## News & Views

##1- intro Os

A typical cancer genome contains thousands of mutations, where 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. 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?

##2 problems of noncoding v coding

Identification of non-coding drivers is significantly 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, 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.

##3 but there's stuff done w noncoding 0sander

Despite these challenges, there has been a great interest in characterizing non-coding drivers in various cancers. Over last few years, several methods have been developed to identify non-coding driver mutations in various cancer cohorts. For instance, previous studies identified recurrent mutations in the TERT promoter for multiple cancer cohorts. Similarly, recurrence based method found driver mutations in upstream regulatory regions of PLEKHS1, WDR74 and SHDH genes in different cancers. Furthermore, pan-cancer analysis of copy number aberrations and gene expression data highlighted the

**Deleted:** the

**Deleted:** large number

**Deleted:** This observation poses two key questions: a) to what extent driver

**Deleted:** in coding region

**Deleted:** yet to be found? and b)

**Deleted:**

**Deleted:**

**Formatted:** Indent: First line: 0"

**Deleted:** to

**Deleted:** determine

**Deleted:** regions of

**Deleted:** also

**Deleted:** (number of samples required)

**Deleted:** 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

**Deleted:** compared to non-

**Deleted:** .

**Deleted:** is potentially analogous to the classic drunk looking under the lamppost problem.

**Deleted:** prior

**Deleted:** have shown that driver

**Deleted:** occur

**Formatted:** Font:Not Italic

**Deleted:** 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,

**Deleted:** studies highlight

role of enhancer hijacking phenomena in regulatory elements of various genes including IRS4, SMARCA1 an TERT. However, these are few examples and at present our understanding of non-coding drivers is incomplete.

#24-Summary – more

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. Furthermore, based on these functional assays, they provide mechanistic insights into driver mutations influencing promoter region of the FOXA1 gene in breast cancer. 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 higher 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. 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 promoter regions in the genome.

#26-3 core why noncoding, how to improve, details & figures

**Deleted:** process in tumorigenesis.

**Deleted:** rudimentary

**Deleted:** high

For a number of reasons, uncovering driver mutations in non-coding elements has been more challenging compared to coding ones. First, aggregating mutation statistics over large non-coding regions compared to their underlying functional territories can severely impact driver discovery. Second, lack of specificity in characterizing non-coding annotations can substantially hinder the power to detect regulatory driver variants. For instance, large false positives in non-coding annotations will inflate the mutational frequency in regulatory regions and increase the number of multiple testing, which will inherently influence driver detection. Third, both coding and non-coding elements (e.g. genes and their regulatory structures) 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 connections 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.

An additional difficulty with identifying non-coding driver mutations is to evaluate their functional impact. Currently, it's unclear whether each nucleotide in a regulatory region is equally important for 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. Nonetheless, much more need to be done to find equivalents of synonymous, nonsynonymous and loss-of-function mutations among non-coding variants. Additionally, 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.

Finally, power to detect low-frequency non-coding driver mutations closely depends on the precise definition of the functional territory and number of non-coding regions. As shown in the figure, increasing the annotation frequency (high N) leads to significantly lower power compared to original

**Deleted:** First, lack of specificity in characterizing non-coding annotations can substantially hinder the power to detect regulatory driver variants. Second

**Deleted:** as well. Third, both coding and non-coding regions

**Deleted:** demarcation

**Deleted:** Finally

power distribution. In contrast, decreasing the annotation (lower N) leads to increase in the overall power. Similarly, restricting the length of functional annotation to the relevant region (core promoters) enhances the power to detect low frequency non-coding driver mutations.

### encode, pcawg & conclusion

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. This effort will generate a comprehensive non-coding somatic variant catalogue, which can be leveraged to detect sparsely mutated regulatory elements 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 setup, large scale annotation compendium such as ENCODE encyclopedia can play a vital role. 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.

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.

**Deleted:** 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.

**Deleted:** 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.

**Deleted:** Finally, identification of driver mutations requires proper estimation of background mutation rates, which are dependent on accurate functional annotation.