# ENCODE Cancer paper

Pazin comments

23 March 2017

High level points:

Interesting idea, interesting work.

With respect to your first question, I think the supplement does a good job explaining the tradeoffs you used to select your pairs. The rationale for K562 with GM12878, based on the amount of available data, the neighboring lineage relationship, and the empirical comparison, are reasonable to me. The MCF7/MCF10 pairing is also well explained, though I think (but don’t know) there are more representative papers to cite on the historical use of this pairing. I understand your case for the A549/IMR90 pairing, yet I am concerned that you will conflate cancer and cell fate with this pairing. The HepG2/liver pairing is straightforward. I have suggested some ideas for HeLa pairing.

With respect to your second question, it seemed clear to me what products you are proposing to generate and share (enhancer calls, enhancer/promoter connectivity, TFs and RBPs predicted to regulate tumorigenesis in particular T/N pairs, TF and RBP regulatory networks from the cell types used). It was not clear where the community would go to find them, and whether they would understand the relationship between them and other data/analyses at the portal, including the ENCODE encyclopedia.

It was unclear what the relationship is, if any, between this work and Shirley Liu’s. Given that you are both part of ENCODE and part of the same DAC award, some explanation of how the present work compares to PMID: 26056275 and PMID: 26518695 would seem to make sense.

It was unclear what the relationship is between this work and the ENCODE encyclopedia and portal. It was also unclear to me whether the DCC, DAC, and other ENCODE groups know that this effort is taking place.

I realize the philosophy of the manuscript is tumor-normal pairs, focused on cell types with as much ENCODE data as possible. I found it more natural to think of what ENCODE cell types might be relevant to the limiting cases of tumor and normal, and in some cases I thought multiple comparisons were interesting, and in some cases that led to cell types that may not have as much data as you are looking for, or the data types you are looking for. If you prefer the TN pair paradigm starting with ENCODE cell types with lots of assays, stay with it, but be aware of other potential comparisons suggested below. PLEASE don’t refer to everything in ENCODE as cell lines (I think this reinforces a negative stereotype some people have about ENCODE); as you can see from the attached screenshot <ENCODE BIosample Categories.pdf>, we have lots of biosamples that are not cell lines. I have also attached a list of some of the biosamples with the most ENCODE data < Cell Types With Many ENCODE Assays.pdf>; both of these screenshots are derived from the ENCODE Data Summary I share quarterly <ENCODE\_Human\_Released\_2017\_02\_03v4.xlsx>. I also added some comments and edits to the files you sent, <EC.Mar.16.2017\_MP.docx> and <17\_03\_23 ENCODE\_Cancer\_Supplementary\_Text\_MP.pdf>.

Details:

1) Given this paper, especially figure 6b, but also figure 6, 7, S6:

<https://www.ncbi.nlm.nih.gov/pubmed/23953118>

I’d be curious about how well the tumor cells cluster with ES cells, as opposed to the mature cell type they appear to be derived from. This fits with an old idea, that cells may de-differentiate during tumorigenesis. We have 131 experiments in H1 ESC, about the same number of experiments as in HeLa. We also have data on other human ESC.

2) I am most concerned about the A549/IMR-90 pairing, because of the epithelial v. mesenchymal/fibroblast comparison. Your comparison across two dimensions (tumor/normal and epithelial/fibroblast) might conflate cell fate with transformed state. I think most human cancers are carcinomas, derived from epithelial cells. Human cancers derived from mesenchymal cells (sarcomas) are less common, and I think their biology is somewhat different. I know in skin the gene expression pattern of epithelial cells (keratinocytes) is quite different from the mesenchymal cells (dermal fibroblasts) and more similar to other epithelial cells, I think Tom G’s ENCODE gene expression clustering is consistent with this idea, and I think this is a common theme in epithelial and mesenchymal biology. Having said this, some cells from epithelial tumors undergo EMT (epithelial to mesenchymal transition), become circulating tumor cells, locate in new sites, then revert back to a more epithelial phenotype as metastases.

