End Sequence Analysis ToolKit (ESAT) expands the extractable from single cell RNA-seq experiments

Alan G. Derr  
*University of Massachusetts Medical School*

Chaoxing Yang  
*University of Massachusetts Medical School*

Rapolas Zilionis  
*Harvard Medical School*

*See next page for additional authors*

Follow this and additional works at: [http://escholarship.umassmed.edu/oapubs](http://escholarship.umassmed.edu/oapubs)

Pocket Part of the [Bioinformatics Commons](http://escholarship.umassmed.edu/oapubs/bioinformatics), [Computational Biology Commons](http://escholarship.umassmed.edu/oapubs/computationalbiology), and the [Genomics Commons](http://escholarship.umassmed.edu/oapubs/genomics)

Repository Citation

Derr, Alan G.; Yang, Chaoxing; Zilionis, Rapolas; Sergushichev, Alexey; Blodgett, David; Redick, Sambra D.; Bortell, Rita; Luban, Jeremy; Harlan, David; Kadener, Sebastian; Greiner, Dale L.; Klein, Allon; Artyomov, Maxim; and Garber, Manuel, "End Sequence Analysis ToolKit (ESAT) expands the extractable from single cell RNA-seq experiments" (2016). *Open Access Articles*. 2781.  
[http://escholarship.umassmed.edu/oapubs/2781](http://escholarship.umassmed.edu/oapubs/2781)

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in Open Access Articles by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
End Sequence Analysis ToolKit (ESAT) expands the extractable from single cell RNA-seq experiments

Authors
Alan G. Derr, Chaoxing Yang, Rapolas Zilionis, Alexey Sergushichev, David Blodgett, Sambra D. Redick, Rita Bortell, Jeremy Luban, David Harlan, Sebastian Kadener, Dale L. Greiner, Allon Klein, Maxim Artyomov, and Manuel Garber

Comments
The version of the article available for download is the authors' final, peer-reviewed accepted manuscript as prepared for publication in Genome Research.

Creative Commons License
This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 License

Rights and Permissions
This manuscript is Open Access. This article, published in Genome Research, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International license), as described at http://creativecommons.org/licenses/by-nc/4.0/.

This article postprint is available at eScholarship@UMMS: http://escholarship.umassmed.edu/oapubs/2781
End Sequence Analysis Toolkit (ESAT) expands the extractable information from single cell RNA-seq data

Alan Derr¹, Chaoxing Yang²*, Rapolas Zilionis³,⁴*, Alexey Sergushichev⁵,⁶, David M. Blodgett⁷, Sambra Redick², Rita Bortell², Jeremy Luban⁸, David M. Harlan⁷, Sebastian Kadener⁹, Dale L. Greiner²,⁸, Allon Klein³, Maxim N. Artyomov⁶, Manuel Garber¹,⁸

1. Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester MA
2. Program in Molecular Medicine, Diabetes Center of Excellence, University of Massachusetts Medical School, Worcester MA
3. Department of System Biology, Harvard Medical School, Boston MA
4. Institute of Biotechnology, Vilnius University, Vilnius, Lithuania
5. Computer Technologies Department, ITMO University, Saint Petersburg, Russia
6. Pathology & Immunology Department, Washington University, St. Louis MO
7. Department of Medicine, Diabetes Center of Excellence, University of Massachusetts Medical School, Worcester MA
8. Program in Molecular Medicine, University of Massachusetts Medical School, Worcester MA
9. Biological Chemistry Department, Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel
Abstract

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 1,000 cDNA libraries, each from an individual pancreatic islet cell. ESAT identified 9 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.
Introduction

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, annotation and, most recently, to measure translational or transcriptional rates (Garber et al. 2011; Ingolia 2010; Rabani et al. 2014). Measuring gene expression from RNA-seq data is complex and presents computational challenges that are unique to RNA-seq: (i) When RNA from a cell population is sequenced, only relative gene or isoform expression can be determined, and (ii) 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 (Roberts et al. 2011; Bullard et al. 2010; 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 (Matsumura et al. 2010; Asmann et al. 2009). 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 polyA 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 (Takahashi et al. 2012; Gu 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 which are masked in bulk cell data (Janes et al. 2010; Apostolou and Thanos 2008; 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
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 mis-annotated 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 multi mapped 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 cell and single cells. We generated 5' and 3' end-sequence data for mouse bone marrow derived dendritic cells (mBMDCs) stimulated with LPS, and compared this data to our previously generated full-length RNA-seq data (Garber et al. 2012). We also applied ESAT to a single cell RNA-Seq from ~1,000 rat pancreatic islet cells using 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 (Matsumura et al. 2010; Asmann et al. 2009). 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 (Figure 1A, Methods) using mouse bone marrow derived dendritic cells (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 filtering out low quality reads (Supplemental Table S1). Variability between biological replicates showed very good reproducibility (R > 0.98, Figure 2A,B, Supplemental Fig S1, Methods). 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).

