpopulations in single-cell transcriptomics data can be induced by the dissociation procedure used to obtain them. (van den Brink et al., 2017). Cross-comparing single-cell data from the same organ produced by different labs will be educative for composing a "no-fly" list of genes whose expression might be dependent on experimental (e.g. dissociation time and enzyme) and analytical (e.g. algorithm used to map the data) conditions. The satellite cell example above shows that is important to always validate curious findings from single-cell data with other complementary techniques. We did this in the last two figures of chapter 3, where we validate a cell surface-marker for alpha cells and also find groups of brightly positive Reg3a+ acinar cells in pancreatic tissue sections. Similar to Ftl in duct cells, this does not provide information about the function this gene might have. The most straightforward follow-up experiment is to quantitatively analyze where these Reg3a+ cells can be found in the pancreas (are they disproportionally often found close to Islets of Langerhans, for example?). Another useful approach would be tot test if they can be found in pancreatic organoids. If they are, a role in response to microbial insult (Reg proteins have been known to do this in the small intestine) is unlikely, since this should not be an issue in organoid culture.As for the cell surface marker that we used to purify alpha and beta cells, it would be interesting to combine it with GatID (described in chapter 5) in order to find FACS gates that purify our cell types of choice to 100% purity.

In **Chapter 4**, we study the development of the mouse embryonic pancreas by using single-cell transcriptomics. We built a resource similar to that described in chapter 3, now covering the second transition of pancreatic development from E12.5 to E18.5. Using StemID, we infer a lineage tree for the various clusters in the dataset and analyze the gene expression dynamics involved in differentiation of various pancreatic cell types. To do so, we first order all the cells into pseudotimelines from the neurogenin+ progenitor cluster to either differentiated alpha or beta cell clusters. Subsequently we compute self-organizing maps of genes along these timelines to provide clusters of genes that peak during different points in development towards differentiated alpha or beta cells. In doing so, we find many genes with Neurogenin3+ cluster-specific expression that have not been previously described to have a function in pancreatic development, but that have a known function in

neuronal development. Since pancreatic development often mirrors that of neuronal cells in terms of gene expression, we hope these genes are valuable targets for understanding pancreatic development. We also find a group of genes (exemplified by Chgb) that peak in a cluster situated between the Neurog3+ and the early alpha and beta cell clusters. They no longer express high levels of Neurog3, but show early markers of endocrine cell fate. It is therefore likely that this cluster consists of late-stage progenitor cells that are more committed towards endocrine cell fate, but do not yet have a defined trajectory. As the path from endocrine progenitor to alpha or beta cell is not completely understood, we hope this dataset will provide a valuable resource for studying this transition. We validated the presence of several novel markers for endocrine progenitor cells as well as for the intermediate progenitor cluster, finding that they are sometimes indeed co-expressed with Neurogenin3 (in other cases being expressed in cells adjacent to Neurog3+ cells). Since this cluster exhibited low expression of mesenchymal markers like Vimentin, it would be interesting to find out if this cluster corresponds to pre-endocrine cells that are delaminating from the ductal epithelium. Again, extensive validation by immunohistochemistry will be the most promising first step to test these hypotheses. On the computational side, it was difficult to find the exact order with which progenitors differentiate to alpha or beta cells. One of the open questions in the field -which cell type appears first?- is hard to answer when the analysis is based on clusters like those that come out of RaceID2. The more we sequence single-cells from developing systems, the clearer it becomes that cells don't tend to fall in precisely defined cell states/types. Transition between generally accepted cell types like progenitors and differentiated cells seem to follow a gradient more than a quick switch from one type to another. New computational methods that rely on the intron to exon ratio of transcripts in a cell and can therefore predict their "age" will likely be useful in tackling this question: with them we can pinpoint which cells are either new or in an actively transcribing transitional phase, hopefully giving us insights into the order of appeareance of cell types in the developing pancreas.

In **chapter 5** we show a practical example of what can be done by combining single-cell transcriptomics with FACS index data. We merged these and used them as input for GateID, an algorithm that allowed us to find FACS gates that use "native" cellular properties like light scatter and mitochondrial or DNA content to enrich cell types that are otherwise impossible to purify (because of lack of fluorescent reporters or antibodies). We use two different organs from two model organisms to test the algorithm: the zebrafish whole kidney marrow and the human pancreas. We were successful in obtaining pure (70-90%) populations of eosinophils, lymphocytes, hematopoietic stem cells and myeloid cells from the zebrafish. Next, we tested GateID's ability to purify alpha and beta cells from the human pancreas, obtaining close to 100% purity for both cell types. The major challenge in applying GateID is in dealing with different kinds of variability between FACS sorts. For example, the intensity of DAPI, the nuclear content marker used in all GateID experiments, can vary between sorts, which means the gate coordinates predicted from data sorted one day cannot be directly used to purify the same population of cells during a different sort. Normalization is required to move the gates along with the shift in points due to technical, day-to-day variability. Another source of variability stems

from the difference in ratios between cell types from a tissue. This variability is harder to deal with, since it not only changes the relative location of the cloud of points in FACS space, but also its shape. We describe several ways of normalizing data across different sorts, but success is limited in the case of rare populations of cells or when the difference in contributing cell types is simply too great. More experiments and a controlled way of dictating cell type contributions will be helpful in further optimizing GateID normalization to make it more robust in the face of experimental variation. The most important idea of this chapter is a very simple one: we have new way of combining transcriptome information with FACS parameters in single cells. We exploit this to offer a systematic way for enrichtment of desired cell types. This can be done using intrinsic cell properties such as size or granularity, without having to resort to pre-existing fluorescent reporters or antibodies.

Outlook

Single-cell transcriptomics adds unprecedented clarity and resolution to cell biology research compared to "traditional" bulk sequencing, where thousands to millions of cells are pooled and analyzed as one. Clarity because it allows us to "purify" cell types in silico: we can cluster all the individually sequenced transcriptomes into groups that can then be linked to the cell types in the tissue. This way, we can describe the gene expression patterns of each cell type without any contamination from other cell types. This is very different to the situation in "traditional" bulk transcriptomics, where many thousands/millions of cells are mixed into one sample and where the most prevalent cell type will dominate data obtained in this way. Single-cell transcriptomics also offers much finer resolution since each individual cell can potentially be identified as a separate (sub) cell type. This way, we can detect populations of cells that behave (slightly) different than their sister cells of the same type. This also means we can look for rare cell types like stem cells, which otherwise would go undetected between the thousands of differentiated cells that surround them.

Of course, there are also some important drawbacks to sequencing single-cells.

