Identifying somatic mutations from tumor-only sequencing using deep learning

Kiran Krishnamachari[a], **Bui Ngyuen Huu An**[a], **Anders Jacobsen Skanderup**[a]

[a] *Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), 60 Biopolis Street, Genome, Singapore 138672, Republic of Singapore* skanderupamj@gis.a-star.edu.sg

* Presenting author

1. Introduction

Somatic variant calling algorithms detect somatic mutations in cancer genomes by analyzing sequencing data from tumor samples. While this is typically performed by comparing against a matched normal sample, such samples are often unavailable in clinical diagnostics or retrospective analyses of archival tumor samples in biobanks. The lack of a matched normal sample significantly affects variant calling accuracy due to the difficulty in distinguishing somatic mutations from germline mutations as well as sequencing artifacts. This can negatively impact cancer treatment decisions, such as using tumor mutation burden to guide immunotherapy recommendations, which depend on accurate variant calling—often derived from tumor-only sequencing. Here, we present VarNet-T, an enhancement over our previously published method, VarNet [1], an end-to-end weakly supervised deep learning framework for accurately identifying somatic variants from aligned tumor reads without a matched normal sample and trained using millions of high-confidence variants. We benchmarked VarNet-T on publicly available whole genome benchmark tumor sequencing datasets, demonstrating performance significantly exceeding existing methods. Notably, VarNet-T achieves an absolute improvement of 20% and 33% in area under the precision-recall curve over the next best method on two publicly available benchmark datasets: the FDA created SEQC2 and COLO829, respectively. Overall, the improved accuracy of our method has the potential to enhance the utility of tumor-only sequencing in precision medicine and cancer genomics.

1.1 Related work

Multiple methods have been developed or adapted for tumor-only somatic variant calling. Mutect2 [2], a commonly used variant caller, can be run in tumor-only mode. Recently, Google released Deepsomatic [3], including models trained for tumor-only somatic variant calling. DeepSomatic uses convolutional neural networks trained on the SEQC2 benchmark dataset. DeepSom [4] is another recently published method that released multiple pre-trained convolutional neural network models for tumor-only whole genome somatic variant calling. DeepSom uses variant calls first generated by Mutect2 (tumor-only mode) as the initial call set for prediction using its models. DeepSom includes multiple models trained on cancer types.

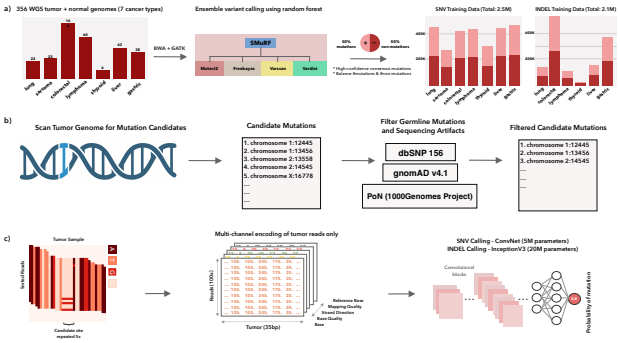

Fig. 1: Overview of VarNet-T. a Training labels were generated from matched tumor/normal genomes using high-confidence calls from four variant callers. b During prediction, VarNet-T scans the tumor genome, filters candidate mutations using public databases, and classifies variants as somatic or non-somatic. c Genomic positions are encoded as multi-dimensional matrices of tumor reads and features, then processed by a CNN.

2. VarNet-T

VarNet-T is a novel framework for somatic variant calling in tumor-only samples without a matched normal. VarNet-T uses deep convolutional neural networks designed to process inputs lacking normal sequencing reads. VarNet-T is trained using over 300 tumor whole genomes of seven cancer types (Fig. 1a). High-confidence somatic mutation calls were generated using an ensemble of mutation callers using matched tumor-normal samples to establish accurate training labels. While tumor-normal pairs were used to generate training labels, the input to the deep learning models do not include normal reads. The training set included 1.25 million high-confidence somatic SNVs and 1.05 million high-confidence somatic indels, along with an equal number of non-somatic sites for each. We trained separate deep learning models for detecting somatic SNVs and indels. In the prediction stage after training, as matched germline samples are not available, VarNet-T uses public germline mutation databases to exclude germline variants. VarNet-T first scans the tumor genome to identify candidate mutations and then filters mutations found in gnomAD v4.1 [5] or dbSNP build 156 [6] (Fig. 1b). Additionally, VarNet-T uses a panel of normals file to filter common sequencing artifacts. Aligned reads overlapping each candidate mutation are encoded in an image-like multi-channel numerical representation including

