University of Massachusetts Medical School

eScholarship@UMMS

Program in Molecular Medicine Publications and Presentations

Program in Molecular Medicine

2016-10-01

# End Sequence Analysis Toolkit (ESAT) expands the extractable information from single-cell RNA-seq data

Alan G. Derr
*University of Massachusetts Medical School*

*Et al.*

## Repository Citation

Derr AG, Yang C, Zilionis R, Sergushichev A, Blodgett D, Redick SD, Bortell R, Luban J, Harlan DM, Kadener S, Greiner DL, Klein A, Artyomov MN, Garber M. (2016). End Sequence Analysis Toolkit (ESAT) expands the extractable information from single-cell RNA-seq data. Program in Molecular Medicine Publications and Presentations. https://doi.org/10.1101/gr.207902.116. Retrieved from https://escholarship.umassmed.edu/pmm_pp/75

# End Sequence Analysis Toolkit (ESAT) expands the extractable information from single-cell RNA-seq data

Alan Derr,[1] Chaoxing Yang,[2,10] Rapolas Zilionis,[3,4,10] Alexey Sergushichev,[5,6] David M. Blodgett,[7] Sambra Redick,[2] Rita Bortell,[2] Jeremy Luban,[8] David M. Harlan,[7] Sebastian Kadener,[9] Dale L. Greiner,[2,8] Allon Klein,[3] Maxim N. Artyomov,[6] and Manuel Garber[1,8]

[1]*Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA;* [2]*Program in Molecular Medicine, Diabetes Center of Excellence, University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA;* [3]*Department of System Biology, Harvard Medical School, Boston, Massachusetts 02115, USA;* [4]*Institute of Biotechnology, Vilnius University, LT 02241 Vilnius, Lithuania;* [5]*Computer Technologies Department, ITMO University, Saint Petersburg, 197101, Russia;* [6]*Department of Pathology and Immunology, Washington University in St. Louis, St. Louis, Missouri 63110, USA;* [7]*Department of Medicine, Diabetes Center of Excellence, University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA;* [8]*Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA;* [9]*Biological Chemistry Department, Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, 91904, Israel*

RNA-seq protocols that focus on transcript termini are well suited for applications in which template quantity is limiting. Here we show that, when applied to end-sequencing data, analytical methods designed for global RNA-seq produce computational artifacts. To remedy this, we created the End Sequence Analysis Toolkit (ESAT). As a test, we first compared end-sequencing and bulk RNA-seq using RNA from dendritic cells stimulated with lipopolysaccharide (LPS). As predicted by the telescripting model for transcriptional bursts, ESAT detected an LPS-stimulated shift to shorter 3′-isoforms that was not evident by conventional computational methods. Then, droplet-based microfluidics was used to generate 1000 cDNA libraries, each from an individual pancreatic islet cell. ESAT identified nine distinct cell types, three distinct β-cell types, and a complex interplay between hormone secretion and vascularization. ESAT, then, offers a much-needed and generally applicable computational pipeline for either bulk or single-cell RNA end-sequencing.

[Supplemental material is available for this article.]

Since it became possible to build and sequence cDNA libraries, RNA-seq has become the most widely used method for genome-wide transcriptome analysis. RNA-seq can be used for many different purposes, from transcriptome quantification to annotation and, most recently, measurement of translational or transcriptional rates (Ingolia 2010; Garber et al. 2011; Rabani et al. 2014). Measuring gene expression from RNA-seq data is complex and presents computational challenges that are unique to RNA-seq: (1) When RNA from a cell population is sequenced, only relative gene or isoform expression can be determined, and (2) statistical models to estimate transcript abundance are confounded by ambiguously mapped reads, uneven transcript coverage, uneven amplification during library construction, low library complexity when initial input is limiting, and many other variables (Bullard et al. 2010; Roberts et al. 2011; Kawaji et al. 2014).

Libraries that generate one tag per transcript give a *digital gene expression* (DGE) measurement. Such libraries target transcript termini rather than the full transcript, and they were introduced soon after full-length RNA-seq library construction methods were first developed (Asmann et al. 2009; Matsumura et al. 2010). DGE libraries have obvious advantages over full-length RNA-seq libraries: They work well for low-quality RNA; PCR duplicates arising during amplification are easily detected by using molecular indices; and since each mRNA molecule is represented by a single tag, quantification is greatly simplified (Asmann et al. 2009; Matsumura et al. 2010; Shiroguchi et al. 2012; Kawaji et al. 2014). While the simple library construction by poly(A) selection or priming has made sequencing the 3′ end of transcripts the most common approach for DGE, 5′ sequencing is also a viable strategy for DGE, and several methods exist that take advantage of the 5′ cap that protects eukaryotic mRNAs to build libraries that target the start of transcripts rather than their ends (Gu et al. 2012; Takahashi et al. 2012).

Until very recently genome-wide transcriptional profiling was relegated to RNA from bulk populations. Many studies of single cells showed critical differences between single cells that are masked in bulk cell data (Apostolou and Thanos 2008; Janes et al. 2010; Zhao et al. 2012; Bajikar et al. 2014). Single-cell RNA-seq techniques have enabled single-cell transcriptomics, and we find that the properties of end-sequencing have made DGE the basis for many single-cell sequencing protocols (Hashimshony et al. 2012; Jaitin et al. 2014; Soumillon et al. 2014; Klein et al. 2015; Macosko et al. 2015).

[10]**These authors contributed equally to this work.**
**Corresponding author: manuel.garber@umassmed.edu**

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.207902.116. Freely available online through the *Genome Research* Open Access option.

Here we describe and apply an End Sequence Analysis Toolkit (ESAT) designed for the analysis of short reads obtained from end-sequence RNA-seq. In this context, we refer to both 3′ and 5′ selective methods as *end-sequencing* and will mostly treat them as similar for all computational matters. ESAT addresses misannotated or sample-specific transcript boundaries by providing a search step in which it identifies possible unannotated ends de novo. It provides a robust handling of multimapped reads, which is critical in 3′ DGE analysis. ESAT provides a module specifically designed for alternative start or 3′ UTR (untranslated region) differential isoform expression. It also includes a set of features specifically designed for the analysis of single-cell RNA-seq data.

As a test case for the utility of ESAT, we first analyzed end-sequence data from both bulk cells and single cells. We generated 5′ and 3′ end-sequence data for mouse bone marrow–derived dendritic cells (mBMDCs) stimulated with LPS, and compared these data to our previously generated full-length RNA-seq data (Garber et al. 2012). We also applied ESAT to single-cell RNA-seq from approximately 1000 rat pancreatic islet cells using a new droplet barcoding method for single-cell transcriptomics (Klein et al. 2015).

## Results

### Accurate mapping of end-sequence libraries presents unique computational challenges

The use of RNA-seq libraries constructed from transcript termini has increased steadily since these powerful methods were first described (Asmann et al. 2009; Matsumura et al. 2010). However, current computational methods for gene quantification are not ideally suited to such end-sequence data. For example, these analytical methods assume that reads originate uniform coverage along the length of the transcript (Li and Dewey 2011; Trapnell et al. 2012; Roberts and Pachter 2013). In spite of the wide adoption of these transcript-end libraries, there has been no concerted effort to evaluate the computational methods that specifically analyze end-sequence data.

To evaluate and test the best approaches for end-sequence data analysis, we generated separate 5′ and 3′ DGE libraries (Methods) (Fig. 1A) using mBMDCs. RNA was isolated after the LPS-stimulation time course. This is a well-studied transcriptional response model for which there is a wealth of expression data available (Garber et al. 2012; Rabani et al. 2014; Jovanovic et al. 2015). It is therefore an optimal experimental system to evaluate computational approaches for gene expression analysis of end-sequence data.

5′ and 3′ end libraries were sequenced to an average depth of 16 million and 19 million reads, respectively, of which 92.1% (5′ libraries) and 88.6% (3′ libraries) aligned to the genome after low-quality reads were filtered out (Supplemental Table S1). Variability between biological replicates showed very good reproducibility ($R > 0.98$; Methods) (Fig. 2A,B; Supplemental Fig. S1). Comparison with previously generated full-length libraries generated by the same laboratory, using the same protocol at two different times (Garber et al. 2012; Jovanovic et al. 2015), showed that the difference between end-sequence and full-length libraries was similar to the difference between full-length libraries from biological replicates made by different people at different times (Supplemental Fig. S2).

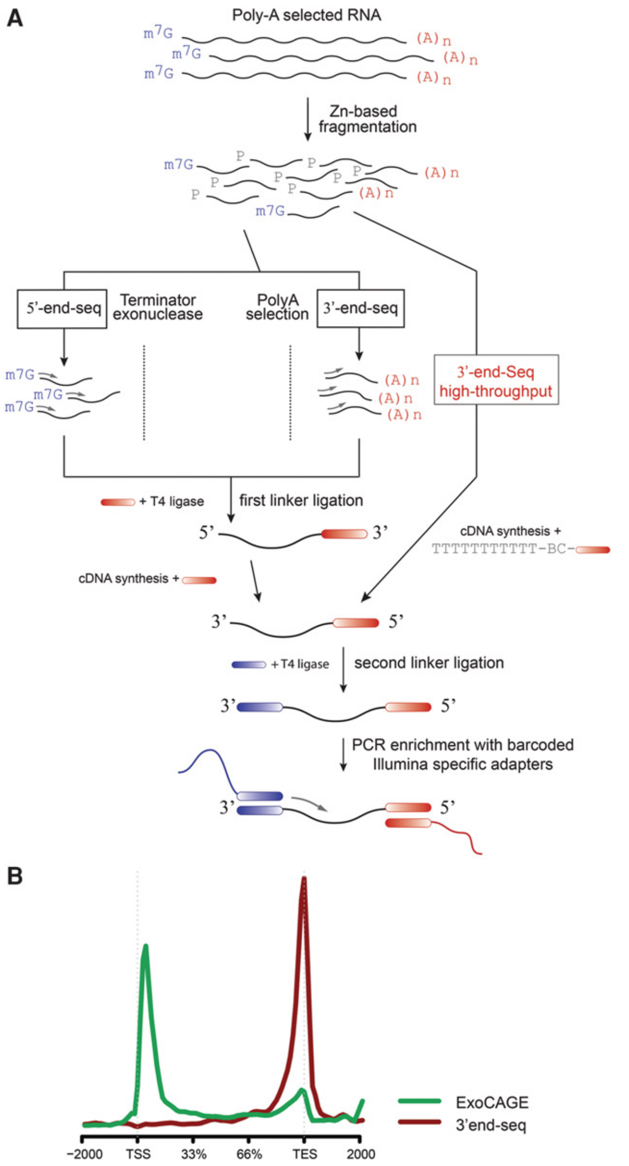


**Figure 1.** End-sequencing libraries for bulk RNA. (*A*) Schematic representation of end-sequencing library methods. (*B*) Aggregation of the location of each aligned read, using the annotated transcription start sites (TSSs) and transcription termination sites (TTSs) as reference, for 5′ (green) and 3′ libraries (red).

