News & Views

A typical cancer genome contains thousands of mutations, where majority occupy non-coding regions of the genome. However, the classical models of cancer posit that only a few of these mutations are under strong positive selection and drive the cancer forward. Currently, large number of these driver mutations have been found in coding regions of the genome. This observation poses two key questions: a) to what extent driver mutations in coding region are yet to be found? and b) whether there are many driver mutations lurking in non-coding regions of the genome?

Identification of non-coding drivers is significantly challenging due to vastness of the non-coding space and the difficulty to accurately determine functional regions of noncoding elements. These issues also confound the power (number of samples required) to detect all non-coding driver mutations in a cancer cohort. In addition, ascertainment also plays a major role in non-coding driver discovery. For instance, due to better understanding of the coding region, lot of effort has been directed toward characterizing driver mutations in coding regions compared to non-coding regions. This is potentially analogous to the classic drunk looking under the lamppost problem.

Despite these challenges, there has been a great interest in characterizing non-coding drivers in various cancers. For instance, prior studies have shown that driver mutations occur in the *TERT* promoter in many cancers. Moreover, functional impact calculation clearly indicates presence of high impact mutations in the non-coding region of various cancer genomes. In addition, multiple studies highlight the role of enhancer hijacking process in tumorigenesis. However, these are few examples and at present our understanding of non-coding drivers is rudimentary.

On page xxx of this issue, Rheinbay et. al. make a foray towards addressing this question. 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. So far, we have seen functional validation for a small number of the non-coding mutations, particularly those related to TERT promoter.

In this study, prediction of driver regulatory elements was based on, identifying non-coding elements that a) harbor significantly high mutation counts relative to expectation, or b) contain clusters of mutations around their regulatory motifs. Furthermore, for driver discovery, patient-specific background mutation rate was utilized, which takes into account of the total mutation frequency and total frequency of bases with sufficient sequencing coverage across all analyzed elements. Moreover, power analyses indicate 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, 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.

For a number of reasons, uncovering driver mutations in non-coding elements has been more challenging compared to coding ones. First, lack of specificity in characterizing non-coding annotations can substantially hinder the power to detect regulatory driver variants. Second, aggregating mutation statistics over large non-coding regions compared to their underlying functional territories can severely impact driver discovery as well. Third, both coding and non-coding regions comprise of discontinuous block of functional territories separated by different genomic elements. These connections are well understood for coding regions, where multiple exons are clearly linked through splice junctions into a transcript. In contrast, we lack such clear demarcation for noncoding regions. For instance, a gene can be connected to the non-coding elements in form of promoters, enhancers or even the entire gene regulatory subnetwork. Finally, coding regions often reside within uniform chromosomal and epigenetic contexts. In contrast, genomic contexts (chromatin state, transcriptional activity and replication timing) of non-coding regions is relatively more heterogeneous. These heterogeneous genomic characteristics make background mutation rate estimation quite challenging, which is key to identifying non-coding driver mutations.

An exhaustive (but exceedingly expensive) approach to deal with these challenges is sequencing a large number of patients in a given cohort. This approach can be feasible only through large-scale

collaborative efforts such as the Pan Cancer Analysis of Whole Genome (PCAWG) project, in which ~2800 tumor-normal samples for 40 different cancer subtypes have been sequenced through WGS. The large number of samples sequenced can provide sufficient power to detect sparsely mutated regulatory elements. However, due to heterogeneous makeup (different subtypes) of cancer cohorts, one would need to maintain sufficient sample frequencies within each cancer type to achieve required power. Furthermore, there is also large disparity in sample frequencies among well-studied and less characterized cancer types.

An alternative approach will be to develop better functional annotations of the non-coding genome with precise definition of functional motifs. In this setup, large scale annotation compendium such as ENCODE encyclopedia can play a vital role. However, the accuracy of non-coding annotation is particularly important because of their large scale. An appreciable false positive rate in defining the annotations will quickly dilute any signal for positive selection in non-coding regions. Furthermore, we will need to link these regulatory regions into distinct modules to better estimate the functional burden of the non-coding genome. Similarly, conservation based annotation such as small blocks of ultra-conserved non-coding elements and ultrasensitive sites in the genome (though a detailed understanding of such elements is often missing) can be very helpful. Finally, identification of driver mutations requires proper estimation of background mutation rates, which are dependent on accurate functional annotation.

In summary, the work by Rheinbay *et al.* underscores the importance of identifying non-coding driver mutations in cancer genome. The falling costs of WGS will further bolster such efforts to comprehensively characterize all clinically significant alterations in cancer genomes. Finally, these comprehensive catalogues of clinically relevant alterations will help us to achieve the goal of cancer precision medicine.