

Manuscript EMBO-2015-40946

Single-cell transcriptomics reveals unique features of human pancreatic islet cell subtypes

Jin Li, Johanna Klughammer, Matthias Farlik, Thomas Penz, Andreas Spittler, Charlotte Barbieux, Ekaterine Berishvili, Christoph Bock, and Stefan Kubicek

Corresponding author: Stefan Kubicek, CeMM Research Centre for Molecular Medicine of the Austrian Academy of Sciences

Review timeline:

Submission date:	29 June 2015
Editorial Decision:	10 August 2015
Revision received:	05 October 2015
Editorial Decision:	08 November 2015
Revision received:	16 November 2015
Accepted:	19 November 2015

Editor: Barbara Pauly

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision

10 August 2015

Thank you very much for the submission of your research manuscript to our editorial office and for your patience while we were waiting to hear back from the referees. We have now received the full set of reviews as well as the referees' feedback on our pre-decision discussions regarding your proposed response to their reports.

As the detailed reports are pasted below and have also been forwarded to you prior to this email, I will only repeat the main point(s) here. As we have discussed before, two referees (1 and 2) raised concerns about the purity and definition of the different cell types based on established marker gene expression and referee 1 still feels that your explanation does not sufficiently address this point (his/her additional comments are also pasted below for your information). S/he suggests using additional, classical marker genes and additional controls to unambiguously rule out cross-contamination and/or mis-classifications. Referee 3 also agrees that double marker analysis would be more informative in this regard.

This being said, I would like to stress again that all reviewers agree on the potential interest of your findings and do, in principle, support publication of your study in our journal. They just feel that some aspects of your data would need to be strengthened to make this a valuable resource for others in the field. Therefore, I would like to give you the opportunity to revise your manuscript, with the understanding that the main concerns of the referees should be addressed. Please also add the MDS data that you have provided during our pre-decision discussion. With regard to the suggestion of referee 1 to extend the downstream analysis, we would not make this a pre-requisite for publication,

but if you have obtained these data already or can obtain them relatively easily, we do, of course, encourage you to add them.

Acceptance of the manuscript will depend on a positive outcome of a second round of review. Please also note that our scientific reports (as opposed to full articles) contain a combined results and discussion section and I would kindly ask you to modify the text accordingly before submitting the revised version. Also, we are using a number-based reference style, the endnote file for which can be found on our website and I would kindly ask you to reformat the citations accordingly. I look forward to seeing a revised form of your manuscript when it is ready. Should you in the meantime have any questions, please do not hesitate to contact me.

REFeree REPORTS

Referee #1:

This is the first paper to provide single cell transcriptome data from human islet cells, and has the potential to be a major reference for investigators in the field. The analysis is very preliminary, and although there are some interesting example genes that are shown, there are few conclusions of biomedical or biological relevance that can be derived from a systematic analysis of the data. This is more of an editorial decision, but I have the feeling that the standard of EMBO-R requires more indepth analysis of the data. In short I believe that this resource would be of great interest if the authors can improve it.

Major Points:

- One of the major interests of this analysis is that it can provide unequivocal evidence for expression of a particular gene in alpha, beta, etc cell types. For example, whilst detection of a gene at significant levels, or even apparent enrichment, in FACS-purified beta cells likely means that the gene is expressed in beta cell cells, there is still the theoretical possibility that the gene is highly very expressed in a minor cellular contaminant. By contrast, if the authors can prove that multiple INS+ cells express this marker this is quite solid evidence for expression in INS+ type.

- What is the discriminative capacity of this analysis? how many cells are double positive for classic single cell type markers (eg hormones). What is the significance of INS reads in single "alpha" cells? Is this likely technical or biological? Can the authors tease this out by looking for lower abundance cell type specific markers that might be exclusive, thereby ruling out cross-contamination of some sort. Without a very careful analysis of these issues it is not possible to determine which markers are expressed at different levels in different cell types, vs. which markers are simply cross-contaminants.

- Is there any evidence to suggest heterogeneity within INS+, PP+ or GCG+ cells? Despite the relatively low number of cells this should be properly analyzed and discussed.

- It might be useful to highlight not just DLK1 and GC as a curiosity, but also systematically display differences between mouse and human cell type specificities. I notice that MAFB is clearly expressed in human beta and alpha (this had been noted before but its unequivocal in this analysis).

- Some suggestions for downstream analysis include:

- (a) Given that the authors allude to T2D in the introduction, they may wish to provide an indication of which T2D susceptibility candidate genes (eg closest genes to lead SNPs from GWAs studies) are expressed in different cell types.

- (b) Authors may wish to compare their findings to published "bulk" FACS-purified islet cell RNAseq datasets (eg Bramswig et al, JCI 2013, Moran et al Cell Metab 2012).

- (c) Authors may want to relate their data to whole islet epigenomic studies, for ex can the authors tease out different cell types that are likely contributing different islet enhancer clusters (pasquali et al, Nature genetics, 2014)?

- (d) The motif analysis is provided in raw format, and it seems that only alpha to acinar comparisons are provided in the excel. It is not very useful and this not valid for publication. It needs to be exploited, processed properly to extract meaningful conclusions.

- The browser shot provided for PDX1 in the supplementary figure is helpful, but the images that are obtained in the online browser for single cells provide a very different picture. This should be clarified, and if it is just about the default color schemes the version provided in the suppl data is clearly more informative. The browser shots of genes are so tiny that they are not readable without zoom in. The fact that this is a resource paper means that the authors should make an effort so that the data is easily utilized.

- Fig2A suggests that ARX is expressed in alpha and pp cells as previously reported. However, in figure 2D this observation is only supported by 2 of the 18 cells classified as pp cells.

