Zheng *Genome Biology* (2017) 18:129
DOI 10.1186/s13059-017-1258-3

Genome Biology

CORRESPONDENCE Open Access

Benchmarking: contexts and details matter

Siyuan Zheng

Please see related Li et al. and Newman et al. correspondence articles: www.dx.doi.org/10.1186/s13059-017-1256-5 and www.dx.doi.org/10.1186/s13059-017-1257-4

Abstract

Benchmarking is an essential step in the development of computational tools. We take this opportunity to pitch in our opinions on tool benchmarking, in light of two correspondence articles published in *Genome Biology*.

Technological advances have profoundly changed cancer research over the past two decades. Increasingly affordable, it is now routine to sequence the genome and transcriptome of a sample. As a result, we are accumulating unprecedented volumes of data. For instance, the Genomic Data Commons hosted by the National Cancer Institute contains 14,551 cancer samples totaling a storage space of 5 petabytes as of May 2017 [1]. The sheer amount suggests data-driven approaches are more viable than ever to provide new insights in cancer biology.

Computational tools are the bridge between these data and insight. Algorithms are designed to answer questions that are either experimentally too laborious or even infeasible. Consequentially, we have tools to nominate driver genes by analyzing patterns of genome-wide somatic mutations [2]; to discover molecular subtypes with distinct expression, copy number, and mutational features [3]; and to predict many fundamental characteristics of cancer samples, such as the average telomere length [4], purity [5], and even the abundance of infiltrating stromal and immune cells [6]. The power of such predictions lies in the amount of data, as associating these characteristics with cancer genomic and clinical parameters in thousands of cases permits the discovery of the subtlest links that otherwise would be buried deep in the data.

Thanks to a vibrant community, we usually have arrays of tools developed to address the same sets of questions. The availability of alternatives is critically important as no tool is guaranteed to grasp the full complexity of cancer genomic data. Take mutation calling, for example: more than a dozen mutation detection tools are available on the market [7]. The choice of which tool to use in a project often relies on popularity, ease of use, and demand for resources. While performance should be the best measure, it is more often that a tool is suggested to outperform competitors in benchmarking experiments conducted by its own developers. In light of this, community efforts such as the ICGC-TCGA DREAM challenges [8] are being established to provide a common benchmark reference to ensure fair and transparent comparisons.

For less widespread applications a consensus benchmark is likely not readily available. This does not suggest in any way that the object of the research is less significant. An example is deconvolution of bulk tumor expression. Pathologists have long learned that tumors are admixtures of malignant cells, fibroblasts, blood vessels, and immune cells. Systematic analysis demonstrates a continuum of tumor impurity, not only within a cancer type but also across cancer tissues of origin [5]. The tools discussed in two correspondence articles [9, 10], CIBERSORT and TIMER, developed by Newman et al. [11] and Li et al. [12], respectively, both aim to delineate admixture at the transcriptome level into immune cell types. While CIBERSORT takes on 22 cell types, TIMER focuses on 6, arguing that including too many variables would introduce statistical collinearity and lead to non-biological associations. Furthermore, Li et al. [12] assert that CIBERSORT was developed based on microarrays and is thus unsuitable for the analysis of RNA-seq data. In response, Newman and colleagues disputed these arguments and suggested that TIMER failed to normalize the immune cell estimates for total amount of leukocytes, which they believe could explain the observed discrepancies between the results generated by the two tools [9].

Correspondence: syzhenger@gmail.com
Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

These correspondence articles clarify important details of both tools, which in our opinion are beneficial to end users. Furthermore, they invite discussions on what are the best strategies and common pitfalls in benchmarking computational tools. Here we summarize, in our view, useful sources for benchmarking experiments.

