Title
Deconvolution of the Human Endothelial Transcriptome

Journal
Cell Systems, 3(3)

ISSN
2405-4712

Authors
Rau, Christoph D
Gao, Chen
Wang, Yibin

Publication Date
2016-09-01

DOI
10.1016/j.cels.2016.09.006

Peer reviewed

HHS Public Access

Author manuscript

Cell Syst. Author manuscript; available in PMC 2017 April 18.

Published in final edited form as:

Cell Syst. 2016 September 28; 3(3): 218–220. doi:10.1016/j.cels.2016.09.006.

Deconvolution of the Human Endothelial Transcriptome

Christoph D. Rau[1], **Chen Gao**[1], and **Yibin Wang**[1,*]

[1]Division of Molecular Medicine, Departments of Anesthesiology, Medicine, and Physiology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

Abstract

A systems approach deconvolutes genes specific to and enriched in endothelium from whole-organ transcriptome data, with applications to other cell types and tissues.

Each human organ consists of many different types of cells working together to carry out complex physiological functions, but understanding cell-type-specific gene expression (Otsuki et al., 2014) is challenging because most published human transcriptome datasets are composite gene-expression profiles generated from mixed cell populations. In this issue of *Cell Systems*, Butler et al. (2016) report a relatively simple but potentially effective way to identify endothelial-cell-enriched genes based on the straightforward correlation of three endothelial-cell-specific reference genes with transcriptome datasets generated from unfractionated human tissues (Butler et al., 2016). They demonstrate that such an approach detects many known and novel endothelial-cell-enriched mRNA species among different human tissue samples. If this concept is extendable to other cell types, this informatics trick could facilitate a rapid, cost-effective deconvolution of whole-tissue gene-expression profiles in order to reveal cell-type-specific features in the otherwise convoluted human transcriptomes.

Current state-of-the-art approaches to identify and analyze cell-type-specific gene expression within a tissue or organ are mechanical in nature, requiring either isolation of a target cell population via cell fractionation or laser capture microdissection prior to RNA analyses (Datta et al., 2015), or employing single-cell RNA sequencing (RNA-seq) on a dissociated cell mixture (Kolodziejczyk et al., 2015). Both methods are resource intensive and technically challenging (Stegle et al., 2015), thus limiting their widespread applications. Furthermore, these methods are not useful for the analysis of datasets acquired from prior studies.

The study reported by Butler et al. (2016) builds on the Human Protein Atlas Project (HPA: http://www.proteinatlas.org) (Uhlén et al., 2015), in which 124 human samples from 32 organs were analyzed by histology and RNA-seq. The authors selected three genes (*CLEC14A*, *vWF*, and *CD34*) as highly reliable endothelial cell reference markers because their mRNA levels were highly correlated with the degree of vascularity across different

*Correspondence: yibinwang@mednet.ucla.edu.

tissue beds, an indicator of the number of endothelial cells in a tissue, and because their expressions were highly correlated to one another.

To demonstrate that these three genes really are endothelial cell specific and can provide adequate sensitivity to detect additional endothelial cell mRNAs based on their combined correlation coefficients, Butler et al. (2016) performed several tests using previously established known endothelial cell mRNAs and other cell-type-specific mRNAs as experimental datasets. These analyses led to the conclusion that the mean correlation value from the three reference genes needed to be set at ≥ 0.5 to detect ~90% of the known endothelial cell mRNAs while discriminating against non-endothelial cell mRNAs. Even with this combined threshold, however, 25% of smooth-muscle-cell-specific genes were also included when they should have been excluded. This result is, perhaps, expected, as intricate interactions are known to exist between vascular smooth muscle and endothelium in almost all vasculature structures, and genes that are correlated to an increase in endothelial cells may also be correlated to an increase in smooth muscle.

