**Cancer genomics**

**Less is more in the hunt for driver mutations**

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

**Sushant Kumar & Mark Gerstein**

A typical cancer genome contains thousands of mutations, the overwhelming majority of which fall outside of protein-coding regions. Classical models of tumour evolution posit that cancer progression is driven by only a few of these variants (often simply termed “drivers”), which are under strong positive selection. But almost all known driver mutations are in coding sequences1,2, raising the question of how many drivers lurk in non-coding regions. Recently published work in *Nature* by Rheinbay *et al.*3 outlines progress towards addressing this question.

The complexity of the genome and the difficulty of characterizing the precise boundaries of non-coding regulatory regions that may harbour drivers (such as promoters and enhancers) greatly complicates the identification of non-coding drivers. Drivers in coding regions are easier to identify, because we have a better understanding of the boundaries of coding regions and the impact that mutations therein might confer on protein functionality. However, this better understanding potentially introduces an ascertainment bias toward coding drivers. Consequently, there has been interest in identifying non-coding drivers using whole-cancer-genome analyses4. 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 360 breast cancer patients. To find the non-coding drivers, they identified non-coding elements that harboured significantly more mutations than expected, in addition to regions containing clusters of mutations around transcription-factor binding sites, to which regulatory proteins bind.

The authors identified putative drivers in nine promoters, and showed that three significantly alter gene-expression levels (these are associated with the genes *FOXA1, RMRP* and *NEAT1*). Their analysis of mutations that recur in many individuals indicated that those in promoters are as common as those in coding genes. Furthermore, they found that the per-base mutation rate of promoters that contain 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. (Here we use the term “functional territory” to suggest the true nucleotide sequences/motif that confer the disease-related activity).

Though this work describes state-of-the-art identification of non-coding drivers, more work remains to be done. The authors’ power analysis — statistical calculations estimating the sample numbers needed to detect an effect on a given size — indicated that 85% of all drivers could reliably be identified if they occurr in at least 10% of the 360 samples studied, but only 70% of drivers in 5% of patients would be identified. Thus, it is worth considering how non-coding elements are defined, and how this plays into statistical power (Fig. 1).

Many non-coding elements are annotated being fairly large genomic segments (~1000 bases). However, this is partly because our techniques for determining the positions of these elements are imprecise — the real functional territory of a regulatory element is often considerably smaller than annotated. For example, consider transcription-factor binding sites. These regions are identified by isolating protein–DNA complexes and sequencing that DNA. Sequences longer than the binding site are often isolated and, when the experiment involves many cells, the resultant signal can be noisy. As such, regions of 1 kilobase can be annotated as binding sites, despite the fact that the true functional site might only be tens of nucleotides long. Analysing recurrent mutations across over-sized regions can thus dilute the true signal of positive selection and hinder driver identification.

One approach to better define functional territories is to identify evolutionarily conserved regions, which are likely to be functionally important. Moreover, non-coding elements, similar to genes, often consist of discontinuous blocks of functional territories. Aggregating these blocks, and skipping over non-functional regions, is also important for maximally enriching the true signal of selection required for driver identification. However, the connections between non-coding elements are less well understood than are those for genes (where coding regions are joined after transcription around well-characterized sequences called splice junctions). Furthermore, they can potentially be 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 is to test for mutational burden (the relative prevalence of mutations in a given region) over many elements. Testing more elements imposes the statistical burden of greater prevalence to be to be considered statistically significant (i.e., as a result of multiply hypothesis testing correction). Thus, one can increase the power of driver detection by making the element set as small and accurate as possible. This suggests that the best way to increase power for non-coding elements is, perhaps non-intuitively, 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, rather than to investigate every base in the genome.

Another difficulty is evaluating the impact of non-coding mutations. It is unclear whether each substitution of a nucleotide in a regulatory region confers an equal impact. In some circumstances, the effect of a mutation can be predicted if it breaks a transcription-factor binding site or creates a new one, for instance8. But 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 functionality. Finally, the power to detect drivers in non-coding regions depends on how uniform the underlying background mutation rate is. Rates are irregular across wide expanses of the genome9, so current approaches will require further refinement.

An 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 uniform cohorts, which will be a challenge owing to the highly heterogeneous nature of cancer. The development of a more compact functional annotation of the non-coding genome represents a compelling alternative. Here, systematicannotation consortia such as the ENCODE project10 have a vital role to play. As evidenced by this study, more drivers can be found by focussing on less of the genome.

**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: mark@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).

***[Thanks for sending your revised figure. I like the simplified graph, but I suggest that we use just this — it will give readers enough to get to grips with, without having to also understand the depiction of non-coding elements and peak signals. Is this OK? Could we change power on the y axis to a percentage, to fit the description you’ve given in the caption?]***

***[I’ve made some more changes to your figure caption to reflect the need to explain all the technical terms again, and to try and outline a bit more about the power calculation used, so that readers can follow the graph easily. Please amend further as needed for accuracy]***

**Figure 1 | Improving the discovery of cancer-driving mutations in non-coding genomic elemenets.** Rheinbay *et. al.*3analysed genomes from 360 patients who had breast cancer and identified cancer-driving mutations in nine non-coding sequences called promoters. They then performed a statistical ‘power analysis’ to determine the percentage of the time (i.e., the power with which) a driver in 5% of patients could be identified using varying sample numbers, given that the authors analysed 20,000 promoters defined as being 650 base pairs long. Their analyses (green curve) reveals a power of about 0.7 to detect driver mutations in 360 samples, meaning that they probably identified 70% of the drivers present in 5% of patients. However, if 100,000 promoters are analysed, the associated power would be decreased, owing to a statistical phenomenon called multiple-testing burden (red curve). By contrast, analyzing 20,000 promoters 450 bases long (but still containing the true binding sites) would enrich for the true signal and increase power (blue curve), suggesting an effective means of identifying more non-coding drivers in the future by using a more compact annotation. (The green curve is an approximate representation of the authors’ analysis taken from fig. 4a of the paper).