**A large scale integrative resource from ENCODE for cancer research**

**Requirement of nature about summary:**

**Articles have a summary, separate from the main text, of up to 150 words, which does not have references, and does not contain numbers, abbreviations, acronyms or measurements unless essential.**

**Long version : 381 words**

[[JZ2MG: I think the current background in the abstract is way too long]]

Cancer is caused by mutations in the DNA which disrupt the normal physiology of cells. While mutations on coding genes have been well characterized, the preponderance of mutations in tumors occur in non-coding regions and are still poorly understood. The new release of the ENCODE data provides an opportunity to bridge these knowledge gaps. For a variety of cancer-derived cell lines, as well as non-cancerous cell lines derived from relevant tissues, ENCODE provides diverse genome-wide assays to assess the potential for non-coding mutations to dysregulate genes.

In this paper, we first developed a regression based model for background mutation rate calibration by removing both internal and external confounding effects. To maximize the statistical power of burden analysis, we then integrated diverse assays for core regulatory elements definition and confined burden tests on a minimum set of high-confidence annotations. To improve functional interpretability, we synthesized multiple evidence to link these regulatory elements to coding genes to define an extended gene neighborhood a whole test unit. This multi-level integrative approach successfully identified novel burdened genes, such as *BCL6* in leukemia, that are associated with patient prognosis.

Besides, we also built up generalized transcription factor (TF) and RNA binding protein (RBP) networks based on thousands of ChIP-seq and eCLIP experiments and formatted them into hierarchies. Intriguingly, we found that the top-level TFs, which tend to regulate other TFs, are enriched with cancer related genes and more significantly drive gene expression. After combining with cancer expression profiles, we pinpointed key regulators such as MYC and SUB1 that significantly drive tumor to normal differential expression and then validated their effects through knockdown experiments.

Furthermore, we built up tissue specific regulatory networks in top tier cell lines to directly measure regulatory changes during tumor transformation and identified highly “rewired” TFs with changed targets and prognostic value, such as IKZF1 and MYC. We also found the massive “rewiring” events in tumor cell lines are often involved in large chromatin and expression changes, but to a less degree of motif gain or loss events.

Finally, we synthesized our results into a companion ENCODE encyclopedia for cancer (EN-CODEC) and proposed a prioritization scheme for key mutations in cancer. We successfully identified active enhancers and seven high impact mutations therein in breast cancer and validated their functional effects through luciferase assays.

**Short version : 243 words**

[[JZ2MG: I think the current background in the abstract is way too long]]

In this paper, we endeavor to collect diverse genome-wide assays in ENCODE to deeply annotation the cancer genome. We first developed background mutation rate calibration model by removing both internal and external confounding effects. To maximize the statistical power, we integrated diverse assays to identify core regulatory elements and confined burden tests only on high-confidence extended gene annotations. Such multi-level integrative approach successfully identified novel burdened genes, such as *BCL6* in leukemia, that are associated with patient prognosis.

Besides, we also built up generalized transcription factor (TF) and RNA binding protein (RBP) networks based on ChIP-seq and eCLIP experiments and placed them into hierarchies. Intriguingly, the top-level TFs are enriched with cancer related genes and more significantly drive gene expression. We also pinpointed key regulators such as MYC and SUB1 that significantly drive tumor-to-normal differential expression and then validated their effects through knockdown experiments.

We further built up tissue specific regulatory networks in top tier cell lines to directly measure regulatory changes during tumor transformation and identified highly “rewired” TFs with changed targets and prognostic value, such as IKZF1 and MYC. We also found the massive “rewiring” events in tumor cell lines are often involved in large chromatin and expression changes, but to a less degree of motif gain or loss effects.

Finally, we synthesized our results EN-CODEC resource and proposed a variant prioritization scheme. We successfully identified seven high impact mutations in breast cancer and validated their functional effects through luciferase assays.

**V1**

Most mutations in cancer occur in the noncoding regions where their impacts are not well characterized. Here we endeavor to create a companion resource to the ENCODE encyclopedia to interpret the noncoding variants.

We first integrated genomic signals to accurate model background mutation rate for burden analysis.  To maximize the statistical power, we built a compact annotation and linked them to coding genes to define extended genes. We identified burdened genes with prognostics values.