Minor points:

The browser should have default auto-scaling to be useful, rather than default =1000 RPKM

Abstract:

"The established marker genes cannot capture cellular heterogeneity in human pancreatic islets" This is a peculiar statement in an introduction for this paper. Established marker genes do capture heterogeneity. Without single cell studies it is a priori unknown whether hormone markers are sufficient to capture all of human islet cell heterogeneity.

Intro:

"While damage to beta cells is causative of type I diabetes and contributes to the later stages of type II diabetes"

The bulk of existing data points to a role of beta cell dysfunction in early stages of T2 diabetes, including genetic predisposition.

Results, first paragraph

"Disassociated": I believe that dissociated is used more often for cells although I am not certain.

Results, first paragraph. When referring to PCA, it not clear if the the authors are referring to different principal components from one PCA, or if they have done iterative PCAs after removing the first PCs from the preceding PCA-this is what appears to have been done in the figure. They should explain clearly how they did this in the text.

Referee #2:

The manuscript of Li et al presents single cell transcriptome data from primary human pancreatic cells from an islet isolation from a 37yr old male with BMI of 22. Of the 70 cells under going RNAseq analysis, 64 passed quality control of an ERCC spikein. Using reiterative principal component analysis based on known cell-specific markers (mainly the specific hormones, SPP1 for ducts, PRSS2 and reg genes for acinar), 12 were beta cells, 18 alpha cells, 11 acinar, 8 duct, 2 delta and 4 undetermined. Then further analysis was able to determine the expression pattern for the PP and delta cells, detect novel transcription factors in some islet cell types and document (with immunostaining for 2 genes DLK1 and GC) opposite expression in human and mouse islets. This valuable resource has also been deposited as a UCSC genome browser track dataset that is available and seemingly user friendly.

This application of new technology for single cell transcriptome analysis contributes a useful database for the field. There are some biological issues that should be clarified here.

1. Even within the same cell type there is considerable variation in both the presence/absence of expression and the level of intensity as seen in the presented biplot visualizations. Is this real or possible a technical issue?

2. It is surprising that the paper by Dorrell et al (Grompe) (Diabetes 2011) in which careful sorting of alpha, beta, acinar and large and small duct cells from 3-4 human donors was not cited. The current study is complementary (particularly adding data on PP an delta cells).

3. It has been well known that REST binding sites are key to islet, particularly beta cell, differentiation (see Bruce et al PNAS 2004 ; DS Johnson Science 2007). This should be acknowledged.

4. What the authors have missed commenting on are their definitive data addressing the question of Maf/MafB in human islets. The Vanderbilt group (Powers & Stein) had reported that in humans unlike rodents, MAFB is expressed in about 50% of the adult beta cells and that MAFA was only expressed in about 50 % of them, but it remained unclear if these were colocalizing or separate populations. The biplots seem to indicate that most insulin-expressing cells expressed more MafB than MafA and their genome browser tracks show 2/12 beta cells with no MafA expression but only 1/12 without MafB but 3 /18 alpha cells with little to no MafB.

4. p 1 Introduction, " damage to beta cells is causative to type 1 diabetes and " contributes to the later stages of type 2 diabetes" does not reflect the current understanding. In both forms of diabetes the lack of adequate functioning beta cells are causative from the early stages of disease. This statement should be revised.

Referee #3:

In this paper Kubicek and colleagues apply single-cell RNAseq to measure the full transcriptome of 70 human pancreatic islet cells. Single-cell RNAseq is an extremely powerful approach for detecting cellular heterogeneity in a complex tissue without the need for biased enrichment for a particular cell type. In this respect the technique is especially suitable for human pancreatic islets, where targeted isolation is not trivial. Although the analysis was only performed on a single individual the paper provides an important resource and also some very interesting analysis, and will prove to be a great resource for those interested in both islet biology and single-cell genomics.

The authors should address the following points:

1) When looking at the track plots in Fig. 2A it seems as if there is a non-uniform coverage of the reads, where most reads are located at the 3' end of the gene. This is common in samples that are highly degraded. When computing RPKM this could lead to biases where long genes are assigned lower expression levels. See PMID 24632678 for a method to correct for this. Another possibility is to simply use the reads or the number of reads in the last 500bps instead of the reads divided by the gene length. The authors should examine whether there is a correlation of expression with gene length and make sure there is no such bias here (and correct for it if exists).

2) While the authors perform very interesting analysis I think they are still missing some of the power of this single-cell measurement technique. For example Fig. 2C shows a very broad distribution of some of the genes. What does this mean? Are there several sub-populations within the islet, e.g. subpopulations of beta cells, each expressing a high/low subset of genes? Is this just a broad expression distribution representing intrinsic variability stemming from stochastic gene expression? It will be interesting to analyze different sub-populations distinguished by e.g. low or high insulin levels within the beta cell population and seek gene expression signatures for these. Some attempt should be done to examine the variability in expression of different genes, if this variability is not swamped by the experimental variability caused by the fact that the RNAseq method only samples a certain fraction of the cellular mRNA. The authors can follow the strategy of Bengtsson et al, 2005, which they cite for such distribution analysis. They now have the ability to do this on a much broader scale with their unique dataset.

3) What about gene pairwise correlations? Bengtsson et al, 2005 showed some very strong correlations (e.g. *Ins1* and *Ins2* in mouse) and also weak correlations for other gene pairs (e.g. *Ins1* and *Actb*), it could be interesting to present some correlation analysis, e.g. what are the genes that are most highly correlated with *Ins*, *Gcg* etc. Are there particular pairs (or larger clusters) of co-regulated genes?