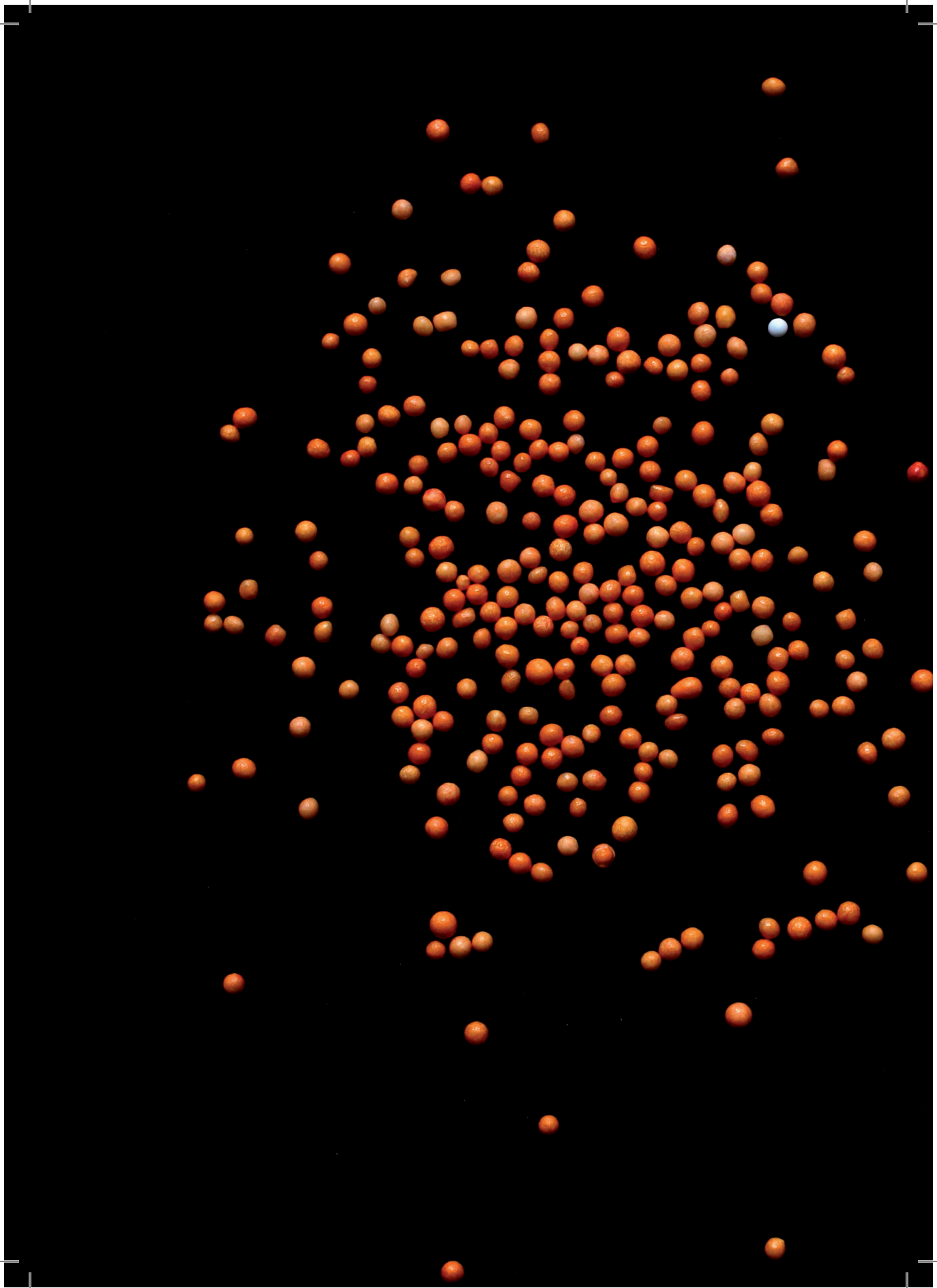
One of these stems from the fact that very low quantities of RNA are used as starting material. A typical mammalian cell has approximately 5 picograms of total RNA, of which only 1%, or 50ng, is mRNA. This brings about two issues that are important to keep in mind when looking at single-cell data of any kind (including techniques to sequence DNA or detect epigenetic marks). First, the conversion of mRNA molecules to sequencing-ready cDNA libraries is never a 100% efficient. In a bulk sample one could lose half the starting molecules and still have an accurate representation of all the mRNA species in a sample. When sequencing mRNA from a single cell, however, one starts losing lowly expressed transcript species entirely from a cell. This is obvious from looking at a table with raw single-cell sequencing transcript counts: zeros are abundant. An example of this problem is the transcription factor *ARX*, which is expressed in alpha cells of the pancreas. In the data described in chapter 3, this transcription factor is detected in only part of the alpha cells. This is a problem that affects many transcription factors, as they are often lowly expressed. While it complicates the analysis, this issue becomes less severe when many cells are processed. Since we are not comparing one alpha to one beta cell (as is often assumed when discussing this issue) the difference in expression of a gene like *ARX* will still be very in a dataset of significant size

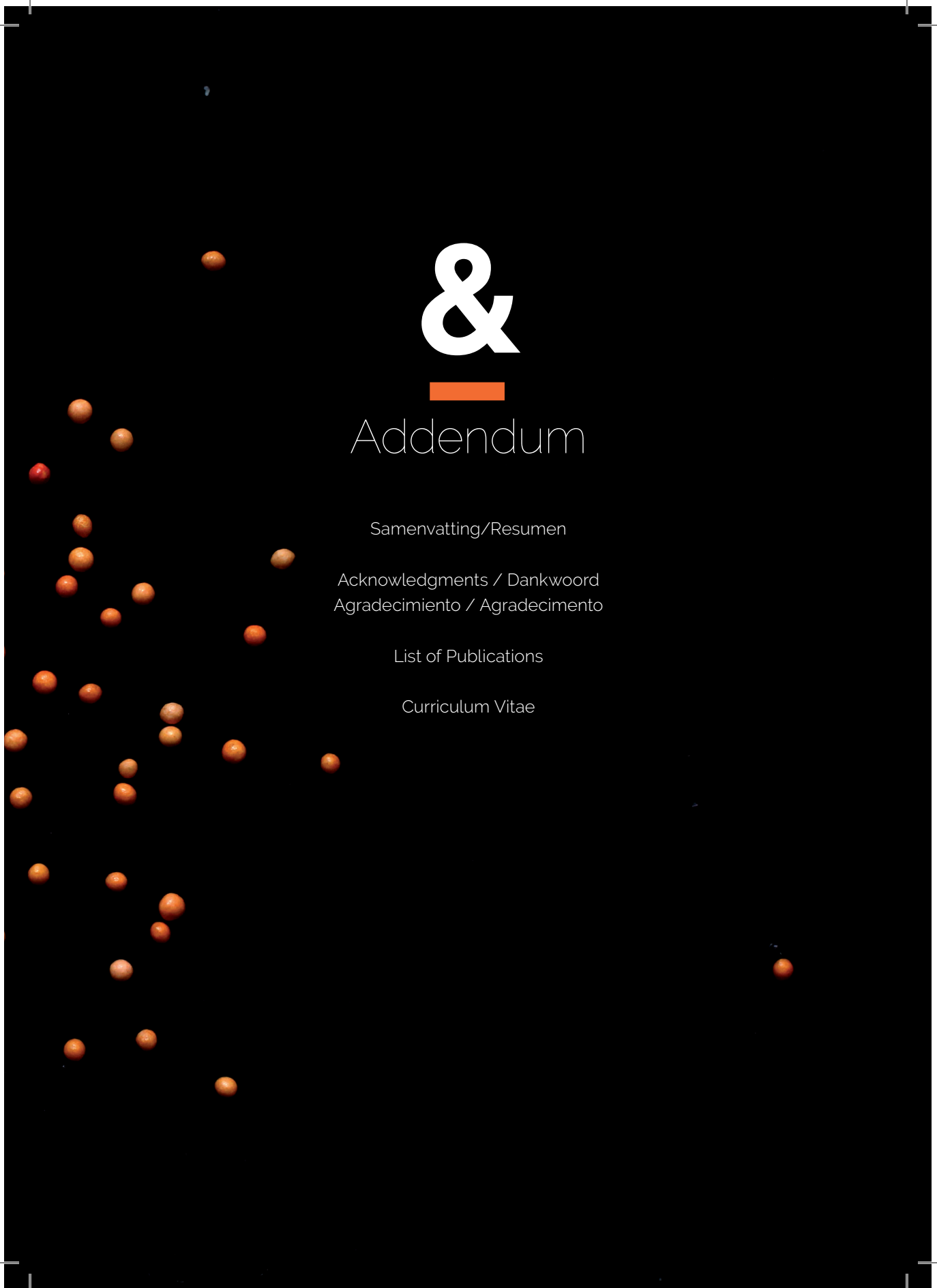
(see the differential gene expression results in chapter 3, figure 1). In other words: the problem is only very severe when dealing with low quantities of cells or with the lowliest expressed genes. Another important piece of information that is lost in most single-cell protocols is the positional information of cells in the tissue. We usually make a single-cell suspension from an organ and sample random cells from this mix, thereby losing any sence of space. Important progress has been recently made into tackling this problem. By annealing fixed tissue on top of a surface with reverse transcription primers that carry positional information, it is possible to obtain 2D information transcriptional information (Stahl et al., 2016). By combining high troughput single-cell sequencing methods with spacial transcriptomics, we can still infer where in a tissue interesting (rare) cell types are located.

Summary

The work described here formed part of the progress that was made in several labs across the world in the "second wave" of single-cell transcriptomics (see introduction). In these last five years our lab moved from manually processing dozens to hundreds of cells per week to routinely sequencing thousands of cells from primary tissue in a single day. On the computational side, we took part in the development of a set of algorithms that allow the user to cluster single-cell transcriptomics data, infer lineages between cell types and predict FACS gates that can be used to purify cell types without the need for fluorescent reporters or antibodies. We applied these methods to the developing mouse and the adult human pancreas, which yielded two resources that can be used to both mine for cell type-specific expression of a gene of choice in the adult pancreas and to see if the expression of this gene changes during pancreatic development. We have validated some of the genes found in these chapters, but more work is required to understand the function of these genes in pancreas biology. For now, I hope others find the progress we made in single-cell sequencing to shine light on pancreas biology, useful. I for one wholeheartedly enjoyed working on it.