Inspection of the data (Fig. 2A; Supplemental Fig. S1) indicated that there were higher rates of variability with the end-sequence libraries—and 3′-end-sequencing methods in particular—than with full-length RNA-seq libraries. First, several genes (marked red in Fig. 2) had poor correlation between 3′-end library replicates. This was surprising since the same genes did not exhibit such variability in 5′-end libraries (Fig. 2A,B; Supplemental Fig. S1). Further inspection showed that all genes with poor replicability had unusually high read pileups on internal poly(A) sequences that tended to be randomly overrepresented from replicate to replicate. Overamplification of low-complexity, adenosine-rich sequences combined with poor handling of those sequences by mapping algorithms may result in the observed variable genes. Second, a sizable number (up to 1800) of RefSeq annotations expressed a 5′ or 3′ end that was different than the annotated end (Fig. 2C–E), as has also been found in previous reports (Carninci et al. 2006;

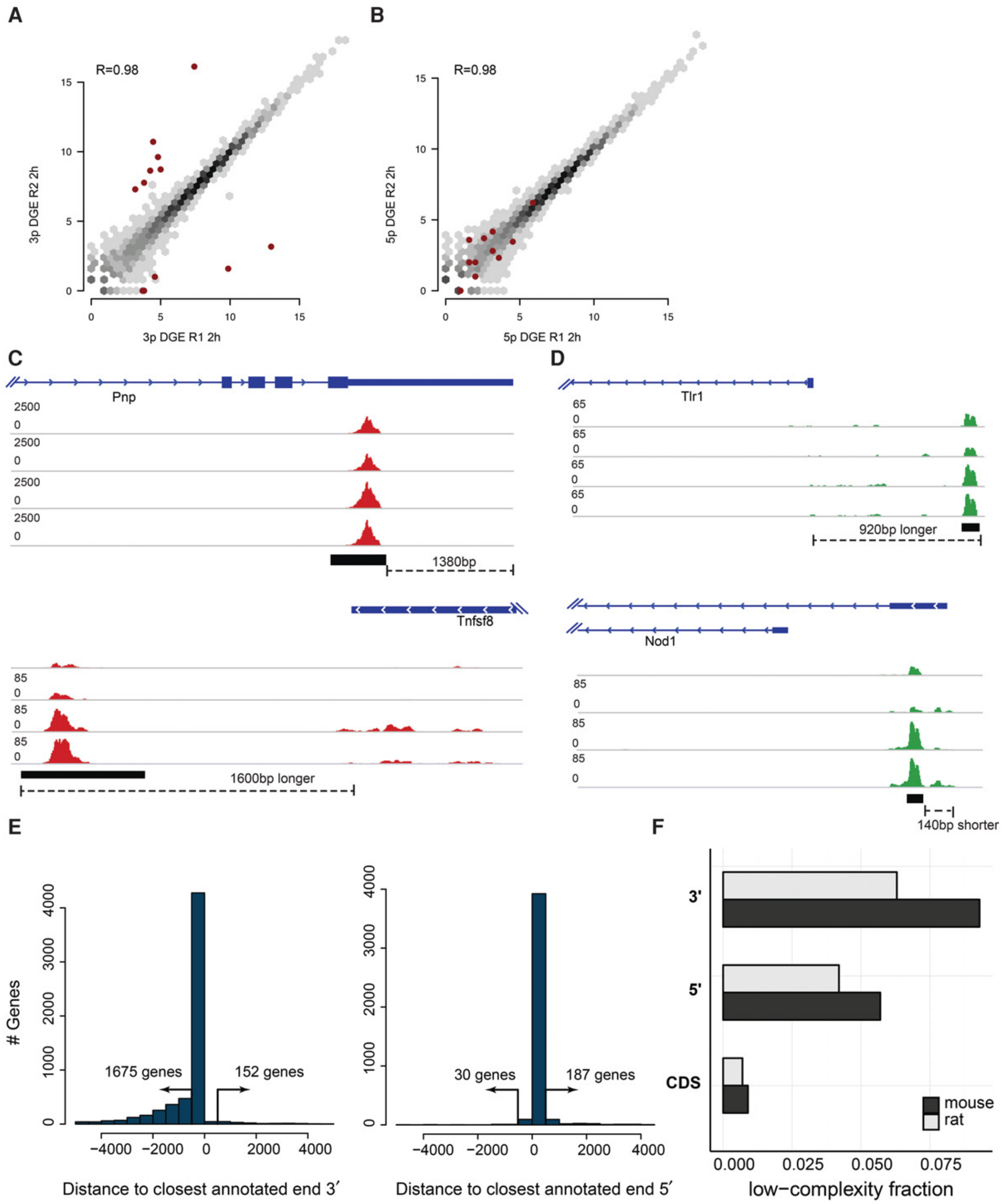
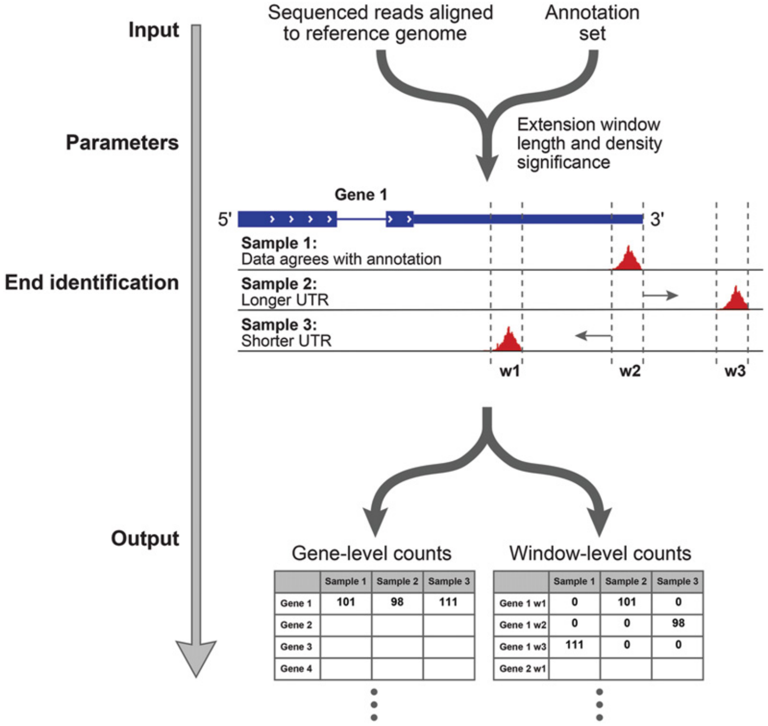


**Figure 2.** Common problems in end-sequence analysis. (*A*,*B*) Scatter plots of gene expression computed from 3′ (*A*) and 5′ (*B*) libraries made from technical replicates of mouse bone marrow–derived dendritic cells (mBMDCs) 2 h after LPS stimulation. Red dots highlight outliers (at least 10-fold difference between replicates in 3′ libraries). (*C*,*D*) Examples of annotated TSSs and TTSs that do not correspond to observed start and end sites in our samples. Read coverage is normalized to library size. (*E*) Distance from the most highly enriched window within each gene to the annotated TTS for the 2-h 3′ library (*left*) and to the TSS for the 5′ library (*right*). (*F*) Fraction of repetitive sequence in 3′ UTRs, 5′ UTRs, and coding sequence (CDS) as estimated by RepeatMasker, downloaded from the UCSC Genome Browser (Smit et al. 2004; Rosenbloom et al. 2015), in mouse (black) and rat (gray) annotated genes.

Shepard et al. 2011; Derti et al. 2012; Hoque et al. 2013). Taken together, these observations suggest that a robust end-sequence quantification method must find transcript ends de novo, rather than rely on gene annotations, as well as filter out any read alignments to internal polyadenylation sites.