base, base qualities, mapping qualities, strand bias and reference base (Fig. 1c). This input representation is fed to the deep learning models for binary classification as somatic vs non-somatic.

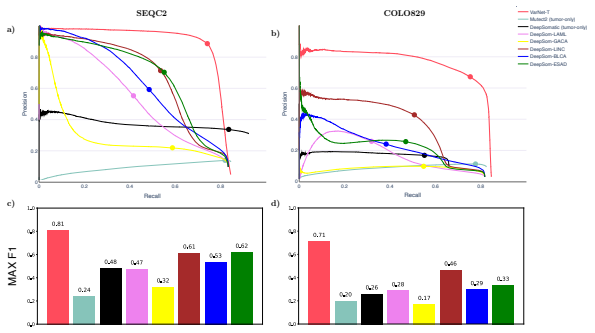

Fig. 2: Tumor-only variant calling accuracy. a,b Precision-recall curves for SNV calling on SEQC2 and COLO829 samples without matched normals, showing VarNet-T achieving the highest AUPRC. c,d Maximum F1 scores for SNV calling on the same samples.

3. Results

We benchmarked the performance of VarNet-T on independent and publicly available benchmark datasets derived from real tumor samples. Ground-truth labels in these datasets were established using both tumor and normal samples. However, we did not use the matched normal samples while running and evaluating the tumor-only variant callers. We benchmarked callers on a somatic reference callset derived from a breast cancer cell line established by the FDA led Sequencing Quality Control Phase 2 (SEQC2) consortium [7], and COLO829 [8], a metastatic melanoma cell line with a multi-institutionally defined reference set of somatic mutations by the Translational Genomics Research Institute (TGEN). These two reference datasets were created by an ensemble approach using data from multiple sequencing and algorithmic pipelines. The SEQC2 dataset was partially validated with targeted sequencing (>2,000-fold coverage) to establish high confidence calls in the original study. On the SEQC2 dataset, VarNet-T achieved the highest Area Under the Precision-Recall Curve (AUPRC) among benchmarked callers by a significant margin (Fig. 2a). VarNet-T achieved AUPRC of 0.773 on SEQC2, compared to the second highest achieved by DeepSom-ESAD (0.577). On COLO829, VarNet-T again achieved the best performing AUPRC of 0.656 (Fig. 2b), followed by DeepSom-LINC (0.318). F1 scores are reported in Figs. 2c and 2d. Notably, VarNet-T approaches the SNV calling performance observed in tumor-normal settings on these samples. On indel calling, VarNet-T again achieved the best AUPRC scores on both SEQC2 (0.527) and COLO829 (0.101). However, all callers performed less accurately on indel calling as it is generally a harder problem even in the tumor-normal setting [1, 9].

4. Analysis

We analysed the features that the VarNet-T models focus on in the input. We computed importance of input pixels using gradients of the model's outputs with respect to its inputs, using a method called guided backpropagation [10]. We visualized the input representation used by VarNet-T (Fig. 3a) and their associated pixel importance scores (Fig. 3b). This analysis highlighted that VarNet-T is most attentive to the candidate mutation column even though it was not explicitly trained to do so (Fig. 3b). Moreover, the data suggests that VarNet-T also pays attention to input pixels surrounding the candidate mutation site, which suggests that the model likely leverages mutational signatures in the sequence context. It is noteworthy that VarNet-T was trained from scratch without including any specialized knowledge of mutations and yet has managed to learn important features such as variant alleles and other relevant mutation signatures in the sequence context to solve the task.