6





Addendum

Samenvatting/Resumen

Acknowledgments / Dankwoord
Agradecimiento / Agradecimento

List of Publications

Curriculum Vitae

Nederlandse samenvatting

Wat zou er gebeuren als we een inventaris kunnen maken van alle soorten cellen in ons lichaam? Is het mogelijk om ons lichaam en de ziektes waar we vatbaar voor zijn beter te begrijpen? Ik denk van wel; ons lichaam bestaat namelijk uit miljarden cellen en honderden verschillende soorten cellen (celtypes) zoals bijvoorbeeld de spiercel, hartcel of zenuwcel. De meeste organen zijn opgebouwd uit verschillende van deze celtypes samen. Vaak hebben we echter geen duidelijk beeld van welke celtypes precies aanwezig zijn in een weefsel. Dit kan erg belangrijk zijn, omdat door het verkeerd functioneren van één celtype of soms zelfs één enkele cel ernstige ziektes kunnen ontstaan. In het geval van kanker, bijvoorbeeld, is er iets misgegaan bij één enkele cel. Bij andere ziektes functioneert juist één bepaald celtype niet goed, wat bijvoorbeeld Diabetes tot gevolg kan hebben. Diabetes wordt veroorzaakt wanneer de beta cellen uit de alvleesklier afsterven of niet goed functioneren. Dit leidt tot misregulatie van de suikerspiegel in het bloed, met ernstige gezondheidsproblemen als gevolg. Voordat we kunnen begrijpen hoe een ziekte op moleculair niveau werkt moeten we eerst echter goed begrijpen hoe een gezond orgaan werkt en hoe het opgebouwd is. Beta cellen zijn bijvoorbeeld niet het enige celtype in de alvleesklier. Er zijn ten minste zeven andere celtypes bekend die samenwerken om de alvleesklier op de juiste manier te laten werken. We hopen dat we door een duidelijk beeld te vormen van de aanwezige celtypes in de pancreas en wat ieder hen kenmerkt, we ook een beter idee hebben over wat er mis gaat bij een ziekte als Diabetes. In mijn thesis heb ik geprobeerd een inventaris te maken van de alvleesklier door te werken aan experimentele technieken en algoritmes die het mogelijk maken om individuele cellen te mRNA-sequencen.

Sequencing is een techniek die ons een lijst geeft van alle messenger RNA (mRNA) moleculen die in een bepaald experimenteel monster aanwezig zijn. mRNA-moleculen zijn de boodschappers tussen de genen in het DNA en de eiwitten die een cel uiteindelijk produceert en die de functie van die cel bepalen. Iedere cel in ons lichaam bevat praktisch dezelfde genetische informatie in de vorm van DNA, maar verschillende celtypes gebruiken verschillende delen of genen van het DNA. Daarmee zijn mRNA-moleculen indicatoren voor cel functie en bij uitstek geschikt om te sequencen.Je dit kunnen vergelijken met de bezorgkoeriers in een stad: Als we de gouden gids (de DNA sequentie) van twee (bijna) identieke steden vergelijken krijgen we géén helder beeld van wat er op dat moment gegeten wordt in deze steden. Als we een lijst hebben van alle bezorgscooters (mRNA-moleculen) die er op dat moment rondrijden komen we er snel achter dat de ene stad bijvoorbeeld een voorkeur voor pizza heeft, waar de andere op dat moment een voorstander is van sushi. Op vergelijkbare manier geeft een lijst van mRNA-moleculen ons inzicht in welke genen er door een orgaan gebruikt worden.

Tot voor kort liet de techniek het enkel toe om sequencing alleen op materiaal van vele duizenden cellen tezamen toepassen. Een typisch sequencing experiment om te onderzoeken wat er misgaat in een ziekte als diabetes zou als volgt gaan: we isoleren het mRNA uit een miljoen cellen uit de alvleesklier van een diabeticus

en vergelijken dat met het mRNA van een miljoen cellen uit de alvleesklier van een gezond persoon. Dit sequencen op macroniveau geeft ons een idee welke genen er belangrijk zijn in de context van deze ziekte, maar omdat er zoveel cellen nodig zijn worden de mRNA-moleculen van verschillende celtypes gemengd tot één gemiddeld mRNA profiel. Hierdoor is het onduidelijk welke genen er in ieder celtype gebruikt worden. Ook wordt hierdoor de bijdrage van zeldzame celtypes, zoals stamcellen, moeilijk te bepalen omdat ze "weggemiddeld" worden door de vaker voorkomende celtypes.

Single-cell sequencing maakt het mogelijk om sequencing op het niveau van individuele cellen te doen, door het mRNA van iedere cel in een experiment eerst een moleculaire barcode te geven en vele keren te kopiëren vóórdat we het samenvoegen voor een sequencing experiment. Op deze manier hebben we genoeg materiaal, maar kunnen we de individuele mRNA-moleculen toch traceren naar de cel waar ze oorspronkelijk in zaten.

Het verschil tussen klassiek en single-cell sequencen is (losjes) vergelijkbaar met het overgaan naar google street view na een tijdlang naar een klassieke stadskaart gestaard te hebben: de bijzondere huizen springen nu plotseling in het oog (zie figuur 1). Door vervolgens alle huizen met elkaar te vergelijken, te groeperen op basis van gelijke eigenschappen en uiteindelijk met de satellietbeelden van de hele stad te vergelijken kunnen we misschien ook nieuwe patronen ontdekken: Huizen met een tuin, bijvoorbeeld, bevinden zich in een bepaalde gedeeltes van de stad, terwijl flats met meer dan 10 verdiepingen zich elders bevinden.

De hoofdstukken van dit proefschrift laten zich als volgt samenvatten:

Hoofdstuk 1 is een introductie waarin de experimentele en computationele voortgang in single-cell sequencing technieken beschreven wordt, als wel de openstaande biologische vraagstukken in alvleesklier-biologie.
In Hoofdstuk 2 beschrijven we StemID, een algoritme dat single-cell sequencing data analyseert om vervolgens een stamboom te maken van alle groepen cellen die in die dataset gevonden zijn. Dit soort algoritmes zijn onmisbaar bij het analyseren van de in de regel complexe datasets die uit single-cell sequencing-experimenten voortkomen.
In Hoofdstuk 3 worden single-cell sequencing en StemID toegepast om de volwassen menselijke alvleesklier in kaart te brengen. Om dit efficiënt te kunnen doen moesten we eerst de bestaande technieken voor single-cell sequencing automatiseren met behulp van FACS (zie uitleg hoofdstuk 5) en pipetteerrobots. Dit geautomatiseerde protocol hebben we SORT-seq genoemd en vormt de basis voor alle data die in de rest van de hoofdstukken wordt beschreven. Voor onze alvleesklier-kaart hebben we duizenden cellen uit vier verschillende orgaandonoren gesequenced. Hierdoor konden we het volledige RNA-profiel van alle aanwezige soorten cellen beschrijven en vele nieuwe genen aanwijzen die celtype-specifiek zijn. Ook vinden we tot dusver onbekende sub-soorten acinaire en beta cellen die zich net iets anders gedragen dan hun zustercellen van dezelfde soort (denk aan het Rietveldhuis). Tot slot verifiëren we een aantal van deze bevindingen door microscopietechnieken toe

te passen op weefselsecties van alvleesklierweefsel (denk aan de vergelijking met satellietbeelden).
De aanpak beschreven in Hoofdstuk 4 lijkt erg op dat in hoofdstuk 3, maar gaat over de embryonale ontwikkeling van de alvleesklier van de muis. Een interessant vraagstuk is hoe de volwassen alvleesklier ontstaat en hoe er nieuwe cellen gemaakt worden tijdens het leven van een individu. Omdat in de volwassen situatie geen stamcellen gevonden zijn die hier verantwoordelijk voor zijn, kijken we hier naar de embryonale ontwikkeling, waar deze stamcellen wel bekend zijn, in de hoop hiermee in de toekomst ooit ook volwassen stamcellen te kunnen identificeren of te produceren. We sequencen hiervoor duizenden cellen tijdens het ontstaan van de alvleesklier in de muis en maken met behulp van StemID een stamboom van het traject van stamcel naar volwassen celtypes zoals alpha- en betacellen. We brengen verschillende groepen genen in kaart die belangrijk zijn voor de ontwikkeling van de verschillende celtypes en vinden een nieuwe celtype die de “missing link” lijkt te zijn tussen de stamcellen en de volwassen celtypes.
Hoofdstuk 5 gaat over GateID, een algoritme waarmee we een celtype naar keuze kunnen purificeren uit een mengsel van verschillende celtypes. Dit doen we door Fluorescence Activated Cell Sorting (FACS) data op te slaan van iedere cel die we sequencen. De FACS is het apparaat dat cellen één voor één uit een mengsel van vele cellen oppakt en daarvan bepaalde waardes meet (b.v. grootte, interne complexiteit en fluorescentie op verschillende golflengtes). Het wordt vaak gebruikt om cellen die een bepaald label hebben gekregen (een fluorescerend antilichaampje bijvoorbeeld) te sorteren uit een grotere gemengde groep cellen. Dit is handig als men een antilichaam heeft dat één celtype kan labelen. Helaas is dit voor veel celtypes niet het geval. Met GateID vergelijken we de FACS informatie van alle verschillende celtypes die we met single-cell sequencing vinden. Hierdoor kunnen we nieuwe selectiemethodes vinden die niet berusten op antilichamen, maar op cel-eigen eigenschappen, zoals grootte en de natuurlijke fluorescentie van een celtype in een bepaalde golflengte. We bewijzen de efficiëntie van deze methode door hiermee verschillende soorten imuuncellen uit de zebravis en alpha en beta cellen uit de menselijke alvleesklier te purificeren.

