Nearing saturation of cancer driver gene discovery

David Hsiehchen¹ · Antony Hsieh²

Received: 26 April 2018 / Revised: 31 May 2018 / Accepted: 31 May 2018 / Published online: 15 June 2018
© The Author(s) under exclusive licence to The Japan Society of Human Genetics 2018

Abstract

Extensive sequencing efforts of cancer genomes such as The Cancer Genome Atlas (TCGA) have been undertaken to uncover bona fide cancer driver genes which has enhanced our understanding of cancer and revealed therapeutic targets. However, the number of driver gene mutations is bounded, indicating that there must be a point when further sequencing efforts will be excessive. We found that there was a significant positive correlation between sample size and identified driver gene mutations across 33 cancers sequenced by the TCGA, which is expected if additional sequencing is still leading to the identification of more driver genes. However, the rate of new cancer driver genes being discovered with larger samples is declining rapidly. Our analysis provides a general guide for determining which cancer types would likely benefit from additional sequencing efforts, particularly those with relatively high rates of cancer driver gene discovery. Our results argue that past strategies of indiscriminately sequencing as many specimens as possible for all cancer types is becoming inefficient. In addition, without significant investments into applying our knowledge of cancer genomes, we risk sequencing more cancer genomes for the sake of sequencing rather than meaningful patient benefit.

The molecular basis of cancer is being unraveled by extensive sequencing efforts of cancer genomes such as those being undertaken by The Cancer Genome Atlas (TCGA). Revealing alterations in cancer genomes has led to a more comprehensive understanding of cancer biology, exposed therapeutic targets, and enhanced the classification of cancer types beyond phenotypes. However, the abundance of cancer mutations has made the identification of bona fide driver genes a nontrivial task, and it remains unclear whether we are close to identifying all driver genes for each cancer [1].

Given that the number of driver genes is bounded by the number of human protein coding genes, it is expected that nearly all clinically significant driver genes will be discovered by sequencing enough cancers and additional efforts would be excessive. In an earlier analysis, Lawrence et al. advocated that between 650 and 5300 samples for

each cancer type will be needed to provide sufficient power to detect genes mutated in at least two percent of cancers [2]. In fact, their down-sampling tests showed that the number of driver cancer genes across cancer types grew nearly linearly with the number of samples sequenced. At the time, their data set reported the identification of up to 254 driver genes based on an analysis of 4742 samples across 21 cancer types available up to 2014. However, the 2018 TCGA analysis on 9423 samples across 33 cancer types revealed only 299 driver genes, falling short of the expected 50% increase in driver genes. This possibly suggests that we may already be approaching saturation in driver gene discovery efforts with much fewer specimens analyzed than previously thought to be needed [3].

To assess the relationship between the intensity of sequencing efforts and the number of driver genes identified, we plotted the number of samples sequenced by TCGA for different cancers against the number of driver genes identified from those samples by a consensus of at least three driver gene identification algorithms (Fig. 1). Driver gene algorithms are listed in Table 1 and discussed in detail by Cheng et al. [4] There was a statistically significant positive correlation across 33 cancer types ($p = 0.001$). A correlation coefficient of 0.53 indicated that 28% of variance in the number of driver genes between different cancer types is potentially explained by the number of samples sequenced. While there is no biological mechanism

✉ David Hsiehchen
gbtwnow@gmail.com

¹ Division of Hematology and Oncology, Department of Medicine, University of Texas Southwestern Medical Center, Dallas, TX, USA

² Division of Gastroenterology, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

for the number of driver genes of different cancers to be related to sample size, it is expected that a correlation would exist when the number of driver genes initially increases with additional sampling. This association would then be expected to become non-significant, as the explained variance of driver genes by sampling diminishes with

Fig. 1 Sample size and the number of cancer driver genes identified for 33 cancer types. The grey line depicts a power law defined by a constant of 4.976 and exponent of 0.56

oversaturation. Thus, the explained variance of the number of driver genes for different cancer types by the number of samples may be informative of the degree of saturation in sequencing. Supporting this notion, the correlation between sample sizes and driver genes using the available TCGA data from 2014 with fewer samples was characterized by a correlation coefficient of 0.67, suggesting that 45% of variance in the number of driver genes between different cancer types was due to sample sizes at the time.