5. Conclusion

Identification of somatic variants in tumor-only mode is common in the setting of clinical diagnostics and archival analysis of tumor banks or cell lines. We have described a novel method for tumor-only somatic variant calling that can substantially outperform existing methods. Our results significantly improve upon existing methods for genome level mutation calling from tumor only samples, which could play a significant role in emerging clinical applications and tumor-based markers, such as tumor mutation burden (TMB) for immunotherapy and DNA mismatch repair (MMR) deficiency for PARP inhibitor therapies.

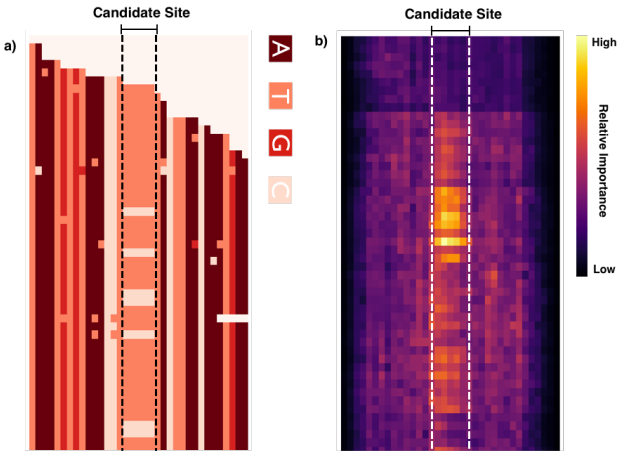

Fig. 3: Input and pixel importance visualization. a VarNet-T encoding (base channel) of a T>C mutation, with the candidate site repeated 5x. b Heatmap of pixel importance scores averaged over 30 mutations, showing VarNet-T focusing on the candidate site and surrounding context. Scores were computed via Guided Backpropagation.

References

[1] Kiran Krishnamachari, Dylan Lu, Alexander Swift-Scott, Anuar Yeraliyev, Kayla Lee, Weitai Huang, Sim Ngak Leng, and Anders Jacobsen Skanderup. Accurate somatic variant detection using weakly supervised deep learning. *Nature Communications*, 13(1):4248, July 2022.

[2] David Benjamin, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, and Lee Lichtenstein. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*, 2019.

[3] Jimin Park, Daniel E. Cook, Pi-Chuan Chang, Alexey Kolesnikov, Lucas Brambrink, Juan Carlos Mier, Joshua Gardner, Brandy McNulty, Samuel Sacco, Ayse Keskus, Asher Bryant, Tanveer Ahmad, Jyoti Shetty, Yongmei Zhao, Bao Tran, Giuseppe Narzisi, Adrienne Helland, Byunggil Yoo, Irina Pushel, Lisa A. Lansdon, Chengpeng Bi, Adam Walter, Margaret Gibson, Tomi Pastinen, Midhat S. Farooqi, Nicolas Robine, Karen H. Miga, Andrew Carroll, Mikhail Kolmogorov, Benedict Paten, and Kishwar Shafin. DeepSomatic: Accurate somatic small variant discovery for multiple sequencing technologies, August 2024. Pages: 2024.08.16.608331 Section: New Results.

[4] Sergey Vilov and Matthias Heinig. DeepSom: a CNN-based approach to somatic variant calling in WGS samples without a matched normal. *Bioinformatics*, 39(1):btac828, January 2023. _eprint: https://academic.oup.com/bioinformatics/article-pdf/39/1/btac828/48731812/btac828.pdf.

