A Deep Learning System Classifies Cancer Origins Using Somatic Mutation Patterns Detected by Plasma Whole-Genome Sequencing

Xindi Zhang[1,2*], Wei Jiao[1], Gurnit Atwal[3], Quaid Morris[3], Lincoln D. Stein[1,2]

1. Ontario Institute for Cancer Research, Toronto, ON, Canada. 2. Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada. 3. Memorial Sloan Kettering Cancer Center, Manhattan, New York, USA

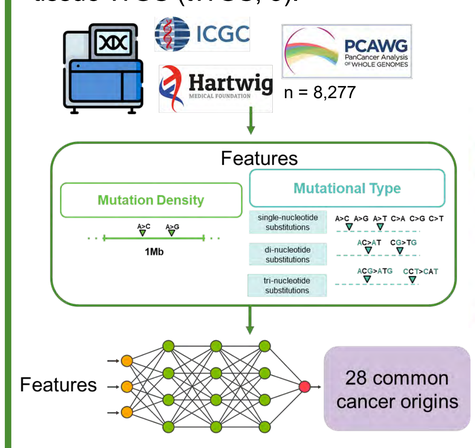
***Email: xindi.zhang@oicr.on.ca**

Introduction

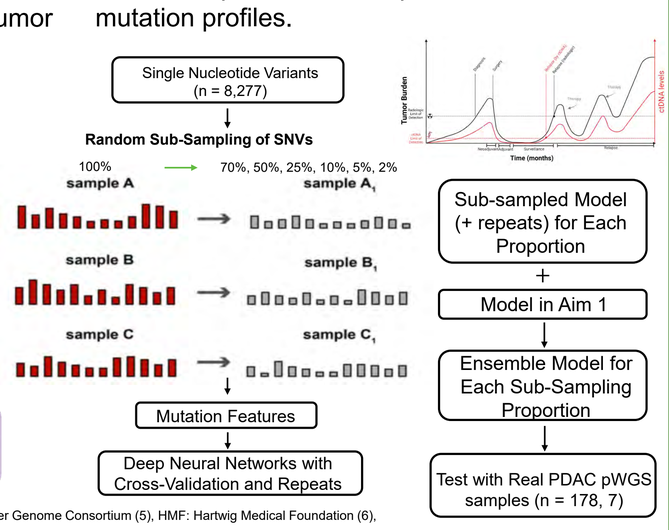
- A tumour's cell of origin is the single major predictor of the natural history of the disease (1).
- However, identifying the tumour's cell of origin can be complex, time-consuming, and error-prone in current clinical settings.
- Somatic passenger mutations capture the epigenetic state of the cell of origin as different cell types have distinctive chromatin profiles (2).
- Tumour cells release circulating tumour DNA into the blood, containing somatic passenger mutations specific to the tumour's cell of origin.
- **Objective: To design a deep learning system to predict cancer origins utilizing passenger mutation profiles detected from circulating tumour DNA collected by plasma whole-genome sequence (pWGS) to aid cancer of unknown primary and early cancer identification.**

Methods

Aim 1. Create a deep learning model using somatic mutation profiles from tumor tissue WGS (tWGS; 3).

