News & Views

A typical cancer genome contains thousands of mutations, where the majority occupy non-coding regions of the genome. 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 of the genome \cite{24071849,24084849}. However, the majority of somatic mutations are located in noncoding regions of the genome. Thus, the 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 accurately finding functional noncoding elements. These issues confound the power to detect all non-coding driver mutations in a cancer cohort. In contrast, identifying driver mutations in coding regions is more intuitive. We have a better understanding of the start and endpoint of different coding regions. In addition, the molecular impact of mutations in coding region is well defined. For instance, does a mutation leads to change in the coded protein(nonsynonymous/synonymous), or it completely knocks out the protein through a loss-of-function mutation? Our better understanding of coding regions potentially creates an ascertainment bias that is leading to identification of larger number of coding driver mutations. This poses the question, 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 characterizing non-coding drivers in various cancers \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, 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 phenomena in regulatory elements of various genes including IRS4, SMARCA1 an 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 evidences suggesting 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 this study, they predicted driver mutation in regulatory elements based on, identifying non-coding elements that a) harbor significantly higher mutation counts relative to expectation, or 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 relatively large cohort size in this study, make 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 majority of driver mutations which are typically present in 3 to 5% of patients in a cohort. Interestingly, their analysis of mutational hotspots indicate that promoters are among different genomic elements with most single-site recurrent mutations. Furthermore, mutation rate of functionally relevant alterations in promoter was found to be very similar to that of well-known coding drivers. This further suggest that smaller frequency of relevant promoter mutations can be attributed to their lower functional territory length.

This work describes the state-of-the-art in identifying non-coding driver mutations present in larger fraction of patients in a cohort. However, detecting low frequency non-coding drivers remains problematic and thus, mandate a comprehensive understanding of key factors influencing their ascertainment. Although, majority of non-coding elements are large in size but their functionally relevant territory is relatively small (Fig1). Aggregating mutation statistics over such large non-coding regions can dilute signal of positive selection and hinder driver identification. As shown in figure1, power

calculations suggest that restricting functional annotation to relevant regions enhances the power. However, accurate definition of these functional territories remains challenging. Both coding and non-coding elements (e.g. genes and their regulatory structures) comprise of discontinuous block of functional territories. These connections are well understood for coding regions, where sequence reads belonging to multiple exons are 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 conserved regions are in principle similar to regulatory motifs (such as TF binding motifs) in the genome. For instance, conservation based annotation such as small blocks of ultra-conserved non-coding elements and ultrasensitive sites in the genome can be very helpful in identifying such non-coding functional territory.

After defining an individual functional territory in the non-coding region, the next step in the driver discovery involves mutation burden testing over multiple regulatory units. Thus, lack of specificity in non-coding annotation will increase the number of multiple testing, which will influence driver detection power. One could obtain necessary power through removing false positives in the annotation set. Furthermore, accurate definition of annotation will also lead to enrichment of relevant functional territories. This is consistent with power analysis (Fig1), where increasing the annotation frequency (high N) leads to lower power, whereas decreasing the annotation (lower N) leads to increase in the overall power.

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. However, functional consequences of mutations in certain regulatory elements such as transcription factor binding sites is more intuitive. 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 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 their underlying background mutation rate. However, heterogeneities in various genomic and epigenomic signals (GC content, chromatin state, transcriptional activity and replication timing) among non-coding elements confound the proper estimation of the background mutation rate \cite{26304545}.

An exhaustive (but exceedingly 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. An alternative approach will be to develop better functional annotations of the non-coding genome with precise definition of functional motifs. In this avenue, large scale annotation compendium such as ENCODE encyclopedia can play a vital role \cite{22955616}. In summary, this work underscores the importance of identifying all clinically relevant non-coding alterations in the genome in order to gain complete insight into cancer progression.