Transcriptional output, cell types densities and normalization in spatial transcriptomics

Manuel Saiselet,[1¶] Joël Rodrigues-Vitória,[1¶] Ligia Craciun,[2] Alex Spinette,[2] Denis Larsimont,[2] Guy Andry,[3] Joakim Lundeberg,[4,5] Carine Maenhaut,[1#] Vincent Detours[1#*]

[¶]Contributed equally

[#]Contributed equally

*Correspondence should be addressed to vdetours@ulb.ac.be

[1] IRIBHM, Université Libre de Bruxelles (ULB), Route de Lennik, 808 - 1070 Brussels, Belgium.

[2] Department of Pathology, Jules Bordet Institute, Université Libre de Bruxelles (ULB), Boulevard de Waterloo, 125 - 1000 Brussels, Belgium.

[3] Department of Head & Neck and Thoracic Surgery, Jules Bordet Institute, Université Libre de Bruxelles (ULB), Boulevard de Waterloo, 125 - 1000 Brussels, Belgium.

[4] Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, SE-106 91 Stockholm, Sweden

[5] Department of Bioengineering, Stanford University, 443 Via Ortega, Shriram Center, Stanford, CA 94305-4245, USA

Spatial transcriptomics measures mRNA at hundreds of 100μm-diameter spots evenly spread across 6.5×6.9mm^2 histological slices. Gene expression within each spot is commonly normalized by total read counts. However we show that the transcriptional output of individual spots reflects the number of cells they contain, hence total read counts per spot reflect relevant biology. Although per-spot read-count normalization reveals important enrichment trends, it may heavily distort cell-type-related absolute local expression and conceal important biological information.

Spatial transcriptomics (ST) bridges untargeted genome-wide mRNA profiling with tissue morphology by making it possible to perform RNA-seq at hundreds of precisely located spots on the surface of an histological slice.[1] Since mRNA diffusion is minimal during tissues permabilization and mRNA capture, the transcriptome of each spot is thought to aggregate the transcriptomes of the cells it contains. The number of cells within a spot and their transcriptional output depend on their type, state and on overall local morphology.

ST data share many limitations of single cell RNA-seq, including low coverage and high dropout rate. So far, ST studies have relied on preprocessing pipelines inspired by single cell RNA-seq studies.[1–7] These include normalization of gene-wise read counts in a cell/spot by the total number of reads collected from that cell/spot. But the number of reads obtained from a spot could reflect its cellular content or technical variation in RNA capture and amplification. Thus, whether read count normalization is warranted in the context of ST remains an open question. We addressed it by quantifying the cellular content of individual spots from image analysis and by comparing it with read counts.

A BRAF V600E-mutated papillary thyroid cancer (PTC) was profiled with ST (Supplementary Material and Methods). Pathology review of the H&E image revealed five major types of morphologies (Fig. 1a). This qualitative approach was complemented by whole-slide machine learning-based localization of nuclei, and their classification within three categories (Fig. 1b): epithelial cells, fibroblasts and 'other cells', which mostly contains immune cells. Among 86,111 detected nuclei (Fig. 1c), 31% were located within ST spots. The mean number of cells per spots varied from 0 to 197 (median 67).

Spot-wise read coverage varied from 356 to 8,749 across the slide (Fig. 1e), a 25-fold variation. It was associated with the number of cells of all types in a multivariate analysis (Fig. 1g), particularly epithelial cells (Fig. 1f). As expected, total read count per spots was highest in dense epithelial areas, and lowest in low cell density fibrotic zones (Fig. 1e,g). We concluded that total read counts per spot reflects relevant quantitative and qualitative features of tissue morphology.

Importantly, the number of nuclei of the different cell types explained together 48% of total read count variance. This is substantial, yet 52% of read count variance is explained by currently unknown factors.

