Dear Orli,

In my last letter to you, I addressed the main points from your phone call. Here I would like to propose a specific revision plan for the ENCODEC (ENCODE and Cancer) paper. We have certainly taken your comments to heart and will be doing a major revision in a short timescale. Your opinion on our proposed outline for revision will be helpful in achieving this. I realize that you are very busy but any comments would be appreciated.

cheers, marK

=================================================================

1. Points related to the data associated with this paper.

Unlike previous roll-outs, ***ENCODE 3 does not associate specific data sets with specific papers***. In addition, there are no dependencies between any of the papers in this package. All the ENCODE data is open to the public and is not associated with, for instance, the encyclopedia paper or a particular companion paper. It is also reasonable that all the papers in the encyclopedia package focus on different aspects of ENCODE (discussed in the next section).

Note that the data used is codified in an agreement with NHGRI and we have made explicit that ***this package is not meant to be structured like previous packages***. Moreover, the integrative analysis is also meant to be spread over many papers such that the encyclopedia package is not perceived as the main integrative analysis paper.

1. The role of the ENCODEC paper in the package.

In this section, we describe how the ENCODEC paper fits uniquely into the encyclopedia package. We think it is entirely justified for Nature to ask how each of the papers in the encyclopedia package fit together and cover a distinct part of the overall ENCODE Portfolio. To make the relationship between the ENCODEC paper and the other papers clear, we have created a grid, shown below. As you can see, the ENCODEC Paper is unique ***in its inclusion of replication timing data, STARR-seq and Hi-C data, rich annotations, and network information***. None of the other papers include these aspects. We will go over each of these aspects below:

Table 1. Summary of data usage in each paper. The red part is unique to the ENCODEC paper

|  |  |  |
| --- | --- | --- |
| Paper | Main data used | Focus |
| Encyclopedia | * DHS * TF ChIP-seq (CTCF and PolII) * Histone ChIP-seq | * Annotation (enhancer like and promoter like) |
| RNA-seq | * RNA-seq * Histone (a little bit) | * Extraction of major cell types from RNA-seq data |
| RBP | * eCLiP * RNA Bind-n-Seq | * RBP binding sites * Motif analysis * RBP functions |
| TF | * 208 TFs in HepG2 * DNA methylation in HepG2 | * Motif analysis * TF co-localization |
| ENCODEC | * TF ChIP-seq (1000+) * Histone ChIP-seq * STARR-seq * Hi-C * ChIA-PET * Replication timing * TF/RBP knockdown | * Deep and accurate Annotation for data rich cancer cell lines * Broad Annotation for more than 20 cancer types * Meta TF/RBP networks * Tissue specific TF/RBP network |

* 2a) Networks

Network is a core aspect of ENCODE that was featured highly in the 2012 roll out. None of the other papers feature networks in the current package. In ENCODEC, in addition to looking at universal (not cell type specific) ChIP-Seq networks, we also look at network changes on a large-scale, tissue-specific manner. We feel that the rewiring of networks is best exemplified in cancer cells.

* 2b) Compact annotation

While the encyclopedia paper considers annotations across cell-types (currently the center-piece of ENCODE), it does not take advantage of the cell lines rich in data. The ENCODEC paper takes a complementary approach by constructing cell-type specific annotations from cell lines rich in assay data. These annotations are important in power calculations related to recurrent mutations. This highly accurate annotation takes advantage of next generation assays such as STARR-seq and elements linked by ChIA-PET and Hi-C. This is not possible obviously in the general and co-annotation but it's extremely useful on the cancer context.

* 2c) 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 selected replication timing data from HeLa cell line due to the limited amount of data available. The wealth of the ENCODE replication timing data will greatly help to parametrize somatic mutation rates. We will highlight this in our revised manuscript.

* 2d) Structural variations

One unappreciated aspect of ENCODE is that next generation functional assays, in addition to characterizing functional elements in the genome, enable one to determine structural variants. This has been the case for the Hi-C experiments, but there are many other experiments done by experimentalists that have given rise to a large number of structural variants. These structural variations of course are most applicable to the cancer cell lines that many of the ENCODE assays have been run on. We have referenced these structural variations in the earlier version of the paper but admittedly have not really highlighted them or talked about them as much. Since ENCODE provides novel SV data and inclusion of SV analysis was suggested by some of the referees, we have 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 specifically for this project.

* 2e) TF/RBP knockdown/knockout experiments

ENCODE has 77 CRISPRi based TF knockout and and 533 shRNA based TF/RBP knockdown experiments, which serves a great resource to investigate network perturbations after disruption of a regulator. The ENCODEc paper is the only paper that focuses on such data. In our current manuscript, we have already used some of such knock down data to validate effects of key regulations in multiple cancer types. We will highlight the usage of such experiments in our revised version.

1. Large scale revision and new data sets

In this revision, we very much take in 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 its original focus.

* ***3a) Structural variants:***

We are adding in some new genome sequencing data sets from a number of the main ENCODE cell lines. These were not previously referenced in our initial submission, and are specifically generated for our revision. We will include some new analysis of SV detection in cancer and put them in new sub-figures.

* ***3b) TF knockout experiments:***

We are expanding our network analysis and using both external and internal (ENCODE) RNA-seq data after TF/RBP knockdowns for validations. We will further investigate the network perturbations after disruption of key regulations.

* ***3c) More STARR-seq data sets:***

We will also add and highlight some additional STARR-seq data sets. In particular, we will take the read out of STARR-seq in prostate cancer (LNCaP) to relate to SVs.

1. BMR calculation.

The BMR calculations in the original manuscripts only occupy two sub figures of six total figures, but received the most criticism (~30%). We tried to document this in the table below. I originally suggested on the phone that we simply cut these out, but at the end of the conversation you felt that was not a good idea. Hence, we propose the following for the BMR calculations.

Table 2. Summary of main comments in the referee report

|  |  |
| --- | --- |
| ***Topic*** | ***Count*** |
| BMR | 19 |
| Power | 4 |
| Presentation | 20 |
| Annotation | 3 |
| Network | 6 |
| Hierarchy | 1 |
| CellLine | 3 |
| Stemness | 3 |
| Validation | 3 |

We will acknowledge criticisms of these calculations in terms of their novelty and how they're placed in context. We will certainly place them correctly in context and knowledge and all the other papers. 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 and it would not have been reasonable for us to cite this paper. However, we admit that it does share some of the methodology in our paper.

We think this only bolsters us from a technical standpoint, but admittedly means our calculations are less novel. However, our goal in using this methodology is not to showcase a novel method of driver detection or model for background mutation rate, but rather to demonstrate how a large range of replication data and other covariates give rise to more accurate estimates of mutation rate. We will try to showcase this more in our revision.

1. Other technical considerations.

These concern things like the power analysis and various burden test. We will rebut these in a point by point fashion and incorporate additional calculations and sub figures as necessary.

* ***5a) Format of the supplement in the paper.*** We will address this from a presentation perspective, and adjust accordingly.