4) For some of the results it seems that bulk analysis might have also uncovered them, e.g. Fig. 2C-F or the markers in Fig. 3. If this is not the case the authors should stress why their technique has the ability to detect such unique markers, e.g. if the marker is expressed only in a subpopulation of beta cells at high levels but very low levels in other beta cells then bulk analysis would miss these cells. Even if this is not the case the population-specific markers are highly interesting, as are the *DLK1* and *GC* genes which are differentially expressed in mouse and human, but if these results could

have only been achieved using single-cell technique it is important to stress this to motivate the use of the single-cell technique.

5) The section "Defining cell type specific gene expression profiles" is not clear, the authors should rewrite it to explain better.

6) The authors should cite PMID 25931473.

Additional comments of referee 1 in response to pre-decision discussion with the authors:

"The response does not actually provide data or conclusive arguments to address the major issues that have been raised, but I think that authors do have the potential to clarify them, despite that some of them are quite challenging. Notably the original manuscript shows that there is frequent co-expression of cell type specific markers. Is this just about 10% mixed lineage cells, as the authors point out, or is there much greater transcriptional "promiscuity" than previously anticipated based on immunolocalization studies? Alternatively, is this "co-expression" simply technical (eg carry-over of abundant RNAs from dead cells, or during subsequent stages of sample processing)? The authors will need to provide controls and quantitative analysis (not simply PCA and MDA) to clarify these questions and others raised in the reviews.

PS I forgot to mention in my previous comment that in figure 3E DLK seems to be expressed mainly in the cytoplasm of insulin positive cells while in mouse it is expressed in the nucleus of glucagon positive cells. This should be clarified."

1st Revision - authors' response

05 October 2015

RESPONSE TO REFEREES

Referee #1:

This is the first paper to provide single cell transcriptome data from human islet cells, and has the potential to be a major reference for investigators in the field. The analysis is very preliminary, and although there are some interesting example genes that are shown, there are few conclusions of biomedical or biological relevance that can be derived from a systematic analysis of the data. This is more of an editorial decision, but I have the feeling that the standard of EMBO-R requires more in depth analysis of the data. In short I believe that this resource would be of great interest if the authors can improve it.

We thank the referee for the positive comments, and in line with the suggestions have now significantly improved the manuscript.

Major Points:

- One of the major interests of this analysis is that that it can provide unequivocal evidence for expression of a particular gene in alpha, beta, etc cell types. For example, whilst detection of a gene at significant levels, or even apparent enrichment, in FACS-purified beta cells likely means that the gene is expressed in beta cell cells, there is still the theoretical possibility that the gene is highly very expressed in a minor cellular contaminant. By contrast, if the authors can prove that multiple INS+ cells express this marker this is quite solid evidence for expression in INS+ type.

We thank the referee for this comment, and agree that single-cell transcriptomics is the method of choice for proving gene expression in these mature human islet cell subtypes. Accordingly, we now provide single cell profiles for several transcription factors, known and newly described marker genes in different representations (Fig. 1C-E, 2, 3, EV4, EV7, EV8). Furthermore, we provide a set of all primary data (Dataset 2) and an easily accessible genome browser implementation (<http://islettranscriptome.computational-epigenetics.org/>).

- What is the discriminative capacity of this analysis? how many cells are double positive for classic single cell type markers (eg hormones). What is the significance of INS reads in single "alpha" cells? Is this likely technical or biological? Can the authors tease this out by looking for lower abundance cell type specific markers that might be exclusive, thereby ruling out crosscontamination

of some sort. Without a very careful analysis of these issues it is not possible to determine which markers are expressed at different levels in different cell types, vs. which markers are simply cross-contaminants.

Single-cell transcriptomics is technically challenging, and we went to great lengths in order to avoid any source of possible cross-contamination. Crosscontamination by sorting two cells into the same well by our FACS method has been ruled out through a series of control experiments that we described previously (Farlik et al. Cell Rep 2015 Suppl. Fig 1), and we do not observe any transcriptomes that are clear mixtures of two cell types.

To further rule out carry-over before PCR barcoding, we have used two control wells per 96-well plate that did not receive any cell by FACS. Sequencing of libraries generating from these wells did not yield any usable reads, suggesting the absence of carry-over.

Furthermore, we have now added a computational approach to estimate the maximum amount of possible cross-contamination of the single cell transcriptomes. To do so, we selected the most extreme alpha cells and beta cells and established virtual transcriptome for different percentages of “contamination” cross alpha and beta cells. For each single cell, we then assigned a virtual profile that best represents the measured transcriptome, and calculated the variance explained by alpha and beta cell contribution (EV2). While none of the beta cells showed any alpha cell contribution, beta cell transcriptomes could explain a small proportion of the variance (<3%) in the measured alpha cell transcription profiles, which provides an upper bound for the observed level of cross-contamination but is more likely explained by partially overlapping regulatory profiles.

Based on all our analyses, we do not observe evidence for significant crosscontamination between alpha and beta cells in this data set. The level of double positive cells expressing both insulin and glucagon is in line with observations with other methods, e.g. immunofluorescence staining of primary human material, where 5-10% of human islet cells are double positive for insulin and glucagon (Mezza et al. Diabetes 2014).

- Is there any evidence to suggest heterogeneity within INS+, PP+ or GCG+ cells? Despite the relatively low number of cells this should be properly analyzed and discussed.

To analyze heterogeneity within the different islet cell types, we have now added PCA performed for each of the cell types separately. The data is presented in Fig. EV3. For example, within alpha cells two cells are separated from the main population. Reassuringly, the factors driving the separation are not markers of other cell types, suggesting the variation is not due to contamination. However, due to the limited cell number we cannot draw conclusions on the abundance and generality of any subpopulations. This needs to be studied when more data is available in future.