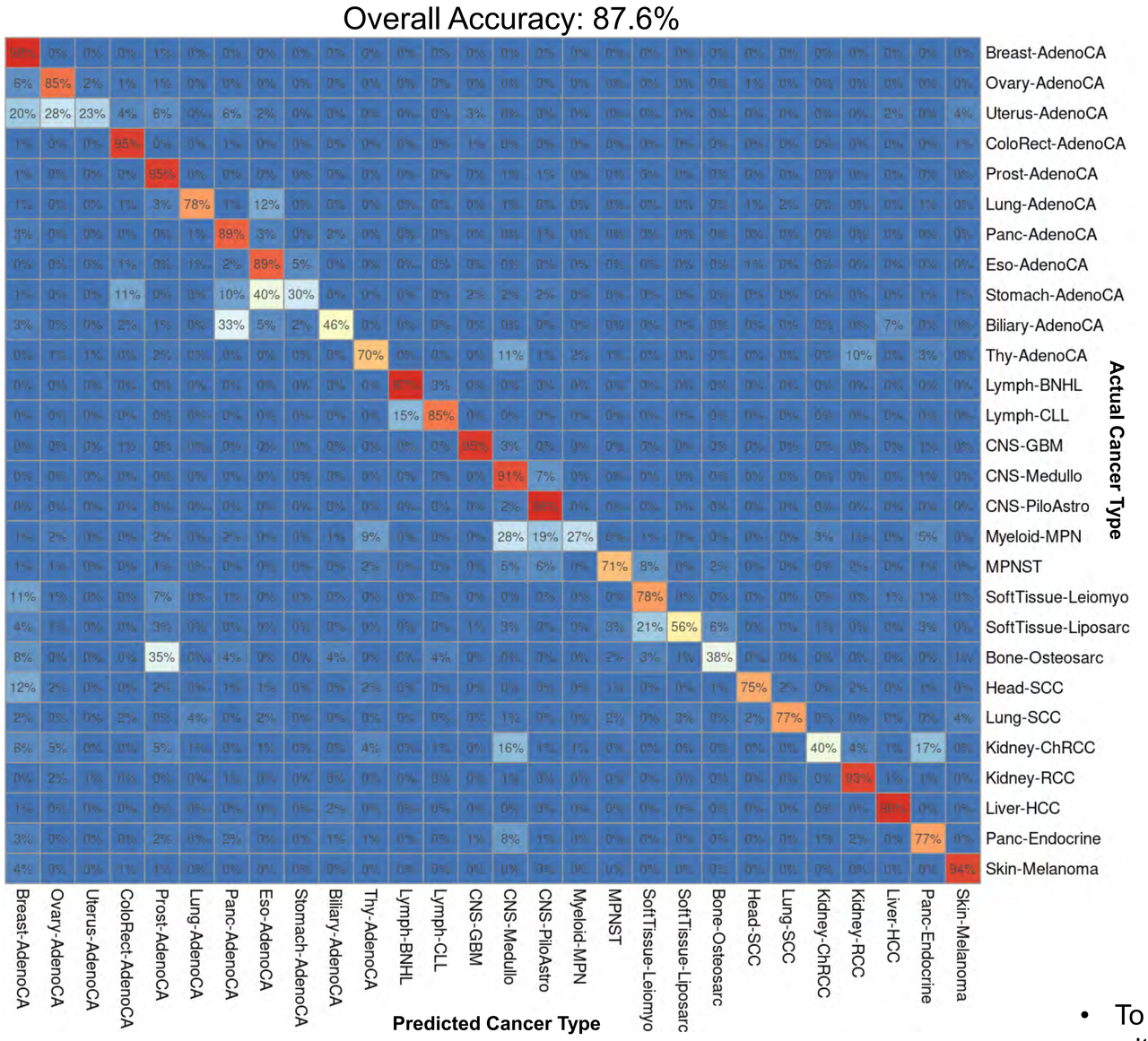


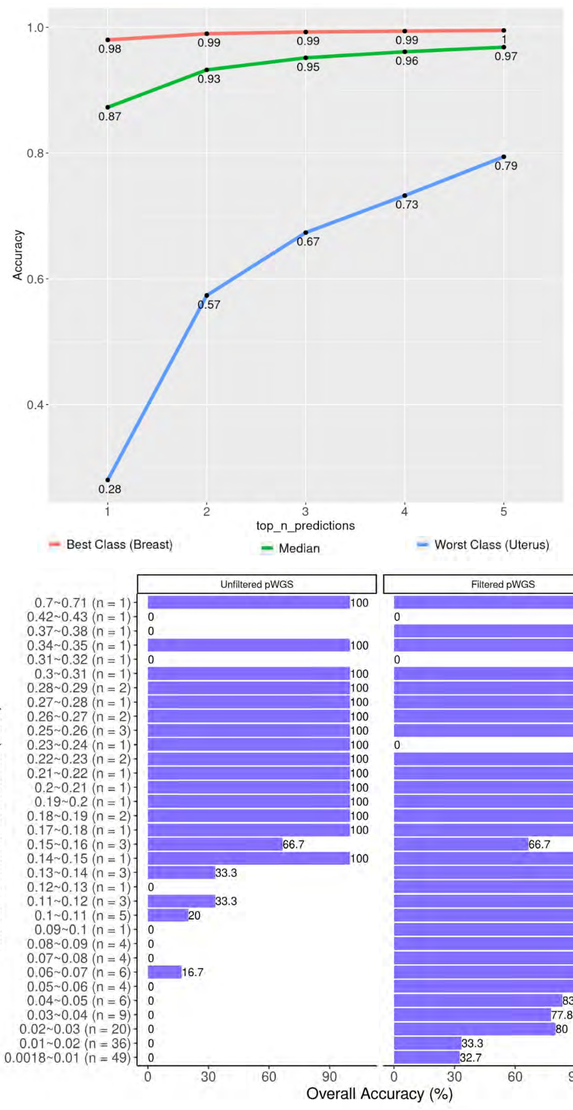
Aim 2. Adapt the model to pWGS somatic mutation profiles.

