IJC
International Journal of Cancer

Using common variants to indicate cancer genes

Lucy F. Stead[1], Helene Thygesen[1], David R. Westhead[2] and Pamela Rabbitts[1]

[1] Leeds Institute of Cancer and Pathology, University of Leeds, St James's University Hospital, Leeds, United Kingdom
[2] Institute of Molecular and Cellular Biology, University of Leeds, Leeds, United Kingdom

The catalogue of tumour-specific somatic mutations (SMs) is growing rapidly owing to the advent of next-generation sequencing. Identifying those mutations responsible for the development and progression of the disease, so-called driver mutations, will increase our understanding of carcinogenesis and provide candidates for targeted therapeutics. The phenotypic consequence(s) of driver mutations cause them to be selected for within the tumour environment, such that many approaches aimed at distinguishing drivers are based on finding significantly somatically mutated genes. Currently, these methods are designed to analyse, or be specifically applied to, nonsynonymous mutations: those that alter an encoded protein. However, growing evidence suggests the involvement of noncoding transcripts in carcinogenesis, mutations in which may also be disease-driving. We wished to test the hypothesis that common DNA variation rates within humans can be used as a baseline from which to score the rate of SMs, irrespective of coding capacity. We preliminarily tested this by applying it to a dataset of 159,498 SMs and using the results to rank genes. This resulted in significant enrichment of known cancer genes, indicating that the approach has merit. As additional data from cancer sequencing studies are made publicly available, this approach can be refined and applied to specific cancer subtypes. We named this preliminary version of our approach PRISMAD (polymorphism rates indicate somatic mutations as drivers) and have made it publicly accessible, with scripts, *via* a link at www.precancer.leeds.ac.uk/software-and-datasets.

Cancer develops *via* the accumulation of somatic mutations (SMs), some of which confer a selective advantage to the tumour, enabling it to proliferate abnormally. Distinguishing such driver mutations, which highlight candidate genes for targeted therapeutics, from passenger mutations (nonpathological by-products of the underlying mutagenic process) is an important task. Two main approaches exist: (*i*) prioritise mutations predicted to detrimentally affect an encoded protein[1] and (*ii*) identify genes repeatedly mutated within, or across, cancer subtypes.[2] The latter results from the hypothesis that SMs in genes causally associated with cancer undergo positive selection in tumours, occurring more often than expected by chance. Scoring this requires determination of the background mutation rate (BMR), given the commonly hypermutated state of cancer genomes, from which to measure the significance of the mutation count in a given gene. Often the rate of synonymous SMs, scaled by the ratio of potential nonsynonymous:synonymous mutations, is used, under the assumption that synonymous mutations are selectively neutral (*i.e.*, phenotypically silent).[2] This is flawed: (*i*) it restricts analysis to protein-coding genes and (*ii*) the assumption of selective neutrality is increasingly hard to justify owing to the prevalence of functional noncoding transcripts. Nonprotein coding genes include microRNAs (miRNAs), long intergenic noncoding RNAs (lincRNAs) and pseudogenes, all shown to have causal associations with various cancers.[3] The functionality of these transcripts results directly from nucleotide sequence, rather than encoded amino acids, based on binding other nucleotides or proteins in a sequence-specific manner.[4] Genetic variation within noncoding transcripts will, therefore, alter their functionality with potential phenotypic consequences, but the notion of nonsynonymous and synonymous variation does not apply. Additionally, synonymous mutations in protein-coding genes can exert a phenotypic effect by altering the resulting mRNA's ability to (*i*) interact with regulatory noncoding RNAs or (*ii*)

Key words: next-generation sequencing, cancer driver genes, somatic mutation

Abbreviations: BMR: background mutation rate; COSMIC: catalogue of somatic mutations in cancer; CP: common polymorphism; lincRNA: long intergenic noncoding RNA; miRNA: microRNA; PRISMAD: polymorphism rates indicate somatic mutations as drivers; RD: rate difference; SM: somatic mutation; UCSC: University of California, Santa Cruz

