Less is More: compact annotation for finding non-coding drivers

Sushant Kumar[a, b] and Mark Gerstein[a, b, c,1]

[a]Program in Computational Biology and Bioinformatics, Yale University
[b]Department of Molecular Biophysics and Biochemistry, Yale University
[c]Department of Computer Science, Yale University
Bass 432, 266 Whitney Avenue, New Haven, CT 06520

[1] Correspondence should be addressed to M.G. (pi@gersteinlab.org)

A typical cancer genome contains thousands of somatic mutations, with the overwhelming majority in non-coding regions. However, classical models of tumor evolution posit that only a few of these are under strong positive selection and drive the cancer forward. Currently, almost all of these "driver mutations" have been found in coding regions[1,2]. Thus, a key question arises: whether many driver mutations lurk in non-coding regions?

Identification of non-coding drivers is challenging due to vastness of non-coding space and the difficulty in characterizing noncoding elements. These issues confound the power to detect non-coding drivers. In contrast, identifying coding drivers is easier: we have a better understanding of the start and end of coding regions and the functional impact of coding mutations -- e.g. whether or not a mutation changes a protein (nonsynonymous/synonymous) or completely knocks it out (loss-of-function)? Potentially, our better understanding of coding regions creates an ascertainment bias and raises the question of whether the paucity of non-coding driver mutations actually reflects a drunk-looking-under-the-lamppost phenomenon.

Nevertheless, there has been great interest in finding non-coding drivers[3], and several methods have been developed specifically to identify them. For instance, previous studies identified recurrent mutations in the TERT promoter[4]. Similarly, a recurrence based method found driver mutations in upstream regulatory regions of the PLEKHS1, WDR74 and SHDH genes[5]. Furthermore, pan-cancer analysis of copy-number aberrations highlighted the role of enhancer hijacking[6]. However, these are few examples and our understanding of non-coding drivers is incomplete.

On page xxx of this issue, Rheinbay et. al. make a further 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. They provide evidence that with uniform ascertainment, one could find as many noncoding drivers as coding ones. Moreover, they predicted mutations within promoters of *FOXA1, RMRP* and *NEAT1* significantly alter transcription and validated this with functional assays.

More specifically, they predicted driver mutations based on identifying non-coding elements that harbor significantly more mutations than expectation and contain clusters of mutations around their regulatory motifs. Furthermore, for driver discovery, they utilized patient-specific background mutation rates. Their power analysis indicates that the cohort size in this study makes possible identifying promotor drivers that are mutated in at least 10% of patients in the cohort. However, they also show that one would need a larger sample size to confidently identify drivers present in ~5% of patients. Interestingly, their analysis of mutational hotspots (single site recurrent mutations) indicates that those in promoters are as common as those in coding genes. Furthermore, the per base mutation rate of promoters with drivers was similar to that of well-known coding regions with drivers. This suggests that the smaller number of driver mutations found in promotors in contrast to coding genes can be attributed to their small amount of functional territory (i.e. they simply occupy less base pairs in the analysis).

This work describes the state-of-the-art in identifying non-coding drivers, but there is still more to do. To understand the directions for improvement, it is worthwhile to briefly review the non-coding annotation process and its interplay with power calculations (Figure). Currently, the majority of non-coding elements are fairly large in size due to the way they are determined from processing noisy functional genomics signals (e.g. 1-kb-sized peak calls). However, their actual functional territory can be considerably smaller, and aggregating mutation recurrence across over-sized regions can dilute the true signal of positive selection and hinder driver identification. Power calculations show that restricting annotation to smaller functionally relevant blocks enhances power. One approach to better define non-coding functional territories is to use conservation. Conserved regions can include regulatory motifs (such as TF binding motifs) and, more generally, ultra-conserved and ultrasensitive sites. 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). These connections are well understood for coding regions, where multiple exons can be clearly linked through splice junctions. In contrast, we lack such clear non-coding connections. For instance, a gene can be connected to multiple promoters and enhancers, and one enhancer and affect multiple genes.

After defining the functional territory of an individual non-coding element, the next step involves mutation burden testing over many elements. Lack of specificity in non-coding annotation will increase the multiple-testing burden, which will decrease driver detection power. One can 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 non-intuitively not an annotation of every base in the genome. Rather it is a compact and highly accurate annotation set with as few elements as possible, where each element corresponds closely to an underlying functional territory, which potentially links discontinuous functional regions in the non-coding genome.

An additional difficulty with non-coding mutations is evaluating their functional impact. Currently, it is unclear whether substitution of each nucleotide in a regulatory region has an equal functional impact. We can see this for certain among well characterized situations in TF binding sites, e.g. some non-coding mutations are considered more disruptive if they break an existing TF motif or generate a new binding motif[7]. Nonetheless, better metrics of functional impact are needed over the whole genome to find the non-coding equivalent of synonymous, nonsynonymous and loss-of-function mutations. Finally, the power to detect drivers in non-coding regions is dependent on the uniformity of the underlying background mutation rate. However, this is far from uniform across wide 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)[8].

An exhaustive (but expensive) approach to deal with some of 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 driver regulatory mutations with sufficient power. However, these large-scale studies require assembling uniform cohorts, which can be challenging due to the highly heterogeneous nature of cancer (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. Here, large scale annotation compendium such as ENCODE can play a vital role[9].

Figure 1: Key factors influencing driver discovery in non-coding regions of the genome: A) The connection between multiple exons (green blocks) in coding regions is well defined. However, connection between promoter and other CREs (cis-regulatory elements) with their underlying gene are more ambiguous. Moreover, the length of functional territory/motif (L) is much smaller compared to the entire CRE peak. B) The power associated with driver discovery is a function of the number (N) and length (L) of promoters: 1) the solid black, 2) dotted black and 3) dashed black lines correspond to power curves for N = 20,000, 100 and 400,000 promoters, respectively. The solid brown line represents the power when mutations are aggregated over promoter cores (i.e., smaller L values) instead of the entire promoter length.

References

1. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–20 (2013).
2. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3,** 2650 (2013).
3. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17,** 93–108 (2016).
4. Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat. Commun.* **4,** (2013).
5. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46,** 1160–1165 (2014).
6. Weischenfeldt, J. *et al.* Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49,** 65–74 (2017).
7. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science (80-.).* **342,** 1235587 (2013).
8. Lochovsky, L. *et al.* LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res.* **43,** 8123–8134 (2015).
9. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).