News & Views

Thousands of somatic mutations accumulate within coding and non-coding regions of a typical cancer genome. However, only handful of these mutations, often termed drivers, play key roles in tumorigenesis. Previous studies have comprehensively characterized coding driver variants in large number of cancer genomes. However, most somatic mutations lie in non-coding regions. Prior studies have shown that driver mutations occur in the *TERT* promoter in many cancer types. However, identifying driver mutations in other non-coding elements has remain a challenge.

On page XXX, Esther *et al*. focus on the mutations in promoter regions of 3 key genes, which play important role in breast cancer progression. They found that mutations leading to gain-of-motif events in *FOXA1* and *RMRP* promoters facilitate enhanced binding of transcriptional activators and increase gene expression. In contrast, mutations in the promoter of NEAT1 disrupted motifs, thereby reducing expression.

To identify these key genes, Esther *et al*. applied an "exome-plus" sequencing strategy (capturing exons, promoters, untranslated regions, and other regulatory motifs in the genome) to identify variants in 360 breast cancer samples. They subsequently developed and applied the MutSigNC tool on this cohort to identify promoter regions of nine genes as driver elements for breast cancer. MutSigNC evaluates non-coding regions associated with specific genes, and predicts that these elements may act as cancer drivers if they either a) harbor significantly high variant counts relative to expectation, or b) contain clusters of mutations around their regulatory motifs. Furthermore, MutSigNC considers the total mutation frequency and total frequency of bases with sufficient coverage across all analyzed elements to compute the patient-specific background mutation rate. Similarly, mutation cluster in a non-coding element is defined based on whether there is cluster of mutation present more than a random expectation. determined in an element compared to a random expectation. Functional assays, which evaluate changes in gene expression and protein binding, were used to assess the effects of mutations on the identified set of driver regions.

Leveraging these assays, they were able to determine that mutations within promoters of *FOXA1, RMRP* and *NEAT1* significantly alter transcription.

Furthermore, these functional assays were used to propose a mechanism of action for the hotspot mutation within the promoter of *FOXA1*, wherein the variant induces a gain-of-motif event, thereby facilitating a canonical binding mode of the E2F transcription factor. The resultant overexpression of *FOXA1* opens chromatin, which further promotes binding of the estrogen receptor (ER) to its target binding sites in the genome.

Major challenge associated with identifying non-coding drivers can be attributed to low patient cohort sizes and inadequate sequencing coverage in promoters (as a result of high GC nucleotide content). Nonetheless, this study is sufficiently powered to identify drivers in promoter elements, which are mutated in at least 10% of patients within the studied cohort. However, power analyses indicate that we would require larger cohort size to detect majority of driver mutations which are typically present in 3-5% of patients. Interestingly, close inspection of mutational hotspot percentages and functional mutation rate of various genes indicate similar abundance of hotspot in coding and promoter region but smaller functional territory for promoters.

This study clearly elucidates the key role of regulatory variants in promoter regions of different genes involved in breast cancer progression. A logical extension will be to determine whether promoter elements of other genes harbor driver mutations in different cancer types. Similarly, we also need to comprehensively investigate the presence of driver mutations in other regulatory elements such as UTRs, enhancers and non-coding RNAs (ncRNAs). However, the exome-plus approach adopted in this study might be limited when trying to identify driver events in other regulatory elements. Thus, it is important to leverage whole-genome sequencing (WGS) to comprehensively characterize driver mutations in non-coding regions of the genome. Furthermore, WGS will also aid in identifying other categories of driver alterations, such as copy number aberrations and large SVs. The relatively uniform coverage across the genome provided by WGS can also help in accurately identifying variants in GC-rich regulatory regions.

Similarly, statistical power is pivotal in identifying coding as well as non-coding driver variants. For a number of reasons, uncovering driver mutations in non-coding elements is far more challenging relative to those in to coding regions. First, non-coding regions are generally much larger than coding regions. Second, we lack information regarding the functional unit or regulatory motif boundaries for most non-coding elements. Third, evaluating the functional impacts of variants in non-coding regions is less straightforward than in coding elements. Furthermore, although both coding and non-coding regions comprise discontinuous blocks of functional annotation, there is clear distinction between them. For instance, exons in a coding region are clearly linked through splice junctions. However, we usually lack such clear demarcation sites for the regulatory units in a given non-coding region. Finally, coding regions often reside within uniform chromosomal and epigenetic contexts. In contrast, the genomic context of non-coding regions is relatively more heterogeneous, thereby making background mutation rate estimation quite challenging. Thus, non-coding driver discovery often requires correction for many covariates, such as chromatin state, transcriptional activity and replication timing, which is non-trivial.

A simplistic (but exceedingly expensive) approach deal with these challenges is to sequence a large number of patients in a given cohort. An alternative approach will be to develop better functional annotations of the non-coding genome with precise definition of functional motifs. Furthermore, we will need to link these functional motifs into distinct modules to better estimate the functional burden of the non-coding genome. We will also need to accurately distinguish between functional consequences of high-impact (such as motif-breaking events) and low-impact mutations. Another approach will be to identify small blocks of ultra-conserved non-coding elements and ultrasensitive sites in the genome (though a detailed understanding of such elements is often missing). Finally, accurate functional annotation will help in proper estimation of background mutation rates (thought the statistical challenges of such estimations should not be understated).

Thus, larger cohorts remain a popular paradigm, and this approach is adopted by 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. In addition, there is a vital need to develop more accurate non-coding genome annotations. The accuracy of non-coding annotation is particularly important because of their scale. An appreciable false positive rate in defining the annotations or inaccurate annotations will quickly dilute any signal for positive selection in non-coding regions.

In summary, the work by Esther *at 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 diagnosis and precision medicine.