Kort samengevat gaat dit proefschrift over experimentele en computationele voortgang die de laatste paar jaar door ons lab is geboekt in het veld van single-cell sequencing. Deze methodes hebben we vervolgens toegepast om alvleesklierbiologie en ontwikkeling te beschrijven en beter te begrijpen.

Figuur 1. Het Gerrit Rietveld huis in Utrecht.
1A: Het huis in standaard kaart perspectief. De zwarte pijl geeft de locatie van het huis aan. 1B: Hetzelfde huis in google street view. De zwarte pijl geeft dezelfde locatie weer.

Resumen

¿Qué pasaría si pudiésemos hacer un registro de todos los tipos de células de nuestro cuerpo? ¿Comprenderíamos mejor a nuestro cuerpo y las enfermedades a las que somos propensos? Creo que sí; nuestro cuerpo está formado por billones de células que se presentan en cientos de tipos diferentes de células (celtypes), como, por ejemplo, las células musculares, cardíacas o nerviosas.

A menudo no tenemos idea, o esta no es muy precisa sobre qué tipos de células están presentes en un órgano o tejido. Que algo funcione mal en una sola célula puede ser la causa de algunas enfermedades, como sucede con el cáncer. En otras, es un determinado tipo de célula que no funciona correctamente lo que provoca enfermedades como la diabetes. Cuando ocurre una destrucción total o parcial de las células beta del páncreas o estas no funcionan bien se da la aparición de la diabetes. Esto lleva por su vez a una mala regulación del nivel de glucosa en la sangre, resultando en graves problemas de salud. Antes de que podamos entender cómo funciona una enfermedad a nivel molecular tenemos que entender cómo está formado y cómo funciona un órgano sano. Por ejemplo, las células beta no son el único tipo de célula que se encuentra en el páncreas. Se conocen por lo menos otros siete tipos de células que trabajan juntas para hacer que el páncreas funcione correctamente. Para comprender cómo funciona un órgano y cómo contribuye a esta función cada tipo de célula, es importante tener un registro inequívoco de todos los tipos de células y sus características específicas. Mi proyecto de tesis se basó en intentar realizar este registro por medio de técnicas experimentales y algoritmos que hagan posible secuenciar células individuales.

La secuenciación es una técnica que nos da una lista de todas las moléculas de ARN mensajero (ARNm) presentes en un determinado órgano. Las moléculas de ARNm son los mensajeros entre los genes en el ADN y las proteínas que finalmente produce una célula y que determinan la función de esa célula. Cada célula de nuestro cuerpo contiene prácticamente la misma información genética en forma de ADN, pero diferentes tipos de células utilizan diferentes partes o genes del ADN. Se podría comparar con los repartidores de una ciudad. Si comparamos las páginas amarillas (el ADN) de dos ciudades (casi) idénticas, no tendremos una idea clara de lo que se está comiendo en un determinado momento en esas ciudades. Si tuviéramos una lista de las entregas (el ARN) que en ese momento se están dando pronto constataríamos que una ciudad, por ejemplo, prefiere la pizza, mientras que otra se inclina por el sushi. Esto es comparable a la lista de moléculas de ARNm que nos permite apreciar qué genes utiliza un órgano. Hasta hace poco, solo podíamos aplicar la secuenciación a una cierta cantidad de ARN de miles de células, porque las máquinas que utilizamos para esto requieren un mínimo determinado de ARNm. Para investigar lo que va mal en una enfermedad como la diabetes un experimento de secuenciación típico sería de la siguiente manera: aislamos el ARNm de un millón de células del páncreas de diabéticos y lo comparamos con el ARNm de un millón de células del páncreas de personas sanas. Esto nos da una idea de qué genes son importantes en el contexto de esta enfermedad, pero debido a que se necesitan tantas células, las moléculas de ARNm de diferentes tipos de células se mezclan y se juntan en un perfil de ARNm

promedio lo que hace que resulte poco claro qué genes son utilizados en cada tipo de célula. También por esto se vuelve difícil encontrar tipos de células más raras, como las células madre, porque estas son “promediadas” con los tipos de células más comunes.
Single-cell sequencing hace posible la secuenciación a nivel de células individuales, dando primero al ARNm de cada célula de un experimento un código de barras y amplificándolo antes de fusionarlo para un experimento de secuenciación. Haciendo esto tenemos suficiente material, pero podemos rastrear moléculas individuales de ARNm hasta la célula en que se encontraban originalmente. Los resultados de estos experimentos, por lo tanto, comprenden información de todos los genes que han sido expresados por esta célula.
Se podría comparar con ver algo en Google Street View después de haber estado un buen rato mirando un mapa de una ciudad. Casas particulares que básicamente tienen la misma función que las casas vecinas a su alrededor, pero con ciertas características especiales, pasan a llamar la atención (ver figura 1). Si a continuación se comparan todas las casas, agrupándolas por sus características y finalmente comparando con las imágenes de satélite de toda la ciudad, también podríamos descubrir nuevos patrones, como, por ejemplo, las casas con jardín están ubicadas en unas partes de la ciudad mientras que los edificios con más de diez pisos se encuentran en otras.

Resumen por capítulo

Capítulo 1- Introducción en que se describe el progreso experimental y computacional en las técnicas de secuenciación de células individuales, así como los problemas biológicos pendientes en biología pancreática.
Capítulo 2 - Describimos StemID, un algoritmo que analiza los datos de secuenciación de células individuales y a continuación genera un árbol genealógico de todos los grupos de células que se encuentran en esos datos. Este tipo de algoritmo es indispensable cuando se analizan los datos complejos que resultan de la secuenciación de células individuales. Como ser humano, estamos limitados a comprender sólo algunas variables en un pequeño número de puntos de medición. En el caso de la secuenciación de células individuales, nos ocupamos normalmente de miles de puntos de medición (de células), cada uno con alrededor de 20.000 variables (los genes).
Capítulo 3 - Sobre el uso de la secuenciación de células individuales y StemID para mapear el páncreas humano adulto. Para poder hacer esto de manera eficiente, primero tuvimos que automatizar las técnicas existentes para la secuenciación de células individuales usando FACS (ver explicación en el capítulo 5) y pipeteo robotizado. Este protocolo automatizado fue llamado SORT-Seq y es la base de todos los datos descriptos en los demás capítulos.
Para monitorizar el páncreas secuenciamos miles de células de cuatro donantes de órganos. Esto nos permitió describir en su totalidad el perfil de ARN de todos los tipos de células presentes e identificar muchos genes nuevos específicos de un tipo de célula. También encontramos subespecies de células acinares y beta desconocidas hasta ese momento que se comportan de forma un poco diferente a sus células hermanas de la misma especie (recuerde el ejemplo de la casa de