## A method specifically designed for end-sequence quantification

In response to our observations, we created ESAT specifically to process end-sequence data. ESAT addressed the two issues described above, as well as other sources of unwanted variance that are specific to end-sequencing: (1) Internal poly(A) sites can result in artificial read alignment pileups and compromise gene quantification (Fig. 2A,B); (2) poor or cell-specific transcript ends can result in missing signal if only reads that align to annotation databases are included, which especially affects 3′-end-sequencing as the 3′ end tends to be more poorly annotated or more cell-type specific (Fig. 2C–E); (3) UTRs, and 3′ UTRs in particular, contain a higher fraction of repetitive or low complexity sequence compared with coding sequence, which adversely affects read mappability (Fig. 2F); (4) single-cell library construction methods are diverse and need different analysis strategies, but the main difference is in how barcodes are added and hence how they are handled computationally (Hashimshony et al. 2012; Jaitin et al. 2014; Soumillon et al. 2014); and (5) alternative transcription start and polyadenylation usage can be measured much more precisely using end-sequencing, but current methods are not ideally suited for this analysis.



|  | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| Gene 1 | 101 | 98 | 111 |
| Gene 2 |  |  |  |
| Gene 3 |  |  |  |
| Gene 4 |  |  |  |

|  | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| Gene 1 w1 | 0 | 101 | 0 |
| Gene 1 w2 | 0 | 0 | 98 |
| Gene 1 w3 | 111 | 0 | 0 |
| Gene 2 w1 |  |  |  |

**Figure 3.** Schematic representation of the ESAT pipeline.

As input, ESAT takes spliced, genomic read alignments from end-sequence libraries together with a gene annotation set (Pruitt et al. 2009; Derrien et al. 2012; Yates et al. 2016). To avoid internal poly(A) sites, ESAT offers the option to discard reads containing continuous stretches of A's or T's longer than a user-defined threshold (10 in this study). To handle incomplete or inaccurate annotations, ESAT scans a user-specified distance beyond the annotated transcript end to find possible transcript ends. ESAT reports all candidate alternative ends by identifying windows with significant read coverage, after controlling for the gene expression level (Methods). After identifying all significant windows, ESAT reports counts at the *gene* and *window* levels, where only windows that exceed a user-specified significance are reported. To reduce the effect of repetitive sequences resulting in ambiguously mapped reads, ESAT has an optional feature that keeps reads that cannot be uniquely mapped genome-wide but that map uniquely to one gene (extended) UTR. To handle different UMI (unique molecular index) strategies, ESAT relies on an external preprocessing step, which appends the cell barcode and UMI to the read ID as a colon-separated string but makes no assumptions about the length of either the cell barcode or UMI, making it simple to support new methods. Finally, ESAT quantifies alternative transcription start sites (TSSs) or transcription termination sites (TTSs) by relying on the significant windows that represent candidate transcript ends (Methods) (Fig. 3).

From the annotations provided, ESAT defines *gene* loci by grouping all isoforms of each gene into a union gene model. The *gene* count is reported as the sum of all reads that mapped to this union gene model. We interpret this count as the sum of the counts for all of the expressed isoforms of the gene. Although using gene counts over the gene body has been shown to result in systematic error (Garber et al. 2011; Trapnell et al. 2012), the nature of end-sequence libraries, which only produce reads for a small portion of each transcript, is not as prone to this systematic error as is full-length RNA-seq library analysis. At the *window* level, ESAT reports the counts for each of the significant windows found. We note that because some 3′-end-sequencing protocols may include internal polyadenylation sites that are not true alternative ends, ESAT allows for the optional removal of any read alignment that contains runs of adenine or thymidine longer than a user-specified length (Methods).

Although there is good correlation between counts obtained at the annotated ends compared with full-length TPMs (transcripts per million) or expected counts (Li and Dewey 2011), using de novo end-finding allowed us to improve the quantification of hundreds of genes without any impact on variability or correlation with full-length libraries (Supplemental Fig. S3A). In all of our analyses, we used an extension of 5 kb for 3′ data and of 1 kb for 5′ data based on the observation that these extensions resulted in the highest correlation with full length (Supplemental Fig. S3B).

## End-sequencing reveals a global change to long 3′ UTR usage upon LPS stimulation

Control of gene expression is complex, often involving many different processes (Moore 2005). Alternative promoter and polyadenylation usage are two key mechanisms by which expression can be controlled transcriptionally or post-transcriptionally (Elkon et al. 2013; de Klerk and 't Hoen 2015). Although the response of mBMDCs to LPS has been extensively characterized (Amit et al. 2009; Garber et al. 2012; Rabani et al. 2014; Jovanovic et al. 2015), the exact contribution of these two mechanisms to the transcriptional response to LPS is largely unknown.

To monitor changes in the expression of transcript ends, we used ESAT's *window* output that represents likely gene TSSs or TTSs that are expressed during the LPS response time course (Methods). To test for changes in TSS and TTS usage independently from changes in total expression, we computed the fraction of reads from each TSS or TTS and tested changes in this distribution (Methods). We found no significant change of either TSS or TTS usage at any specific locus. We observed, however, a significant global trend toward short isoform usage when looking at all genes that

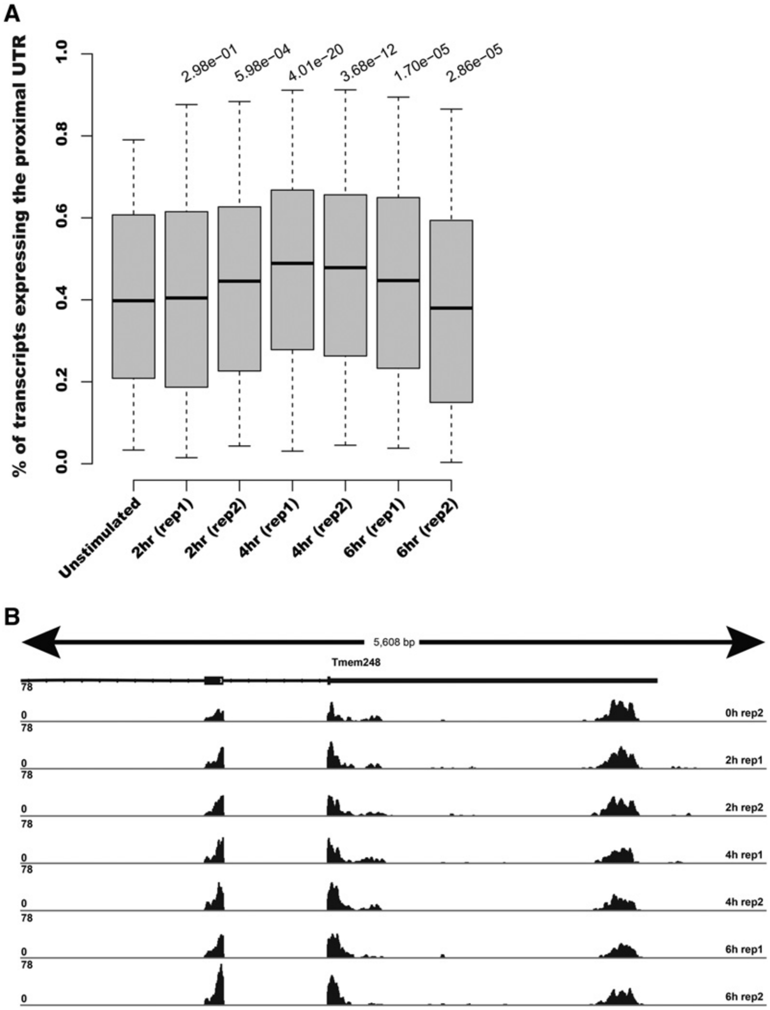


**Figure 4.** Global switch to shorter 3′ UTR expression in stimulated DCs. (*A*) Boxplots of the fraction of transcripts expressing the shortest UTR for genes with detectable expression of at least two distinct 3′ UTRs in unstimulated DCs (total of 1807). *P*-values were computed using a Mann-Whitney rank sum test between the unstimulated distribution and each of the time points shown. (*B*) Illustrative example showing the subtle yet reproducible increase in the expression of the shorter isoform of *Tmem248* in stimulated DCs. Read coverage is normalized to library size.

have multiple alternative TTSs (Fig. 4). The trend is detectable at 2 h after LPS stimulation and peaks at the 4-h time point. No such trend is observed when using 5′ data (Supplemental Fig. S4), suggesting that the TTS effect was not an artifact of our analysis. Although the exact mechanism or consequence of this global phenomena is unclear, it is consistent with a recent report describing a "telescripting" process where, as a result of a transcriptional burst, protected poly(A) sites become accessible, resulting in a trend toward shorter UTRs (Berg et al. 2012) that is similar to what we observe. Whether this trend has a role in the response of mBMDCs to LPS remains an open question.

## Single-cell islet sequencing reveals a complex cellular composition

Given the suitability of 3′ sequencing protocols for both multiplexing and low input, it is no surprise that it is the basis for three recently reported single-cell RNA-seq methods (Hashimshony et al. 2012; Jaitin et al. 2014; Soumillon et al. 2014). We reasoned that our computational approach will easily apply to the analysis of these and other 3′-based single-cell protocols.

To fully evaluate the usability and relevance of ESAT for single-cell RNA sequencing, we used inDrop (Klein et al. 2015) to map islet cell composition in a rat model for type 1 diabetes (T1D) (Zipris et al. 2005) that we have extensively studied in the past (Greiner et al. 2001; Bortell and Yang 2012). This system is ideal for two reasons. First, all previous descriptions of islet composition have relied on known markers of cells (Pechhold et al. 2009; Dorrell et al. 2011; Hrvatin et al. 2014; Blodgett et al. 2015) or sequenced a relatively small number of cells (70) (Li et al. 2015). inDrop allows the profiling of thousands of cells, and hence, rare cell populations of even <1% are likely to be well represented. Detection of rare cells is a powerful use of single-cell RNA sequencing (Grün et al. 2015), and its application to T1D disease progression will likely transform our current understanding of this disease. For example, detection of activated macrophages is critical as they may be indicators of autoimmune disease and T1D progression in particular. Second, rat genome annotation quality lags far behind the human and mouse genomes, and 3′ sequencing quantification is therefore expected to be particularly challenging, providing a good test for ESAT's transcript quantification approach.

We harvested islets from a biobreeding diabetes-resistant rat (BBDR) that is phenotypically normal but can be induced by environmental perturbation to rapidly develop an autoimmune T1D-like disease (Bortell and Yang 2012). We used nondiabetic BBDRs as islet donors and isolated and dissociated the islets into single-cell suspensions as previously described (Blodgett et al. 2015). Cells were flowed through the inDrop system to obtain single-cell RNA-seq libraries, which were then pooled together for sequencing as previously described (Fig. 5A; Klein et al. 2015). In all, we sequenced 1063 cells at an average 147 thousand reads per cell, which we processed using an ESAT feature specially designed to handle barcoded cells and UMIs (Methods) (Supplemental Table S2). UMIs allow robust identification of PCR duplication artifacts that are common in low-input libraries (Shiroguchi et al. 2012).