These tests revealed the limitation of such correlation analysis in terms of robustness and specificity, perhaps due to the limited number of reference genes used in the test. Nonetheless, the outcome of applying it to the 20,000-plus mRNA transcripts detected in the human tissues from the Human Protein Atlas Project dataset is quite encouraging. In total, 332 mRNA transcripts were detected to be endothelial-cell-enriched genes after excluding possible smooth muscle cell “spill-over” by testing against a number of smooth-muscle-cell-specific marker genes. Notably, ~70% of them were re-identified when applying the same approach to independent human transcriptome datasets generated from Genotype-Tissue Project (GTEx) (http://www.gtexportal.org/home) (Carithers et al., 2015).

These identified mRNAs include known, as well as many unknown and un-characterized, endothelial-cell-enriched mRNAs. For several novel endothelial cell genes, Butler et al. (2016) provided additional experimental validation by targeted analysis of mRNA and protein levels in human tissues or purified endothelial cells. In the end, this approach led to the detection of a substantial number of endothelial-cell-enriched mRNAs from transcriptome profiles generated using human tissues without the need to perform extensive cell purification, isolation, or single-cell RNA-seq.

The endothelial-cell-expression profile established by this study can be useful to researchers in future functional studies of other specific genes. The same approach can also be used by researchers to detect endothelial-cell-enriched mRNAs from other datasets. In comparison to microarray microdissection with analysis of differences (MMAD) (Liebner et al., 2014), the current informatics approach is significantly simpler and relies on fewer assumptions.

Beyond the proof-of-concept demonstration reported here to identify endothelial-cell-enriched and endothelial-cell-specific genes from human tissues, this approach can also be useful in several other potential applications. As suggested by Butler et al. (2016), the same method may be applied to detect other cell-type-specific genes in intact organs, such as neuron versus glia cells in the brain, muscle versus fibroblasts in the heart, or epithelium

versus endothelial in the lung. The method should be applicable as long as a cell-type-specific reference gene set can be identified and established in each dataset (Figure 1).

Furthermore, if gene-expression profiles can be established under a basal state for different cell types within a tissue or organ, these identified genes can be used to monitor how cell-type-specific gene expression is altered under different stressors, global phenotypic changes, or tissue types. This would be particularly valuable since the extensive datasets of human transcriptome deposited in the public domains could be further interrogated in light of cell-type-specific contributions. Finally, there is no reason why the same principle cannot be extended to other RNA species, such as long non-coding RNAs and microRNAs.

Despite the simplicity and wide applicability of their method, Butler et al. (2016) clearly demonstrate that their approach also has shortfalls and limitations. The most pervasive difficulty with methods based on correlation to deconvolute datasets is the need to balance false positives against false negatives. In their study, they showed that a 10% false-negative rate for known endothelial-cell-specific markers was linked to a 25% false-positive rate for smooth-muscle-cell-specific markers. The false-positive rate for other cell-type markers remained to be assessed. An attempt to reduce this false-positive rate would involve either increasing the number of reference genes or increasing the stringency of correlation to these genes, both of which would, by necessity, increase the false-negative rate.

Further research into this topic should explore the use of more sophisticated analysis methods such as unsupervised network construction algorithms like WGCNA (Langfelder and Horvath, 2008) to identify genes that cluster with known reference genes. For example, previous work applied machine learning to condition-specific gene-expression datasets to identify genes specific to podocytes in the kidney (Ju et al., 2013). Alternatively, principle component analysis may provide an additional means to increase the sensitivity of the method to genes that may be enriched in multiple cell types. Clearly, the resolution of such analysis also precludes its detection of rare cell types or subtypes of cells undergoing potential physiological or pathological changes. Finally, the current approach may be vulnerable if the reference gene expression is significantly altered under pathological disturbance independent from other cell-type-specific genes within the targeted tissues. This problem may have to be resolved by incorporating more diverse tissue samples that take into account these alterations due to pathological disturbances to increase the analysis power.

In short, as we rapidly expand our data generation capacity through evermore powerful sequencing and mass spectrometry techniques, the current capacity to annotate and interrogate these datasets to extract useful knowledge and information is increasingly becoming rate limiting. The study from Butler et al. (2016) highlights the utility of datasets such as the Human Protein Atlas and the Genotype-Tissue Project, provides a valuable resource for endothelial biologists, and adds another simple but potentially useful method in our toolbox to achieve a deconvoluted human transcriptome with better resolution for individual cell and tissue types.