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Yorkshire Cancer Research; **Grant number:** L341PG
DOI: 10.1002/ijc.28951
History: Received 29 Sep 2013; Accepted 2 Apr 2014; Online 5 May 2014

Correspondence to: Lucy F. Stead, Leeds Institute of Cancer and Pathology, University of Leeds, St James's University Hospital, Leeds LS9 7TF, United Kingdom, Tel.: +44-113-2065-627, Fax: +44-113-3438-601, E-mail: l.f.stead@leeds.ac.uk

Short Report

What's new?
Somatic mutations are important drivers of the cancerous process but identifying the key "driver" mutations remains a challenging question. The authors hypothesize that the variation level in healthy tissue represents a transcript's tolerance to mutation and that if the number of mutations in tumors exceeds this level, positive selection might have occurred that point to this transcript as a major driver in carcinogenesis. They tested their program with a large dataset of somatic mutations and obtained a ranked list of genes significantly enriched in known cancer-associated genes. Their program called PRISMAD is publicly available and could help identify new driver mutations in various tumors.

fold stably, directly affecting translation.[5] We hypothesise that the amount of variation observed within a transcript in nondiseased tissue is a measure of that transcript's tolerance to mutation, and that if the number of observed tumour-specific mutations exceeds this level it suggests positive selection in the tumour and a possible role for that transcript in carcinogenesis. This hypothesis is best tested in a tissue-specific manner, as mutational signatures in cancer vary by subtype owing to different mutagen exposure and disease processes.[6] However, this requires a database of known SMs in nondiseased tissue of origin of the cancer in question. Such a database does not currently exist as blood is most commonly used as the matched normal for genomic sequencing. However, these data will likely be available in future owing to RNA sequencing, which often includes a matched nondiseased tissue of origin to provide an expression baseline.

As the most ideal datasets for testing our hypothesis are not available, we opted to investigate whether the rate of common human germline variation [*i.e.*, the common polymorphism (CP) rate] could provide an alternative BMR for scoring SM rates in cancer genomes. Most oncogenes and tumour suppressor genes are highly conserved within mammals, indicating the important physiological roles of those genes. Similarly, noncoding regions from which functional transcripts are transcribed are often conserved.[7] A single mutation within any evolutionary constrained region could be responsible for detrimental phenotypic changes and is, therefore, unlikely to be commonly observed within the human germline. Our adapted hypothesis is that any genomic region in tumours that harbours SMs more often, relatively, than it harbours CPs is a candidate for carcinogenesis. To test this, we ascertained the CP rate within humans and compared it to the SM rate, using tumour-specific SMs identified from sequencing studies.

Material and Methods

Genome annotation

Annotations for human reference genome GRCh37 were downloaded from GENCODE14[8] and transcript records merged, *via* a bespoke perl script, creating a single annotation per gene ID with nonredundant exons delineated.

Population data

A bespoke script (available online) accessed Ensembl69,[9] *via* its perl programming interface, and extracted the total number of basepairs, and the number of commonly polymorphic loci, within each exon. A commonly polymorphic locus is a variant position sequenced in germline samples at least 20 times with a minor allele frequency of 5–50% (see Supporting Information for justification of the chosen allele frequency). The rate (commonly polymorphic alleles per kilobase) is calculated and output per gene. Ensembl69 contained information from dbSNP137, including all data from the 1000 Genomes Project phase 1 and HapMap phase 3.

Somatic mutations

We use SM to mean a tumour-specific substitution or indel involving less than 500 bp. Genome coordinates were converted, where necessary, to GRCh37 using the University of California, Santa Cruz (UCSC) liftOver tool. SMs were downloaded from catalogue of somatic mutations in cancer (COSMIC)62 *via* Biomart or extracted from Supporting Information in additional publications (with no study overlap).[10] All COSMIC SMs were validated from primary tumours and identified using whole genome sequencing. All manually extracted data were from whole genome or exome sequencing studies only. Supporting Information Table 1 outlines all references for the SMs collated. A bespoke perl script (available online) was used to ascertain the SM rate using our amended genome annotation files. Analysis was restricted to exon regions.