- It might be useful to highlight not just DLK1 and GC as a curiosity, but also systematically display differences between mouse and human cell type specificities. I notice that MAFB is clearly expressed in human beta and alpha (this had been noted before but its unequivocal in this analysis).

We agree with the reviewer that MAFB is an interesting example for the difference between human and mouse. To systematically analyze the mouse vs. human differences, we used a published mouse data set (Benner et al. 2014) in which the authors defined 100 alpha cell genes and 100 beta cell genes. The expression pattern of the top 60 of these genes in our data set is plotted in Fig. EV10. Consistency between human and mouse islet transcription is observed in multiple well-studied genes, including INS, GCG, PDX1, PAX4, MAFA, ARX. But for the majority of the alpha or beta cells genes we do not observe the same cell type specificity in human cells. Some of these differences might be explained by the different technologies used and possible contaminating cell types in the mouse data set which was generated from FACS-purified populations. In line with these arguments, transcriptomes of the commonly used mouse cell lines aTC1 and bTC3 also do not correlate very well with the mouse islet data set. Again, some of the differences will certainly be biological (primary cells vs. cell line), other by co-purifying populations (e.g. PP cells co-purifying with alpha cells) in FACS-sorted islets. The observed species differences will require careful validation, as we did for two of the differentially regulated genes.

- Some suggestions for downstream analysis include:

(a) Given that the authors allude to T2D in the introduction, they may wish to provide an indication of which T2D susceptibility candidate genes (eg closest genes to lead SNPs from GWAs studies) are expressed in different cell types.

We thank the reviewer for the suggestion, and now provide an assignment of diabetes GWAS genes to the different cell types. We downloaded diabetes-related genes identified by GWAS study from NIH-GWAS website (<https://www.ebi.ac.uk/gwas/>). For these genes, the expression level and cell type specificity from our data set is presented in Fig. EV7.

(b) Authors may wish to compare their findings to published "bulk" FACS-purified islet cell RNA-seq datasets (eg Bramswig et al, JCI 2013, Moran et al Cell Metab 2012).

As suggested by the referee, we now included a comparison of our data set to published FACS-purified bulk samples and the whole transcriptomes (Fig. EV1D). Indeed the FACS-purified bulk beta cells are highly similar to beta cell defined in our data set. Additionally, we compared the transcriptome of whole islets from a published data set to the bulk samples (500 cells) of our human islet in Fig. EV1D. Again, our bulk samples are highly similar to published whole islet transcriptomes.

(c) Authors may want to relate their data to whole islet epigenomic studies, for ex can the authors tease out different cell types that are likely contributing different islet enhancer clusters (pasquali et al, Nature genetics, 2014)?

It will be very interesting to relate single cell transcriptomes to whole islet epigenome studies. Such an approach has the potential to assign epigenetic changes to changes in the cell type composition and function. Furthermore, the availability of single cell methylome technology will enable simultaneous studies of transcriptional and DNA methylation differences. These analyses and experiments will be a future focus, but go beyond the scope of the current manuscript.

(d) The motif analysis is provided in raw format, and it seems that only alpha to acinar comparisons are provided in the excel. It is not very useful and this not valid for publication. It needs to be exploited, processed properly to extract meaningful conclusions.

We now provide a revised Dataset3 that includes gene set enrichment results for the comparisons between all the different cell types observed.

- The browser shot provided for PDX1 in the supplementary figure is helpful, but the images that are obtained in the online browser for single cells provide a very different picture. This should be clarified, and if it is just about the default color schemes the version provided in the suppl data is clearly more informative. The browser shots of genes are so tiny that they are not readable without zoom in. The fact that this is a resource paper means that the authors should make an effort so that the data is easily utilized.

We have now revised the supplemental figure for better readability (EV4).

- Fig2A suggests that ARX is expressed in alpha and pp cells as previously reported. However, in figure 2D this observation is only supported by 2 of the 18 cells classified as pp cells.

It is true that we only observe robust ARX expression in two PP cells, out of the nine PP cell we observed. As of now, we can only speculate on the biological significance of this observation.

Minor points:

The browser should have default auto-scaling to be useful, rather than default =1000 RPKM

We find the setting of 1000 RPKM useful for highly expressed marker genes, but the scale can be changed to auto-scaling in the genome browser, simply by clicking the track link and selecting auto-scaling.

Abstract:

"The established marker genes cannot capture cellular heterogeneity in human pancreatic islets"

This is a peculiar statement in an introduction for this paper. Established marker genes do capture heterogeneity. Without single cell studies it is a priori unknown whether hormone markers are sufficient to capture all of human islet cell heterogeneity.

We have now rewritten this sentence to "However, the established marker genes cannot capture the complete spectrum of cellular heterogeneity in human pancreatic islets, ..."

Intro:

"While damage to beta cells is causative of type I diabetes and contributes to the later stages of type II diabetes" The bulk of existing data points to a role of beta cell dysfunction in early stages of T2 diabetes, including genetic predisposition.

We have now rewritten this sentence to "While damage to beta cells causes diabetes, the other pancreatic cell types may also contribute in ways that are not well understood."

Results, first paragraph

"Disassociated": I believe that dissociated is used more often for cells although I am not certain.

Disassociated has been used previously in the context on pancreatic islets (first, PMID: 1098053, 185 Pubmed Entries for disassociated islet) and we have decided to keep it.

Results, first paragraph. When referring to PCA, it not clear if the the authors are referring to different principal components from one PCA, or if they have done iterative PCAs after removing the first PCs from the preceding PCA-this is what appears to have been done in the figure. They should explain clearly how they did this in the text.

We have now rewritten that paragraph, to make clear that we have done iterative PCAs.

