[Strapline: One or two words to describe the general subject area. What about this? Or feel free to suggest an alternative]

Cancer genomics

[Main title: It should be no more than 40 characters, including spaces, should not include punctuation (including colons), and should be easily understandable for non-specialists How about this shortened, variation on yours, which just fits? Or feel free to suggest another]

Less is more in the hunt for driver mutations

[Standfirst: 190-225 characters, including spaces, to outline the new results for a general audience. The aim is to entice readers to reader on. How about this? Please amend as needed, bearing the aforementioned restrictions in mind]

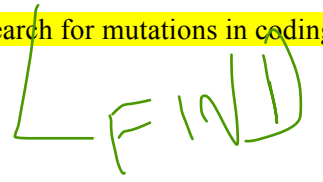
An analysis of 360 breast-cancer genomes has identified nine cancer-driving mutations in non-coding DNA sequences that regulate gene expression, hinting at the prevalence of such drivers in cancer genomes. See Article p.XXX

Sushant Kumar & Mark Gerstein

[This opening paragraph is great. I've just made a few tweaks to fit our house style, which dictates that the opening sections should be simple and provide a teaser of the new results — is this OK?]

A typical cancer genome contains thousands of mutations ***[OK? To avoid having to define somatic]***, 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 ***[Ok to add? To clarify what non-coding elements are and to introduce the roles of promoters and enhancers, which need defining at some point]***. 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 ***[Simplification OK?]***. It is possible that our understanding of coding regions creates an ascertainment bias and makes it more likely that researchers search for mutations in coding regions ***[Simplification OK?]***. Nevertheless,



there has been great interest in finding non-coding drivers[4] ***[OK to shorten? As we don't go on to discuss methods]***. Previous studies have provided a few examples[5–7], but our understanding of non-coding drivers is far from complete. ***[Changes to shorten this section OK? Interested specialists can look to the papers to find the gene names]***.

Rheinbay *et al.* set out to identify coding and non-coding driver mutations in an unbiased fashion, using cells from a cohort of 360 people who had breast cancer. To find non-coding drivers, the researchers measured the rate at which mutations typically arose across the whole genome, then searched for 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 ***[OK to move up? To begin the results by explaining how they set out to find non-coding drivers. Also, expanded explanations OK? Please amend as needed for accuracy]***.

[OK to delete a sentence? To avoid repetition with that about hotspots later in this paragraph] 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 nucleotides that are mutated in multiple patients ***[Expansion OK?]***) 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 ***[OK?]*** can ~~simply~~ be attributed to the fact that they are smaller — they account for fewer nucleotides in the analysis.

This work is state-of-the-art, but there is more still to do. ***[OK to move the following sentence down? To lead into the discussion on power]*** The authors' power analysis — a statistical calculation that predicts the sample numbers needed to detect an effect of a given size ***[Definition of a power analysis OK?]*** — 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) ***[OK? Or please could you provide a number to give an idea of what fairly large means in this context?]***. 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 or DNA-packaging signatures — are typically noisy. ***[I was a bit confused about how exactly functional genomics is used to annotate these elements – I've attempted to include a simple explanation, but please amend as needed for accuracy]***. The functional territory of a regulatory element can therefore be considerably smaller than is annotated. Calculations of mutation rates that take into account oversized regions can hinder driver identification ***[Simplification OK?]***. Power calculations show that restricting annotation to smaller, functionally relevant regions enhances power.

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 ***[I've rephrased to explain why, rather than to give examples; is this OK? Please amend if this explanation is inaccurate?]***. 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 during processing of messenger RNA, making links readily apparent ***[Simplification OK?]***. 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 ***[Does this shortened phrasing still capture your meaning? Please could you add a few words to spell out how understanding the connections between them improves annotation of functional territories?]***.

[OK to delete a section here and replace with the simplified description highlighted below? It's a bit complex for our format, and would need quite a bit of unpacking Please amend further as needed — I'm sure I've made mistakes here.] 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. More information might also be gained by analysing discontinuous functional regions that regulate one gene, increasing statistical power by enabling testing of just one hypothesis — that a mutation alters regulation of that gene. In this way, rarer drivers can be uncovered ***[OK? To finish this section by reminding readers of why we want to increase power]***.

Once driver mutations are identified ***[correct?]***, the next challenge is to evaluate their effect. 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 currently dependent on a uniform background mutation rate. However, this is not the case for wide expanses of the genome[9], so the approach will require further refinement ***[OK to shorten? To avoid defining chromatin and explaining how the signals causes mutational changes]***.

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 ***[OK to shorten? As it's not clear why one in particular should be singled out here]***. 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 ***[What do you mean by large scale in this context? That they look at many tissues? Are genome-wide? Or something else?]*** annotation compendiums such as the ENCODE project[10] have a vital role to play.

[The current ending feels a bit negative — could you add a couple of sentences to return to bring the discussion back to the current paper, highlighting the advance that it makes and explaining how it puts us on the road to overcoming these hurdles?]

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: 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).

[Thanks for your figure suggestion. I like part a, and I think that visualizing the genome in this way will help readers get to grips with the concepts in the piece. However, I propose that we use just part a. To explain, the power analysis is a bit complex for a News & Views — the graph would require quite a bit of explanation and readers don't need to understand how the graphs work to follow the piece as a whole. Instead, we can simply state in the caption how power can be increased. In addition, I propose that we omit the CRE peak — again, we'd need to explain to readers what it represents, which isn't needed to follow this piece. The zoom-in you've included that highlights the size of the functional territory will be enough for readers to get the idea. Is this OK?]

[I've made changes to your figure caption to take into account my proposed modifications and our house style. This dictate that captions should stand alone from the main text and mention everything depicted. Please amend further as needed, bearing these restrictions in mind]

- **Figure 1| Improving discovery of cancer-driving mutations in the non-coding genome.** Genes contain coding sequences called exons, the links between which are well established — the messenger RNA that they encode is amalgamated after transcription. Gene expression is regulated by non-coding elements, including nearby promoters and distant enhancers. The links between these regulatory elements and genes are less well understood. Rheinbay *et al.*[3] conducted a systematic, unbiased analysis of 360 breast-cancer genomes to identify genetic mutations in non-coding sequences that drive cancer progression, and found nine such mutations in promoters. In the future, more non-coding drivers could be found by analysing more sequences, or by better understanding the links between non-coding elements and genes. In

addition, regions annotated as non-coding elements are often much larger than the actual regulatory sequence within the element. Limiting the regions analysed could improve driver identification.