Statistical analysis

CP and SM rates were analysed in R. Attempts to ascertain the best way to amalgamate the CP rate and SM rate information into a single metric are given in Supporting Information. Functional analysis was performed using the DAVID Bioinformatics Resources, release 6.7.[11] Statistical tests were performed in R.

Comparison with other programmes

The Supporting Information contains information on a comparison between PRISMAD (polymorphism rates indicate somatic mutations as drivers) and another programme that is applicable to noncoding regions.

miRNA folding predictions

The fasta sequence for the wild-type and mutant miRNA precursor, hsa-mir-99b, were input to RNAfold.[12] Resulting predictions are those according to the minimum free energy and partition function. Free energy values were output for

Short Report

Table 1. Highlighting genes that contain candidate somatic driver mutations in different functional classes

Class of gene	Total	Mean RD (variants/kb)	Median RD (variants/kb)
Protein-coding	20,036	−1.49	−1.16
lincRNA	6,296	−2.80	−2.24
miRNA	3,110	−2.67	0
lncRNA	786	−2.43	−1.96
Pseudogene	13,004	−2.69	−1.83

RD: rate difference (somatic mutation rate minus common polymorphism rate).

each structure and used to ascertain the change between wild-type and mutated sequences.

PRISMAD web server

The web server was written in PHP and html.

Results

We define a CP as one with a minor allele frequency of at least 5% at a locus genotyped at least 20 times. This threshold performed best amongst those tested (Supporting Information). The CP rate per gene is the number of exonic CPs divided by the number of exonic kilobases. Using genome annotation files, we separated genes into functional classes: protein-coding, antisense, long-intergenic noncoding (linc)RNA, long noncoding (lnc)RNA, micro (mi)RNA and pseudogene. The lncRNAs are distinct from lincRNAs in that they are located within genes; they are on the sense strand, making them distinct, also, from antisense genes.

Inspecting rates of variation in cancer genes

We hypothesise that germline mutations in cancer-associated genes are more likely to be phenotypically detrimental (owing to effects at the RNA as well as protein level) and are, thus, more likely to be selected against leading to a reduced CP rate in cancer-associated genes compared with noncancer-associated genes. This is similar to the notion that driver SMs will undergo positive selection within the tumour leading to a higher SM rate in cancer-associated genes compared with noncancer-associated genes in the tumour. To test this we ascertained the list of 483 known, protein-coding cancer genes from the Cancer Gene Census.[13] The median CP rate was 1.21 CP/kb for cancer genes and 1.61 CP/kb for noncancer genes. In agreement with our hypothesis, the CP rate for cancer genes was significantly lower (Wilcoxon, p: 1.28×10^{-12}). The median SM rate was 0.40 SM/kb for cancer genes and 0.31 SM/kb for noncancer genes. As expected the SM rate is significantly higher in cancer genes (Wilcoxon, p: 1.63×10^{-5}) but, interestingly, the effect size is not as large as for CP rate. We wished to use this information to rank somatically mutated genes with respect to the likelihood that they are causally associated with cancer. We attempted several statistical modelling approaches (Supporting Information), concluding that the best results were obtained using created a metric we called the rate difference (RD), obtained by subtracting the CP rate from the SM rate:

$$\mathrm{RD} = \mathrm{SM\ rate} - \mathrm{CP\ rate}. \quad (1)$$

The median RD for cancer genes was −0.83 variants/kb and for noncancer genes −1.17 variants/kb. The RD is significantly higher for cancer genes and the effect is greater than that of both SM rate and CP rate in isolation (Wilcoxon, p: 1.04×10^{-14}).

