News & Views

A typical cancer genome contains thousands of somatic mutations, where the overwhelming majority occupy non-coding regions. However, classical models of cancer posit that only a few of these mutations are under strong positive selection and drive the cancer forward. Currently, almost all of these "driver mutations" have been found in coding regions \cite{24071849,24084849}. Thus, a key question arises, whether there are many driver mutations lurking in non-coding regions of the genome?

Identification of non-coding drivers is challenging due to vastness of the non-coding space and the difficulty in characterizing functional noncoding elements. These issues confound the power to detect non-coding driver mutations in a cancer cohort. In contrast, identifying coding driver is easier: We have a better understanding of the start and endpoint of different coding regions, and the functional impact of mutations in coding region is well defined. For instance, does a mutation leads to change in the coded protein(nonsynonymous/synonymous), or does it completely knock out the protein through a loss-of-function mutation? Potentially, this better understanding of coding regions creates an ascertainment bias and raises the question of whether driver mutations are primarily in coding region or it's just that we don't know where to look for the non-coding drivers.

Despite these challenges, there has been a great interest in finding non-coding drivers \cite{26781813}. Over last few years, several methods have been developed to identify non-coding driver mutations. For instance, previous studies identified recurrent mutations in the TERT promoter for multiple cancer cohorts \cite{23887589}. Similarly, a recurrence based method found driver mutations in upstream regulatory regions of PLEKHS1, WDR74 and SHDH genes in different cancers \cite{25261935}. Furthermore, pan-cancer analysis of copy number aberrations and gene expression data highlighted the role of enhancer hijacking related to IRS4, SMARCA1 and TERT \cite{27869826}. However, these are few examples and at present our understanding of non-coding drivers is incomplete.

On page xxx of this issue, Rheinbay et. al. make a foray towards addressing this question \cite{}. For a cohort of 360 breast cancer patients, they attempt to look for coding and non-coding driver mutations, in an unbiased fashion. In this study, they provide evidence that in case of uniform

ascertainment in a cancer genome, one could find as many noncoding driver mutations as coding ones. Moreover, they predicted that mutations within promoters of *FOXA1, RMRP* and *NEAT1* significantly alter transcription. These findings were further validated using functional assays measuring changes in gene expression and protein binding.

In particular, they predicted driver mutations based on, identifying non-coding elements that a) harbor significantly higher mutation counts relative to expectation, and b) contain clusters of mutations around their regulatory motifs. Furthermore, for driver discovery, they utilized patient-specific background mutation rate, which takes into account of the total mutation frequency and total frequency of bases with sufficient sequencing coverage across all analyzed elements. Moreover, their power analysis indicates that the relatively large cohort size in this study makes it possible to identify driver mutations in promoter regions, which are mutated in at least 10% of patients in the cohort. However, they also show that one would need even larger sample size to identify confidently drivers which are typically present in ~5% of patients in a cohort. Interestingly, their analysis of mutational hotspots indicates that promoters have as many single-site recurrent mutations as coding genes. Furthermore, the per base mutation rate of driver promoters were found to be very similar to that of well-known coding drivers. This further suggest that smaller frequency of relevant promoter mutations in contrast to coding genes can be attributed to their small amount of functional territory (i.e. they simpler occupy less base pairs in the analysis).

This work describes the state-of-the-art in identifying non-coding driver mutations. However, there is still more to be done. To understand the directions for improvement it is necessary to review aspects of the non-coding annotation process and its interplay with power calculations. Currently, due to the way they are determined from processing noisy functional genomics signals, the majority of non-coding elements are fairly large in size (e.g. calling 1kb sized peaks). However, their real functional territory maybe considerably smaller (Fig1). Aggregating mutation recurrence over such large non-coding regions can dilute the true signal of positive selection and hinder driver identification. As shown in figure1, power calculations suggest that restricting annotation to smaller functionally relevant blocks enhances the power. However, accurate definition of these functional territories remains challenging.

Moreover, both coding and non-coding elements (e.g. genes and their regulatory structures) comprise of discontinuous block of functional territories (and this discontinuous nature becomes more apparent as the functional blocks shrink in size). These connections are well understood for coding regions, where multiple exons can be clearly linked through splice junctions into a transcript. In contrast, we lack such clear connections for noncoding regions. For instance, a gene can be connected to non-coding elements such as promoters, enhancers or even the entire gene regulatory subnetwork. Furthermore, same gene can be connected to multiple distal elements or vice versa. One approach to better define non-coding functional territories is to complement functional genomics based definition with the underlying conservation signals. These regions can be conserved regulatory motifs (such as TF binding motifs) and, more generally, ultra-conserved and ultrasensitive sites.

After defining the functional territory of an individual non-coding element, the next step in the driver discovery involves mutation burden testing over many regulatory elements. Thus, lack of specificity in non-coding annotation will increase the number of multiple testing, which will decrease driver detection power. One could increase specificity through removing as much false positives as possible in the annotation set. Thus, overall the best annotation for increasing power for driver detection is a compact and highly accurate annotation set with as few elements as possible, where each element correspond to an underlying functional territory, which potentially links discontinuous functional regions in the non-coding genome (low L and N in figure XXX).

An additional difficulty with identifying non-coding driver mutations is to evaluate their functional impact. Currently, it's unclear whether substitution of each nucleotide in a regulatory region has an equal impact on its function. We can see this for certain among well characterized situations in TF binding sites: For instance, some non-coding mutations are considered more disruptive if they break an existing or generate a new binding motif for transcription factors \cite{24092746}. Nonetheless, better metric of functional impact is needed over the whole genome to find equivalents of synonymous, nonsynonymous and loss-of-function mutations among non-coding variants. Finally, the power to detect drivers in non-coding regions is dependent on the uniformity of the underlying background mutation rate.

However, this rate is far from uniform across the expanses of the genome and is known to co-vary in a complex way with various genomic and epigenomic signals (chromatin state, transcriptional activity and replication timing) \cite{26304545}.

An exhaustive (but expensive) approach to deal with these challenges is sequencing a large number of patients. This approach can be feasible only through large-scale collaborative efforts such as the Pan Cancer Analysis of Whole Genome (PCAWG) project. These efforts will generate comprehensive non-coding variant catalogue, which can be leveraged to detect regulatory mutations with sufficient power. However, these large-scale studies assume to develop a uniform cohort, which can be challenging due to highly heterogeneous cancer samples (e.g. different breast cancer subtypes). An alternative approach will be to develop a more compact functional annotation of the non-coding genome with precise definition of functional territory. In this avenue, large scale annotation compendium such as ENCODE can play a vital role \cite{22955616}.