Referee #2:

The manuscript of Li et al presents single cell transcriptome data from primary human pancreatic cells from an islet isolation from a 37yr old male with BMI of 22. Of the 70 cells under going RNAseq analysis, 64 passed quality control of an ERCC spikein. Using reiterative principal component analysis based on known cell-specific markers (mainly the specific hormones, SPP1 for ducts, PRSS2 and reg genes for acinar), 12 were beta cells, 18 alpha cells, 11 acinar, 8 duct, 2 delta and 4 undetermined. Then further analysis was able to determine the expression pattern for the PP and delta cells, detect novel transcription factors in some islet cell types and document (with immunostaining for 2 genes DLK1 and GC) opposite expression in human and mouse islets. This valuable resource has also been deposited as a UCSC genome browser track dataset that is available and seemingly user friendly.

This application of new technology for single cell transcriptome analysis contributes a useful database for the field.

We would like to thank the referee for the positive assessment of our work.

There are some biological issues that should be clarified here.

1. Even within the same cell type there is considerable variation in both the presence/absence of expression and the level of intensity as seen in the presented biplot visualizations. Is this real or possible a technical issue?

We have now added a computational approach to estimate the maximum amount of possible cross-contamination of the single cell transcriptomes (EV2). This analysis has excluded possible cross-contamination as a source of variability. Furthermore, we observed higher variance in a subset of alpha cells, whereas all beta cells transcriptomes were free from alpha cells' contribution explaining their variance. These observations are in line with previously published hypothesis of greater

plasticity of alpha cells (Bramswig et al. 2013). In summary, these data make us believe that most variation seen is real not technical.

2. It is surprising that the paper by Dorrell et al (Grompe) (Diabetes 2011) in which careful sorting of alpha, beta, acinar and large and small duct cells from 3-4 human donors was not cited. The current study is complementary (particularly adding data on PP and delta cells).

We thank the referee for pointing out this oversight and now cite this publication in the revised manuscript.

3. It has been well known that REST binding sites are key to islet, particularly beta cell, differentiation (see Bruce et al PNAS 2004 ; DS Johnson Science 2007). This should be acknowledged.

Thank you for this comment, we now discuss this point in the revised manuscript.

4. What the authors have missed commenting on are their definitive data addressing the question of Maf/MafB in human islets. The Vanderbilt group (Powers & Stein) had reported that in humans unlike rodents, MAFB is expressed in about 50% of the adult beta cells and that MAFA was only expressed in about 50 % of them, but it remained unclear if these were colocalizing or separate populations. The biplots seem to indicate that most insulin-expressing cells expressed more MafB than MafA and their genome browser tracks show 2/12 beta cells with no MafA expression but only 1/12 without MafB but 3 /18 alpha cells with little to no MafB.

We believe that this comment underscores the utility of the resource, and have now included a sentence on MAFA and MAFB.

4. p 1 Introduction, " damage to beta cells is causative to type 1 diabetes and "contributes to the later stages of type 2 diabetes" does not reflect the current understanding. In both forms of diabetes the lack of adequate functioning beta cells are causative from the early stages of disease. This statement should be revised.

We revised the sentence accordingly in the revised manuscript.

Referee #3:

In this paper Kubicek and colleagues apply single-cell RNAseq to measure the full transcriptome of 70 human pancreatic islet cells. Single-cell RNAseq is an extremely powerful approach for detecting cellular heterogeneity in a complex tissue without the need for biased enrichment for a particular cell type. In this respect the technique is especially suitable for human pancreatic islets, where targeted isolation is not trivial. Although the analysis was only performed on a single individual the paper provides an important resource and also some very interesting analysis, and will prove to be a great resource for those interested in both islet biology and single-cell genomics.

We thank the referee for the positive comments and appreciation of our work.

The authors should address the following points:

1) When looking at the track plots in Fig. 2A it seems as if there is a non-uniform coverage of the reads, where most reads are located at the 3' end of the gene. This is common in samples that are highly degraded. When computing RPKM this could lead to biases where long genes are assigned lower expression levels. See PMID 24632678 for a method to correct for this. Another possibility is to simply use the reads or the number of reads in the last 500bps instead of the reads divided by the gene length. The authors should examine whether there is a correlation of expression with gene length and make sure there is no such bias here (and correct for it if exists).

We have now included an analysis of the correlation between expression and gene length (Fig. EV1C), showing prove that there is no bias due to transcript length.

2) While the authors perform very interesting analysis I think they are still missing some of the power of this single-cell measurement technique. For example Fig. 2C shows a very broad

distribution of some of the genes. What does this mean? Are there several sub-populations within the islet, e.g. subpopulations of beta cells, each expressing a high/low subset of genes? Is this just a broad expression distribution representing intrinsic variability stemming from stochastic gene expression? It will be interesting to analyze different sub-populations distinguished by e.g. low or high insulin levels within the beta cell population and seek gene expression signatures for these. Some attempt should be done to examine the variability in expression of different genes, if this variability is not swamped by the experimental variability caused by the fact that the RNAseq method only samples a certain fraction of the cellular mRNA. The authors can follow the strategy of Bengtsson et al, 2005, which they cite for such distribution analysis. They now have the ability to do this on a much broader scale with their unique dataset.

We agree with the reviewer that the detection of variation with a cell type is one of the advantages of single cell RNA-seq assay. To further address this point, we performed PCA within the individual cell (Fig. EV3). This figure shows some cells deviating from the bulk of their sub-type. Expanding our analysis to more cell in the future will reveal whether distinct subpopulations of cell exist.