Using RD to rank genes, genome-wide

Our approach is applicable genome-wide as it uses the CP rate, which can be ascertained for any given genomic region, as a baseline for interpreting SM rates. We calculated the RD for each of 20,036 protein-coding genes, 6,296 lincRNAs, 3,110 miRNAs, 786 lncRNAs and 13,004 pseudogenes (Table 1 and Supporting Information Tables 2 and 3). The top 1,000 protein-coding genes (*ca*. 5%), ranked by descending RD, included significantly more known cancer genes than expected by chance (χ^2, p: 0.00028), whereas the top 1,000 ranked by descending SM rate did not (χ^2, p: 0.38). This indicates that RD is a more powerful predictor than SM rate alone. This enrichment was not observed if the datasets were separated into synonymous and nonsynonymous variants (χ^2, $p > 0.01$, Supporting Information).

Functional analysis of the top 1,000 protein-coding genes according to RD revealed significant enrichment in the pathways of cadherin signalling (PANTHER P00012, adjusted $p < 0.05$) in which 21 members were highlighted (Supporting Information Table 4), and Wnt signalling (PANTHER P00057, adjusted $p < 0.05$), with 34 members highlighted (Supporting Information Table 5).

The top-ranking noncoding transcripts mostly lacked a single exonic CP, with only 76 containing more than one SM (Supporting Information Table 3). Literature searches revealed a dearth of information regarding the functionality of the top-ranking noncoding transcripts according to RD except in the case of *MIR99B*. This is a miRNA with an RD of 14.5 variant/kb owing to an SM identified in a gastric tumour.[14] The *MIR99B* gene produces two mature miRNAs (hsa-miR-99b-3p and hsa-miR-99b-5p); the dysregulation of both has been associated with carcinogenesis.[15] Mutations within miRNAs can have causal associations with cancer.[16] The SM, NC_000019.9:g.52195904G>A, highlighted by our approach resides within a predicted base-paired portion of the hsa-mir-99b precursor hairpin from which the two mature miRNAs are excised (Fig. 1*a*). The mutation is predicted to alter precursor folding in such a way that removes local base pairing and causes a predicted reduction in folding stability by 1.24 kcal/mol program (Fig. 1*b*). This altered configuration and change in stability could alter the processing of the hairpin, required to excise the mature miRNAs.

