# An integrative ENCODE resource for cancer: interpreting non-coding mutations and gene regulation

Most somatic mutations in cancer are non-coding while the characterized drivers are predominantly located in coding regions, creating a conundrum as to whether non-coding regions are important for oncogenesis. Here we endeavor to address this issue through creating a companion resource to the main ENCODE encyclopedia. In particular, we integrate diverse ENCODE data to precisely calibrate background mutation rates. We utilize advanced functional-genomic assays, especially STARR-seq and Hi-C, to develop compact annotations and accurate, extended gene models (linking enhancers to coding regions), allowing us to achieve better statistical power for mutational burden analysis. We also construct detailed regulatory networks to interpret tumor gene expression and mutation profiles, pinpointing effects of key regulators such as the transcription-factor MYC and the RNA-binding-protein SUB1 and then validating them. We build cell-type specific networks to directly measure the degree of "rewiring" during oncogenesis, classifying changes as either moving toward or away from a stem-like state. Finally, we use our overall resource -- comprising the compact annotations, networks, and burdened regions -- to prioritize non-coding elements and their mutations, and we validate a subset of them through targeted experiments.

[JZ2everyone: I would like to use real numbers to say the scale of the integration and will update once DL has these numbers. Also, I would like to make it clear at the very beginning what does ENCODEC has]

# Title A: An integrative ENCODE companion resource to interpret cancer genome

# Title B: Dissecting cancer genomes by integrating ENCODE data

[JZ2everyone: I like the 2nd title because it sounds more powerful. I don’t want to emphasize even in the title that this is a resource. I believe that this would go way too far… it is easy to get the comments like “OK why not merge it into the main encyclopedia paper”, it is a double-edged sword]. I think to emphasize ENCODE itself is enough…

################## new abstract #####################

The new release of ENCODE datasets contains thousands of functional characterization data from numerous cancerous cells and are thus valuable to dissect cancer genomes. Here, we endeavor to tailor them into a targeted resource, called ENCODEC, by providing deep annotations and high-confident regulatory networks. Specifically, by integrating dozens of epigenetic features with novel assays, such as STARR-seq , ENCODEC significantly refines and compactifies generic annotations to maximize the computational power for mutation recurrence analysis. The integration of 475 ENCODE signal data, particularly that for replication timing, further allows us to build more accurate background mutation rates models than previous methods for reliable burdened region detection. Furthermore, ENCODEC synthesizes other novel assays, such as Hi-C and RNA-binding protein assays (i.e., eCLIP), in addition to large-scale TF ChIP-seq, to build cancer-specific regulatory networks to investigate regulatory target rewiring during oncogenesis and compare them with STEM-like state. More generally, ENCODEC also provides reconciled networks to explain cancer-specific expression patterns and pinpoint key regulators that driver large-scale tumor-to-normal expression changes. Combining these with recurrence analysis, we develop a stepwise prioritization scheme for somatic mutations and demonstrate how this can be instantiated in practice. We further validated the functional consequences of our prioritized variants through luciferase assays in breast cancer.