3) What about gene pairwise correlations? Bengtsson et al, 2005 showed some very strong correlations (e.g. *Ins1* and *Ins2* in mouse) and also weak correlations for other gene pairs (e.g. *Ins1* and *Actb*), it could be interesting to present some correlation analysis, e.g. what are the genes that are most highly correlated with *Ins*, *Gcg* etc. Are there particular pairs (or larger clusters) of coregulated genes?

We now added pairwise correlations of all measured genes with the main marker genes (*INS*, *GCG*, *SST* and *PPY*). While for *INS* and *SST* several highly correlated genes could be identified (e.g. *DLK1*, *CD72* for *INS* and *NKX6-3*, *ZNF430* for *SST*), the expression profiles of *GCG* and *PPY* did not show high correlation with any of the other genes.

4) For some of the results it seems that bulk analysis might have also uncovered them, e.g. Fig. 2C-F or the markers in Fig. 3. If this is not the case the authors should stress why their technique has the ability to detect such unique markers, e.g. if the marker is expressed only in a subpopulation of beta cells at high levels but very low levels in other beta cells then bulk analysis would miss these cells. Even if this is not the case the population-specific markers are highly interesting, as are the *DLK1* and *GC* genes which are differentially expressed in mouse and human, but if these results could have only been achieved using single-cell technique it is important to stress this to motivate the use of the single-cell technique.

It is true that the difference in *DLK1* and *GC* gene expression should also have been detected in bulk transcriptomes from purified alpha and beta cells. We checked two different publications (Dorrell et al *Diabetologia* 2011 and Bramswig et al *JCI* 2013) that used the same purification strategy and were performed by different groups. The cell-type specific expression of *GC* in human islets was identified by both publications but not highlighted. The expression pattern of *DLK1* was variable. The Dorrell et al paper showed *DLK1* is among the top 50 beta cell specific genes but the specificity is much smaller (<4 fold) in the Bramswig et al paper. One possible explanation of this variance is the bulk sample experiment relies on antibody-based FACS-sorting, which is not the case in our assay. Beside this, with single cell RNA-seq we learned that *GC* is highly expressed in both alpha and pp cells, which is very difficult to detect in bulk-sample based studies.

5) The section "Defining cell type specific gene expression profiles" is not clear, the authors should rewrite it to explain better.

We have rewritten this section in the revised manuscript.

6) The authors should cite PMID 25931473.

We now cite this publication in the revised manuscript.

Thank you for your patience while we have reviewed your revised manuscript. As you will see from the reports below, the referees are now all positive about its publication in EMBO reports. I am therefore writing with an 'accept in principle' decision, which means that I will be happy to accept your manuscript for publication once the few minor issues that were still raised by the reviewers (please find their reports copied below) have been addressed.

Since these concerns are mostly related to further clarifications or textural changes I would not expect this to be a major issue.

On a more formal note, you currently have 12 expanded view (EV) figures. Unfortunately, our publisher can only accommodate 5 EV figures and unless you are able to combine some (which I think will be hard as they are already pretty large), I would unfortunately need to ask you to identify some that would be displayed as appendix figures. This is basically a third level of figure display (after the main figures and the EV figures) and would make sense for data that is not directly related or linked to any of the main figures. I am sorry for this inconvenience and hope for your understanding and cooperation.

If all remaining corrections have been attended to, you will then receive an official decision letter from the journal accepting your manuscript for publication in the next available issue of EMBO reports. This letter will also include details of the further steps you need to take for the prompt inclusion of your manuscript in our next available issue.

Thank you again for your contribution to EMBO reports.

REFEREE REPORTS

Referee #1:

This is an excellent revision that has improved in clarity content and focus. The resource will be of great value to the community. I have a few comments I hope the authors can deal with very easily

1. Intro:

"We obtained an average of 12.7 million high quality reads per single cell, of which 62.9% aligned to the human reference genome"

I suggest to rephrase. As stated that would mean that nearly 40% were not human sequences but I assume that authors mean 62% are aligned after filtering for quality, uniqueness etc

2. Results

"we found evidence for the hypothesis that alpha cells are more plastic than beta cells"

This conclusion is overstated given that it is only based on the observation that there is high expression variance/heterogeneity in a few alpha cells

3. GWAS/Figure EV9

Overall this is a very useful analysis of the data.

Panel A. Can you please explain further in the figure legend how this analysis was carried out, including criteria that was used to represent a locus/gene in each column. What does "diabetes not specified mean?"

In panel B the authors have decided to only show genes that have a stronger cell type specificity when comparing between pancreatic cell types. This means that information on major Type 2 diabetes loci like TCF7L2 is missing -it is actually known to be relatively ubiquitously expressed. Can the authors make this point more clear in the fig legend. What does the scale in panel B represent: "Mean In(RPKM)"?

In the text in the results section the authors provide a somewhat negative message, by highlighting

that only some "GWAS genes" have a strong cell type specificity pattern when comparing to other pancreatic cell types. However, most "GWAS genes" appear to be expressed in some pancreatic endocrine cells or in duct cells. Obviously a gene can play an important role in beta, delta or alpha cells even though it is also expressed in several other pancreatic cell types.

DRB1 is an established Type 1 diabetes locus so I am not sure why it is classified with Type 2 (I am guessing there might be some data somewhere on T2 diabetes). It is worth noting, eg in figure legend, what set of GWAS loci were interrogated -was it the GWAS catalog for Type 2 diabetes, all loci with genome wide significance P values? Are obesity loci also being included?

Referee #2:

The revised manuscript of Li et al has provided additional clarifications and important analyses. These data should provide a excellent resource for the islet biology community.

Two things linger:

- 1) it would be useful to define GC as vitamin D binding protein the first time used since the reader may think it is a typo and actually means glucagon (GCG).
2. My previous question re MAFA and MAFB in the same cell was not addressed. A sentence was added that "there was robust MAFB expression in alpha, beta and delta cells" but that MAFA was only in beta cells. This is a question of interest to the islet biology community and would be helpful to address more fully since the data are there.