Inspection of the data (Figure 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 Figure 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 (Figure 2A,B and Supplemental Fig S1). Further inspection showed that all genes with poor replicability had unusually high read pileups on internal polyA sequences that tended to be randomly over-represented from replicate to replicate. Over-amplification 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 1,800) of RefSeq annotations expressed a 5' or 3' end that was different than the annotated end (Figure 2C-E), as has also been found in previous reports (Derti et al. 2012; Shepard et al. 2011; Hoque et al. 2013; Carninci et al. 2006). 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 the End-Sequence Analysis Toolkit (ESAT) specifically designed 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: (I) Internal polyA sites that result in artificial read alignment pileups and compromise gene quantification (Figure 2A,B). (II) 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 (Figure 2C-E). (III) UTRs, and 3' UTRs in particular, contain a higher fraction of repetitive or low complexity sequence compared to coding sequence, which adversely affects read mappability (Figure 2F). (IV) Single-cell library construction methods are diverse and need different analysis strategies, however 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). (V) 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.

As input, ESAT takes spliced, genomic read alignments from end-sequence libraries together with a gene annotation set (Pruitt et al. 2009; Yates et al. 2016; Derrien et al. 2012). To avoid internal polyA sites, ESAT offers the option to discard reads containing continuous stretches of As or Ts 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 can’t 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 site (TSS) or transcription termination site (TTS)
by relying on the significant windows that represent candidate transcript ends (Figure 3, Methods).

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 to 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 1kb 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 mouse bone marrow derived dendritic cells (mBMDDCs) to LPS has been extensively characterized (Amit et al. 2009; Garber et al. 2012; Jovanovic et al. 2015; Rabani et al. 2014), 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 TSS or TTS 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 towards short isoform usage when looking at all genes that have multiple alternative TTS (Figure 4). The trend is detectable at 2 hours after LPS stimulation and peaks at the 4 hour time point. No such trend is observed when using 5′ data (Supplemental Fig 4), 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 towards shorter UTRs (Berg et al. 2012) that is similar to what we observe. Whether this trend has a role in the response of mBMDDCs 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; Soumillon et al. 2014; Jaitin 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 I diabetes (T1D) (Zipris et al. 2005) that we have extensively studied in the past (Bortell and Yang 2012; Greiner et al. 2001). 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 less than 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 non-diabetic BBDRs as islet donors, and isolated and dissociated the islets into single cell suspensions as previously described (Bladgett 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 (Klein et al. 2015) (Figure 5A). 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).

Similarly 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 *lapp* (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 2 UMIs in at least 3 cells (Methods). In all, 991 cells (93%) and 8264 genes passed our validation criteria and were used in the analysis (Methods).

Although islet cell composition is dominated by hormone secreting cells, many other cell types contribute to these complex organoids. 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 9 distinct groups (Figure 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 (Figure 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 (B-H adjusted p-value < 0.01) and at least 2-fold different between any two sets of cells (Methods). Clustering of these 942 genes revealed clear signatures of the remaining three clusters (Figure 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 cells in the C9 cluster express typical angiogenesis markers (Shih et al. 2002) (Figure 6A,B). 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 contains neither acinar nor ductal cells. This is likely a result of the much simpler islet isolation protocols in the rat model system compared to 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 to 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 suggests 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 heterogenous 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. Among the different cell types we observe an inverse correlation between Insulin and amylin (Iapp) (Figure 6C,D). Since dysregulation of amylin has
been associated with \(\beta\)-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, that a small fraction of \(\alpha\)-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 does not provide strong evidence for or against this phenomena in rat; 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 \(\beta\)-cells genes at high levels, their expression of \(\alpha\)-cell genes is low, perhaps pointing to \(\beta\)-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 experimentalists generating such data. Our rat islet data highlighted the need for specialized handling of this data, as for example, 24% of genes would be unusable without searching for unannotated UTRs. Single cell data shows
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 rat where onset of T1D can be monitored (Yang et al. 2013b; Bortell and Yang 2012). 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 to 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 datasets that are currently being produced.