Similar to bulk end-sequencing, the UTR extension step greatly improves inDrop usability. After extending UTRs, we recover 24% more highly expressed genes, and 14 more cells pass our threshold criteria (Methods) (Supplemental Table S2). Importantly, key genes may be missed without proper UTR handling. For example, expression of genes such as *Iapp* (also known as amylin) would be grossly underestimated without properly handling their 3′ UTR. We removed genes with unreliably low expression by selecting only genes with at least two UMIs in at least three cells (Methods). In all, 991 cells (93%) and 8264 genes passed our validation criteria and were used in the analysis (Methods).

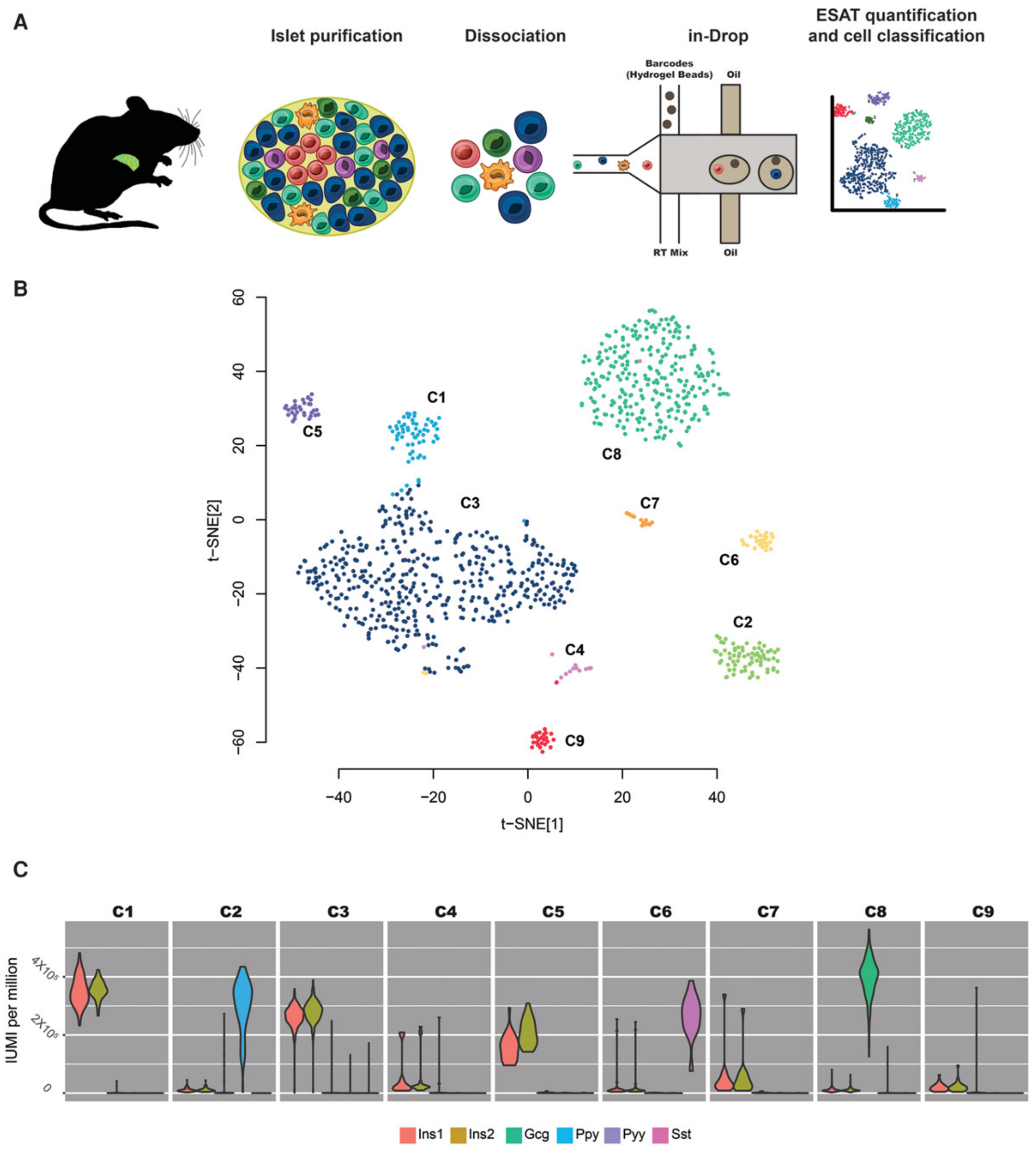


**Figure 5.** Single-cell analysis of rat pancreatic islets. (*A*) Summary of the study. (*B*) Two-dimensional view of a nine-component independent component analysis (ICA) projection using t-distributed stochastic neighbor embedding (t-SNE). Cells are colored according the clusters obtained after spectral clustering (Methods) of the nine-component ICA projection. (*C*) Violin plots showing the distribution of hormone expression across cells in each cluster.

Although islet cell composition is dominated by hormone-secreting cells, many other cell types contribute to these complex organoids. By using independent component analysis (ICA) to reduce the high dimensional nature of single-cell transcriptomic data, followed by spectral clustering (Methods), we found that the cells fell into nine distinct groups (Fig. 5B). In six of these groups, accounting for 93% of the cells, hormone expression is dominant: Three clusters (C1, C3, C5) are dominated by insulin (encoded by the *Ins1* and *Ins2* genes in rat, 57% of cells), while glucagon (*Gcg*, 27% of cells), somatostatin (*Sst*, 3% of cells), and pancreatic polypeptide (*Ppy* 6% of cells) dominate C8, C6, and C2, respectively (Fig. 5C).

We next sought to classify the remaining three cell clusters, in which expression was not dominated by hormone secretion transcripts. To this end, we carried out differential gene expression analysis between the different clusters. We obtained 942 genes that were both significant (Benjamini-Hochberg adjusted *P*-value <0.01) and at least twofold different between any two sets of cells (Methods). Clustering of these 942 genes revealed clear signatures of the remaining three clusters (Fig. 6A; Supplemental Fig. S5): While gene ontology analysis shows strong enrichment in vascular development for clusters C4 and C9 (Supplemental Table S3), manual inspection of cluster-specific genes revealed that C4 cells specifically express pericyte markers (Paquet-Fifield et al. 2009), while

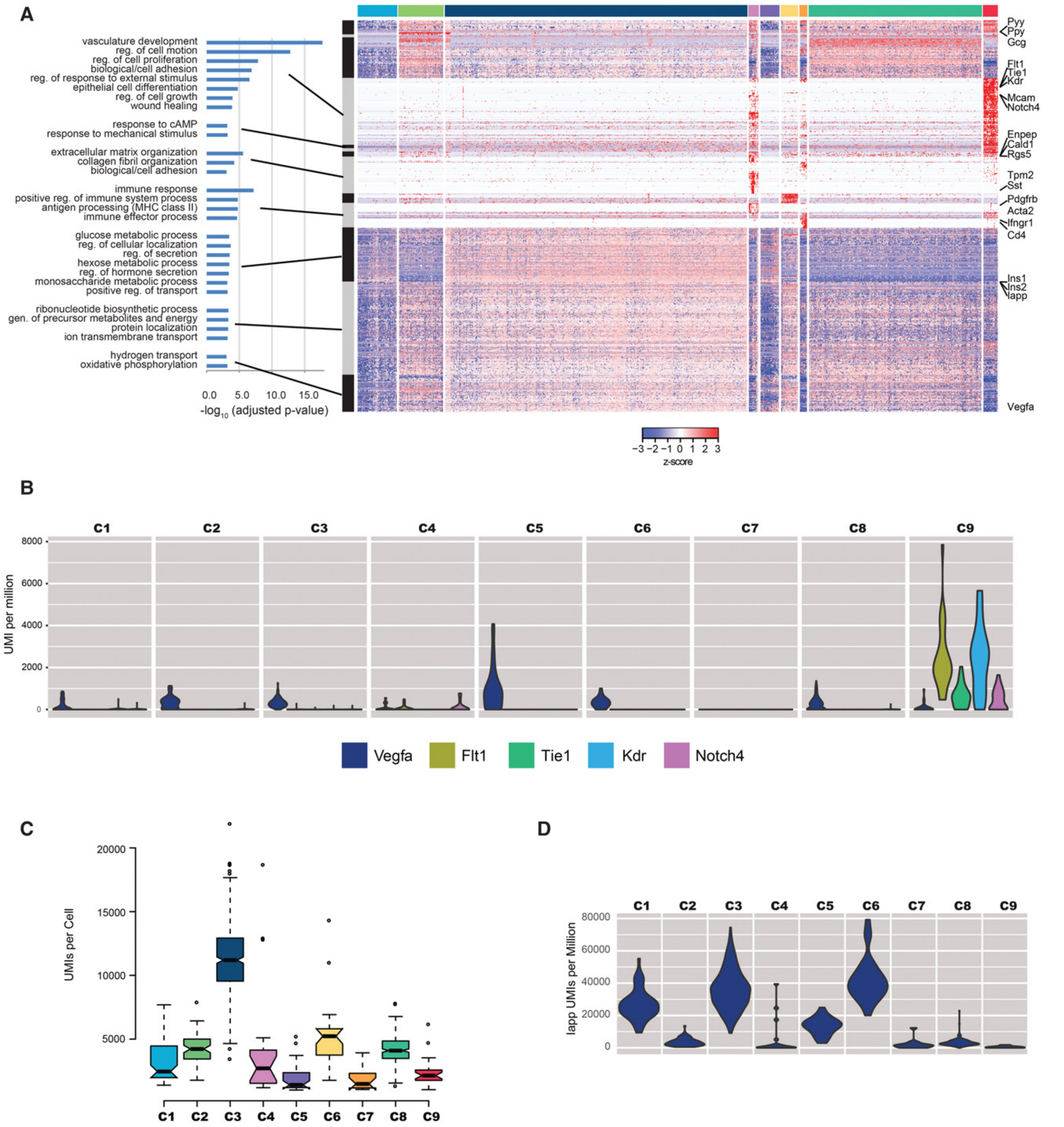


**Figure 6.** Classification of 1000 islet cells. (*A*) Hierarchical clustering of the 940 genes that are differentially expressed between any pair of the nine-cell clusters. Gray and black rectangles indicate major gene clusters identified by hierarchical clustering, and Gene Ontology terms significantly enriched (adjusted *P*-value <0.001) for each group are indicated (data available in Supplemental Table S3). Each of the cell clusters (same order and color as in Fig. 5B) are indicated by the rectangles at the *top* of the figure. Hand-picked genes of interest are highlighted at the *right* of the figure. (*B*) Violin plot showing *Vegfa* expression and that of its receptors. (*C*) Boxplots showing the differences in the distribution of total UMIs per cell for each of the cell clusters. (*D*) Violin plots showing the normalized expression of amylin (*Iapp*) in UMIs per million across the cells in each cluster.