Referee #3:

The authors have done a good job in addressing most of the referees' comments. Unfortunately they have not done a good job in addressing my main concern about RPKM normalization and gene length. EVIC, which they claim addresses this point, is not clear. What are the red and blue curves? Why not simply show the correlation between RPKM and gene length? Also, the authors do not elaborate on the normalization they used: "We calculated RNA expression levels using RPKM normalization, which corrects for non-uniform read distribution (e.g. 3-prime bias) along the transcripts.", please elaborate on this procedure. This step is fundamental and must be explained better, if a length bias does not exist a simple correlation of the RPKM with gene length will show it.

Additional comment:

Correlation analysis - very interesting but what is the criterion to show only these genes and not others? Is there any merit of statistical significance of the correlation? Why not cluster the full correlation matrix of all genes?

2nd Revision - authors' response

16 November 2015

Referee #1:

This is an excellent revision that has improved in clarity content and focus. The resource will be of great value to the community. I have a few comments I hope the authors can deal with very easily

We thank the referee for the positive assessment and appreciation of the improvements in the revised version.

1. Intro:

"We obtained an average of 12.7 million high quality reads per single cell, of which 62.9% aligned to the human reference genome"

I suggest to rephrase. As stated that would mean that nearly 40% were not human sequences but I assume that authors mean 62% are aligned after filtering for quality, uniqueness etc

On average actually "only" 62.9% of the high quality reads (after quality filtering and adapter trimming) aligned to the human genome. In addition, 9% of the high quality reads align to the

ERCC spike-in controls. A rate of 20-30% unalignable reads is not uncommon in low-input sequencing experiments. We tried to identify the origin of these unalignable reads using blast but without any systematic results.

2. Results

"we found evidence for the hypothesis that alpha cells are more plastic than beta cells"
This conclusion is overstated given that it is only based on the observation that there is high expression variance/heterogeneity in a few alpha cells

We thank the referee for this comment and have now rephrased this sentence.

3. GWAS/Figure EV9

Overall this is a very useful analysis of the data.

Panel A. Can you please explain further in the figure legend how this analysis was carried out, including criteria that was used to represent a locus/gene in each column. What does "diabetes not specified mean?"

We have updated the figure legend to clarify how this analysis was carried out. GWAS results for the search word "Diabetes" were downloaded from the GWAS Catalog (<https://www.ebi.ac.uk/gwas/>). Each gene was classified as "Type 1 Diabetes", "Type 2 diabetes" or "not specified" (including diseases like "Cystic fibrosis related diabetes" that could not easily be assigned to T1D or T2D) according to the reported GWAS patient group.

In panel B the authors have decided to only show genes that have a stronger cell type specificity when comparing between pancreatic cell types. This means that information on major Type 2 diabetes loci like TCF7L2 is missing -it is actually known to be relatively ubiquitously expressed. Can the authors make this point more clear in the fig legend. What does the scale in panel B represent: "Mean ln(RPKM)"?

As suggested, we have now revised the legend accordingly. We display expression values as Mean ln(RPKM), the mean natural logarithm of the RPKM values of all cells assigned to a particular type.

In the text in the results section the authors provide a somewhat negative message, by highlighting that only some "GWAS genes" have a strong cell type specificity pattern when comparing to other pancreatic cell types. However, most "GWAS genes" appear to be expressed in some pancreatic endocrine cells or in duct cells. Obviously a gene can play an important role in beta, delta or alpha cells even though it is also expressed in several other pancreatic cell types.

We thank the referee for this comment and have rephrased the discussion accordingly.

DRB1 is an established Type 1 diabetes locus so I am not sure why it is classified with Type 2 (I am guessing there might be some data somewhere on T2 diabetes). It is worth noting, eg in figure legend, what set of GWAS loci were interrogated -was it the GWAS catalog for Type 2 diabetes, all loci with genome wide significance P values? Are obesity loci also being included?

Thank you for calling this to our attention. We realized that we had inadvertently changed the assignment of genes and trait/disease as downloaded from GWAS Catalog. We have fixed this mistake now and updated the Figure. DRB1 is now classified as Type 1 Diabetes gene. The displayed genes remain the same but their classification as T1 or T2 diabetes has changed for some genes.

Referee #2:

The revised manuscript of Li et al has provided additional clarifications and important analyses. These data should provide a excellent resource for the islet biology community.

We thank the referee for the positive comment.

Two things linger:

1) it would be useful to define GC as vitamin D binding protein the first time used since the reader may think it is a typo and actually means glucagon (GCG).

As suggested, we have now clearly defined GC as vitamin D binding protein at the first occurrence.

2. My previous question re MAFA and MAFB in the same cell was not addressed. A sentence was added that "there was robust MAFB expression in alpha, beta and delta cells" but that MAFA was only in beta cells. This is a question of interest to the islet biology community and would be helpful to address more fully since the data are there.