Methods

5’ Sequencing

5’ RNA sequencing libraries were generated following a recently developed protocol (Afik et al., manuscript under review). Briefly, RNA was fragmented, and then enriched for 5' ends using Terminator Exonuclease (Epicentre). Reaction mixture was cleaned up with 2.5X of SPRI beads and then dephosphorylated with FastAP (Fermentas), cleaned (2.5X SPRI, Agencourt) and then ligated to a linker1 (5Phos/AXXXXXXXXXAGATCGGAAGAGCGTGTGAG/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. RT was then performed for the pooled sample, with a specific primer (5’-CCTACGACGCTCTTCC-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 second ligation was performed, where 3’end of cDNA was ligated to linker2
(5Phos/AGATCGGAAGAGCACACGTCTG/3ddC/) using T4 RNA ligase I. Reaction Mixture was cleaned up (Silane beads) and PCR enrichment was set up using enrichment primers 1 and 2 (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3' , 5'-CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3', where XXXXXXX is barcode sequence) and Phusion HF MasterMix (NEB). After clean-up with 0.8X 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 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

50ng-1ug of total RNA is polyA selected using Oligo-dT beads (Invitrogen) and then fragmented using Ambion Fragmentation buffer using incubation time of 1 min 50sec at 70C. Fragmented RNA is then cleaned-up using 2.5X volume on SPRI beads (Agencourt) and then subjected to second polyA 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 second round of polyA 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) 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 polyA selection is preferred when constructing annotation library suit to achieve best compatibility with full-length and/or 5'end sequencing.

### Sequencing

Constructed libraries were sequenced with 2x25bp paired-end Illumina pipeline that is part of standard sequencing pipeline at Broad Institute. For all analysis we used only 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, Figure 1A). In 5’ end (ExoCAGE) libraries, this read should match the transcription start site (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 polyA 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 Figs 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) and GSE36104 (Garber et al), remapped and re-quantified using RSEM (Li and Dewey 2011) (v 1.2.7) to obtain expected counts.

To select the an appropriate UTR extension we iterated through different extensions (0, 500, 1,000, 2,000, 5,000 and 10,000 bp) and computed the correlation between full and end-sequenced data 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 5,000 bp extension for 3’ libraries and 1,000 bp for 5’ libraries after which we see the first drop (Supplemental Fig 3A). We therefore selected 5,000 bp for 3’ end-sequence analysis and 1,000 for 5’ analysis.

**BBDR Rat Islet Preparation**

Pancreatic islets from BBDR rats 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, New York), and counted with 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 5,000 islet cells were collected and processed following the InDrop protocol (Klein et al. 2015). A library containing about 1,000 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 156M 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 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 retaining 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: In this dataset, a total of only 5 UMI-filtered reads mapped to Ghrl - 1 in one cell and 4 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 3 UMI-filtered reads in at least 2 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.) 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 independent component analysis (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 9 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 (Van der Maaten and Hinton 2008), using the R *tsne* package) to map the ICA projection for visualization. The 2-D t-SNE mapping of the result is shown in **Figure 5B**, colored to identify the clusters resulting from the spectral clustering. We found that 9 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 log2 fold-change with an absolute value >1.0 (i.e. 2-fold 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≤z≤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 (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* (**Figure 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 to the pre-stimulation (t0) sample using a paired Wilcoxon rank-sum test.

For the 5’ data analysis presented in the supplementary 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 1. If this flag is not provided in the genomic
alignments, multimapped reads can not 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 7A)
- **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 7B)
- **ignore** - multimapped reads are discarded, since there is no way to definitively determine the ‘true’ mapping location. (Supplemental Fig 7C)
- **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 none, it is discarded. (Supplemental Fig 7D) 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 1 (i.e., uniquely mapped), and the original temporary file is deleted. In the second step, the temporary file containing the uniquely mapped reads is re-processed 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 'meta-transcript', 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:

- 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
- name2: gene symbol

When collapsing gene transcripts into a single meta-transcript, the first transcript for a
gene specifies the chromosome and strand for the meta-transcript. 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 
meta-transcript 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 unique molecule identifiers (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 3 colon-separated fields are not found, the read is skipped and the ‘Improper read name: …’ warning is raised.