Rietveld). Por último, verificamos algunos de estos hallazgos aplicando técnicas de microscopía a secciones de tejido pancreático (recuerde la comparación con imágenes de satélite).
Capítulo 4 – El enfoque es muy similar al del Capítulo 3, pero trata del desarrollo embrionario del páncreas en ratones. Un problema interesante es cómo se origina el páncreas adulto y cómo se producen nuevas células durante la vida de un individuo. Dado que en adultos no se han encontrado células madre que fueran responsables de esto, nos restringimos a observar el desarrollo embrionario, donde las células madre sí están presentes, con la esperanza de que en el futuro se llegue a poder identificar las células madre adultas o producirlas. Secuenciamos miles de células durante el inicio del páncreas en embriones de ratón, y con la ayuda de StemID armamos un árbol genealógico de la trayectoria de las células madre a células maduras de tipo alfa y beta. Encontramos diversos grupos de genes que son importantes para el desarrollo de los diferentes tipos de células y también hallamos un nuevo tipo de célula que parece ser el "eslabón perdido" entre las células madre del páncreas y los tipos de células adultas.
Capítulo 5 - Sobre GateID, un algoritmo con el que podemos purificar selectivamente un tipo de célula en una mezcla de diferentes tipos de células. Lo hacemos por medio de la técnica de citometría de flujo (Fluorescence Activated Cell Sorting, FACS, por su sigla en inglés) almacenando los datos de cada célula que secuenciamos. El FACS es el aparato que aisla células una por una de una mezcla de muchas células y mide ciertos valores (por ejemplo, tamaño, densidad y fluorescencia en diferentes longitudes de onda). A menudo se usa para clasificar células que han recibido una determinada etiqueta fluorescente (un anticuerpo emisor de luz verde, por ejemplo) de un grupo mixto más grande de células. Esto es útil si se tiene un anticuerpo que pueda etiquetar un tipo de célula. Desafortunadamente, no es el caso para muchos tipos de células. Con GateID comparamos la información FACS de todos los tipos de células diferentes que encontramos con la secuenciación de células individuales. Esto nos permite encontrar nuevos métodos de selección que no se basen en anticuerpos, sino en propiedades específicas de la célula, como tamaño y fluorescencia natural de un tipo de célula en una cierta longitud de onda. Comprobamos la eficacia de este método al purificar diferentes tipos de células inmunológicas del pez cebra y células alfa y beta del páncreas humano.

En pocas palabras, esta tesis doctoral trata del progreso experimental y computacional realizado en los últimos años por nuestro laboratorio en el campo de la secuenciación de células individuales. Estos métodos se han aplicado para describir y comprender mejor la biología y el desarrollo del páncreas.

Figura 1. La casa de Gerrit Rietveld en Utrecht.
1A: la casa en perspectiva en un mapa estándar. La flecha indica la ubicación de la casa. 1B: la misma casa en Google Street View. La flecha indica la misma casa.

Acknowledgements/Dankwoord/Agradecimiento/ Agradecimento

First of all, I would like to thank **Alexander**. Thanks for responding so quickly when I sent you an application letter as a master student with little proven experience. Somehow you saw something in there and decided to invite me to MIT and offer me a spot in your new lab at the Hubrecht. I'm seldom nervous for moments like this, but I can tell you that I was very nervous before applying to your lab! I think it has worked out very nicely, and I can tell you it's been a privilege to work in your group. I have learned a lot observing the way you approach scientific questions and how you stay ahead of the curve by constantly looking for new fields to study. It's been equally impressive to see how quickly you analyze and understand new, complicated data. I'll try to carry some of that enthusiasm and adaptability with me. I've really felt at home in the lab culture of constantly hacking protocols and trying new things. You manage to hire a seemingly endless stream of interesting, varied and nerdy people that somehow fit quite well together. Most of all I'd like to thank you for all the freedom you give everyone in your lab. It's been a pleasure.

I would also like to thank all the members of the reading and examination committees, **Françoise**, **Jop**, **Edwin**, **Jacco**, **Niels**, **Harry**, **Eelco** and **Frank**: Thanks a lot for all the time and interest you put into reading this thesis. I'm looking forward to your questions already.

Dylan, of the people I met during my job application at MIT I remember you the best. Partly because you were the first one I met that later also moved to Utrecht, but mostly because you directly asked me with a annoyed face why I thought it was a good idea to apply in a suit. I remember we didn't much like each other. The beauty of this is we later found out –first trough discussions about science an later about a myriad of other things- that we turned out to have much in common. Like you mentioned in your thesis: it has been refreshing to find out you share important common values with someone on the other end of the political spectrum and even end up becoming friends. This doesn't happen often and I want to thank you for it. Also thanks for your eternal irony, black humor and being a co-member of the axis of evil. And for all BJJ sparring sessions, shared dinners and being a walking science encyclopedia. I look forward to seeing each other again during my defense and to all our future sparring sessions, both intellectually and on the mat.
Lennart, you stated your PhD only shortly before me, but when I arrived you had set up so much in the lab and already knew so much that it seemed you had been there for years. I admire the way you always seem sure about your path and how you set the bar for PhD students in Alexander's lab very high from day one. Thanks for your whisky savvy, for shared mates and for exchanging stories about camping and trekking journeys. In short, thanks for five years of blood, sweat but no tears (It seems we are both a bit too "nuchter" for that last part). Strangely enough we have never co-authored a paper. I hope we get the chance to mend this in the future.
Kay, flaco, we also overlapped for the biggest part of our PhD's and you too made my start at Alexanders lab much smoother than it would have been otherwise. We shared many memorable moments, like the conference in Keystone and our road trip after it. Thanks for your bioanalyzer whispering and for your inexhaustible talent

to exploit one specific, irritating part of a videogame until you drive everyone to the brink of smashing their controller's trough the TV screen. And for forming part of the Axis of evil, your inappropriate jokes and amazing language skills. And for, as a wise man once said: "simply being Kay". It can buy me a boat. **Abel**, thanks for all the PhD-project progress discussions and for all the memorable parties and hospitality at your house(s). It's been a great couple of years and I'm happy to see that multiple of your projects started working out recently. Thanks for the dinners with Corina and for being my masterclass and PhD retreat roomie. And thanks for your epic Hungarian birthday cake dance, I'll always remember it! **Chloe**, thanks for first unofficially and now officially being part of the tight-nit group that is our lab during all these years. And for all the shared evenings, dinners and masterclass hangovers and being the lab outing organizer. Thanks for being a crucial member of the noisy office and as such always inventing new ways of playing office-tennis. And for your awesome baking skills. You've gone through some rough patches in the past and I'm looking forward to being at your defense. You've definitely earned it. **Adi**, first of all, thanks for being my paranimf. We've had a good bond since you joined the lab a few years ago and I'd like to thank you for all the shared indian smokes, your bash skills, guitar playing and for Samahan. And thanks for embarking on this company adventure with me. Even though the future of it is still insecure, it already has been exciting and I'm looking forward to the rest of it. Thanks to **Shalu** for all the grappa shots and showing us who the real anti-Sid is. And to **Aira** for transmitting her cuteness into our lab life through your stories. **Dominic**, thanks for being our R wizard when we needed one the most. You lead the way in those early frantic days of trying to understand single-cell sequencing data and without you this thesis would have been considerably thinner. I wish you all the best with your family and group in Freiburg. I hope we get to hike the black forest sometime soon. Fa'ka **Jean-Charles,** alles abong? Thanks for all your well-placed "hoezo?"s, for playing squash in the empty rooms of the new Hubrecht building and for testing paper airplane models in the hallway. And thanks for your precise, creative and sometimes incredibly blunt way of doing science. And for your awesome OG hipsterness and the adventures of your stolen trash can. Box, ouwe. **Philipp**, thanks for all the geeky riddles and for educating us about the big RNAse conspiracy. And for all the Monty Python puns and our pipette tip-box Pac man masterpiece. It was truly the best way to deal with bead drying waiting steps. **Sid**, you were part of the original cast of characters when I joined the lab and I've enjoyed working alongside you a lot. Thanks for all the discussions about science and new protocols. And for your contagious laugh when someone made a nerdy enough joke. **Christoph**, thanks for all the BJJ and boxing sparring sessions and your weird humor. I'll miss the Royale dinners with cherry buquet Fernandez and for our discussions about life while on the bike to work or grappling class. **Thom**, bedankt voor al je harde werk tijdens je stage en voor je onbreekbare doorzettingsvermogen. Ik hoop dat we snel een publicatie waar jij ook op staat uit kunnen sturen! Thanks to **Judith** for taking a huge load of our labs shoulders by single-handedly taking care of the Institute's single-cell sequencing demand. You're doing an awesome job at it. And you also thanks for embarking on our joint company adventure! **Anna**, thanks for being a constant source of experimental expertise in the lab. And for your animated stories about sports, kids, food and for sharing wilderness adventure stories. I believe you

are the only one who heard my starving-in-Norway story without flinching and in whom it brought out a story of your own about starving in Siberia. **Anna Catalana!** Thanks for introducing us to Éle and for being my laptop charger supplier in need when I needed to analyze data the night before group meeting. But mostly thanks for being your ever enthusiastic sparkling self. **Susanne**, thanks for all your sudden inappropriate comments, being your eccentric self, and for conversations about eating lab items. **Maya**, thanks for the cookies and your pottery birthday presents. I'll take my mini-fruit bowl with me to all my future desks. **Josi**, bedankt voor het scheppen van hoognodige orde in de chaos van ons oude lab en voor het delen van je enorme kennis wat betreft moleculaire biologie. **Nico**, thanks for sharing mates with me and for your enjoyable grumpiness. Even though you are an Argentinian we've gotten along pretty well ;). **Maria**, thanks for attempting to tame Kay and for your epic eye rolls. **Buys**, thanks for your witty skepticism and the spaghettograms. **Frederic** and **Wouter**, thanks for the boxing lessons and a shared love for martial arts and BBQ. I'm looking forward to our mid-winter grill! **Peter**, **Marloes**, and **Emma**: even though we did not overlap for very long I'd like to thank you for being an awesome new addition to the lab. It's been great seeing how quickly you guys are setting up new protocols. I'm looking forward to reading the papers! Of course thanks to previous lab members like **Nick**, **Magda**, **Emiliano**, **Sandy**, **Sophie**, and all others for your all your help and kindness.