cells in the C9 cluster express typical angiogenesis markers (Fig. 6A, B; Shih et al. 2002). Cluster C7 is enriched in genes with immune function and likely represents innate cells such as macrophages or dendritic cells (Supplemental Table S3), which have been previously reported to reside in islets and are critical for proper development (Xiao et al. 2014; Morris 2015). Our data also revealed an interplay between hormone-producing cells and vascular cells (cluster C9). All hormone-producing cells, and in particular α and δ cells, express high levels of vascular endothelial growth factor A (*Vegfa*), while cells in cluster C9 express its receptors (e.g., *Flt1*, *Kdr*) and the transcription factors induced upon VEGFA signaling (e.g., *Jun*, *Fos*, *Egr1*, gene cluster 3a in Supplemental Table S3). This highlights the power of single-cell analysis, not only as a cell classification method but also as a tool to uncover more complex signaling among cells within an organ.

We note that in contrast to a recent report on single-cell mRNA sequencing of human islet cells (Li et al. 2015), our data contain neither acinar nor ductal cells. This is likely a result of the much simpler islet isolation protocols in the rat model system compared with the more-involved islet isolation from human samples, resulting in a pure islet population with no detectable contamination of cells that are not present in the islet proper. The comparison also highlights the importance of sequencing a large number of cells. With inDrop, we analyzed 991 cells compared with the 70 human cells previously reported (Li et al. 2015). The larger number of cells sequenced not only revealed important low abundance cells (immune and vascular) but also showed subtle differences between hormone-secreting cells.

Interestingly, our data suggest that there is great heterogeneity in hormone-producing cells. For example, three clear clusters of insulin-producing cells emerge differing mainly in the total number of transcripts detected. As with previously published RNA-seq data in heterogeneous cell populations, different cell types yield a different number of transcripts, likely due to their different raw number of mRNA molecules but also possibly due to different efficiency in library construction. We observe among the different cell types an inverse correlation between insulin and amylin (*Iapp*) (Fig. 6C,D). Since dysregulation of amylin has been associated with β-cell death in type 2 diabetes (Huang et al. 2007; Guardado-Mendoza et al. 2009), it will be interesting to know in which cells and the exact portion of cells where amylin becomes dysregulated during diabetes onset.

We had recently reported that after sorting human islet cells using intracellular staining of insulin, glucagon, and somatostatin, a small fraction of α-cells appeared to also transcribe insulin, although without any detectable protein observed (Blodgett et al. 2015). This was surprising since it is commonly thought that islet hormone-secreting cells express a single hormone. Our BBDR data do not provide strong evidence for or against this phenomenon in rats; we detected a small number of cells (16) that express high levels of both *Gcg* and *Ins1/2* (Supplemental Fig. S6A). However, this is not unexpected given that 14 is the number of expected doublets assuming a 4% doublet rate (Klein et al. 2015). Nevertheless, it is notable that while these 16 cells express β-cell genes at high levels, their expression of α-cell genes is low, perhaps pointing to β-cell-like cells expressing low levels of glucagon (Supplemental Fig. S6B; Supplemental Table S3). Whether a fraction of hormone-producing cells express more than a single hormone will be better addressed with a larger number of sequenced cells.

## Discussion

End-sequencing techniques have become commonplace. We have presented a unified perspective of these techniques, both computational and experimental. By providing a computational framework specially designed for analysis of data of this type, we hope to increase the accessibility and usefulness of these methods.

Our study of bulk RNA in LPS-stimulated DCs revealed a surprising shift in 3′ usage. Although its functional consequence remains to be tested, previous reports suggest that this shift may result in regulatory changes of transcripts produced after stimulus, perhaps reducing its stability and increasing its turnover. Such observations are only possible with end-sequencing libraries, since subtle changes in UTR usage are not readily observable with standard RNA-seq libraries.

Since many single-cell methods rely on end-sequence libraries, we also attempted to provide a computational tool that is readily usable by investigators generating such data. Our rat islet data highlighted the need for specialized handling of these data as, for example, 24% of genes would be unusable without searching for unannotated UTRs. Single-cell data show that islet analysis can now be extended to include all cells within the islet rather than just those for which there is a readily available antibody. Our analysis shows islets are composed of at least nine distinct cell types, including multiple β-cell subtypes, as well as it helped us detect a complex interplay between the different islet cells types that culminates in VEGF signaling by β-cells that drives vascularization through an ERK-dependent pathway in endothelial cells. It would be fascinating to extend these analyses to model organisms such as the NOD mouse or the BBDR where onset of T1D can be monitored (Bortell and Yang 2012; Yang et al. 2013b). In these models, single-cell analysis would reveal not only changes in the exact composition of hormone-producing cells but also subtle differences within each group. We also believe that our observation of VEGFA signaling between hormone-producing cells and vascular cells points to the potential of using single-cell sequencing to study not only changes in organ composition but also changes in signaling between cells in the islet prior to T1D development.

We finally note that our islet data should provide an important complement to similar human and mouse single-cell RNA data sets that are currently being produced.

## Methods

### 5′ Sequencing

5′ RNA sequencing libraries were generated following a recently developed protocol (S Afik, O Bartok, M Artyomov, A Shishkin, S Kadri, X Zhu, M Gutman, P McDonel, M Garber, S Kadener, in prep.). Briefly, RNA was fragmented and then enriched for 5′ ends using Terminator Exonuclease (Epicentre). Reaction mixture was cleaned up with 2.5× of SPRI beads and then dephosphorylated with FastAP (Fermentas), cleaned (2.5× SPRI, Agencourt), and then ligated to a linker1 (5Phos/AXXXXXXXXA GATCGGAAGAGCGTCGTGTAG/3ddC/; XXXXXXXX is an internal barcode specific for each sample) using T4 RNA ligase I (NEB). Ligated RNA was cleaned-up by Silane beads (Dynabeads MyOne, Life Technologies) and pooled into a single tube. Reverse transcription was then performed for the pooled sample, with a specific primer (5′-CCTACACGACGCTCTTCC-3′) using an AffinityScript multiple temperature cDNA synthesis kit (Agilent Technologies). Then, RNA-DNA hybrids were degraded. The reaction mixture was cleaned up using Silane beads and a second ligation was performed, where the 3′ end of the cDNA was ligated to linker2 (5Phos/AGATCGGAAGAGCACACGTCTG/3ddC/) using T4 RNA ligase I. Reaction mixture was cleaned up (Silane beads), and PCR enrichment was performed using enrichment primers 1 and 2 (5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA CGACGCTCTTCCGATCT-3′, 5′-CAAGCAGAAGACGGCATACGA GATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGA TCT-3′, where XXXXX XX is barcode sequence) and Phusion HF MasterMix (NEB). After clean-up with 0.8× volume of SPRI beads, the library was sequenced. The barcoding-first strategy is easily incorporated into all three protocols reported here. Specifically, this is achieved by using linker1-XXXX(barcode) sequences for 5′/full-length protocols or the linker1-XXXX-T15 primer for 3′-end RNA-seq. However, from the experimental perspective, the most streamlined, least time consuming and high-throughput protocol is the barcoding-first variation of 3′-end RNA-seq.

## 3′-end RNA-seq

Fifty nanograms to 1 µg of total RNA is poly(A) selected using Oligo-dT beads (Invitrogen) and then fragmented using Ambion fragmentation buffer using an incubation time of 1 min 50 sec at 70°C. Fragmented RNA is then cleaned-up using 2.5× volume on SPRI beads (Agencourt) and then subjected to a second poly(A) enrichment step. Specifically, we used cDNA synthesis with custom oligo-dT priming (CCTACACGACGCTCTTCCGATCT-T15). The reaction mixture is then processed in the same way as in the 5′ protocol starting with hybrid degradation and so on. Alternatively, when using a second round of poly(A) selection after RNA fragmentation for 3′-end targeting, enriched RNA is processed in the same way as described for 5′-end-sequencing. While both methods show good correlation with full-length RNA-seq (0.95 in both cases), the choice of specific 3′ end enrichment method should be made depending on the goals of the experiment. Specifically, for 3′-end-based quantitation, custom oligo-dT priming provides more streamlined design, while secondary poly(A) selection is preferred when constructing an annotation library to achieve the best compatibility with full-length and/or 5′-end-sequencing.

## Sequencing

Constructed libraries were sequenced with 2×25-bp paired-end Illumina pipeline that is part of standard sequencing pipeline at Broad Institute. For all analyses, we used only a single mate out of the pair to mimic single end-sequencing available in most standard sequencing core facilities. In our analysis, we used one mate read at the 3′ end of each cDNA fragment (5′ end in the original RNA fragment) (Fig. 1A). In 5′ end (ExoCAGE) libraries, this read should match the TSS; in 3′end libraries, this read should map within ∼200–300 bp upstream of the transcript's 3′ end, thus avoiding the generation of reads from the poly(A) track. The read density distributions for the two methods are shown in Figure 1B.

