doi:10.1038/nature23086

CANCER GENOMICS

Less is more in the hunt for driver mutations

An analysis of 360 breast-cancer genomes has identified cancer-driving mutations in 9 non-coding DNA sequences called promoters, which regulate gene expression. The result hints at the prevalence of such drivers.

SUSHANT KUMAR & MARK GERSTEIN

A typical cancer genome contains thousands of mutations, the overwhelming majority of which occur in sequences that do not encode proteins, but can still influence gene expression ***[OK? To help people follow the next sentence]***. Classical models of tumour evolution posit that cancer progression is driven by only a few of these mutations, which are under strong positive selection and so are preferentially maintained in the cancer-cell population ***[OK? To spell out positive selection in this context]***. But almost all known driver mutations lie in coding sequences[1,2], 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 into this issue.

The identification of non-coding drivers is challenging, owing to the vastness of the genome and the difficulty of precisely locating non-coding elements that might contain drivers. These elements can be regulatory regions such as promoters and enhancers, which modulate gene expression. Drivers in coding regions are easier to identify, because we understand the boundaries of coding regions and how mutations in these regions might affect protein production and function, potentially leading to cancer. However, the resulting greater focus on coding drivers can create a bias towards their identification. Consequently, there has been interest in identifying non-coding drivers by analysing whole cancer genomes[4]. Previous studies have provided a few examples[5–7], 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 with breast cancer. To find non-coding drivers, the researchers searched for promoters and enhancers that harboured significantly more mutations than expected, or that contained clusters of mutations, because these can identify ***[OK?]*** transcription-factor binding sites, at which regulatory proteins bind.

The authors found putative drivers in nine promoters, and showed that three of these (those associated with the *FOXA1*, *RMRP* and *NEAT1* genes) significantly altered gene-expression levels. Their analysis of the subset of mutations that recur in many individuals indicated that those in promoters are as common as those in protein-coding genes. Furthermore, the authors found that the per-base mutation rate of promoters containing drivers was similar to that of coding regions with drivers. This suggests that fewer drivers have been found in promoters than in coding regions simply because promoters' functional territories — the nucleotide sequences that actually confer disease-related activity — are smaller.

This work describes state-of-the-art identification of non-coding drivers, but there is more still to do. The authors' power analysis (statistical calculations estimating the sample numbers needed to detect an effect of a given size) indicated that 85% ***[Please can you say where this figure is given in the paper?]*** of all drivers could be reliably identified if they occurred in at least 10% of the 360 samples studied, but only 70% of drivers present in 5% of patients would be identified (Fig. 1). To understand directions for improvement, it is worth considering how non-coding elements are defined, and how this plays into statistical power.

Many non-coding elements are annotated (their locations identified) ***[OK?]*** as fairly large sequences (about 1 kilobase long). However, this is partly because our techniques for determining the positions of these elements are imprecise — their real functional territory is often considerably smaller than that annotated. For example, transcription-factor binding sites 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, 1-kilobase regions can be annotated as binding sites when the actual functional site might be only tens of nucleotides long. Analysing recurrent mutations across oversized regions can thus dilute the true signal of positive selection and hinder driver identification.

One approach to better defining functional territories is to identify evolutionarily