[5] Siwei Chen, Laurent C. Francioli, Julia K. Goodrich, Ryan L. Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, Nicholas A. Watts, Christopher Vittal, Laura D. Gauthier, Timothy Poterba, Michael W. Wilson, Yekaterina Tarasova, William Phu, Riley Grant, Mary T. Yohannes, Zan Koenig, Yossi Farjoun, Eric Banks, Stacey Donnelly, Stacey Gabriel, Namrata Gupta, Steven Ferriera, Charlotte Tolonen, Sam Novod, Louis Bergelson, David Roazen, Valentin Ruano-Rubio, Miguel Covarrubias, Christopher Llanwarne, Nikelle Petrillo, Gordon Wade, Thibault Jeandet, Ruchi Munshi, Kathleen Tibbetts, Maria Abreu, Carlos A. Aguilar Salinas, Tariq Ahmad, Christine M. Albert, Diego Ardissino, Irina M. Armean, Elizabeth G. Atkinson, Gil Atzmon, John Barnard, Samantha M. Baxter, Laurent Beaugerie, Emelia J. Benjamin, David Benjamin, Michael Boehnke, Lori L. Bonnycastle, Erwin P. Bottinger, Donald W. Bowden, Matthew J. Bown, Harrison Brand, Steven Brant, Ted Brookings, Sam Bryant, Sarah E. Calvo, Hannia Campos, John C. Chambers, Juliana C. Chan, Katherine R. Chao, Sinéad Chapman, Daniel I. Chasman, Rex Chisholm, Judy Cho, Rajiv Chowdhury, Mina K. Chung, Wendy K. Chung, Kristian Cibulskis, Bruce Cohen, Kristen M. Connolly, Adolfo Correa, Beryl B. Cummings, Dana Dabelea, John Danesh, Dawood Darbar, Phil Darnowsky, Joshua Denny, Ravindranath Duggirala, Josée Dupuis, Patrick T. Ellinor, Roberto Elosua, James Emery, Eleina England, Jeanette Erdmann, Tõnu Esko, Emily Evangelista, Diane Fatkin, Jose Florez, Andre Franke, Jack Fu, Martti Färkkilä, Kiran Garimella, Jeff Gentry, Gad Getz, David C. Glahn, Benjamin Glaser, Stephen J. Glatt, David Goldstein, Clicerio Gonzalez, Leif Groop, Sanna Gudmundsson, Andrea Haessly, Christopher Haiman, Ira Hall, Craig L. Hanis, Matthew Harms, Mikko Hiltunen, Matti M. Holi, Christina M. Hultman, Chaim Jalas, Mikko Kallela, Diane Kaplan, Jaakko Kaprio, Sekar Kathiresan, Eimear E. Kenny, Bong-Jo Kim, Young Jin Kim, Daniel King, George Kirov, Jaspal Kooner, Seppo Koskinen, Harlan M. Krumholz, Subra Kugathasan, Soo Heon Kwak, Markku Laakso, Nicole Lake, Trevyn Langsford, Kristen M. Laricchia, Terho Lehtimäki, Monkol Lek, Emily Lipscomb, Ruth J. F. Loos, Wenhan Lu, Steven A. Lubitz, Teresa Tusie Luna, Ronald C. W. Ma, Gregory M. Marcus, Jaume Marrugat, Kari M. Mattila, Steven McCarroll, Mark I. McCarthy, Jacob L. McCauley, Dermot McGovern, Ruth McPherson, James B. Meigs, Olle Melander, Andres Metspalu, Deborah Meyers, Eric V. Minikel, Braxton D. Mitchell, Vamsi K. Mootha, Aliya Naheed, Saman Nazarian, Peter M. Nilsson, Michael C. O'Donovan, Yukinori Okada, Dost Ongur, Lorena Orozco, Michael J. Owen, Colin Palmer, Nicholette D. Palmer, Aarno Palotie, Kyong Soo Park, Carlos Pato, Ann E. Pulver, Dan Rader, Nazneen Rahman, Alex Reiner, Anne M. Remes, Dan Rhodes, Stephen Rich, John D. Rioux, Samuli Ripatti, Dan M. Roden, Jerome I. Rotter, Nareh Sahakian, Danish Saleheen, Veikko Salomaa, Andrea Saltzman, Nilesh J. Samani, Kaitlin E. Samocha, Alba Sanchis-Juan, Jeremiah Scharf, Molly Schleicher, Heribert Schunkert, Sebastian Schönherr, Eleanor G. Seaby, Svati H. Shah, Megan Shand, Ted Sharpe, Moore B. Shoemaker, Tai Shyong, Edwin K. Silverman, Moriel Singer-Berk, Pamela Sklar, Jonathan T. Smith, J. Gustav Smith, Hilkka Soininen, Harry Sokol, Rachel G. Son, Jose Soto, Tim Spector, Christine Stevens, Nathan O. Stitziel, Patrick F. Sullivan, Jaana Suvisaari, E. Shyong Tai, Kent D. Taylor, Yik Ying Teo, Ming Tsuang, Tiinamaija Tuomi, Dan Turner, Teresa Tusie-Luna, Erkki Vartiainen, Marquis Vawter, Lily Wang, Arcturus Wang, James S. Ware, Hugh Watkins, Rinse K. Weersma, Ben Weisburd, Maija Wessman, Nicola Whiffin, James G. Wilson, Ramnik J. Xavier, Anne O'Donnell-Luria, Matthew Solomonson, Cotton Seed, Alicia R. Martin, Michael E. Talkowski,