Thanks to the de **Koning and Carlotti** groups. Without all the work your groups have done the research described in this book would have been impossible. **Eelco**, thanks for supervising a big part of the experiments described in this thesis and all the brainstorming sessions on the (sometimes strange) results we got. **Gita**, thanks a lot for all your hard work on our many joint projects (I've counted, there are at least 9 different ones!). I could not have wished for a better partner in crime for combing single-cell sequencing with pancreas biology. I loved your enduring enthusiasm and toughness, even in the face of the thousands of trizol extractions and all hundreds of plates we processed together. I've especially admired how you fought your way trough a difficult time during Vivaans first few months while we were busy with submitting a paper and wish you and your family all the best in the future. Thanks also to both you and **Siddhart** for inviting me for dinners and poker. **Leon**, ook erg bedankt voor alle trizol extracties (helaas kwamen we er na een jaar achter dat we het beter over konden doen met de robot) en voor alle gezamenlijke experimenten in het embryonic pancreas project. Ik wens ook jou en je familie het beste toe in Zwitserland! **Erik** en **Tim**, bedankt voor alle stainings. And many thanks to all other members of the group for their help and discussions. **Françoise** thanks for always thinking along with all the pancreas projects, for always asking how things were going and for all the nice Keystone memories. **Nathalie**, thanks for working together on your project, and being one of the last of the trizol-tribe.

To our neighbors, the **van Rheenen group**: I enjoyed sharing lab and office space with you guys and It's too bad we had to do without you these last few months. **Jacco**, bedankt voor alle korte gang-gesprekken en je eeuwige opgewektheid. **Co**, bedankt voor het samenwerken aan je project en voor de keren dat we samen naar de nieuwe data keken. Thanks to **Carrie** (for your contagious boisterousness and

cat pictures), **Laura**, **Ari** (happy waving from a distance), **Lotte**, **Frank**, **Jessica** (emergency smokes), **Pim** (spontane praatjes bij het koffie apparaat), **Sander**, **Andy**, **Daan** and all others.

I'm equally happy with our new neighbors from the **Kind group**: **Jop**, thanks for supervising my PhD progress. And to **Corina** for all the dinners, Tennis and for sharing food when you were a master student in our group and my desk-neighbor. Also for our joint taste for movie soundtracks and for all the hugs. **Kim**, thanks for being the pioneer of a group that now harbours many people, and with whom our group has many shared projects. Also for your own brand of enjoyable grumpiness. **Sara** for all the "Ciao!"s and many microwave-queue conversations, **Samy** for testing my knowledge on Greek mythology and to all the other members of the group.

I'd also like to thank all members from the **Holstege group** for our collaborations on the SORT-seq protocol. First, thanks to **Frank** for showing an interest in our way of doing CEL-Seq2 and dedicating quite some effort and resources to making the protocol better. **Thanasis**, I've really enjoyed setting up the mosquito and other parts of the protocol with you. I have learned a lot from your tenacious approach and eye for detail when doing experiments. Thanks also to **Philip** and **Tito** for all their work on the computational side of this project.

As I move on to thank **all others** I've worked and bonded with these last 5 years I'm sure I will forget a few. I apologize in advance, but in a sense this is a compliment to you all, since there are so many of you to thank: You have made the Hubrecht a great place to work in. I'm a worried I'll never find anther institute with so much comradery between scientists who are also doing cutting-edge research and know how to throw a party.

Javi, thanks for always running your heart out in the last minutes of our football matches, for all the never-ending video game nights and co-ordering steaks online. Good luck with the last "loodjes" of your PhD. **Tim**, also thanks for the videogame nights and for being my sole source of comfort for NFL talk and crying about the Giants. **Maartje and Pieterjan**, bedankt voor jullie enthousiasme tijdens onze voetbalwedstrijden en de vele borrels. En voor het zetten van een voorbeeld voor wat hard werk is. To all members of our football team(s): **PJ**, **Javi**, **Saman** and **Wim** thanks for organizing. Thanks to **Nico**, to **Samu** for your epic dribbles, **Enric**, it was a pleasure watching you torture the opposing strikers, **Axel** (I'll mis your dry humor), **Geert** (also for all the Brazil-talk), **Abel**, **Frank**, **Ollie**, **Tim**, **Lucas**, **Deepak**, **Lucas Kaaijman**, **Wouter** and all others that I played with. Thanks to my PhD committee buddies **Lars**, **Ingrid**, **Sanne, Eelco** and **Mitchell**.To **Manda** (For sharing an Amsterdam posse background), **Britta** (for laying the scientific tracks for me, I promise I'm not stalking you), **Annabel** (for hopping trough our lab on regular basis and all your delicious cakes), **Christa** for our conversations about food and marathons and being my Lombok neighbor. **Bas** voor het hakken (en pas op voor die wijnvlekken), andere **Bas** voor Pubquizzes en gezelligheid. **Rob** voor alle gesprekken en de lekkere kaas en het uitnodigingen van iedereen voor je verjaardag. **Lolo** and **Euclides** for all the miniconversations in the lunch queue.

Or outside when you guys are lizarding up the last rays of sun and enjoying your cigarettes. And for the parties and limoncello. **Eirinn** and **Rowena** for your genuine kindness and bright Australian optimism. **Erica** for all the dances. Thanks to **Spiros** (I know you're reading this) and **Bana**. Thanks also to **Eirinn**, **Alex** and **Ajit** for all the borrels and to all the people that attend them like **Christian, Sven, Anna, Laurent, Hesther, Monika, AK, Marta, Charlotte, Melanie, Dennis, Kim, Jens, Yorick and Joep** (thanks for the retreat football matches, guys)**, Kadi, Caro, Nicolas, Sasja, Juri, Juliette, Carien and Carlo**. Thanks to **Menno** for his dry introductions to lunchmeetings and to **Catherine** for all her critical questions during the same. **Eva**, thank you for supervising my PhD progress and for help with our company-related questions. **Niels**, **Wouter** and **Hans**, also thanks for the latter.

Van het **ondersteunend personeel** wil ik ook een aantal mensen bedanken:
Annemiek en **Litha**: heel erg bedankt voor al jullie hulp tijdens het klaarmaken van dit boekje en voor alle andere keren dat jullie mijn verwarde vragen met een glimlach op een geduldige en efficiënte manier beantwoordden. **Reinier** en **Stefan**, bedankt voor het flexibel zijn alle keren dat er weer uit het niets een sort gepland moest worden. **De Utrecht Sequencing Facility** voor sequencing experimenten en in het bijzonder **Ewart** en **Mark** voor hulp met het opzetten van een sequencing protocol dat bij onze experimenten paste. Ook bedankt aan het **Islet Isolation Team** in Leiden voor al hun harde werk. **John** voor je onderhandelingen voor de vele apparaten we nodig hadden de afgelopen jaren. **Thea**, dankjewel voor je warme begroetingen iedere dag en het streng zijn wanneer het nodig is (als men kamers niet gereserveerd heeft bijvoorbeeld). Ook bedankt aan de leden van de **civiele dienst**, in het bijzonder **Romke** en **Elroy** voor al hun hulp en aan de **IT guys**, in het bijzonder **Peter-Erik**, **Jimmy** en **Arjan** voor het hunne.

From my **Mount Sinai** colleagues, I'd first like to thank **Michael** for giving me the opportunity and freedom to run a odd-one-out project in your lab and teaching me some valuable lessons I still use every day (like stacking experiments smartly and thinking in terms of a story/publication layout as quickly as possible). **Anita**, thanks for all the endless hours we spent in cell culture together and teaching me how to do iPSC experiments. **Miguel**, thanks for many shared cell culture weekends and for your tip on Alexander's lab when I came to you for PhD-career advice. **Amelie**, **Laura**, **Carlos**, **Francesco**, **Su-Yi**, **Roland**, **Sara**, **Alice**, **Rita**, **Arven**, **Brittany**, **Sara**, **Wissam**, **Orit**: thanks for may espressos, for making New York a home away from home and the dinners at moustache.

