A typical cancer genome contains thousands of somatic mutations. Classical models of tumor evolution posit that only a few of these mutations are under strong positive selection and drive cancer progression. Currently, almost all of these "driver mutations" have been found in coding regions 1,2. However, the overwhelming majority of cancer mutations lie in non-coding regions. Thus, a key question arises: how many driver mutations remain to be discovered in non-coding regions?

Identification of non-coding drivers is challenging due to the vast scale of the non-coding genome, and the difficulty of characterizing noncoding elements. These issues decrease the power to detect non-coding drivers. In contrast, identifying coding drivers is more straightforward: we have good understanding of where coding regions start and end, and of what functional impact coding mutations may produce -- e.g. change in a protein sequence (synonymous or non-synonymous mutation) or gene knock-out (loss of function mutation). Our better understanding of coding regions may create a form of ascertainment bias. The paucity of identified non-coding drivers may actually reflect our preference to search for driver mutations in coding regions, where we are best able to evaluate the effect of mutation.

Nonetheless, there has been great interest in non-coding drivers3. Several methods have been developed to identify them. A recurrence based method found driver mutations in upstream regulatory regions of the PLEKHS1, WDR74 and SHDH genes4. Similarly, previous studies identified recurrent mutations in the TERT promoter5. Pan-cancer analysis of copy-number aberrations has highlighted the role of enhancer hijacking in cancer6. However, these are few examples and our understanding of non-coding drivers is incomplete.

On page xxx of this issue, Rheinbay et. al. report progress towards identifying non-coding drivers \cite{}. For a cohort of 360 breast cancer patients they looked for both coding and non-coding driver mutations in an unbiased fashion. They identified non-coding driver mutations by locating non-coding elements that harbor significantly more mutation than expected and that contain clusters of mutations around their regulatory motifs. They also utilized patient-specific background mutation rates for driver discovery. With uniform ascertainment, they found as many noncoding drivers as coding drivers. They predicted that mutations within the promoters of *FOXA1, RMRP* and *NEAT1* significantly alter transcription and then validated these predictions with functional assays.

Their power analysis indicates that the size of the cohort in their study makes it possible to identify promoters with drivers if those promoters are mutated in at least 10% of patients. 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 hotspots 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 small number of driver mutations found in promoters in comparison to coding genes can be attributed to their small functional territory (i.e. they simply occupy fewer base pairs in the analysis).

The work of Rheinbay et. al. describes the state-of-the-art in identifying non-coding drivers, but there is still more to do. To understand possible directions for improvement, it is worthwhile to briefly review the non-coding annotation process and its effect on power calculations (Figure). Currently, the majority of annotated non-coding elements are fairly large in size, because they result from processing of noisy functional genomic signals (e.g. 1-kb-sized peak calls). However, their actual functional territories maybe considerably smaller than annotated. 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. Related to power and to the size of functional elements, both coding and non-coding elements (e.g. genes and their regulatory structures) may span multiple discontinuous blocks of functional territories. This discontinuous nature becomes more apparent as the functional block size shrinks. The connections between functional blocks are well understood for coding regions, where multiple exons can be clearly linked through splice junctions. In contrast, we lack such clear understanding of connections for non-coding regions. For instance, a gene can be connected to multiple promoters and enhancers, and one enhancer can affect multiple genes. Non-coding functional territories can be better defined through use conservation. Conserved regions can be regulatory motifs (such as TF binding motifs) and, more generally, ultra-conserved and ultrasensitive sites.

After defining the functional territory of individual non-coding elements, the next step in driver identification involves mutation burden testing over many elements. Lack of specificity in non-coding annotations will increase multiple-testing burden, which will decrease driver detection power. Specificity can be increased by removal of as many false positives as possible within the annotation set. Thus, the best annotation for increasing power for driver detection is not an annotation of every base in the genome. Rather it is a compact and highly accurate annotation set with as few elements as necessary, where each element corresponds closely to an underlying functional territory, and where discontinuous functional regions in the non-coding genome are linked together.

Even with a well-defined annotation set, it is still difficult to evaluate the functional impact of non-coding mutations affecting annotated regulatory regions. Currently, it is unclear whether all nucleotides in a regulatory region have equal functional impact when mutated. We can see this most clearly for well characterized TF binding sites, where some non-coding mutations are considered more disruptive if they break an existing TF motif or generate a new binding motif7. Better metrics of functional impact are needed over the whole genome to find the non-coding equivalents of synonymous, nonsynonymous and loss-of-function mutations. The power to detect drivers in non-coding regions is also dependent on an underlying background mutation rate that varies across the genome according to complex genomic and epigenomic signals (chromatin state, transcriptional activity and replication timing)8.

An exhaustive (albeit costly) approach to deal with some of these challenges is sequencing a large number of patients. This approach is feasible only through large-scale collaborative efforts such as the Pan Cancer Analysis of Whole Genome (PCAWG) project. These efforts will generate large non-coding variant catalogues, which can be leveraged to detect 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 is to develop a more compact functional annotation of the non-coding genome with precise definitions of functional territories. Here, a large scale annotation compendium such as ENCODE can play a vital role9.

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. 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).

5. Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat. Commun.* **4,** (2013).

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).

\cite{22955616}.