# Proposed Revision for the ENCODEC paper

# A. General thoughts on the place of ENCODEC in the package

In your letter, you write: *“the referees mentioned that the current manuscript … does not report separate new datasets, methods, or clear novel findings. Some referees also recommended that this may be more suitable as Perspective in a specialized journal.”*

To be clear, in previous emails, we have made explicit that (1) this paper is to be considered as a "resource" paper, not a novel biology paper and (2) that the current Encyclopedia package is not meant to be structured like previous packages (ie '12 ENCODE). The integrative analysis is meant to be spread over a number of papers and not centered on a single one. (Also, note that the ENCODE 3 "data" is not explicitly tied to any paper. Unlike previous roll-outs, ENCODE 3 does not associate particular data sets with specific papers and make use of these data contingent on that paper's publication. This is codified in an agreement with NHGRI.)

Therefore, it is reasonable that all the papers in the encyclopedia package highlight different aspects of ENCODE and this is what the project has strived to do in the Encyclopedia package. In this section, we describe how ENCODEC fits uniquely into the package. To help make this clear, we have created the grid below.

|  |  |  |
| --- | --- | --- |
| Paper | Main data used | Analysis Focus |
| Moore et al | o DHS & Histone ChIP-seq  o 2 TF ChIP-seq (CTCF & PolII) | o Broad Annotation (enhancer- and promoter- like)  o Sub-classification of annotations & GWAS application |
| Breschi et al | o RNA-seq  o Histone (a little bit) | o Extraction of major cell types from RNA-seq  o Identify cellular composition |
| Nostrand et al | o eCLiP, RBP knockdown  o RNA Bind-n-Seq | o RBP binding sites & functions  o Motif analysis |
| Partridge et al | o 208 TFs in HepG2  o DNA methylation in HepG2 | o Motif analysis  o TF co-localization |
| Zhang et al | o ChIP-seq, DHS  o STARR-seq/ Rep. timing/ Hi-C/ TF/RBP knockdown | o Integrative annotation for data rich cell lines & application to cancer  o TF/RBP networks & their change |

ENCODEC is unique in its highlighting a number of ENCODE assays (e.g. replication timing, TF knockdowns, STARR-seq and Hi-C), its deep, integrative annotations combining a wide variety of assays in specific cell types, and its analysis of networks. None of the other papers include these aspects. Note also that while we do NOT feel ENCODEC is a cancer genomics paper, we feel that cancer is the best application to illustrate certain key aspects of ENCODE data and analysis - particularly deep annotations and network change.

Some more detail on this points:

1) Networks. These are a core aspect of ENCODE that was featured in the 2012 roll out. None of the other papers feature networks in the current package. In ENCODEC, in addition to looking at universal ChIP-Seq networks, we also look at network changes ("rewiring") for specific cell types. We feel that this is best exemplified in oncogenesis.

2) Deep, integrative annotation – different & complementary to the Encyclopedia. While the encyclopedia paper considers broad, "universal" annotations across cell-types (currently the centerpiece of ENCODE), it focuses the data common to most cell types (DHS, 2 histone marks and 2 TFs). It does not take advantage of the cell types much richer in assays -- the other dimension of ENCODE (diagrammed in the paper's first figure). The ENCODEC paper takes a complementary approach, constructing a more accurate annotation using a large battery of histone marks (>10), next generation assays such as STARR-seq and elements linked by ChIA-PET and Hi-C.

2) Replication Timing. Although a major feature of ENCODE is replication timing, none of the other papers use it. Previous work on mutation burden calculation usually selects replication timing data from the HeLa cell line due to the limited data availability. The wealth of the ENCODE replication timing data greatly helps to parametrize somatic mutation rates.

3) SVs. One unappreciated aspect of ENCODE is that next-generation assays, in addition to characterizing functional elements in the genome, enable one to determine structural variants.

4) Knockdowns. ENCODE has 77 CRISPRi based TF knockout and and 533 shRNA based TF/RBP knockdown experiments.

# B. Substantial revisions planned highlighting the paper's ENCODE role

In this revision, we very much take the editors’ comments to heart and are trying to do a large-scale revision of the paper, adding in new elements to make it stronger, while still keeping its original perspective. Here is an overview of what is planned:

* 2a) Highlight more SVs & their effect on the regulatory network

We have referenced structural variations in the earlier version of the paper but admittedly did not highlighted them enough. Since ENCODE provides novel SV data and inclusion of SV analysis was suggested by some of the referees, we will greatly expanded our analysis of SVs in the context of cancer. We will include some new figures as well as add a variety of new data sets that have been designed for this purpose (eg some new genome sequencing data sets for a number of the main ENCODE cell lines). These were not previously referenced in our initial submission, and are specifically generated for our revision.

* 2b) Highlight more the ENCODE STARR-seq data sets:

We will add and highlight some additional STARR-seq data sets -- in particular, STARR-seq in prostate cancer (LNCaP) to relate to SVs.

* 2c) Adding in further aspects of network analysis, with expanded use of the TF knockout experiments

We will look for network similarity between cell types on a global scale and use the robust statistics in our rewiring analysis incorporated with more TF knockdown data as validation. We are expanding our network analysis and using both external and internal (ENCODE) RNA-seq data after TF/RBP knockdowns for validation. We will further investigate the network perturbations after disruption of key regulators.[[rewrite!!]]

# C. How we will address the BMR criticisms







Note that the BMR calculations in the original manuscripts only occupies two sub figures of six total figures, but received the overwhelming amout of the technical criticism. For example, we have 42 itemized technical comments from all referees (in addition to 20 on overall presentation), and 19 of them are all focused on the BMR calculation part.Following the very helpful discussion with the editor, we will acknowledge criticisms of these calculations in terms of their novelty and how they are placed in context. In connection with this, we want point out that the key reference (Martincorena et al 2017) that many of the referees cited appeared in November 2017 -- two months after our paper was submitted (!). Thus, it would not have been reasonable for us to cite. However, we admit that it does share some of the methodology in ENCODEC. We think this bolsters us from a technical standpoint, but admittedly means our calculations are less novel. Note, however, our goal is not to showcase a novel computational method of driver detection (which is the case in Martincorena et al 2017), but rather to demonstrate how a wide range of replication timing and other data give rise to more accurate estimates of mutation rate. We will try to make this clearer in our revision.

# 4. How we will address other technical considerations

These concern things like the power analysis and various statistical tests on our networks. We are optimistic that will be able to address all of these easily. We will rebut these in a point by point fashion and incorporate additional calculations and sub figures as necessary.