Dear Editor,

We wish to submit a manuscript titled "Tradict enables high fidelity reconstruction of the eukaryotic transcriptome from 100 marker genes" for your consideration. The transcriptome, the intermediary between DNA and protein, represents a critical node of regulation for all life, and consequently, the ability survey the entire transcriptome through RNA-seq is revolutionizing our understanding of how cells and organisms grow, develop, and respond to the environment. The current required effort and cost entailed in generating read counts for every (or most) transcript are limiting for scaling transcriptome analyses.

We developed a method we call **Tradict (transcriptome predict).** Tradict is 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). Tradict was trained using a representative sampling of over 23,000 *Arabidopsis thaliana* and *Mus musculus* RNA-Seq datasets 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. To our knowledge, Tradict is 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 scalable transcriptome-wide screening/high-throughput profiling, but also simultaneously affords readily interpretable mechanistic insight that monitoring a single phenotype cannot. This unique coupling should greatly facilitate genetic dissection (e.g. forward genetic screening, breeding, QTL mapping) and drug discovery (e.g. narrowing in on the mode-of-action of a small molecule during screening itself). We believe Tradict compares favorably with previous studies in this area as these have not modeled the expression of transcriptional programs, do not work transcriptome-wide, and have been based on increasingly obsolescent technology (microarrays)[1,2]. We further provide easy-to-use software that users can use to 1) build their own transcriptome databases from personal/custom or publicly available sources and 2) train and apply Tradict for their own applications. Taken together, we suggest that Tradict offers unparalleled advantages in large-scale transcriptomics, and therefore has the potential to be a rapidly disseminated, breakthrough technology.

Yours sincerely,

Surojit Biswas[a]
surojitbiswas@g.harvard.edu

Philip A. Wigge[b]
Philip.Wigge@slcu.cam.ac.uk

References

1. Donner, Y., Feng, T., Benoist, C. & Koller, D. Imputing gene expression from selectively reduced probe sets. *Nat. Methods* **9,** (2012).
2. Ling, M. H. T. & Poh, C. L. A predictor for predicting Escherichia coli transcriptome and the effects of gene perturbations. *BMC Bioinformatics* **15,** 140 (2014).

As potential reviewers, we would like to request: 1) Dana Pe'er (Comp. Sys. Bio; Dept. of Biological Sciences, Columbia University, dpeer@biology.columbia.edu), 2) Joeseph Ecker (Agricultural genomics/breeding; HHMI & Salk Institute, ecker@salk.edu) 3) John Marioni (Comp. & Evo. genomics; EMBL-EBI, marioni@ebi.ac.uk) 4) Jennifer Listgarten (Machine Learning in Comp Bio.; Microsoft Research New England, jennl@microsoft.com), 5) Oliver Stegle (Statistical Genomics; EMBL-EBI, stegle@ebi.ac.uk)

[a]Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck St. 4th floor, Boston, MA 02115.
[b]Sainsbury Laboratory, University of Cambridge, Bateman St, Cambridge, CB2 1LR, UK.