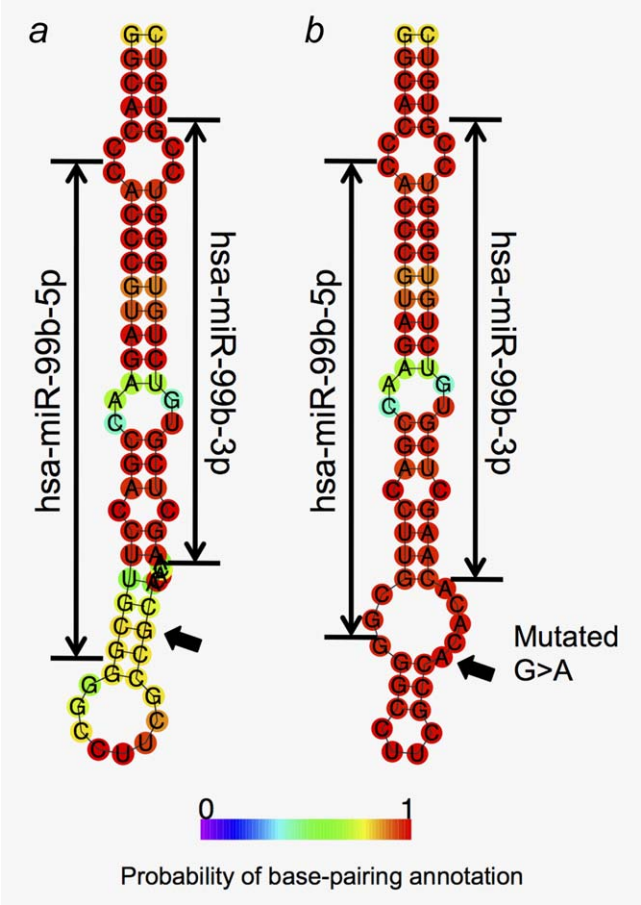

Figure 1. Predicted folding of the hsa-mir-99b precursor in wild-type (*a*) and somatically mutated (*b*) form. The locations of the mature miRNAs (had-miR-99b-3p and had-miR-99b-5p) that are excised from the precursor are annotated. The colouring indicates the probability of base pairing as indicated by the scale bar. The location of the variant position is given by the block arrow, with the change in the mutant sequence labelled on the figure. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Discussion

Machine learning methods that identify disease-causing nonsynonymous mutations reveal that evolutionary conservation is an, if not the most, important predictor variable.[1,17] This is because genetic variants that detrimentally alter the function of an encoded protein undergo negative selection throughout evolution. It follows that genetic variation that detrimentally alters the noncoding function of a transcript will also undergo negative selection. We tested the hypothesis that the number of commonly occurring germline polymorphisms within a genomic region (protein coding and/or noncoding) can be used as a BMR from which to score SM rates in tumours and identify potential cancer-driving genes. In support of this theory, exonic SNP density (number of polymorphic loci) is one of the most informative predictive features in a machine-learning tool to predict cancer-driving mutations.[1] We developed a parsimonious method for ranking genes/genomic regions using the RD [Eq. (1)]. Our method is not restricted to protein-coding regions and makes no prior assumptions regarding which mutations are phenotypically silent.

General cancer pathways

The top ranked genes highlighted by PRISMAD were enriched for Wnt signalling and cadherin signalling pathways. Wnt signalling is involved in cell–cell communication and its study is becoming increasingly widespread in cancer research.[18] Similarly, the role of cadherins in various types of cancer continues to be an area of active research.[19] The elucidation of cancer-related pathways by our approach further indicates its merit.

Application to noncoding genes

Attempts to investigate noncoding SMs thus far have been on the level of specific mutations within single samples, without reproducibility, or have been anecdotal. In those cases, though, it has been stressed that it is likely that some drivers will lie within noncoding regions.[20,21] We applied our method to several types of noncoding genes implicated in carcinogenesis: lincRNAs, miRNAs, lncRNAs and pseudogenes. We revealed few CPs in many of these genes. This is expected given that interest in noncoding regions, and the ability to sequence them to the required depth, has only increased in the last decade, meaning there is a dearth of information on variation rates therein. The 1000 genomes pilot constituted whole genome sequencing, but thereafter the project focused on protein-coding regions.[22] We believe that although economically understandable, negation of noncoding regions may be detrimental to cancer research. Many whole genomes have been sequenced, with SM data deposited in relevant databases. Similar deposition of genome-wide germline variants into dbSNP would facilitate the creation and use of approaches such as ours.

Our approach is to highlight some noncoding genes as potentially harbouring driver SMs, but a lack of functional information makes these difficult to verify. Rather, we hope that validating our approach in protein-coding genes suggests the noncoding genes highlighted are worthy of prioritisation or, at least, when additional noncoding germline variation is present in online databases, ours is an approach worth applying. We highlighted one known cancer-associated miRNA gene, *MIR99B*, using PRISMAD, and predicted how the SM identified within it may result in altered processing and expression of two mature miRNAs.

Many methods exist to specifically identify nonsynonymous cancer-driving mutations. Our approach can be used alongside these, potentially highlighting distinct genes, but we do not propose our method replace them if the goal is to highlight nonsynonymous variants.