To assess the effect of normalization we compared raw read counts with raw counts adjusted for total counts and with scale-normalized expression estimates generated by DCA,[8] a recent neural autoencoder-based algorithm developed in the context of single cell RNA-seq. Figs. 2a and 2b show expression of thyroglobulin (TG, a thyroid differentiation marker), of vimentin (VIM, a mesenchymal intermediate filament), respectively. Raw counts and normalized expression of TG all closely followed epithelial density and total counts (compare Fig. 2a to Figs. 1a and 1e). VIM raw counts were substantial in the epithelial areas, but also in the cellular fibrosis and immune foci. Normalization, however, revealed a dramatically different picture, particularly for DCA: while remaining high in cellular fibrosis and immune foci, VIM expression was lower in epithelium (Fig. 2b).

Thus, normalization affected the spatial expression pattern of VIM, but not TG. The absolute numbers of epithelial cells and fibroblasts per spot were weakly associated (Fig. 2c), while their relative proportions, *i.e.* their number divided by the total number of cells within a spot, were massively anti-correlated (Fig. 2d). The raw counts of TG and VIM are positively correlated (Fig. 2e), while their normalized values are negatively correlated (Fig. 2f). The positive correlation between TG and VIM raw counts (Fig. 2e) suggests that the

tumoral epithelium expresses VIM and could undergo an epithelial-mesenchymal transition. This transition has been reported in BRAF V600E-mutated tumors[9] such as this one. VIM is also expressed in primary cultures of normal thyrocytes treated with Epidermal Growth Factor, which inhibits differentiation, but not of thyrocytes treated with Thyroid Stimulating Hormone, which promotes thyroid differentiation—while both are mitogenic.[10] Alternatively, VIM expression by fibroblast could be promoted by nearby epithelial cells. Overall, raw counts seemed more related to the number of cells of a given cell type, while normalized expression captured cell types' relative proportions.

To gain insights on the global effect of normalization, we calculated the distribution of correlations between genes across spots for all three expression metrics (Fig. 2g). Raw counts were positively correlated for any pair of genes. By contrast, normalized expressions correlations were centered on 0. This implies that genes tend to show a similar expression pattern that reflects total transcriptional output when raw counts are considered, while normalized expression highlights contrasts between genes.

Note that denoising and missing data imputation are not technically tied to total read counts normalization. For example, DCA estimates read counts scale factors, but users are free to use its scale-adjusted or un-adjusted outputs.[8]

These findings echo an earlier study pointing at potentially misleading interpretations arising from the comparison of normalized bulk transcriptomes.[11] ST is different, however: by visualizing the cellular context of morphology we may gauge the cellular correlates of normalization factors. In the context of single cell RNA-seq, transcriptome-matched cell imaging allowing independent identification of cell morphology may help evaluate the impact of total read count normalization in the future.

In summary, raw read counts inform about the absolute density of cell types, while normalized expression informs about their relative proportions. Which one is most relevant depends on the biological question being addressed. It is remarkable that normalized expression better detects specific morphologies such as cellular fibrosis (Figs. 2a and 2b, right panels), while raw counts do reflect actual cell-type local densities and may highlight atypical expression patterns, as exemplified here for VIM in regions of high epithelial density.