Acknowledgments

This work is supported in part by NIH, HL114437 for Y.W. and American Heart Association Postdoctoral Fellowship Award to C.D.R.

References

Butler LM, Hallström BM, Fagerberg L, Pontén F, Uhlén M, Renné T, Odeberg J. Cell Syst. 2016; 3 this issue, ■■■–■■■.

Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, De-Luca DS, Peter-Demchok J, Gelfand ET, et al. Biopreserv Biobank. 2015; 13:311–319. [PubMed: 26484571]

Datta S, Malhotra L, Dickerson R, Chaffee S, Sen CK, Roy S. Histol Histopathol. 2015; 30:1255–1269. [PubMed: 25892148]

Ju W, Greene CS, Eichinger F, Nair V, Hodgin JB, Bitzer M, Lee YS, Zhu Q, Kehata M, Li M, Jiang S, et al. Genome Res. 2013; 23:1862–1873. [PubMed: 23950145]

Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. Mol Cell. 2015; 58:610–620. [PubMed: 26000846]

Langfelder P, Horvath S. BMC Bioinformatics. 2008; 9:559. [PubMed: 19114008]

Liebner DA, Huang K, Parvin JD. Bioinformatics. 2014; 30:682–689. [PubMed: 24085566]

Otsuki L, Cheetham SW, Brand AH. Wiley Interdiscip Rev Dev Biol. 2014; 3:429–443. [PubMed: 25174322]

Stegle O, Teichmann SA, Marioni JC. Nat Rev Genet. 2015; 16:133–145. [PubMed: 25628217]

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. Science. 2015; 347:1260419. [PubMed: 25613900]

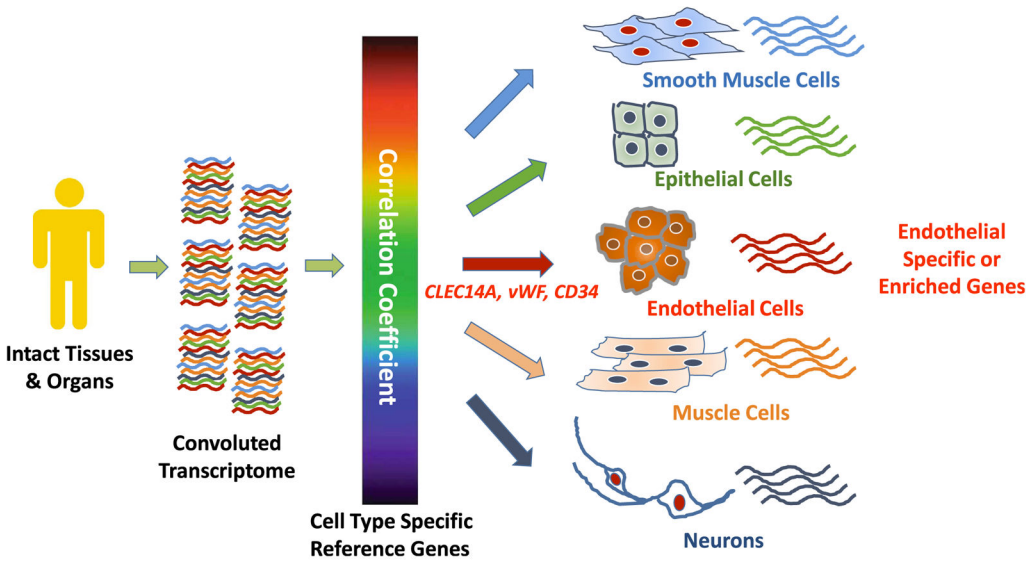

Figure 1. Reference-Gene-Based Human Transcriptome Deconvolution

Illustration of the approach employed by Butler et al. (2016) to identify endothelial-specific or -enriched genes from unfractionated human tissues using combined correlation coefficient analysis with three endothelial-cell-specific reference genes. Other cell types are included to indicate potential applications to deconvolute the human transcriptome for other cell-type-specific genes.