Van mijn collegas uit de **Nuclear Organization Group** wil ik graag vooral **Pernette** en Lisette bedanken voor hun begeleiding en omdat ze me voluit de kans gaven om te experimenteren binnen hun lab/project. Ik heb er veel van geleerd! Ook bedankt aan **Roel** voor alle kritische vragen tijdens meetings en aan **Mannus** voor het werken aan de review over reprogramming. And thanks to **Maike**, **Mariliis**, **Anne**, **Diewertje** and all others for help, discussions and coffee breaks in the alien autopsy room on the floor above us.

&

Cristina, Se puede decir que tu clase fue mi introducción al mundo académico y vos siempre te ocupaste que fuese en un ambiente lindo, seguro e interesante. Recuerdo el primer día de escuela, nos habían sentado en filas, esperábamos que nos dijesen quién sería nuestra maestra. También me acuerdo que con sólo verte ya había sido suficiente y estaba deseando que me tocase ir con vos. Eras alegre, simpática y estabas mucho más atenta a todo que las demás maestras y eso era evidente, hasta para un gurí de 4 años. ¡Por suerte me tocó ir a tu clase!
Karel, Ik herinner me jou geanimeerde nabootsingen van romeinse veldslagen tijdens geschiedenis lessen en je enthousiasme om nieuwe dingen te leren als de dag van gisteren. Bedankt voor 2 mooie jaren! **Peter en Adelheid**. Zonder jullie was mijn middelbare schoolcarrière tot een veel minder voorspoedig einde gekomen. Aangezien deze het startpunt voor mijn studie was, en die weer het startpunt was voor dit boekje, is de inhoud hiervan voor een belangrijk deel aan jullie te danken. Bedankt voor jullie vertrouwen in mijn kunnen en voor het dapper de discussie aangaan met jullie eigen collega's tijdens personeelsvergaderingen. Het kan niet makkelijk zijn geweest, en jullie hebben dit niet eens maar meermalen en consequent gedaan. Ik ben jullie enorm dankbaar. **Jan en Juliëttte**. Bedankt voor jullie blijvende interesse in jullie leerlingen. **Jan**, vooral bedankt voor het verhogen van mijn Peano-cijfer omdat je doorhad dat ik de stof echt begreep. Uiteindelijk ging het jou daar uiteindelijk om (ook al had ik zoals gebruikelijk niet genoeg mijn best gedaan om alle axioma's uit mijn hoofd te leren). **Juliëtte**, dankjewel voor de 1C etentjes en voor alle mooie verhalen en lessen. Moge er nog vele volgen.

Wilfred, allereerst bedankt dat je naast me zult staan als paranimf. Sinds we naast geplaatst werden in de eerste klas en erachter kwamen dat we dezelfde Garfield agenda hadden zijn we eigenlijk altijd vrienden geweest. We hebben veel gedeeld: schaaklessen, Tekken, beugels verstoppen op zeilkamp, de dood van Just, Rome reis, avonden in de Doos. Bedankt voor het delen van al deze momenten, en vooral voor al onze gesprekken over wat we willen en gaan doen met ons leven.Ik waardeer ze enorm. **Stephan**, wat mij betreft had jij ook naast me gestaan als paranimf, maar helaas moest een coinflip de keuze maken. Volgens mij zorgde allereerst ons gelijksoortige gevoel voor humor ervoor dat we snel vrienden werden tijdens een random wiskunde les. Tnx voor alle lameheid, hangen bij Youthie, Dissa en het up to date houden van mijn Sranan Tongo en hiphop kennis. Maar vooral voor alle reizen. Het was mooi om me af en toe helemaal af te sluiten van alle moderne gemakken/ afleidingen en deze om te ruilen voor lange avonden aan kampvuurtjes en slopende tochten door moerassen, bergen en bossen. Ik kijk nu al uit naar de volgende tocht.
Gary, Bedankt voor alle lange gesprekken na doos avonden en voor het uitwisselen van de laatste belevenissen via voicechat midden in de nacht. En daarmee ook voor het delen van hetzelfde verknipte bioritme. Tnx voor het introduceren van MMA. BJJ heeft me tijdens het laatste jaar van mijn PhD veel goed gedaan en dat heb ik voor een groot deel aan jou te danken. Ush. **Wanga**, ook bedankt voor alle lame humor, RTW struggles en het uitwisselen van onze PhD ervaringen. Ook al woon je al een tijd redelijk ver weg heb ik niet het gevoel dat onze vriendschap verwaterd is. Ik vind het mooi om te zien dat we na al die jaren en afstand uiteindelijk toch binnen 1 week allebei onze PhD halen. **Dojo**, tnx voor je onuitputtelijke kennis over

politicologie, sonologie, recepten, hallucinogenen, natuurkunde, linux en al die andere onderwerpen waar je verborgen kennis over blijkt te hebben. Ik waardeer je kijk op het leven zeer en ik hoop er nog vele “wohoo’s” plus vingersnaps te mogen horen als we het over iets lijps hebben. **Merel**. Jij was mee naar boston toen ik voor deze PhD solliciteerde, en het is passend dat je me geholpen hebt met de kaft en layout van dit boekje. Heel erg bedankt dat je dit zo last-minute kon doen!. En bedankt voor het rennen, A&M, Choquequirao en Champignon. En voor al je hulp en onze avonturen toen we in New York woonden.
Anne, Gracias por, desde los primeros momentos, haber sido parte de mi vida en Holanda. Siempre nos has ayudado mucho a Lílian y a mí. ¡Y muchas gracias por tu receta de Chimichurri!. **Anke**, Muchas gracias por toda tu ayuda todos estos años -en especial cuando yo estaba en el Barlaeus- y por ir a mi defensa de tesis directo del aeropuerto. **Melby y Charles**, Gracias por haber estado ahí desde nuestros primeros años en Holanda y por compartir una historia muy parecida a la nuestra. Nos hemos llevado muy bien siempre y aprecio mucho nuestra amistad. **Job Sr**., **Job**, **Gustavo**, **Miriam** y **Noelia**, Gracias por todos los cumpleaños, asados y tantos momentos lindos. Espero compartir muchos más con todos ustedes.

Kat, je was een ontzettend belangrijk deel van mijn leven tijdens deze hele PhD. Je leefde altijd mee en was als geen ander altijd echt op de hoogte van hoe het met mij (en mijn onderzoek) ging. Bedankt ook voor je geduld als ik weer eens een vakantie korter inplande dan we gewild hadden vanwege paper deadlines. Het samen maken van mega-hamburgers, reizen naar Sicilië, Marokko, Frankrijk, Gent en het uitlachen van kleine hondjes heeft veel van de moeilijkere momenten tijdens mijn PhD veel lichter gemaakt. Bedankt voor al deze dingen, en voor alle liefde.

Elena y **Peio**, Gracias por habernos invitado a vuestro casamiento y por venir a Utrecht para la defensa de mi tesis. ¡Qué bien que podamos pasar otro día festivo juntos!

Monica, Tabaré, Sofía e Irene, aunque no hemos tenido mucho contacto estos últimos años: gracias por los asados, los mates y por siempre hacerme sentir bienvenido y en casa las pocas veces que nos vimos. No sé por qué, pero escribir esto me hace recordar a cuando los visitaba en la casa de Colón. Eran días lindos y menos complicados.

Don, bedankt voor vele etentjes, voor de gesprekken na het eten als de kinderen naar bed zijn. En voor je eeuwige enthousiasme en vragen naar mijn onderzoek. Ik ben alleen bang dat ik nog steeds geen steek verder ben wat het betreft het vinden van een perfect lokferomoon voor vissen. **Lisa**, ¿Sabés que de mi familia sos la única que me ha ayudado con un experimento en el laboratorio? Me visitaste en el Hubrecht cuando todavía eras chiquita y me ayudaste con mucha seriedad. Era un día de invierno y cuando terminamos ya había oscurecido. Te portaste muy bien durante el para vos tan largo viaje en tren al Hubrecht y después de bicicleta a mi casa. Porque hacía mucho frío fuiste empaquetada en mi buzo que te quedaba muy grande. Pero no nos importó y nos divertimos mirando conejos.