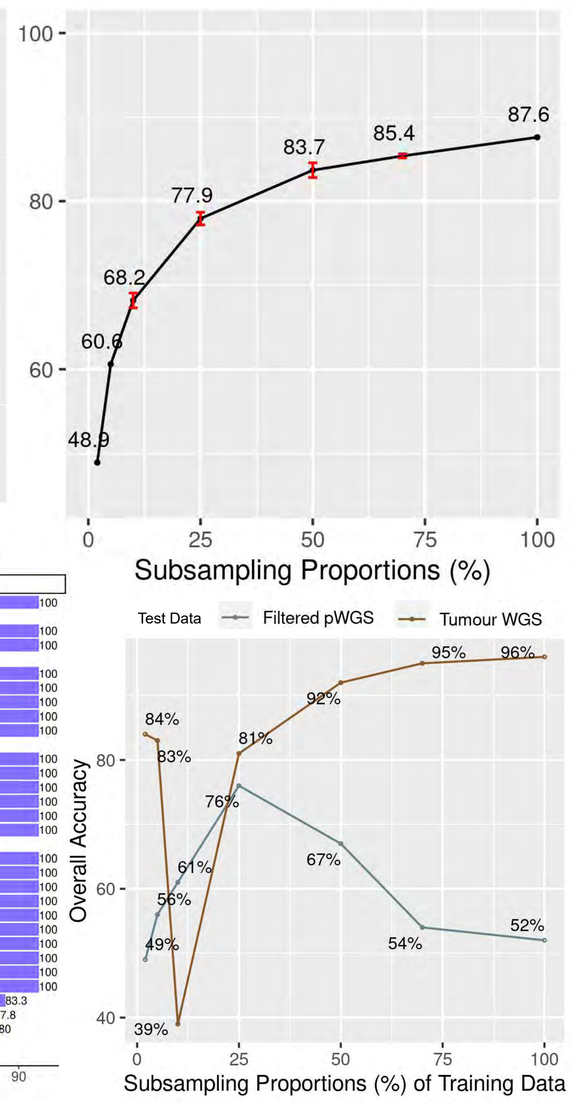


PCAWG: PanCancer Analysis of Whole Genome (4), ICGC: International Cancer Genome Consortium (5), HMF: Hartwig Medical Foundation (6), SNV: Single Nucleotide Variants, PDAC: Pancreatic Adenocarcinoma

Results







Aim 1. Create a deep learning model using somatic mutation profiles from tWGS:

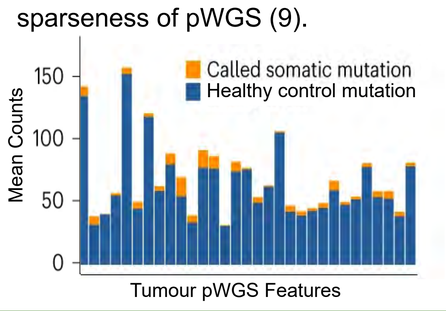
- We created a feed-forward neural network that predicted 28 common cancer types, with an overall held-out accuracy of 87.6%.
- Top N-rank analysis indicates that the model could be valuable for differential diagnosis.

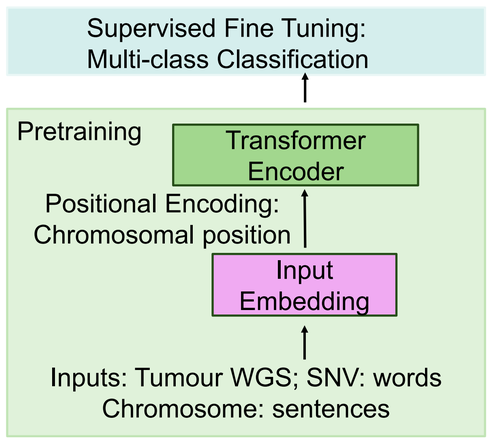
Aim 2. Adapt the model to pWGS somatic mutation profiles:

- The overall accuracy exceeded 78% when using more than 25% of the tWGS mutation profile of a patient.
- To validate the model in aim 1, classification accuracy in predicting PDAC (n = 178; 7) with tWGS was 96%. When using pWGS data from the same samples, filtered variants matched those found in tWGS (filtered pWGS), as the test data, we observed accuracies of 80-100% for samples with TRD ≥ 0.05. This indicates that the algorithm's performance is greatly improved when non-tumour variant calls are removed.
- Testing PDAC samples on our ensemble models showed that accuracies for tumour sample prediction were consistently above 80%, except for the model trained with 10% sub-sampling. For filtered pWGS samples, the highest classification accuracy was 76% with the model trained using 25% of each patient's tWGS mutation profile.

Future Proposal

- Estimate and remove non-tumour variant noise by subtracting variant profiles from a panel of normal controls (8).
- Use Bidirectional Encoder Representation Transformers to accommodate the sparseness of pWGS (9).





Conclusion

Our study presents the development of a feed-forward neural network trained with tWGS, capable of predicting 28 common cancer types with an overall accuracy of 87.6%. When training models with simulated pWGS somatic variant profiles by randomly sub-sampling a proportion of tWGS mutation profiles for each sample, the overall accuracy consistently surpassed 78% when using more than 25% of the mutation profiles. In the validation phase using real PDAC samples as the test data, the model achieved 96% accuracy with tWGS data, and the removal of non-tumour variants significantly improved accuracy in making prediction for pWGS data from the same samples. Our ensemble models, which combined the tWGS-trained model with models trained on simulated pWGS mutation profiles (and repeats), consistently exhibited accuracies above 80% for predicting PDAC tumour samples, except for the model trained with 10% of sub-sampling. