Dear Dr. Bahcall,

We would like to submit our manuscript entitled “XXX” for consideration for publication in Nature. This paper will serve as a companion encyclopedia for ENCODE project with a specific focus on Cancer research, dubbed as ENCODEC. We would like to summarized several key points in the paper.

Around 70 percent of the ENCODE cell lines are cancerous, with the top-tier ones enriched with hundreds of experimental assays. Hence, we first organized these cell lines into tumor-normal pairings for five major cancer types. After large-scale data integration, we created our ENCODEC resource, including 1) uniformly processed signal tracks; 2) accurate non-coding annotations; 3) precise regulatory region and gene linkages; 4) TF and RBP networks. The accuracy of the resource we released here is far beyond any previous efforts because we integrated a serial of really novel assays, such as STARR-seq, Hi-C, RAMPAGE, and eCLIP.

With such resource, we could better dissect the cancer mutational landscape and prioritize key positions on the genome that are associated with tumorigenesis. Specifically,

• We could synthesize signals from hundreds of assays to jointly estimate the background mutation rate at an unprecedented accuracy. We further reconcile comprehensive regulatory elements with coding regions for an overall mutation burden evaluation and discovered novel candidate cancer drivers genes with significant prognostic value. We also prioritized key regulatory regions with higher than expected mutations.

• We collected hundreds of ChIP-seq experiments to set up cell-type specific TF networks. We defined the notion rewiring index, in which TF regulatory logic units change in relation to chromatin state, and in response to many mutations. We get a clear sense of what types of TF network changes occur in oncogenesis, and how these changes relate to known events in oncogenesis.

• We reconciled the cell-type specific TF and RBP networks to identify key regulators that drive tumor-to-normal differential expression in 15 cancer types. For the discovered key regulators, such as MYC, we demonstrated how our network information could help to reveal how it works with other regulators in multiple cancer types.

Altogether, we prioritized our prioritized key regulators, high-impact regulatory regions, and deleterious SNVs therein through a walk-down scheme by considering the regulation powers, hierarchies in networks, regulatory changes, mutational burdens, and local context effect. We further validated them separately by different small-scale assays. Particularly, we confirmed the role of MYC and SUB1 to up-regulate their targets in breast and lung cancers by knockdown experiments. We validated the enhancer activities through luciferase assays and demonstrated the effect of our selected SNV in these regions by introducing mutations into wild-type sequences.

Finally, we consolidated our annotations, pipelines, analysis results, and experimental validations into this comprehensive ENCODEC resource. This work constitutes a powerful resource that fully utilizes the richness of deep sequencing functional assays in top-tier ENOCDE cell lines, and it can be easily extended to analyze a variety of other cancer types. Furthermore, we have reconciled ENCODEC annotations with the main ENCODE encyclopedia resource for easy retrieval. Our work shed light on the path to achieve better cancer genomes interpretation by incorporating high dimensional function assays in a large-scale. We believe this would be of great interest to the general readership of Nature.

We would like to suggest the following investigators as potential referees:

* A
* B
* C

Thank you for your consideration.

Sincerely yours,

XXX