We constructed regulatory networks to interpret the expression and mutation profiles of cancer patients. Specifically, we prioritized key regulators, such as MYC and SUB1, and validated them to knock downs experiments. We further directly measured the rewiring status of TFs in cell type specific networks during normal-to-tumor transformation and also categorize it to either moving toward or away stem-like state.

Finally, we integrated these analyses to prioritize key features in cancer and validated them in MCF-7.

**V2**

Most mutations in cancer occur in the noncoding regions, yet most of the characterized drivers in cancer are in coding regions. It is important to investigate whether changes in the noncoding regions are functional in cancer progression. Here, we reformulate the ENCODE data into a companion resource to address this question.

In particular, we integrate the ENCODE data to calibrate precise background mutation rate models and demonstrate how the more advanced essays from ENCODE, particularly STARR-seq and Hi-C, can help to develop a more compact annotation for better statistical power to find burdened regions.

Next, we build ENCODE regulatory networks to pinpoint key regulators during the oncogenic transformation and validate them by knock-down experiments. We also cancer associated cell-type specific networks to measure regulatory changes during oncogenesis and catalog them to either moving toward or away from a stem-like state.

We further integrate the overall ENCODE resource to prioritize noncoding elements and mutations and then to validate them in further small-scale studies.

**V3**

Most of the mutations in cancer genomes occur in the noncoding regions while almost of the characterized driver mutations occur in the coding parts. This creates a central conundrum whether the non-coding regions, the genome, have functionally important mutations during oncogenesis. The ENCODE data set provides a platform for addressing this question. In particular, we develop a companion resource based on this data that addresses this question. First, we show how the ENCODE data can be used to develop accurate models for background mutation rates, to accurately find recurrent mutations in cancer genome mixed datasets. Then we show how the advanced encode assays, particularly the [StarSeek 00:01:27] and the Chia Pet enable us to develop highly accurate definitions of enhancers and the linkage in enhancers to genes to make for a compact annotation that is more powered for finding recurrent regions. Then to link these non-coding regions into extended gene, to which also that are able [to burdening 00:01:56] calculations.

Next, we show how the encode data can be used to build extended regulatory networks in cancer associated cell types and to see how these regulatory networks change between tumor and normal states. We observe in these contexts extensive rewiring of various sub-networks with many of the transcription factors that tend to be losing edges, moving in a more stem cell-like direction. We find that most of these rewiring changes are better attributed to chromatin changes than direct effects of a mutation. Finally, we show how the encode regulatory network can be used to prioritize key transcription factors and also RNA-binding proteins based on gene expression and cancer tissue. We can validate this prioritization with knockout experiments. We further use the encode, our resource to prioritize non-coding elements and mutations and show that these also can be validated in small-scale experiments.

New abstract.

V4 **(149 words)**

Most somatic mutations in cancer are noncoding while the characterized drivers are predominantly located in coding regions, creating a central conundrum whether the noncoding regions are functionally important during oncogenesis. Here we endeavor to create a companion resource to the ENCODE encyclopedia to address this question.

In particular, we integrate the ENCODE data to precisely calibrate background mutation rates and synthesize advanced essays, especially STARR-seq and Hi-C, to develop compact annotations and accurate gene linkages to achieve better statistical power for burden analysis.

We also construct regulatory networks to interpret cancer expression and mutation profiles and pinpoint key regulators such as MYC and SUB1. We build cell-type specific networks to directly measure regulatory changes during oncogenesis and classify them to either moving toward or away from a stem-like state.

Finally, we integrate the overall ENCODE resource to prioritize noncoding elements and mutations and then validate them through small-scale studies.

V5**(150 words)**

Most mutations in cancer occur in the noncoding regions where their impacts are not well characterized. Here we endeavor to create a companion resource to the ENCODE encyclopedia to interpret the noncoding variants.

In particular, we integrate the ENCODE data to precisely calibrate background mutation rates and synthesize advanced essays, especially STARR-seq and Hi-C, to develop compact annotations and accurate gene linkages for better statistical power in burden analysis. As a result, we identified burdened genes with prognostics values. We constructed regulatory networks to interpret cancer expression and mutation profiles. Specifically, we prioritized key regulators, such as MYC and SUB1, and validated them through knock-down experiments. We directly measured the rewiring status of TFs in cell-type specific networks during oncogenesis and categorize them to either moving toward or away from stem-like state. Finally, we integrate the overall resource to prioritize noncoding elements and mutations and validate them through small-scale studies.