1. Ground truth. A dataset produced by gold standard experimental approaches is the strongest evidence for validation. To name a few: fluorescence in situ hybridization (FISH) for the validation of absolute copy number predictions, and the quantitative PCR-based telomeric repeat amplification protocol (TRAP) assay for the validation of telomerase activity prediction, etc. In some circumstances the property of interest is challenging to quantify and experiments can be designed to create an artificial reference. For example, a dilution series is able to provide an artificial sample with known proportions of mixtures, which can be compared to computational estimates such as purity. When using an experimental dataset as the benchmarking reference, it is important to keep in mind the context in which computational and experimental data align. As an exaggerated example, one would be ridiculed to validate a novel driver mutation found in breast cancer in a glioma cell line. In reality, however, we are tempted to make extrapolations based on limited sets of experiments. Acknowledging that we cannot practically test all possible scenarios, we stress that context should be carefully considered when applying or designing experimental data for benchmarking.
2. Simulation. Often times a gold standard is simply impossible to get, and this dilemma applies to many cancer genome analyses. For instance, we do not have a complete list of bona fide DNA structural rearrangements (SVs) or somatic mutations that allow us to evaluate the sensitivity and specificity of calling tools, albeit it is straightforward to validate the returned candidate events. In cases like this, we resort to other means, such as simulation. Generating a good simulation dataset takes deep understanding of the simulated object and addition of appropriate noise levels. Details are important as to how well the simulated data reflect the real case scenario. We encourage all authors to share their code on public platforms such as GitHub or SourceForge to enhance transparency and reproducibility regardless of requirements from publishers.
3. Literature and public resources. Literature represents a wealth of information. However, it is not uncommon to find contradictory evidence from the literature to either prove or disprove a conclusion. For this reason, citing one or a few references does not significantly strengthen benchmarking, in our view. Knowledge bases such as the Cancer Gene Census [13] are expert curations of publications and are more reliable sources. The same rationale might be extended to databases that collect data through text mining, which replaces human subjectivity with sematic parsing and machine learning. Benchmarking carries more weight as long as it is done with systematically collected data rather than hand-picked examples.
4. Other tools. Competing tools are usually compared with rather than benchmarked to. However, consensus results returned by multiple tools are presumably more reliable if they diverge in their premise and technical details. A consensus can be intuitively defined when two tools are enlisted but the complexity rises sharply upon inclusion of multiple tools, in which case careful evaluation is warranted to ensure proper stringency. In another setting, a new tool can be benchmarked to an existing tool addressing the same question but with different modalities of input data. This is particularly useful as such comparisons may provide insights beyond benchmarking.

Despite its importance, benchmarking is only a part of the tool development cycle. End users should be invited to the dialogue as they are the ones that apply and test tools in projects where they possess intimate knowledge that tool developers may not have. In this sense benchmarking of a tool occurs not only before but also after its publication.

Acknowledgements
We thank Dr Xiaojing Wang from Baylor College of Medicine and Dr Zechen Chong from the University of Alabama at Birmingham for critical reading of the manuscript.

Funding
SZ is supported in part by the NIH grant P50CA127001 Career Enhancement Award at MD Anderson Cancer Center.

Competing interests
The author declares that he has no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 2 June 2017 Accepted: 12 June 2017
Published online: 05 July 2017

References

1. NIH: Cancer Institute. Genomic Data Commons. 2017. https://gdc.cancer.gov/. Accessed 31 May 2017.
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. Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, et al. Comprehensive pan-genomic characterization of adrenocortical carcinoma. Cancer Cell. 2016;29:723–36.
4. Barthel FP, Wei W, Tang M, Martinez-Ledesma E, Hu X, Amin SB, et al. Systematic analysis of telomere length and somatic alterations in 31 cancer types. Nat Genet. 2017;49:349–57.
5. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30:413–21.
6. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.
7. OMIC TOOLS. Somatic SNV detection. 2017. https://omictools.com/somatic-snp-detection-category. Accessed 31 May 2017.
8. Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods. 2015;12:623–30.
9. Newman AM, Gentles AJ, Liu CL, Diehn M, Alizadeh AA. Data normalization considerations for digital tumor dissection. Genome Biol. 2017;18: doi: 10.1186/s13059-017-1257-4
10. Li B, Liu JS, Liu XS. Revisit linear regression based deconvolution methods for tumor gene expression data. Genome Biol. 2017;18: doi: 10.1186/s13059-017-1256-5
11. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7.
12. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17:174.
13. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4:177–83.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central