ENCODE now comprises thousands of functional genomics data sets; it is possible to tailor these into a targeted resource for interpreting cancer genomes. In particular, this resource can be used, firstly, to measure the impact of non-coding mutations, constituting the bulk of the somatic variants. Moreover, by integrating advanced assays (e.g. STARR-seq) with many epigenetic features, we can make a more focused and refined genome annotation, increasing the power for detecting recurrent somatic mutations in cohorts. Second, ENCODE signal data, especially replication timing, allows constructing precise, cell-type-matched models for background mutation rates, considerably more accurate than previous models. Third, ENCODE data, incorporating new assays, such as Hi-C and RNA-binding protein assays, in addition to large-scale transcription-factor ChIP-seq, allows the construction of extensive regulatory networks. In some contexts, these networks reveal how connections "rewire" during oncogenesis, as well as how these changes relate to a stem-cell state. More generally, one can use ENCODE networks to prioritize regulators most associated with large-scale expression changes in cancer. Combining the networks with the refined annotations and background mutation models, one can develop a step-wise scheme for prioritizing non-coding mutations. Here, we show how this can be instantiated, and we perform a number of focused, experimental validations (i.e., luciferase assays and shRNA knockdowns) to demonstrate how the resource can highlight mutations with significant consequences in cancer.

==================== old ====================

ENCODE comprises thousands of functional genomics data sets, related to many types of cancer; it is possible to tailor these into a targeted resource for interpreting cancer genomes. In particular, this resource can be used to measure the impact of non-coding mutations, constituting the bulk of the somatic variants. Moreover, by integrating advanced assays (e.g. STARR-seq) with many epigenetic features, we can make a more focused and refined genome annotation, increasing the power for detecting recurrent somatic mutations in cohorts. Second, ENCODE signal data, especially replication timing, allows constructing precise, cancer-matched models for background mutation rates considerably more accurate than previous models. Third, ENCODE data, incorporating new assays, such as Hi-C and RNA-binding protein assays, in addition to large-scale transcription-factor ChIP-seq, allows the construction of extensive regulatory networks. In some contexts, these networks reveal how connections "rewire" during oncogenesis, as well as how these changes relate to a stem-cell state. More generally, one can use ENCODE networks to prioritize regulators most associated with large-scale expression changes in cancer. Combining the networks with the refined annotations and background mutation models, one can develop a step-wise scheme for prioritizing non-coding mutations. Here, we show how this can be instantiated, and we perform a number of small-scale validations (i.e., luciferase assays and siRNA knockdowns) to demonstrate how the resource can highlight mutations with significant consequences in cancer.