Notably, the scaling relationship between sample size and cancer driver genes followed a power law distribution with a scaling exponent of 0.561. Similar results were obtained when the number of cancer driver genes was determined by at least two to four published computational algorithms widely used to annotate likely functional cancer mutations [4, 5]. A scaling exponent less than one indicates that with an increase in the number of samples sequenced across different cancer types the rate of driver gene discovery increases at a slower pace. This implies that sequencing efforts are already encountering decreasing rates of return in terms of discovering new driver mutations. Consistent with this, the power law relationship between

Table 1 Driver gene identification algorithms

Driver gene algorithm	Reference
ActiveDriver	Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. <i>Mol Syst Biol.</i> 2013;9:637
CoMDP	Zhang J, Wu LY, Zhang XS, Zhang S. Discovery of co-occurring driver pathways in cancer. <i>BMC Bioinformatics</i> 2014;15:271
DawnRank	Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. <i>Genome Med.</i> 2014;6:56
Dendrix	Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. <i>Genome Res.</i> 2012;22(2):375–85
DriverNet	Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. <i>Genome Biol.</i> 2012;13:R124
e-Driver	Porta-Pardo E, Godzik A. e-Driver: a novel method to identify protein regions driving cancer. <i>Bioinformatics</i> 2014;30(21):3109–14
iPAC	Ryslik GA, Cheng Y, Cheung KH, Modis Y, Zhao H. Utilizing protein structure to identify non-random somatic mutations. <i>BMC Bioinformatics</i> 2013;14:190
MDPFinder	Zhao J, Zhang S, Wu LY, Zhang XS. Efficient methods for identifying mutated driver pathways in cancer. <i>Bioinformatics.</i> 2012;28:2940–7
MEMo	Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. <i>Genome Res.</i> 2012;22:398–406
MSEA	Jia P, Wang Q, Chen Q, Hutchinson KE, Pao W, Zhao Z. MSEA: detection and quantification of mutation hotspots through mutation set enrichment analysis. <i>Genome Biol.</i> 2014;15:489
MutSigCV	Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutation heterogeneity in cancer and the search for new cancer-associated genes. <i>Nature.</i> 2013;499:214–8
NetBox	Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. <i>PLoS ONE</i> 2010;5:e8918
OncodriveCLUST	Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveClust: exploiting the positional clustering of somatic mutations to identify cancer genes. <i>Bioinformatics.</i> 2013;29:2238–44
OncodriveFM	Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. <i>Nucleic Acids Res.</i> 2012;40:e169
Simon	Youn A, Simon R. Identifying cancer driver genes in tumor genome sequencing studies. <i>Bioinformatics</i> 2011;27:175–81

sample sizes and driver genes using the 2014 TCGA data demonstrated a scaling exponent of 1.09. This supralinear scaling indicates that early sequencing efforts were associated with increasing rates of returns in the number of driver genes discovered. Although further genomic investigation of uncharacterized cancers and perhaps even some prevalent cancers would be fruitful, the decline in both the scaling exponent and explained variance with an increasing number of cancers sequenced supports the hypothesis that sequencing efforts are approaching saturation.

The analysis of different rates of cancer driver gene discovery between different cancers may aid the determination of which cancer types may benefit most from greater sample sizes. Cancer types lying above the plotted best fit line have a relatively faster driver gene discovery rate, implying that current sequencing efforts are likely not close to saturation. This must be cautiously interpreted as it is not expected that different cancer types will have the same number or distribution of driver genes which would impact the rate of discovery. Nonetheless, our plot suggests that outliers including diffuse large B cell lymphoma, lung adenocarcinoma, lung squamous cell carcinoma, and adrenocortical carcinoma may benefit from additional sequencing of specimens, which is congruent with prior power analyses and the greater tumor mutation burden of these cancers [2, 6].

It is undeniable that large scale genomic projects are valuable, but they are associated with very significant investments of money, time, and expertise. The opportunity costs of such investments will likely remain underestimated as research policies and funding in the past decade has been driven by the prevailing belief that more data is better [7, 8]. Even if the cost of genome sequencing is no longer prohibitive, the deluge of data from sequencing efforts is expected to mushroom requiring additional time, hardware, and expertise to further analyze and store. Our results argue that past strategies of indiscriminately sequencing as many specimens as possible for all cancer types is inefficient. In this time of austerity in research funding, it makes sense that we should be evaluating whether advancements from further sequencing efforts are truly innovative or merely

incremental. In addition, unless it can be demonstrated that cancer genomics can alter or improve clinical practice, we run the risk of sequencing for the sake of sequencing rather than for meaningful patient benefit. Thus, it is essential that future research assess how genomic data can be broadly applied such as stratifying patients to targeted therapies, particularly as the sparse efforts to do so in large clinical studies thus far have resulted in ambiguous conclusions [9].

Acknowledgements DH designed the study. DH and AH analyzed the data and wrote the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science* 2013;339:1546–58.
2. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 2014;505:495–501.
3. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* 2018;173:371–85.
4. Chung IF, Chen CY, Su SC, Li CY, Wu KJ, Wang HW, et al. DriverDBv2: A database for human cancer driver gene research. *Nucleic Acids Res.* 2016;44(D1):D975–9.
5. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci USA* 2016;113:14330–5.
6. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* 2017;9:34.
7. Weinberg R. Point: Hypotheses first. *Nature* 2010;464:678.
8. Yaffe MB. The scientific drunk and the lamppost: massive sequencing efforts in cancer discovery and treatment. *Sci Signal* 2013;6:pe13.
9. Le Tourneau C, Delord JP, Goncalves A, Gavoille C, Dubot C, Isambert N, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *Lancet Oncol.* 2015;16:1324–34.