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

Jing Zhang\*, Donghoon Lee\*, Vineet Dhiman\*, Peng Jiang\*, William Meyerson, Matthew Ung, Shaoke Lou, Patrick Mcgillivray, Declan Clarke, Lucas Lochovsky, Lijia Ma, Grace Yu, Arif Harmanci, Mengting Gu, Koon-kiu Yan, Anurag Sethi, Qin Cao, Daifeng Wang, Gamze Gursoy, Jason Liu, Xiaotong Li, Michael Rutenberg Schoenberg, Joel Rozowsky, Lilly Reich, Juan Carlos Rivera-Mulia, Jie Xu, Jayanth Krishnan, Yanlin Feng, Jessica Adrian, James R Broach, Michael Bolt, Vishnu Dileep, Tingting Liu, Shenglin Mei, Takayo Sasaki, Su Wang, Yanli Wang, Hongbo Yang, Chongzhi Zang, Feng Yue, David M. Gilbert, Michael Snyder, Kevin Yip, Chao Cheng, Robert Klein, X. Shirley Liu, Kevin White, Mark Gerstein

**Abstract**

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 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 functional-genomic assays, especially STARR-seq and Hi-C, to develop compact, accurate and extended annotations of gene models (linking enhancers to coding regions), allowing us to achieve better statistical power for burden analysis. We also constructed detailed regulatory networks to interpret tumor gene expression and mutation profiles, and predicted the effects of key regulators, such as the transcription-factor MYC and the RNA-binding-protein SUB1, in shaping tumor specific gene expression patterns, which were experimentally validated. 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 validated a subset of them through targeted experiments.

2nd generation

recently developed assays