Dear Editor,

We wish to submit a manuscript titled “Tradict enables low-dimensional sampling and high fidelity reconstruction of the eukaryotic transcriptome” for consideration as an Article at Nature Methods. The transcriptome, the intermediary between DNA and protein, represents a critical node of regulation for all life. It is therefore no surprise that RNA-Sequencing has revolutionized basic and applied biological research by affording quantitative insight into how an organism or cell is growing, developing, and responding to its environment. Finding ways to obtain equivalent information at higher throughput, but lower cost is a high-priority, ongoing effort that will benefit all those in the biological and biomedical sciences, basic and applied, alike.

Toward this end, we develop Tradict (transcriptome predict) a novel, robust-to-noise, and probabilistically sound algorithm for inferring the transcriptome using only the expression measurements of a single, context-independent, machine-learned subset of 100 marker genes (~0.05% of the transcriptome). Using a representative sampling of over 23,000 *Arabidopsis thaliana* and *Mus musculus* transcriptomes that we assembled from publicly available sources, we train Tradict to prospectively reconstruct gene expression, and to predict, with a high degree of accuracy, the expression of a comprehensive, but quickly interpretable collection of transcriptional programs that represent the major biological processes and pathways of the cell. These capacities, make Tradict the first method to:

1) propose and use a novel, large-scale data model capable of directly modeling the non-negative outputs of sequencing-based expression measurement assays -- the current state-of-the-art.

2) learn, by virtue of the size and comprehensiveness of its training dataset, a marker panel that can be used independently of most (if not all) contexts and applications.

3) define and accurately model the expression of a comprehensive, but interpretable list of a few hundred transcriptional programs in a supervised manner.

The latter point is, in our view, especially important. It suggests that Tradict not only enables cheap and highly scalable transcriptome-wide screening/high-throughput profiling, but also simultaneously affords readily interpretable mechanistic insight that monitoring a single phenotype cannot. Consequently, this unique coupling should greatly facilitate genetic dissection (e.g. screening, breeding, QTL mapping) and drug discovery (e.g. finding the mode-of-action of a small molecule at the time of screening). When compared to previous methods, which do not model the expression of transcriptional programs, do not work transcriptome-wide, and are for use on microarrays, we believe that Tradict fills a large, rapidly growing niche and therefore has great potential to be widely adopted.

Warmly,

Surojit Biswas

As potential reviewers, we would like to request: 1) Dana Pe’er (Dept. of Biological Sciences, Columbia University, [dpeer@biology.columbia.edu](mailto:dpeer@biology.columbia.edu)), 2) Jennifer Listgarten (Microsoft Research New England, [jennl@microsoft.com](mailto:jennl@microsoft.com)), 3) Oliver Stegle (EMBL-EBI, [stegle@ebi.ac.uk](mailto:stegle@ebi.ac.uk))