It has recently been shown that additional factors, *i.e.*, gene expression level and stage of replication, affect the number of tumour-specific SMs that a gene acquires, irrespective of involvement in carcinogenesis.[23] This is thought to result from DNA repair and replication effects: genes expressed at

low levels are less exposed to transcription-coupled repair, and those replicating late succumb to error owing to a reduction in the concentration of free nucleotides within the cell. These factors will equally affect mutation rates in noncancer cells, though to a lesser degree because these cells are not aberrantly proliferating. Our original hypothesis better incorporates these recent findings: an RD calculated from commonly polymorphic sites specifically within matched normal tissue (which will include germline and nondiseased tissue SMs) to the tumour in question will factor in the aforementioned biases, assuming that expression profiles and replication timing of such cells are similar to the cancer cells that originated from them. Unfortunately, there is currently insufficient publicly available appropriate sequencing data to test this extended hypothesis, but it provides an avenue for future research.

Whilst this manuscript was under review, an article indicating how patterns of polymorphisms within noncoding regions can be used as a basis for identifying cancer-driving SMs has been published in *Science*.[24] The authors describe a method that can be used in conjunction with ours: FunSeq. This approach filters out germline polymorphisms and then prioritises variants according to their location in conserved regions, binding motifs and (for protein-coding genes) hubs of gene networks. This work further highlights that, as the number of tumour-specific SMs increases, approaches aimed at scoring variation in noncoding regions are sorely needed.

Acknowledgements

This work was supported by Yorkshire Cancer Research (grant number L341PG to P.R.). The authors thank Cyriac Kandoth with his help in providing data for them to run the comparison outlined in the Supporting Information.

References

1. Carter H, Chen S, Isik L, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* 2009;69:6660–7.
2. Youn A, Simon R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* 2011;27:175–81.
3. Costa FF. Non-coding RNAs: new players in eukaryotic biology. *Gene* 2005;357:83–94.
4. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature* 2012;482:339–46.
5. Zur H, Tuller T. Strong association between mRNA folding strength and protein abundance in *S. cerevisiae*. *EMBO Rep* 2012;13:272–7.
6. Alexandrov L, Nik-Zainal S, Wedge D, et al. Signatures of mutational processes in human cancer. *Nature* 2013;500:415–21.
7. Washietl S, Hofacker I, Lukasser M, et al. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol* 2005;23:1383–90.
8. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012;22:1760–74.
9. Flicek P, Amode MR, Barrell D, et al. Ensembl 2011. *Nucleic Acids Res* 2011;39:D800–D806.
10. Forbes SA, Bindal N, Bamford S, et al. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res* 2011;39:D945–D950.
11. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2008;4:44–57.
12. Hofacker I. Vienna RNA secondary structure server. *Nucleic Acids Res* 2003;31:3429–31.
13. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. *Nat Rev Cancer* 2004;4:177–83.
14. Wang K, Kan J, Yuen ST, et al. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat Genet* 2011;43:1219–23.
15. Sun D, Lee YS, Malhotra A, et al. miR-99 family of microRNAs suppresses the expression of prostate-specific antigen and prostate cancer cell proliferation. *Cancer Res* 2011;71:1313–24.
16. Wu M, Jolicoeur N, Li Z, et al. Genetic variations of microRNAs in human cancer and their effects on the expression of miRNAs. *Carcinogenesis* 2008;29:1710–16.
17. Stead LF, Wood IC, Westhead DR. KvSNP: accurately predicting the effect of genetic variants in voltage-gated potassium channels. *Bioinformatics* 2011;27:2181–6.
18. Polakis P. Wnt signaling in cancer. *Cold Spring Harb Perspect Biol* 2012;4:2008052.
19. Heuberger J, Birchmeier W. Interplay of cadherin-mediated cell adhesion and canonical wnt signaling. *Cold Spring Harb Perspect Biol* 2010;2:2002915.
20. Pleasance ED, Cheetham RK, Stephens PJ, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 2010;463:191–6.
21. Mardis ER, Ding L, Dooling DJ, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *New Engl J Med* 2009;361:1058–66.
22. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061–73.
23. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;499:214–18.
24. Khurana E, Fu Y, Colonna V, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 2013;342:1235587.

Short Report