Heidi L. Rehm, Mark J. Daly, Grace Tiao, Benjamin M. Neale, Daniel G. MacArthur, Konrad J. Karczewski, and Genome Aggregation Database Consortium. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625(7993):92–100, January 2024.

[6] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, January 2001.

[7] Li Tai Fang, Bin Zhu, Yongmei Zhao, Wanqiu Chen, Zhaowei Yang, Liz Kerrigan, Kurt Langenbach, Maryellen de Mars, Charles Lu, Kenneth Idler, Howard Jacob, Yuanting Zheng, Luyao Ren, Ying Yu, Erich Jaeger, Gary P. Schroth, Ogan D. Abaan, Keyur Talsania, Justin Lack, Tsai-Wei Shen, Zhong Chen, Seta Stanbouly, Bao Tran, Jyoti Shetty, Yuliya Kriga, Daoud Meerzaman, Cu Nguyen, Virginie Petitjean, Marc Sultan, Margaret Cam, Monika Mehta, Tiffany Hung, Eric Peters, Rasika Kalamegham, Sayed Mohammad Ebrahim Sahraeian, Marghoob Mohiyuddin, Yunfei Guo, Lijing Yao, Lei Song, Hugo Y. K. Lam, Jiri Drabek, Petr Vojta, Roberta Maestro, Daniela Gasparotto, Sulev Kõks, Ene Reimann, Andreas Scherer, Jessica Nordlund, Ulrika Liljedahl, Roderick V. Jensen, Mehdi Pirooznia, Zhipan Li, Chunlin Xiao, Stephen T. Sherry, Rebecca Kusko, Malcolm Moos, Eric Donaldson, Zivana Tezak, Baitang Ning, Weida Tong, Jing Li, Penelope Duerken-Hughes, Claudia Catalanotti, Shamoni Maheshwari, Joe Shuga, Winnie S. Liang, Jonathan Keats, Jonathan Adkins, Erica Tassone, Victoria Zismann, Timothy McDaniel, Jeffrey Trent, Jonathan Foox, Daniel Butler, Christopher E. Mason, Huixiao Hong, Leming Shi, Charles Wang, and Wenming Xiao. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. *Nature Biotechnology*, 39(9):1151–1160, September 2021. Number: 9 Publisher: Nature Publishing Group.

[8] David W Craig, Sara Nasser, Richard Corbett, Simon K Chan, Lisa Murray, Christophe Legendre, Waibhav Tembe, Jonathan Adkins, Nancy Kim, Shukmei Wong, Angela Baker, Daniel Enriquez, Stephanie Pond, Erin Pleasance, Andrew J Mungall, Richard A Moore, Timothy McDaniel, Yussanne Ma, Steven J M Jones, Marco A Marra, John D Carpten, and Winnie S Liang. A somatic reference standard for cancer genome sequencing. *Scientific Reports*, 6(1):24607, 2016.

[9] Anne Bruun Krøig\aard, Mads Thomassen, Anne Vibeke Lænkholm, Torben A. Kruse, and Martin Jakob Larsen. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PLoS ONE*, 11(3):1–15, 2016. Publisher: Public Library of Science.

[10] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for Simplicity: The All Convolutional Net. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.