Reads from single-cell libraries are pre-processed to remove PCR duplicates resulting from the library preparation process. Since each read is tagged with a barcode to indicate the source cell, and a unique molecule identifier (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 pre-processing 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
Custom python scripts used for single cell analysis are included in the Supplemental_Data_S1.

Data access

Raw sequence data and ESAT outputs reported in this paper are available from Gene Expression Omnibus under accession GSE79651 at http://www.ncbi.nlm.nih.gov/geo/.

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.G and D.H. were supported by the NIH (grant DK105837), M.G. and J.L. by NHGRI (grant U01 HG007910), J.L. by NIDA (grant DP1DA034990), and D.G. and D.H. by HIRN (grant UC4DK104218).


Figure Legends

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 (TTS) as reference, for 5’ (green) and 3’ libraries (red).

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 (mBMDDCs) 2 hours 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 highly enriched window within each gene to the annotated TTS for the 2hr 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 (Rosenbloom et al. 2015; Smit et al. 2004) in mouse (black) and rat (gray) annotated genes.

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

**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.

**Figure 5. Single cell analysis of rat pancreatic islets.** A) Summary of the study. B) two dimensional view of a 9-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 9-component ICA projection. c) Violin plots showing the distribution of hormone expression across cells in each cluster.

**Figure 6. Classification of 1000 Islet cells.** A) Hierarchical clustering of the 940 genes that are differentially expressed between any pair of the 9 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 Figure 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 cells 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.
References


Takahashi H, Lassmann T, Murata M, Carninci P. 2012. 5[prime] end-centered expression


A

Poly-A selected RNA

\[ \text{m}^7\text{G} \quad \text{m}^7\text{G} \quad \text{m}^7\text{G} \]

\[ (\text{A})^n \]

Zn-based fragmentation

\[ \text{m}^7\text{G} \quad \text{m}^7\text{G} \quad \text{m}^7\text{G} \]

5'-end-seq

Terminator exonuclease

PolyA selection

3'-end-seq

\[ (\text{A})^n \]

\[ (\text{A})^n \]

\[ (\text{A})^n \]

3'-end-Seq high-throughput

+ T4 ligase

first linker ligation

\[ \text{m}^7\text{G} \quad \text{m}^7\text{G} \quad \text{m}^7\text{G} \]

\[ (\text{A})^n \]

\[ (\text{A})^n \]

\[ (\text{A})^n \]

cDNA synthesis +

TTTTTTTTTTT-BC-

5’

3’

+ T4 ligase

second linker ligation

3’

5’

PCR enrichment with barcoded Illumina specific adapters

B

ExoCAGE

3’end-seq

-2000 TSS 33% 66% TES 2000
Figure 2

A

B

C

D

E

F

R=0.98

R=0.98

Pnp

Tnfsf8

Tlr1

Nod1

CDS

5' DGE R1 2h

3p DGE R2 2h

3p DGE R1 2h

5p DGE R2 2h

1380bp longer

920bp longer

1600bp longer

140bp shorter

1675 genes

152 genes

30 genes

187 genes

0.000 0.025 0.050 0.075

low-complexity fraction

# Genes

Distance to closest annotated end 3'

Distance to closest annotated end 5
Sequenced reads aligned to reference genome

Annotation set

Extension window length and density significance

Parameters

End identification

Output

Input

Gene 1

Sample 1: Data agrees with annotation

Sample 2: Longer UTR

Sample 3: Shorter UTR

Gene-level counts

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>101</td>
<td>98</td>
</tr>
<tr>
<td>Gene 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Window-level counts

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1 w1</td>
<td>0</td>
<td>101</td>
</tr>
<tr>
<td>Gene 1 w2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gene 1 w3</td>
<td>111</td>
<td>0</td>
</tr>
<tr>
<td>Gene 2 w1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A

Islet purification

Dissociation

in-Drop

ESAT quantification and cell classification

B

C

C1

C2

C3

C4

C5

C6

C7

C8

C9

UIM per million

Ins1 Ins2 Gcg Ppy Pyy Sst
End Sequence Analysis ToolKit (ESAT) expands the extractable
from single cell RNA-seq experiments

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

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

Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

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

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

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

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

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

Published by Cold Spring Harbor Laboratory Press