**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**

**Sushant Kumar & Mark Gerstein**

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. But almost all known driver mutations are in coding sequences1,2, raising the question of how many drivers lurk in non-coding regions. 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 genome and the difficulty of characterizing the precise location of non-coding elements (eg regulatory regions that modulate gene expression), which might contain drivers. Drivers in coding regions are easier to identify, because we have a better understanding of the boundaries of these regions and of the impact that mutations in them might on the production and function of proteins. However, our better understanding potentially creates an ascertainment bias toward coding drivers, the drunk-looking-under-the-lamppost phenomenon in cancer genomics. Consequently, with the whole-genome analysis of cancers there has been interest in identifying non-coding drivers4. Previous studies have provided a few examples5–7, but our understanding is far from complete.

Rheinbay *et al.* set out to identify coding and non-coding driver mutations in an unbiased fashion, using samples from a cohort of 360 people with breast cancer. To find the non-coding ones, they identified non-coding elements harbouring significantly more mutations than expected, or that contained clusters of mutations around transcription-factor binding sites (known locations for regulating proteins to bind to).

The authors identified putative driver mutations in nine promoters, and showed that three of these significantly altered gene-expression levels (those associated with the *FOXA1, RMRP* and *NEAT1* genes). Their analysis of mutational hotspots (recurrent mutations at a single site) indicated that those in promoters are as common as those in coding genes. Furthermore, they found that the per-base mutation rate of promoters with drivers was similar to that of coding regions with drivers. This suggests that that fewer drivers have been found in promoters than in coding regions simply because their "functional territory" is smaller.

This work describes the state-of-the-art in identifying non-coding drivers, but there is more to do. The authors’ power analysis — statistical calculations estimating the sample numbers needed to detect an effect of a given size— indicated that their sample 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, many non-coding elements are annotated as being fairly large. However, this is partly because our techniques for determining the positions of these elements are imprecise, and the real functional territory of a regulatory element is considerably smaller than annotated. For instance, transcription-factor binding sites are often called as 1-kb "peaks" from a noisy cross-genome binding signal, when in fact the actual "functional" site of factor binding might only measure in tens of nucleotides. Thus, aggregating mutational recurrence across over-sized regions instead of actual functional territories can dilute the true signal of positive selection and hinder driver identification.

 One approach to better define the precise functional territory of a non-coding element is identifying evolutionary conserved portions, which are likely more functionally important. Moreover, non-coding elements, like genes, often consist of discontinuous blocks of functional territories. The connections between these are well understood for genes. That is, coding exons are joined up around splice junctions during processing of messenger RNA. But the connections between non-coding elements and between these and the genes are less well understood, and potentially complex — genes can be connected to multiple promoters and enhancers, and one enhancer can affect multiple genes.

After defining the functional territory of a non-coding element, the next step involves testing for mutational burden over many elements. The more elements one tests the larger, the multiple-testing penalty will be on the resulting statistics. Thus, one can increase power through making the element set as small and accurate as possible. This suggest that 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 as closely as possible to an underlying functional territory.

An additional difficulty with non-coding mutations is evaluating their functional impact. Currently, it is unclear whether each potential substitution of a nucleotide in a regulatory region has an equal 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 instance8. 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 depends on how uniform the underlying background mutation rate is. However, this is not the case across wide expanses of the genome9, so the approach will require further refinement.

An exhaustive but effective 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 give us better statistics that 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 is developing a more compact functional annotation of the non-coding genome by better and more precisely defining functional territories. Here, systematicannotation compendiums such as the ENCODE project10 have a vital role to play. Thus, in the pursuit of more drivers we may be actually be served by less.

**Sushant Kumar** *and* **Mark Gerstein** *are in theProgram in Computational Biology and Bioinformatics and in the Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA. M.G. is also in the Department of Computer Science, Yale University*

*e-mail: pi@gersteinlab.org*

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. Rheinbay *et al. Nature* XXX (2017).

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).

**Figure 1| Improving discovery of cancer-driving mutations in the non-coding genome.** The power to identify regulatory driver elements is closely dependent on the annotation of these elements on genomic level.

An approxate representation of the graph from figure XX of Rheinbay is show in green... this has . Note has a has size of 450 bp for protomotors and it's over 20K promotors. For the 360 individuals in the study this gives a power of 70 percent to detect a driver in XXX percent of the population.

If we use more more promotors for instance increasing to 100K then this will inc. the mult. testing burden and descrease the povwer. Conversely, if we shrink the size or hte lements while keeping the same number this will increase the power shwon in the red curve .

The Presence of high false positive in annotation set leads to increased number of elements (many without signal) and this increases the multiple testing burden for driver discovery, thus decreases the power (*red line, N=100K promoters*) compared to more accurate annotation set (*green line, N=20K promoters*). Similarly, for a given annotation set (*N = 20K promoters*), restricting the definition of the regulatory element to the underlying functional territory (*blue line, L= 450*) increases the power compared to annotation definition spanning larger regions (green line, L=650).