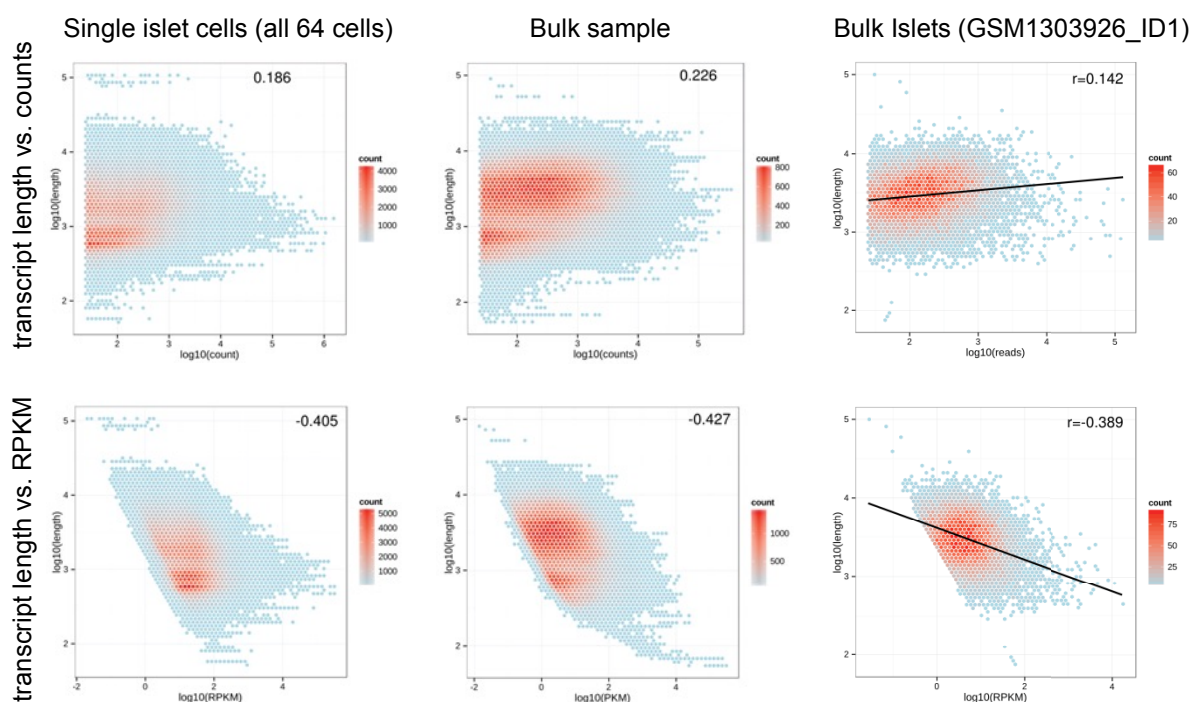
Of the twelve beta cells studied, six cells express MAFA and MAFB concomitantly, five express only MAFB and in one beta cell we detect neither of the two factors. We have now further emphasized these findings in the main text.

Referee #3:

The authors have done a good job in addressing most of the referees' comments. Unfortunately they have not done a good job in addressing my main concern about RPKM normalization and gene length. EV1C, which they claim addresses this point, is not clear. What are the red and blue curves? Why not simply show the correlation between RPKM and gene length? Also, the authors do not elaborate on the normalization they used: "We calculated RNA expression levels using RPKM normalization, which corrects for non-uniform read distribution (e.g. 3-prime bias) along the transcripts.", please elaborate on this procedure. This step is fundamental and must be explained better, if a length bias does not exist a simple correlation of the RPKM with gene length will show it.

We have relied on ERCC spike-in controls (for which the number of molecules in the reaction (aka "expression level") as well as the "transcript" length is known) to address the correlation of RPKM values with gene length. As indicated in the figure legend, EV1c (now EV1d) plots raw counts and RPKMs in dependence of ERCC length colored by coverage (≤ 25 reads in red, > 25 reads in blue), confirming the absence of a length bias. As described in the method section, RPKM normalization was done using the tool BitSeq, which has been previously published [1]. We further confirmed with the authors that BitSeq accounts for read-distribution biases (Antti Honkela, personal communication, 25.2.2015).

At the request of the referee, we have now added the correlation plot of RPKM with transcript length and detected a negative correlation of -0.405. We now provide this plot as new Fig. EV1c. The higher expression of shorter transcripts has previously been reported using an independent method [2]. To confirm that these effects are not specific to single-cell RNA-Seq we here provide additional plots from a bulk RNA-Seq sample processed with a different pipeline (tophat/cufflinks) and from a randomly selected human islet sequence from the literature (GSM1303926_ID1) [3]. Both data sets show negative correlation of RPKM values and transcript length comparable to single-cell data.



We would like to further point out that in our entire study we refrain from any claims of comparing expression levels of different transcripts. Rather, we only compare the expression of the same transcript between cells, thereby eliminating any transcript length bias. Since normalization for comparing expression levels of different transcripts is an unsolved issue in the community, we now also include a sentence in the main manuscript that raises awareness to this potential bias.

Additional comment:

Correlation analysis - very interesting but what is the criterion to show only these genes and not others? Is there any merit of statistical significance of the correlation? Why not cluster the full correlation matrix of all genes?

As indicated in the figure legend, we selected the endocrine markers INS, GCG, PPY, and SST and plotted all transcripts correlated with $r > 0.9$. Font size limitations prevent including transcripts with lower correlation or additional marker genes in a single figure. However, we provide all raw data as supplement, allowing readers to generate correlation plots for any gene of interest to desired cutoffs.

1. Glaus P, Honkela A, Rattray M (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* **28**: 1721-8
2. Chiaromonte F, Miller W, Bouhassira EE (2003) Gene length and proximity to neighbors affect genome-wide expression levels. *Genome Res* **13**: 2602-8
3. Cnop M, Abdulkarim B, Bottu G, Cunha DA, Igoillo-Esteve M, Masini M, Turatsinze JV, Griebel T, Villate O, Santin I, *et al.* (2014) RNA sequencing identifies dysregulation of the human pancreatic islet transcriptome by the saturated fatty acid palmitate. *Diabetes* **63**: 1978-93

3rd Editorial Decision

19 November 2015

I am very pleased to accept your manuscript for publication in the next available issue of EMBO reports. Thank you for your contribution to our journal.