## Assessment of library reproducibility and selection of optimal extension

We plated DCs in a 96 well-plate and stimulated independently. Technical replicates shown in Figure 2, and Supplemental Figures S1–S3 represent libraries made from different wells after LPS stimulation. Full-length libraries were downloaded from the Gene Expression Omnibus under accession number GSE59793 (Jovanovic et al. 2015) and GSE36104 (Garber et al. 2012), remapped, and requantified using RSEM (v 1.2.7) (Li and Dewey 2011) to obtain expected counts.

To select an appropriate UTR extension, we iterated through different extensions (0, 500, 1000, 2000, 5000, and 10,000 bp) and computed the correlation between full and end-sequenced data at genes that had at least 10 expected counts (Li and Dewey 2011) for full-length libraries, or reads mapped, for end-sequence libraries. Correlations with or without proper multimapper handling showed a small, yet consistent increase up to the 5000-bp extension for 3′ libraries and 1000 bp for 5′ libraries after which we see a drop (Supplemental Fig. 3A). We therefore selected 5000 bp for 3′ end-sequence analysis and 1000 bp for 5′ analysis.

## BBDR islet preparation

Pancreatic islets from BBDRs were isolated by collagenase digestion as previously described (Yang et al. 2013a). Islets were dissociated using TrypLE (Invitrogen) used a published protocol (Blodgett et al. 2015). After all the washes, the dissociated islet cells were suspended in PBS, filtered through 15 micron CellMicroSieve (BioDesign), and counted with a hemocytometer. The cells were further diluted to the final concentration (100,000 cells/mL) with OptiPrep (Sigma) and PBS, and the final concentration of OptiPrep was 15% vol/vol. About 5000 islet cells were collected and processed following the InDrop protocol (Klein et al. 2015). A library containing about 1000 cells was constructed and sequenced.

## Analysis of single-cell rat islet RNA-seq data (ICA + clustering)

Sequencing was performed on an Illumina HiSeq sequencer, resulting in 156 million paired-end reads. With this protocol, the cell barcode and unique molecule identifier (UMI) are contained in the forward read (R1) and the transcript sequence information is contained in the reverse (R2) read. The steps in the single-cell analysis pipeline are as follows:

1. A custom Python script identifies reads with valid barcodes and UMIs by examining the R1 reads. For protocols that allow for barcode error correction such as inDrop, barcode errors are corrected in this step. UMIs are considered 'valid' if they do not contain any "N's." Valid barcode and UMIs are appended to the read name in the corresponding R2 read with the colon-separated format ":<barcode>:<UMI>." This information is carried with the read through all subsequent steps in the pipeline and is used by ESAT for cell demultiplexing and PCR duplicate removal. An output file is also created, containing the total number of times each barcode is observed. This file is used in the next pipeline step.
2. A custom R script reads the barcode count file from the previous step and determines which barcodes are likely to correspond to single cells based on the total number of reads. Generally, these will be a relatively small number of barcodes that account for a large percentage of the reads. When the reads are sorted by read counts, a "knee" in the curve can be identified indicating a minimum number of reads required for a barcode to be considered a true cell. This threshold value is used in step 4.
3. Genomic alignment of all reads with valid barcodes and UMIs (from step 1) is performed with TopHat. For this analysis, we used TopHat v2.0.9 (Kim et al. 2013), with the rat genome (rn5) as a reference.
4. To reduce the memory requirements for ESAT, we applied an additional custom Python script to remove any reads from low-count barcodes from the alignments. The script takes the barcode counts file from step 1 and the count threshold calculated in step 2 as inputs, as well as the genome-aligned reads (BAM format). A new BAM file is created from each alignment file, which contains only reads from barcodes with sufficient read counts.
5. Finally, the barcode-filtered, genome-aligned reads are processed by ESAT to obtain gene-level quantification. In ESAT's single-cell preprocessing step, PCR duplicates are identified and removed using the barcode/UMI information appended to the read name in step 1. By performing this step, the counts provided in ESAT's output files can be interpreted as the total number of original transcript molecules sampled from each cell for each gene. These will be referred to as "UMI-filtered reads." For this analysis, we ignored multimapped reads

(-*multimap ignore*) and extended the search region to up to 5000 bases past the end of the annotated gene transcripts (-*wExt 5000*). As with the mouse DC analysis below, we used RefSeq gene annotations (rn5), using only transcripts whose ID begin with "NM," in order to remove small RNAs and noncoding RNAs that tend to have poorer annotations. Although the window-level output was not used for this analysis, additional window-related parameter settings used were a scanning window length of 100 bases (-*wLen 100*) with a window overlap of 50 bases (-*wOlap 50*), and the retention of all windows containing at least one read (-*sigTest 1.0*).

There are five known types of islet cells, each with primary marker genes whose expression can be used to classify the cell types. Alpha cells produce glucagon (*Gcg*), beta cells produce insulin (*Ins1* and *Ins2* in rat), gamma or PP cells produce pancreatic polypeptide (*Ppy*), delta cells produce somatostatin (*Sst*), and epsilon cells produce ghrelin (*Ghrl*). (Note that in this data set, a total of only five UMI-filtered reads mapped to *Ghrl*—one in one cell and four in another—so epsilon cells will not be considered in this analysis.)

We first removed any cells with fewer than 1000 total UMI-filtered reads. The next step in the analysis of the ESAT results is removing genes with too few UMI-filtered reads to be considered useable. We used raw (unnormalized) UMI-filtered read counts and selected only genes with at least three UMI-filtered reads in at least two cells for our analysis (8254 genes). Our intention in this step was only to remove genes with expression levels that are so low that they are likely to be uninformative. Genes with the most discriminatory power are selected using the statistical analysis in the following steps. We note here that, without applying the 5000-base extension to the transcripts with the -*wExt 5000* parameter, only 6652 genes passed these filter criteria. Thus, including a 5000-base extension results in 24% more genes available for this analysis.

For all of the further analysis steps, we use log-transformed, normalized UMI counts, using a simple UMIs/million normalization (i.e., divide UMI counts for each gene by the total for the cell, then multiply by 1 million). By using the selected genes, we next performed PCA to help determine the number of principal components and the subset of the most variable genes to be used with ICA. We selected only the principal components that accounted for the first 10% of the total variance and noted that each additional PC contributed negligibly to further explaining the variance. To select the genes used for ICA, we used the rotation (loadings) from the PCA for these components and selected only genes whose Mahalanobis distance from the origin are in the top 20th percentile. Our selection criteria resulted in selecting the first nine principal components and 1654 genes. The selection criteria eliminated one of the known islet cell marker genes (*Ppy*), so we augmented the list of selected genes with this gene in order to allow us to later validate the ICA and clustering results.

We next applied ICA (R function *fastICA*) (Langlois et al. 2010) to the filtered data to attempt to separate the cells into subgroups, followed by spectral clustering on the ICA projection to identify the cells in each subgroup. We then used t-distributed stochastic neighbor embedding (t-SNE; using the R *tsne* package) (Van der Maaten and Hinton 2008) to map the ICA projection for visualization. The 2D t-SNE mapping of the result is shown in Figure 5B, colored to identify the clusters resulting from the spectral clustering. We found that nine clusters separated the groups without further subdividing some of the larger groups into multiple clusters.

To identify the most significant genes in each cluster, we applied DESeq2 (Love et al. 2014), contrasting the cells in each cluster with the cells in each of the other clusters. We then selected all genes with an MLE-adjusted $\log_2$ fold-change with an absolute value >1.0 (i.e., twofold difference) and an adjusted *P*-value <0.01 to identify genes that had significantly differential expression between the clusters. We used the online resource DAVID (Huang et al. 2009) to identify GO categories and KEGG pathways that showed enrichment for each cluster.

The heatmap in Figure 6A was constructed using the R heatmap.2 function in the *gplots* package. Before clustering the genes (rows), we computed the *z*-score for each gene and clipped the *z*-score to $-3 \leq z \leq 3$ to enhance the readability of the heatmap. We used the default hierarchical clustering in heatmap.2, using a distance metric of (1-cor(<gene row data>))/2. To select the major gene groups, we cut the resulting row dendrogram (data not shown) at a height that appeared to identify the major visual features of the heatmap (0.585). Several smaller subsets of genes that formed distinct narrow bands were identified by further cutting the major branches at a height of 0.5, such as the small group of genes showing significant enrichment in the GO "response to Camp" biological process category.

We found that the cells in clusters C2, C6, and C8 showed relatively high expression of *Ppy*, *Sst*, and *Gcg*, respectively, corresponding to delta, gamma, and alpha cells. Clusters C1, C3, and C5 all show relatively high expression of beta-cell markers *Ins1* and *Ins2* (Fig. 5C). The remaining three clusters show increased expression of immune (cluster C2) or vascularization-related (clusters C6 and C7) genes.

## 'Telescripting' detection

For the analysis results presented in Figure 4A, we processed the 3′ LPS-stimulated mouse genomic alignments using the following ESAT parameters: ignore multimappers (-*multimap ignore*), transcript extension of 1000 bases (-*wExt 1000*), window *P*-value significance <0.05 (-*sigTest 0.05*), a scanning window length of 100 bases (-*wLen 100*) with a window overlap of 50 bases (-*wOlap 50*), removing any reads containing continuous stretches of As or Ts ≥ 10 in length (-*filtAT 10*). We used RefSeq mm10 transcript mappings, filtered to remove any transcript IDs that did not begin with "NM" to remove all small RNA transcripts. For this analysis, we used both the gene- and window-level outputs of ESAT. Before computing the fraction of reads in each significant window for each gene for each sample, we applied the following filters to the "no treatment" (t0) sample to select the genes for evaluation: We removed any genes with fewer than 250 reads total to remove genes with very low expression, removed genes with only one significant window, required at least 20% of the reads for a gene to occur in the final window, and limited the ratio of maximum to minimum reads in any two windows to 1.5 to ensure that reductions in the reads after treatment would be measurable. These filtering criteria resulted in a total of 1807 genes for the Figure 4A results. We tested a wide variety of filtering parameters, and all gave similar results, with only minor changes in the distributions and *P*-values.