ENCODE has some data from untransformed lung epithelial cells:

Primary bronchial epithelial cells:

https://www.encodeproject.org/search/?searchTerm=bronchial+epithelial&type=Experiment&award.project=ENCODE

Primary alveolar epithelial cells:

https://www.encodeproject.org/search/?searchTerm=lung+epithelial&type=Experiment

There is also EN-TEx data from lung:

https://www.encodeproject.org/search/?searchTerm=GTEx&type=Experiment&biosample\_type=tissue&y.limit=&biosample\_term\_name=upper+lobe+of+left+lung

ENCODE has a lot of data from untransformed epithelial cells from other organs:

Primary skin epithelial cells (keratinocytes):

https://www.encodeproject.org/reference-epigenomes/ENCSR193SZM/

https://www.encodeproject.org/search/?searchTerm=keratinocyte&type=Experiment&award.rfa=ENCODE2&award.rfa=ENCODE3

REMC keratinocytes:

https://www.encodeproject.org/search/?searchTerm=keratinocyte&type=Experiment&award.rfa=Roadmap

Primary breast epithelial cells:

ENCODE HMEC:

https://www.encodeproject.org/search/?searchTerm=HMEC&type=Experiment&award.rfa=ENCODE2&award.rfa=ENCODE3

REMC HMEC:

https://www.encodeproject.org/search/?searchTerm=HMEC&type=Experiment&award.rfa=Roadmap

3) I am next most concerned about the K562/GM12878, because of the erythro/myeloid v. lymphoid comparison and because they are both transformed. Myeloid and lymphoid leukemias are somewhat different. Your comparison across two dimensions (tumor/”normal” and myeloid/lymphoid) might conflate cell fate with transformed state (similar to the A549/IMR90 comparison). I am also concerned because the manuscript appears to describe K562 as a model for CLL, yet K562 are derived from a CML patient and are myeloid, not lymphoid. CML and CLL appear to respond to different therapies. CD34 cells are an interesting alternative, as they are hematopoietic stem cells, thus they are somewhat “de-differentiated” with respect to a common myeloid progenitor cell (with which K562 shares some properties).

4) For breast cancer, one could consider MCF7 and T47D (transformed) v. MCF10 (on the way) v. HMEC (untransformed) v. ESC. (I don’t understand whether MCF7, from a patient with an adenocarcinoma, is a better model for ductal or glandular carcinoma.) I understand the historical use of MCF7 v. MCF10 comparisons, and agree that is one reason the comparison has value.

Primary breast epithelial cells:

ENCODE HMEC:

https://www.encodeproject.org/search/?searchTerm=HMEC&type=Experiment&award.rfa=ENCODE2&award.rfa=ENCODE3

REMC HMEC:

https://www.encodeproject.org/search/?searchTerm=HMEC&type=Experiment&award.rfa=Roadmap

T47D Mammary Tumor cell line (ductal carcinoma; most common breast cancer?):

https://www.encodeproject.org/matrix/?searchTerm=T47D&type=Experiment

There is also EN-TEx data:

https://www.encodeproject.org/search/?searchTerm=GTEx&type=Experiment&biosample\_type=tissue&y.limit=&biosample\_term\_name=breast+epithelium

5) For HepG2, ENCODE has data from hepatocytes differentiated in culture from H9 ESC:

https://www.encodeproject.org/search/?searchTerm=hepatocyte&type=Experiment&replicates.library.biosample.donor.organism.scientific\_name=Homo+sapiens&biosample\_type=in+vitro+differentiated+cells&limit=all

there is also EN-TEx liver data:

https://www.encodeproject.org/search/?searchTerm=GTEx&type=Experiment&biosample\_type=tissue&y.limit=&biosample\_term\_name=right+lobe+of+liver

We also have H1 and H9 ESC data, if that is a useful comparison to you.

6) For HeLa, one could compare it to any primary epithelial cell type, or perhaps a few epithelial cell types. Of course it is thought that HPV integration in the HeLa genome was a critical event in transformation (PMID: 3028716, PMID: 23925245), so viral TF, protein coding genes, and regulatory DNA may play a role in addition to host genes.