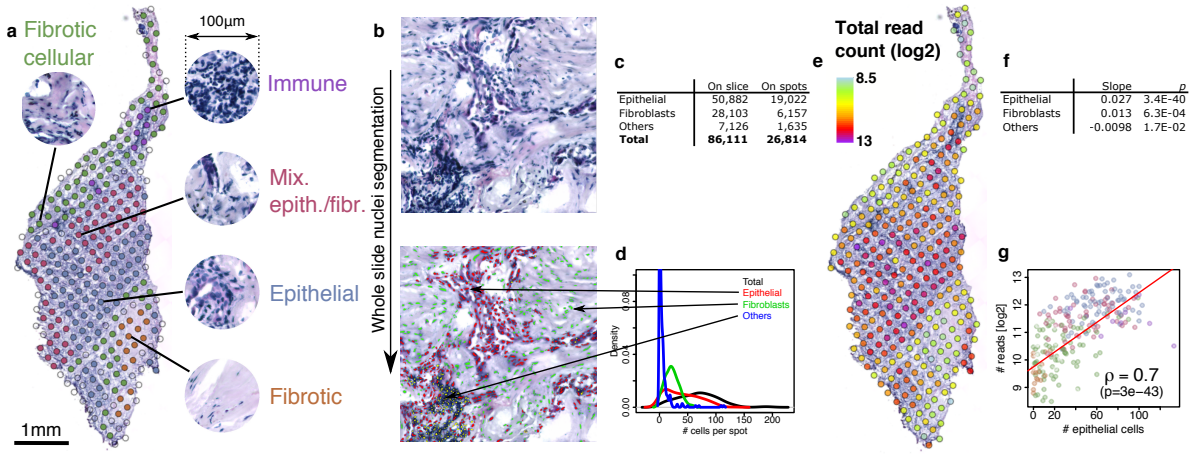

Figure 1. **Variation of total read counts is related to morphology and number of cells of different types**. **a**, Five types of morphological regions were determined from pathology. The transcriptome was determined for each spot with ST. **b**, The nuclei on the H&E image were segmented and classified with a machine learning-based algorithm (see Supplementary Material and Methods). **c**, Nuclei counts. **d**, Distributions of the number of cells per ST spot. **e**, Total read count per spot. **f**, Multivariate analysis of association between cell numbers and the $\log_2$ of total read count per spot. **g**, Each point represents a ST spot with same color code as in panel **a**, ρ denotes the Spearman's correlation.