For each of the selected genes for each timepoint, we then computed the total fraction of reads that were *not* in the final (most 3′) window. The barplots in Figure 4A show the distribution of these fractions for each sample. Finally, we computed the significance of the differences of each distribution compared with the prestimulation (t0) sample using a paired Wilcoxon rank-sum test.

For the 5′ data analysis presented in the Supplemental Material (Supplemental Fig. S4), we applied the same filters and window selection criteria, resulting in a total of only 26 genes. We then computed the total fraction of reads *not* in the first (most 5′) window.

## ESAT methodology

A schematic view of ESAT is shown in Figure 3. As inputs, ESAT takes a set of alignment files (SAM or BAM) with genome alignment coordinates, a file containing transcript coordinates (BED or text file), and various command-line parameters that specify the details of how analysis should be performed.

In the first pass of the analysis, all reads in all alignment files are combined, and the total number of alignments starting at each genomic coordinate is counted. The end of the transcript specified by the *-task* parameter (3′ or 5′) is extended by the number of bases specified by the *-wExt* parameter in the appropriate direction for each transcript. The extension is truncated if it overlaps with a neighboring transcript on the same strand. Next, the exonic regions specified in the transcript coordinates file (plus any extension) are scanned with a fixed-length window that sums the total number of reads in each window. For each window position, a *P*-value (based on a SCAN statistic) is computed. The positions of all windows with a *P*-value less than or equal to a specified *P*-value threshold are saved for the second pass analysis. The window width, overlap from one window to the next, *P*-value threshold, number of bases to search past the annotated transcription end, and which end to apply the extension (5′ or 3′) are all supplied as command-line parameters. In the case where a series of overlapping or adjacent windows all contain a significant number of reads (i.e., all of the windows have *P*-values below the specified threshold value), the contiguous region containing all of the windows is scanned in one-base steps to locate the window position containing the highest number of alignment starts. The location of this window is then used to represent this region.

In the second pass of analysis, the total number of reads within each window from the first pass are separately summed for each input file, and all files for each experiment are combined. The locations of the windows, the associated transcript identifier (gene or transcript ID), and the total number of reads in that region from each input experiment are reported in the "window-level" output file. In addition, the total number of reads for each transcript or gene for each experiment are reported in the "gene-level" output file.

## Multimapped reads

ESAT provides four methods for processing reads marked as "multimapped." Multimapped reads are identified as having the value of the NH tag (in the optional flags section of SAM file format) set to a value greater than one. If this flag is not provided in the genomic alignments, multimapped reads cannot be identified. The method used for processing multimapped reads is selected with the *-multimap <method>* command-line argument, where *<method>* is one of the following:

- *normal*—Multimapped reads are processed the same as uniquely mapped reads. If a read is mapped to N locations, this will have the effect of counting up to N times, once for each valid transcript location to which it maps (Supplemental Fig. S7A).
- *scale*—If the read is mapped to N genomic locations, the read is assigned a fractional count of 1/N for each valid transcript location to which it maps (Supplemental Fig. S7B).
- *ignore*—Multimapped reads are discarded, since there is no way to definitively determine the "true" mapping location (Supplemental Fig. S7C).
- *proper*—If a read is mapped to multiple genomic locations but only one of those locations falls within the boundaries of a transcript (after splicing out introns and extension), the read is assigned to that location. If the read maps to multiple valid transcript locations or to none, it is discarded (Supplemental Fig. S7D). This method is performed in two steps. In the first step, all reads that are marked as multimapped are written out to a temporary file during the first pass of the analysis. After all reads have been processed, the temporary files holding the multimapped reads are first sorted by read ID (using the SortSam function from the picard package) and then processed to determine whether they map to valid transcript locations. Reads that uniquely map to a valid transcript location are then saved to a new temporary file with the multimapping flag for the read set to one (i.e., uniquely mapped), and the original temporary file is deleted. In the second step, the temporary file containing the uniquely mapped reads is reprocessed through ESAT's first pass so that these reads are included when identifying significant window locations. In the second pass of analysis, these files are simply treated as extra BAM files for each experiment.

## Additional command-line arguments/parameters

### Input alignment files

*-input <alignment file>*: To process a single genomic alignment file, use -input, followed by the name of the alignment file (SAM or BAM format).

*-alignments <alignment list file>*: To process multiple genomic alignment files, use -alignments, followed by the name of a file containing the list of alignment files to be processed. Each line of the file should contain an experiment identifier followed by a tab, followed by the full path name of the alignment file (SAM or BAM format). Multiple alignment files can be provided for each experiment, and all reads will be assigned to the experiment as if a single, concatenated input file was provided. Note that a tab is used to separate the experiment ID from the input file name to allow spaces in both the experiment ID and file name.

### Input transcript annotation files

*-annotations <annotation BED file>*: When transcript annotations are provided with this parameter, each transcript in the input file is processed independently, and gene- and window-level outputs are produced for each transcript in the file. Note that when providing the transcript annotations in this way, multiple isoforms for the same gene are treated as if they came from separate genes. If, for example, multiple transcripts in the file share a common exon, any reads mapping to that exon will be assigned to each of the transcripts.

*-geneMapping <annotation table file>*: When transcripts are provided with this parameter, all transcripts for each gene are collapsed into a single "metatranscript," which contains all exonic and UTR regions from all transcripts for each gene. The *<annotation table file>* is a simple tab-delimited file containing transcript mapping information as provided by the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables). The minimum table columns required by ESAT are as follows:

- name—the transcript ID;
- chrom—chromosome identifier;
- strand—strand (+ or −);
- txStart—transcript start location;
- txEnd—transcript end location;
- exonStarts—comma-separated list of exon start positions;
- exonEnds—comma-separated list of exon end positions; and
- name2—gene symbol.

When collapsing gene transcripts into a single metatranscript, the first transcript for a gene specifies the chromosome and strand for the metatranscript. If additional transcripts for that gene are on a different strand or chromosome, those transcripts are discarded, and a warning message is generated indicating a "New isoform mismatch for <gene symbol>…," followed by the genomic coordinates of the conflicting transcript. When providing transcript coordinates in this way, a single line in the gene-level output file is provided for each gene symbol (rather than each transcript), and any significant windows falling within the metatranscript are reported with the gene symbol in the window-level output file.

### Window specifications

-wLen <*length*>: scanning window length, in bases (≥1, default = 50).

-*wOlap* <*overlap*>: number of bases overlap from one window to the next (≥0, default = 0).

-*wExt* <*extension*>: number of bases to extend beyond the end of the annotated transcript (≥0, default = 0).

-*sigTest* <*P-value*>: window significance threshold (0.0 ≤ *P-value* ≤1.0, default = 1.0). Only windows with a SCAN statistic *P*-value less than or equal to this value will be reported in the window-level output file.

-*task* <*end*>: specifies whether to process this as a 3′-end library (*score3p*) or 5′-end library (*score5p*). (Default = *score3p*.)

### Low-complexity filtering

-*filtAT* <*n*>: Discard any reads with continuous stretches of As or Ts ≥ <n> in length.

### Output file

-*output* <*file prefix*>: specifies the prefix for the output file name. Two output files will be created, <*file prefix*>.gene.txt, containing the gene-level results, and <*file prefix*>.window.txt, containing the window-level results.

### Single-cell parameters

-*scPrep*: indicates that the reads are from a single-cell protocol which uses cell barcodes and UMIs.

-*bcMin* <*barcode count*>: This optional parameter specifies the minimum number of UMI-filtered reads that must be associated with a barcode to have it be considered valid, as described above.

Reads are assumed to have the cell barcode and UMI appended to the read name field in the following format: <read name>:<cell barcode>:<UMI>. If at least three colon-separated fields are not found, the read is skipped and the 'Improper read name: …' warning is raised.

Reads from single-cell libraries are preprocessed to remove PCR duplicates resulting from the library preparation process. Since each read is tagged with a barcode to indicate the source cell and with a UMI to tag the source mRNA molecule, PCR duplicates can be identified as reads with the same cell barcode and UMI that map to the same gene transcript. A single instance of a read with a given cell barcode and UMI mapping to a transcript is sufficient to count as an observation of a transcript molecule, and all further instances of the same cell barcode and UMI mapping to the same transcript are discarded as PCR duplicates. The PCR duplicate removal process is implemented as a preprocessing step, in which the filtered reads (i.e., reads corresponding to specific barcode/UMI/transcript) are written out to temporary files, and those files replace the original input files for ESAT processing. As the reads are being filtered, the number of filtered reads that are associated with each barcode (cell) are counted. The -*bcMin* <*barcode count*> parameter allows the user to set a minimum number of filtered reads that each barcode must have. Cell barcodes with fewer than <*barcode count*> filtered reads are marked as 'invalid,' and reads with these barcodes are ignored in all further processing. Since memory is required to count every observed barcode/UMI/transcript, removal of aligned reads with low-count barcodes is also part of the preprocessing pipeline for preparing single-cell data for ESAT. The additional -*bcMin* parameter allows the initial low-count barcode filtering to be lenient (to reduce the memory footprint of the PCR duplicate removal step) but then allows more aggressive filtering to remove any barcodes (cells) with an insufficient number of UMI-filtered reads. This both removes cells with poor coverage and reduces the memory requirement for the second pass of analysis.

### Software availability

Our analysis toolkit and its source code are freely available from https://github.com/garber-lab/ESAT. The exact source used in this work is tagged as GR_release and is also available in the Supplemental Material. Custom python and R scripts used for single-cell analysis are included in Supplemental Data S1.

## Data access

Raw sequence data and ESAT outputs from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE79651.