Anna, ¡Mi sobrina seria! Gracias por compartir de verdad, como única, mi amor por los bastognekoekjes con leche, baklava, papeles, libros, mapas y por el gato Garfield. Desde que eras chiquita que nos entendemos muy bien y estoy muy curioso por ver la linda persona que vas a ser de grande. **Daan**, Gracias por jugar al fútbol conmigo en el pannakooi durante tus visitas a Utrecht y las mías a Assendelft. También gracias por las partidas de ajedrez y por tratar de ganarme. Me voy a sentir muy contento el día que me ganes. También por hacer asados junto conmigo, ayudarme con mi jardín en el verano y por tus dibujos. Sobre todo gracias por ser un mimoso y –como lo saben expresar tan bien en Uruguay– tan buena gente. **Claudia**, Gracias por siempre preocuparte por mí y demostrar interés por mi trabajo. Sos la única persona que tuvo el mismo padre ausente que yo. Y aunque no crecimos juntos creo que esto ha creado (paradójicamente) una conexión de hermanos entre nosotros. Para algo sirvieron nuestro pasados complicados que fueron paralelos por mucho tiempo, pero que ahora se han juntado en este lindo presente :)

Mãe, é difícil descrever as coisas por que te agradeço em poucas linhas. Precisaria ser um livro aparte, mas vou tentar: Obrigado principalmente por ter sido uma mãe presente e dedicada e por sempre me ter oferecido casa e carinho (em todos os lugares que moramos). Não pode ter sido fácil ser mãe e pai ao mesmo tempo (lembras da canção Cubana?) viajar a outro continente e começar tudo de novo. Até o dia de hoje às vezes me sinto criança quando visito tua casa, que já faz tempo não é a casa aonde cresci. Mas quando reconheço a limpeza que sempre domina a tua cozinha e banheiro, a rede na sala ou os papéis e livros que sempre tapam as tuas mesas me sinto em casa de verdade. Falando de livros, obrigado por todos os domingos que passamos na biblioteca central de Amsterdã e por me mostrar o importante que e ler e pensar (criticamente). Foram atributos que me serviram e me continuam servindo muito. Vejo esta tese como testemunha disto. Obrigado por sempre ter estado ao meu lado e por me ter exigido, especialmente pelos momentos aonde teria sido mais fácil deixar um assunto de lado. Sempre o fizeste quando achavas que era necessário. E por me ajudar mesmo quando faço algo com que tu não estas imediatamente de acordo. Como o assunto do negócio. Notei que não gostaste da idéia, mas te mordeste a língua, me fizeste perguntas críticas e o dia seguinte me mandaste um link sobre negócios na Holanda. Mesmo que não leio esses mails com a freqüência de que gostaria, os aprecio muitíssimo. Agradeço muito como sempre conseguiste misturar tanto carinho com um pouco de caráter espartano quando se trata do dever e "meter para adelante" (Deves ser uma das poucas meninas que lêem sobre os espartanos na escola primaria e gostam deles). É passado por alguns momentos difíceis na minha vida e durante este PhD, mas sabendo que posso contar contigo e com as coisas que me ensinaste muitos desses momentos tornaram-se em problemas temporários. Ou seja: te quero agradecer pela forma em que me formaste e a tua forma de ser, que é como dizem que disse o Che: "Hay que endurecerse siempre, pero sin perder la ternura jamás".

Mauro

List of publications

* equal contribution
# co-corresponding author

Published

M. J. Muraro, H. Kempe, and P. J. Verschure, Concise review: the dynamics of induced pluripotency and its behavior captured in gene network motifs. **Stem Cells**, 2013.

D. Grün, **M. J. Muraro**, J.-C. Boisset, K. Wiebrands, A. Lyubimova, G. Dharmadhikari, M. van den Born, J. van Es, E. Jansen, H. Clevers, E. J. P. de Koning, and A. van Oudenaarden. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data, **Cell Stem Cell**, 2016.

M. J. Muraro*, G. Dharmadhikari*, D. Grün, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M. A. Engelse, F. Carlotti, E. J. P. de Koning#, and A. van Oudenaarden#, A Single-Cell Transcriptome Atlas of the Human Pancreas.**Cell Systems**, 2016.

P. Dierickx, M. W. Vermunt, **M. J. Muraro,** M. P. Creyghton, P. A. Doevendans, A. van Oudenaarden, N. Geijsen, and L. W. Van Laake, Circadian networks in human embryonic stem cell-derived cardiomyocytes. **EMBO Reports**, 2017.

C. L. G. J. Scheele*, E. Hannezo*, **M. J. Muraro**, A. Zomer, N. S. M. Langedijk, A. van Oudenaarden, B. D. Simons#, and J. van Rheenen#, Identity and dynamics of mammary stem cells during branching morphogenesis. **Nature**, 2017.

Benedetta Artegiani, Anna Lyubimova, **Mauro Muraro**, Johan van Es, Alexander van Oudenaarden, Hans Clevers. An unbiased single-cell RNA sequencing study reveals cellular and molecular dynamics of the hippocampal neurogenic niche. **Cell Reports**, 2017

Submitted

Onur Basak, Teresa G. Krieger, **Mauro J. Muraro**, Kay Wiebrands, Daniel E. Stange, Marc van de Wetering, Johan H. van Es, Alexander van Oudenaarden, Benjamin D. Simons, Hans Clevers. Active Troy+ neural stem cells sense niche size to regulate their number. **PNAS**

Jean-Charles Boisset, Judith Vivié, Dominic Grün, **Mauro J Muraro,** Anna Lyubimova, Alexander van Oudenaarden. Mapping the physical network of cellular interactions identifies new niches in the mouse bone marrow. **Nature methods**

Fabian Kruse, Cilia de Heus, Laurence Garric, **Mauro J Muraro**, Wendy Noort, Dennis E.M. de Bakker, Federico Tassadori, Joshua Peterson, George Posthuma, Dominic Grün, Willem J. van der Laarse, Judith Klumperman, Richard T. Jaspers, Alexander van Oudenaarden, Jeroen Bakkers. Single-cell transcriptomics identifies a cardiac progenitor cell with a distinct metabolism in the regenerating zebrafish heart. **Nature**

In preparation

Léon van Gurp*, **Mauro J. Muraro***, Tim Dielen, Lina Seneby, Gitanjali Dharmadhikari, Gerard Gradwohl, Alexander van Oudenaarden#, Eelco J. P. de Koning#. The dynamics of pancreas development resolved by single cell transcriptomics

Aditya Barve*, Chloé S Baron*, **Mauro J Muraro***, Gitanjali Dharmadhikari, Reinier van der Linden, Eelco J. P. de Koning, Alexander van Oudenaarden. Cell sorting trained by single-cell transcriptome data allows cell type purification without using fluorescent markers

Tito Candelli*, Philip Lijnzaad*, **Mauro J Muraro**, Alexander van Oudenaarden,Thanasis Margaritis#, and Frank Holstege#.A versatile preprocessing and QC pipeline for Single Cell RNA-seq

Curriculum Vitae

Mauro Muraro was born in Porto Alegre, Brazil on the 15th of July 1985. On his first birthday, his family moved to Montevideo, Uruguay, where he grew up and attended the first years of primary school up to 1992. He moved to Holland with his mother in May 1992, where they briefly lived in Nieuwegein, Vlaardingen and Rotterdam before settling in Amsterdam. He attended the Joop Westerweel primary school and the Barleaus Gymnasium secondary school in Amsterdam, where he graduated in 2004. The same year, he began studying Bio-Exact at the University of Amsterdam (UvA), where he started appreciating systems biology and the value of combining experimental with computational work. To wrap up his bachelors, he did an internship at the Pasteur Institute in Paris at the group of Pedro Alzari, where he built a 3D protein model of a mycobacterium tuberculosis protein from X-ray crystallography data. He then enrolled into a masters program in molecular life sciences at the UvA in 2009, during which he did two internships: The first was done at the Nuclear Organization Group in the group of Pernette Verschure at the UvA, focused on epigenetic gene regulation control. The second internship took him abroad to New York, where he worked on reprogramming of mouse skin cells to induced pluripotent stem cells in Michael Rendels group at Mount Sinai Hospital. This internship was a valuable introduction into stem cell and developmental biology and led him to decide to do a PhD that would combine the fields of stem cells and systems biology. The search for a lab that would truly combine the two led him to Alexander van Oudenaarden, who was moving from MIT to the Hubrecht Institute, where he planned to apply his expertise in systems biology to stem cell and developmental biology. He started his PhD in the van Oudenaarden group at the Hubrecht Institue in 2012. The results of this PhD are described in this thesis.
Mauro is currently exploring the possibilities of applying the skills learned at the AvO lab in the form of a company that will help biologists with single-cell sequencing experiments and analysis.