Review of: "Precise identification of cancer cells from allelic imbalances in single cell transcriptomes"

Harmen van de Werken[1]

1 Erasmus MC

Potential competing interests: The author(s) declared that no potential competing interests exist.

Precise identification of cancer cells from allelic imbalances in single cell transcriptomes

As tumors consist of cancer but also non-cancer cells such as immune cells, stroma cells and endothelial cells, one of the challenges in single cell sequencing RNA from tumor cells is to make a distinction between cancer cells and non-cancer cells. Trinh and colleagues claim that gene expression values or specific marker gene expression levels will not always distinguish cancer from non-cancer cells. Therefore, they infer cancer vs. non-cancer cells by assessing copy number variations (CNVs) in single cell transcriptome data from allelic imbalances instead of shifts in average expression in two different cancer types: Renal Cell Carcinoma (RCC) and neuroblastoma. Although both methods are capable of detection CNVs, they show that gene expression alone will result in many false positive estimations of CNVs. This is particularly shown for normal tissues which are not expected to harbour CNVs. This paper serves as a great example for other researcher who wish to improve on current methods to distinguish cancerous cells from normal cell when analysing single cell transcriptome data. Although, the amount of work is limited, it shows the potential of using allelic imbalances to assess if cells are non or true cancer cells using single cell RNAseq. I would like to compliment the authors on their interesting work. Nevertheless, I have some remarks.

Minor issues.

1. The manuscript could be clearer if the authors state, in the beginning, that they developed a new algorithm which is implemented in R called alleleIntegrator which they compare to the current standard CopyKAT algorithm.
2. How does alleleIntegrator coop with homozygotic deletions and with epigenetic changes? Both cannot be inferred with their method but could be indirectly inferred from the expression profiles. Could the authors, at least discuss, both changes in their manuscript.
3. This paper does not have standard format for publication. For example, there are no sections with headers and the methods and materials section is added as a supplementary. Perhaps, the authors could work on an extended version of this paper.
4. In the 5th section they describe Figure 1 and the number of reads needed to infer genomic variations, however the explanations are short and therefore the numbers mentioned are not clear. Please use more words to make your concept clear. In addition, it is not clear what is mentioned with “(range 0 to 260)”.
5. AlleleIntegrator is explained in more detail in de supplementary methods, however a short explanation of the method

before going into the results will improve the readability of the manuscript.

6. Can the authors suggest the minimum coverage needed for a single cell sequencing experiment to make CNV inference possible based on allelic imbalance in the way that they proposed?
7. What is the genomic window (bp) for which allelic imbalances were called?
8. How do doublets affect the estimation of CNVs? Which method is better able to deal with doublets? Or would both methods suffer (average shift in expression or allelic imbalance)?
9. Please add link to the github directory of alleleIntegrator.
10. Please change the names in figure 2 from “CopyKat” and “Allele Integrator” to “CopyKAT” and “alleleIntegrator”.
11. Please increase the size of figure 1b and 1C vs 1a and 1D as these are important results from the paper.