Cancer genomics

Less is more in the hunt for driver mutations

An analysis of 360 breast-cancer genomes has identified nine cancer-driving promoters in non-coding DNA sequences that regulate gene expression, hinting at the prevalence of such drivers in cancer genomes. See Article p.XXX

A typical cancer genome contains thousands of mutations, the overwhelming majority of which are in non-protein-coding sequences. Classical models of tumour evolution posit that cancer progression is driven by only a few of these mutations. But almost all known driver mutations are in coding sequences[1,2], raising the question of how many drivers lurk in non-coding regions of the genome. In a paper online in *Nature*, Rheinbay *et al.*[3] make a foray towards the answer.

Identification of non-coding drivers is challenging, owing to the vastness of the non-coding genome and the difficulty of characterizing the positions of specific non-coding elements (regulatory regions such as promoters and enhancers that modulate gene expression), which might be predicted to contain driver mutations. Coding drivers are easier to identify, because we have a better understanding of the start and end of coding regions, and of the impact that coding mutations might have on production and function of the protein encoded in that region. It is possible that our understanding of coding regions creates an ascertainment bias toward driver mutations identification in coding regions, making it a drunk-looking-under-the-lamppost phenomenon. Nevertheless, there has been great interest in finding non-coding drivers[4]. Previous studies have provided a few examples[5–7], but our understanding of non-coding drivers is far from complete.

Rheinbay *et al.* set out to identify coding and non-coding driver mutations in an unbiased fashion, using tissues from a cohort of 360 people who had breast cancer. To find non-coding drivers, the researchers identified non-coding elements that harboured significantly more mutations than would be expected, or that contained clusters of mutations around sequences such as transcription-factor binding sites, which regulate the element's activity.

The authors identified putative driver mutations in nine promoters, and showed that three of these (those regulating expression of the genes *FOXA1, RMRP* and *NEAT1*) significantly altered gene-transcription levels. Their analysis of mutational hotspots (single site recurrent mutations) indicated that those in

promoters are as common as those in coding genes. Furthermore, the per-base mutation rate of promoters that contained drivers was similar to that of coding regions known to contain drivers. This suggests that the reason that fewer drivers have been found in promoters than in coding regions can be attributed to the fact that their functional territories are smaller — they account for fewer nucleotides in the analysis.

This work describes the state-of-the-art in identifying non-coding drivers, but there is more still to do. The authors' power analysis — a statistical calculation that predicts the sample numbers needed to detect an effect of a given size— indicated that their sample size of 360 could be used to reliably identify drivers only if they occurred in at least 10% of patients in the cohort. To understand the directions for improvement, it is worth considering how non-coding elements are defined, and how this plays into statistical power (Fig. 1).

Currently, most non-coding elements are annotated as being fairly large (about one kilobase long). However, this is at least partly attributable to the fact that the techniques used to determine the positions of non-coding elements — which involve looking for characteristic features, such as specific molecular modifications, bound proteins and chromatin accessibility — are typically noisy. The functional territory of a regulatory element can therefore be considerably smaller than is annotated. Thus, aggregating mutation recurrence across over-sized regions instead of functional territories can dilute the true signal of positive selection and hinder driver identification.

One approach to better define the functional territories of non-coding elements is to identify evolutionary conserved regions, which are likely to be functionally important and so are more likely to contain driver mutations. It should also be noted that non-coding elements, like genes, consist of discontinuous blocks of functional territories. The connections between these territories are well understood for genes, because coding regions are joined up around splice junctions during processing of messenger RNA, making links readily apparent. But the connections between non-coding elements and between these elements and the genes they regulate are less well understood, and are complex — genes can be connected to multiple promoters and enhancers, and one enhancer can 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. Power calculations show that

restricting annotation to smaller, functionally relevant regions enhances power. Thus, the best way to increase the power of driver detection in non-coding elements is, perhaps non-intuitively, not to investigate every base in the genome. Rather, it is to analyse a compact and highly accurate annotation set containing as few elements as possible, in which each element corresponds closely to an underlying functional territory and 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. In some circumstances, it is clear what effect a mutation will have — if it breaks a transcription-factor binding site or creates a new one, for instance[8]. Nonetheless, better metrics of functional impact are needed over the whole genome to find non-coding equivalents of the coding mutations known to alter protein production or behaviour. Finally, the power to detect drivers in non-coding regions is dependent on the uniformity of the underlying background mutation rate. However, this is not the case for wide expanses of the genome[9], so the approach will require further refinement.

An exhaustive but expensive approach to deal with some of these challenges is sequencing many patients. This approach is feasible only through large-scale collaborations. Such efforts will generate comprehensive catalogues of non-coding variants, which can be leveraged to detect more driver mutations. However, these large-scale studies require the assembly of uniform cohorts, which can be challenging owing to the highly heterogeneous nature of cancer. An alternative approach would be to develop a more compact functional annotation of the non-coding genome by precisely defining functional territories. Here, large-scale annotation compendiums such as the ENCODE project[10] have a vital role to play. In summary, in the current work Rhienbay et. al. highlight how less is more when it comes to identifying non-coding driver mutations for cancers.