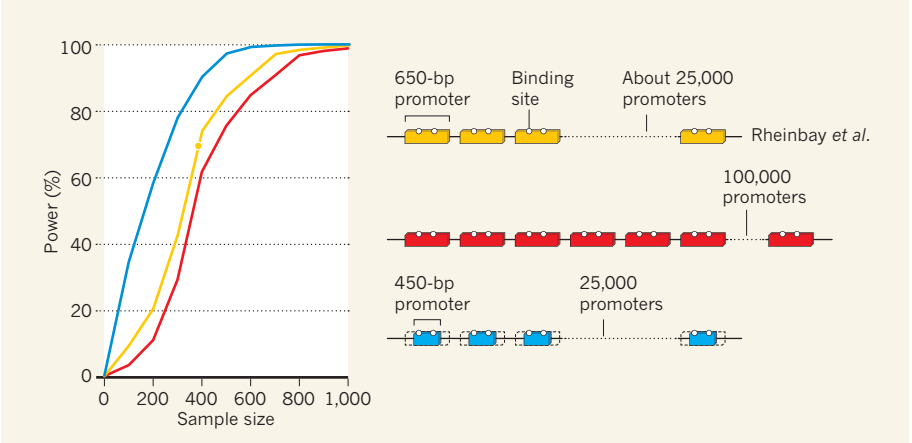

Figure 1 | Improving discovery of cancer-driving mutations in the non-coding genome. Rheinbay *et al.*[3] analysed genomes from 360 people with breast cancer, and identified cancer-driving mutations in 9 non-coding sequences called promoters — probably in binding sites for transcription factors, which regulate gene expression. They performed a power analysis (yellow curve) to determine the percentage of the time a driver present in 5% of people could be identified using varying sample sizes, given that they analysed 25,000 promoter sequences, each 650 base pairs (bp) long. Their power analysis reveals that, when using a sample size of 360, they can identify 70% of the drivers present in 5% of people (yellow dot). How can power be improved? As an example, if 100,000 promoters were analysed, power would decrease owing to a statistical phenomenon called the multiple testing penalty (red). By contrast, analysing promoters 450 bp long would increase power (blue), as long as they still contained the binding sites. This points to a way of identifying more non-coding drivers. (The yellow curve is an approximation of the authors' analysis in Fig. 4a of the paper[3].)

conserved regions, which are likely to be functionally important. Moreover, non-coding elements, like genes, often consist of discontinuous blocks of functional territories. Linking up these blocks, and skipping over non-functional regions, is another way to maximize the potential for driver identification. However, the way in which non-coding elements are connected is less well understood than for genes (in which coding regions are joined up after transcription around well-characterized sequences called splice junctions). Furthermore, the connections can be complex — genes can be linked to multiple promoters and enhancers, and one enhancer can affect many 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. The more elements one tests, the higher must be the prevalence of a given driver before it can be considered statistically significant, owing to a statistical approach called the multiple testing penalty. Thus, one can increase the power of driver detection by making the element set as small and accurate as possible (Fig. 1). This suggests that the best way to increase power for non-coding elements is, perhaps counter-intuitively, to analyse a compact and highly accurate annotation set containing as few elements as possible, rather than to investigate every base in the genome.

Another difficulty is evaluating the effect of non-coding mutations. It is unclear whether each substitution of a nucleotide in a regulatory region has an equal impact. In some circumstances, a mutation's effect can be predicted — if it breaks a transcription-factor binding site or creates a new one, for instance[8]. But better metrics of functional impact are needed over the whole genome if we are 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 genome[9], so current approaches will require further refinement.

An effective approach to dealing with some of these challenges is to sequence the genomes of many patients. This approach is feasible only through large-scale research 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, systematic annotation compendiums, such as the ENCODE project[10], have a vital part to play. As Rheinbay and colleagues' study neatly demonstrates, more drivers can be found by focusing on less of the genome. ■

Sushant Kumar *and* **Mark Gerstein** *are in the Program 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. The Cancer Genome Atlas Research Network. *Nature Genet.* **45,** 1113–1120 (2013).
2. Tamborero, D. *et al. Sci. Rep.* **3,** 2650 (2013).
3. Rheinbay, E. *et al. Nature* http://dx.doi.org/10.1038/nature22992 (2017).
4. Khurana, E. *et al. Nature Rev. Genet.* **17,** 93–108 (2016).
5. Vinagre, J. *et al. Nature Commun.* **4,** 2185 (2013).
6. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. *Nature Genet.* **46,** 1160–1165 (2014).
7. Weischenfeldt, J. *et al. Nature Genet.* **49,** 65–74 (2017).
8. Khurana, E. *et al. Science* **342,** 1235587 (2013).
9. Lochovsky, L., Zhang, J., Fu, Y., Khurana, E. & Gerstein, M. *Nucleic Acids Res.* **43,** 8123–8134 (2015).
10. The ENCODE Project Consortium. *Nature* **489,** 57–74 (2012).