**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 *[OK? to explain a bit more what promoters are]***

**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, which are under strong positive natural selection. 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 that might contain drivers — for example, regulatory regions such as promoters and enhancers that modulate gene expression. Drivers in coding regions are easier to identify, because we have a better understanding of the boundaries of coding regions and of the impact that mutations in them might have on the production and function of proteins. However, our better understanding potentially creates an ascertainment bias toward coding drivers ***[I’m afraid we can’t include the “drunk-looking-under-the-lampost phenomenon” — if we were to include this we’d have to give it its proper name (which I understand is the streetlight effect) and then explain what it means, which seems a bit too much for this piece. OK to remove?]***. 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 people who had breast cancer. To find the non-coding drivers, the researchers identified non-coding elements that harboured significantly more mutations than expected, or that contained clusters of mutations around transcription-factor binding sites, at which regulatory proteins bind.

The authors identified putative drivers 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 mutations that recur in many individuals ***[OK?]***  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 contained 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 (the number of nucleotides that confer their activity ***[OK? We should provide readers with a clear definition of a functional territory, to help them follow the rest of the discussion]***) is smaller.

This work describes state-of-the-art identification of 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 drivers could be reliably identified if they occurred in at least 10% of the 360 samples studied, but only about 70% of drivers present in 5% of patients would be identified ***[Ok to add back in? To tie into your figure]***. To understand directions for improvement, 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 as being fairly large ***[Please give a size range (for instance, could we say hundreds of base?). Physicists, for example won’t know whether 100bp, 1kb or 1mb is fairly large in this scenario]***. 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. As an 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 result can be noisy. As such, regions of 1 kilobase can be annotated as binding sites when the actual functional site might be only tens of nucleotides long ***[Many of our readers won’t be familiar with techniques such as ChIP and won’t know what peaks, calling or signals refer to in this context. I’ve had a go at simplifying this again; is this OK?]***. Analysing recurrent mutations ***[OK? I’m not quite sure what aggregating recurrence means]*** 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 evolutionary conserved regions, which are more functionally important than non-conserved regions. Moreover, non-coding elements, similar to genes, often consist of discontinuous blocks of functional territories. The connections between these blocks are well understood for genes — coding regions are joined up around sequences called splice junctions after gene transcription. But the connections between non-coding elements and between these elements and genes are less well understood, and potentially complex. Genes can be connected to multiple promoters and enhancers, and one enhancer can affect multiple genes. ***[Please could you add a few words to spell out how understanding the connections between them improves annotation of functional territories?]***

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 ***[OK? Or please replace with a simple definition]***) over many elements. The more elements one tests, the higher the prevalence of a given driver will need to be to be considered statistically significant, owing to a statistical phenomenon called the multiple testing penalty ***[I’ve attempted to explain this for non-specialists — please replace with a more accurate explanation if needed]***. 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 has an equal impact. In some circumstances, what effect a mutation will have 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 behaviour. 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 Rheinbay and colleagues’ approach 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 compendiums such as the ENCODE project10 have a vital role to play. If the innovative approach taken by Rheinbay and colleagues can be refined ***[OK to add? To end with an upbeat sentence explicitly about the current study]***, more drivers can be pursued 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: marki@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 discovery of cancer-driving mutations in the non-coding genome.** 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 ‘power analysis’ to determine the percentage of the time (the power with which) a driver present 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. This analysis (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. If 100,000 promoters were analysed, power would be decreased, owing to a statistical phenomenon called multiple burden testing (red curve). By contrast, analyzing 20,000 promoters 450 bases long would increase power (blue curve), pointing to a way to identify more non-coding drivers in the future. (The green curve is an approximate representation of the authors’ analysis taken from fig. 4a of the paper.)