## Acknowledgments

We thank Mitchell Guttman for his help in writing the manuscript and the many discussions we had during the course of this project. We also thank Nathan Lawson for his help interpreting the clusters enriched in vascular markers. We thank members of the Garber Laboratory for helpful discussions and careful reading of the manuscript. M.G., D.L.G., and D.M.H. were supported by the National Institutes of Health (NIH, grant DK105837), M.G. and J.L. by NIH, National Human Genome Research Institute (grant U01 HG007910), J.L. by NIH, National Institute on Drug Abuse (grant DP1DA034990), and D.L.G. and D.M.H. by NIH, Human Islet Research Network (grant UC4DK104218).

*Author contributions:* M.N.A., M.G., and S.K. conceived the ESAT and its applications to bulk (population) RNA-seq. M.N.A. and S.K. designed the bulk RNA end-sequencing protocol. A.D. implemented ESAT. M.N.A. generated the dendritic cell bulk RNA libraries. A.D., M.N.A., A.S., J.L., and M.G. performed the dendritic cell analysis. C.Y., R.Z., and S.R. generated the single-cell rat islet data. D.M.B., R.B., A.D., M.G., D.L.G., A.K., and D.M.H. conceived and analyzed the data corresponding to the rat islet part of the study. M.G. and A.D. wrote the paper with help from M.N.A., D.L.G., D.M.H., D.M.B., R.B., S.K., A.K., J.L., and C.Y. All authors read and approved the final manuscript.

## References

Amit I, Garber M, Chevrier N, Leite AP, Donner Y, Eisenhaure T, Guttman M, Grenier JK, Li W, Zuk O, et al. 2009. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* **326:** 257–263.

Apostolou E, Thanos D. 2008. Virus infection induces NF-κB-dependent interchromosomal associations mediating monoallelic *IFN*-β gene expression. *Cell* **134:** 85–96.

Asmann YW, Klee EW, Thompson EA, Perez EA, Middha S, Oberg AL, Therneau TM, Smith DI, Poland GA, Wieben ED, et al. 2009. 3′ tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC Genomics* **10:** 531.

Bajikar SS, Fuchs C, Roller A, Theis FJ, Janes KA. 2014. Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. *Proc Natl Acad Sci* **111:** E626–E635.

Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D, Zhang Z, Cho S, Sherrill-Mix S, Wan L, et al. 2012. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150:** 53–64.

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

Bortell R, Yang C. 2012. The BB rat as a model of human type 1 diabetes. *Methods Mol Biol* **933:** 31–44.

Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* **11:** 94.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38:** 626–635.

de Klerk E, 't Hoen PAC. 2015. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet* **31:** 128–139.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22:** 1775–1789.

Derti A, Garrett-Engele P, MacIsaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22:** 1173–1183.

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

Elkon R, Ugalde AP, Agami R. 2013. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet* **14:** 496–506.

Garber M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* **8:** 469–477.

Garber M, Yosef N, Goren A, Raychowdhury R, Thielke A, Guttman M, Robinson J, Minie B, Chevrier N, Itzhaki Z, et al. 2012. A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol Cell* **47:** 810–822.

Greiner DL, Rossini AA, Mordes JP. 2001. Translating data from animal models into methods for preventing human autoimmune diabetes mellitus: caveat emptor and primum non nocere. *Clin Immunol* **100:** 134–143.

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

Gu W, Lee H-C, Chaves D, Youngman EM, Pazour GJ, Conte D, Mello CC. 2012. CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as *C. elegans* piRNA precursors. *Cell* **151:** 1488–1500.

Guardado-Mendoza R, Davalli AM, Chavez AO, Hubbard GB, Dick EJ, Majluf-Cruz A, Tene-Perez CE, Goldschmidt L, Hart J, Perego C, et al. 2009. Pancreatic islet amyloidosis, β-cell apoptosis, and α-cell proliferation are determinants of islet remodeling in type-2 diabetic baboons. *Proc Natl Acad Sci* **106:** 13992–13997.

Hashimshony T, Wagner F, Sher N, Yanai I. 2012. CEL-Seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep* **2:** 666–673.

Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G, Tian B. 2013. Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat Methods* **10:** 133–139.

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

Huang C-J, Lin C-Y, Haataja L, Gurlo T, Butler AE, Rizza RA, Butler PC. 2007. High expression rates of human islet amyloid polypeptide induce endoplasmic reticulum stress mediated β-cell apoptosis, a characteristic of humans with type 2 but not type 1 diabetes. *Diabetes* **56:** 2016–2027.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4:** 44–57.

Ingolia NT. 2010. Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol* **470:** 119–142.

Jaitin DA, 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.

Janes KA, Wang C-C, Holmberg KJ, Cabral K, Brugge JS. 2010. Identifying single-cell molecular programs by stochastic profiling. *Nat Methods* **7:** 311–317.

Jovanovic M, Rooney MS, Mertins P, Przybylski D, Chevrier N, Satija R, Rodriguez EH, Fields AP, Schwartz S, Raychowdhury R, et al. 2015. Dynamic profiling of the protein life cycle in response to pathogens. *Science* **347:** 1259038.

Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, Shin JW, Kojima-Ishiyama M, Kawano M, Murata M, et al. 2014. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res* **24:** 708–717.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14:** R36.

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161:** 1187–1201.

Langlois D, Chartier S, Gosselin D. 2010. An introduction to independent component analysis: InfoMax and FastICA algorithms. *Tutor Quant Methods Psychol* **6:** 31–38.

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

Li J, Klughammer J, Farlik M, Penz T, Spittler A, Barbieux C, Berishvili E, Bock C, Kubicek S. 2015. Single cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Rep* **17:** 178–187.

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

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161:** 1202–1214.

Matsumura H, Yoshida K, Luo S, Kimura E, Fujibe T, Albertyn Z, Barrero RA, Krüger DH, Kahl G, Schroth GP, et al. 2010. High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. *PLoS One* **5:** e12010.

Moore MJ. 2005. From birth to death: the complex lives of eukaryotic mRNAs. *Science* **309:** 1514–1518.

Morris DL. 2015. Minireview: emerging concepts in islet macrophage biology in type 2 diabetes. *Mol Endocrinol* **29:** 946–962.

Paquet-Fifield S, Schlüter H, Li A, Aitken T, Gangatirkar P, Blashki D, Koelmeyer R, Pouliot N, Palatsides M, Ellis S, et al. 2009. A role for pericytes as microenvironmental regulators of human skin tissue regeneration. *J Clin Invest* **119:** 2795–2806.

Pechhold S, Stouffer M, Walker G, Martel R, Seligmann B, Hang Y, Stein R, Harlan DM, Pechhold K. 2009. Transcriptional analysis of intracytoplasmically stained, FACS-purified cells by high-throughput, quantitative nuclease protection. *Nat Biotechnol* **27:** 1038–1042.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. 2009. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37:** D32–D36.

Rabani M, Raychowdhury R, Jovanovic M, Rooney M, Stumpo DJ, Pauli A, Hacohen N, Schier AF, Blackshear PJ, Friedman N, et al. 2014. High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell* **159:** 1698–1710.

Roberts A, Pachter L. 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* **10:** 71–73.

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biol* **12:** R22.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2015. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* **43:** D670–D681.

Shepard PJ, Choi E-A, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17:** 761–772.

Shih S-C, Robinson GS, Perruzzi CA, Calvo A, Desai K, Green JE, Ali IU, Smith LEH, Senger DR. 2002. Molecular profiling of angiogenesis markers. *Am J Pathol* **161:** 35–41.

Shiroguchi K, Jia TZ, Sims PA, Xie XS. 2012. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci* **109:** 1347–1352.

Smit A, Hubley R, Green P. 2004. *RepeatMasker Open-3.0. 2004*. Institute for Systems Biology, Seattle, WA. http://www.repeatmasker.org/.

Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. 2014. Characterization of directed differentiation by high-throughput single-cell RNA-seq. *bioRxiv*. doi: http://dx.doi.org/10.1101/003236.

Takahashi H, Lassmann T, Murata M, Carninci P. 2012. 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc* **7:** 542–561.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2012. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31:** 46–53.

Van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *J Mach Learn Res* **9:** 2579–2605.

Xiao X, Gaffar I, Guo P, Wiersch J, Fischbach S, Peirish L, Song Z, El-Gohary Y, Prasadan K, Shiota C, et al. 2014. M2 macrophages promote β-cell proliferation by up-regulation of SMAD7. *Proc Natl Acad Sci* **111:** E1211–E1220.

Yang C, Diiorio P, Jurczyk A, O'Sullivan-Murphy B, Urano F, Bortell R. 2013a. Pathological endoplasmic reticulum stress mediated by the IRE1 pathway contributes to pre-insulitic β cell apoptosis in a virus-induced rat model of type 1 diabetes. *Diabetologia* **56:** 2638–2646.

Yang C, Jurczyk A, diIorio P, Norowski E, Brehm MA, Grant CW, Guberski DL, Greiner DL, Bortell R. 2013b. Salicylate prevents virus-induced type 1 diabetes in the BBDR rat. *PLoS One* **8:** e78050.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. 2016. Ensembl 2016. *Nucleic Acids Res* **44:** D710–D716.

Zhao M, Zhang J, Phatnani H, Scheu S, Maniatis T. 2012. Stochastic expression of the interferon-β gene. *PLoS Biol* **10:** e1001249.

Zipris D, Lien E, Xie JX, Greiner DL, Mordes JP, Rossini AA. 2005. TLR activation synergizes with Kilham rat virus infection to induce diabetes in BBDR rats. *J Immunol* **174:** 131–142.

*Received April 4, 2016; accepted in revised form July 27, 2016.*

CSH PRESS GENOME RESEARCH

# End Sequence Analysis Toolkit (ESAT) expands the extractable information from single-cell RNA-seq data

Alan Derr, Chaoxing Yang, Rapolas Zilionis, et al.

*Genome Res.* 2016 26: 1397-1410 originally published online July 28, 2016
Access the most recent version at doi:10.1101/gr.207902.116