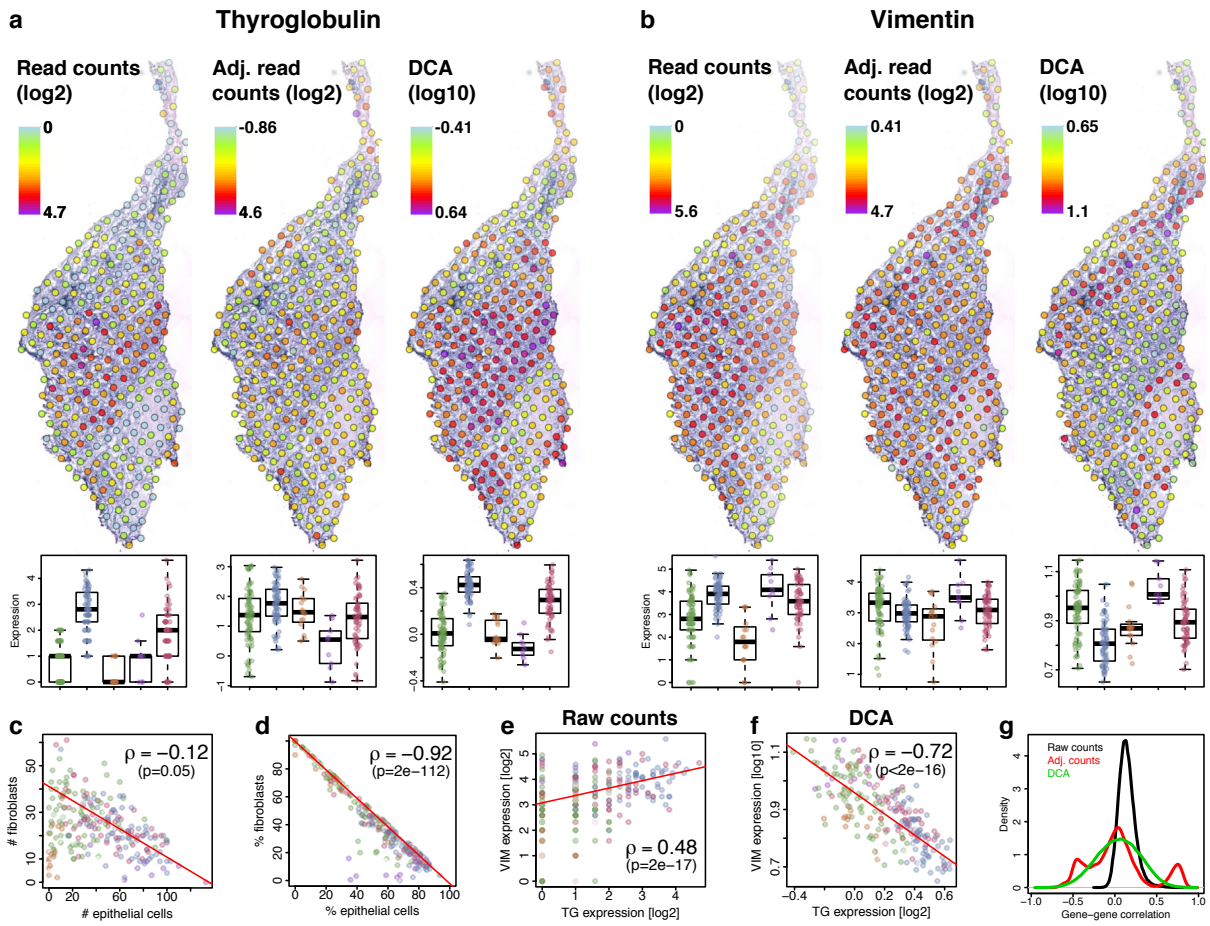

Figure 2. Effect of normalization. **a,** Expression of thyroglobulin without normalization (left), with adjustment for total read counts (center), and with a neural autoencoder-based normalization (DCA, right). Boxplots represents the expression of spots in the regions shown Fig. 1a (same color code). **b**, Same as panel **a**, except that vimentin expression is shown. **c**, Points represent spots (same color code as Fig. 1a) with the absolute number of epithelial cells (*x*-axis) and the fibroblasts (*y*-axis). **d**, Same as **c** except that cell types proportions are depicted. The large negative correlation stems from the low number of 'other' cells (Fig. 1c). **e** and **f**, Expression of TG and VIM are compared using raw counts or auto-encoder-based normalization. **g**, Distribution of gene vs. gene correlation across spots for raw counts and normalized data.

Code and data availability

https://github.com/vdet/st-normalization

References

1. Ståhl, P. L. *et al. Science* **353**, 78–82 (2016).
2. Asp, M. *et al. Sci Rep* **7**, 12941 (2017).
3. Giacomello, S. *et al. Nat Plants* **3**, 17061 (2017).
4. Berglund, E. *et al. Nat Commun* **9**, 2419 (2018).
5. Lundmark, A. *et al. Sci Rep* **8**, 9370 (2018).
6. Salmen, F. *et al. bioRxiv* 358937 (2018). doi:10.1101/358937
7. Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. & Lundeberg, J. *Cancer Res.* **78**, 5970–5979 (2018).
8. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. *bioRxiv* 300681 (2018). doi:10.1101/300681
9. Knauf, J. A. *et al. Oncogene* **30**, 3153–3162 (2011).
10. Coclet, J., Lamy, F., Rickaert, F., Dumont, J. E. & Roger, P. P. *Mol. Cell. Endocrinol.* **76**, 135–148 (1991).
11. Lovén, J. *et al. Cell* **151**, 476–482 (2012).

Acknowledgments

We thank Annelie Mollbrink and Jose Navarro for help with the ST protocol and bioinformatics. This work was supported by the Belgian Cancer Foundation, the Naets Foundation and 'Les Amis de l'Institut Bordet'. M.S. is supported by FNRS, J.R. by the Fonds National de la Recherche, Luxembourg (#11587122).

Author contributions

M.S. performed experiments with support of J.L.. J.R.V. and V.D. performed computational analysis, L.C., A.S. and D.L. handled sample banking and pathology review, G.A. resected the tumor. C.M. and V.D. supervised the research and wrote the manuscript.

Competing interests

No competing interest.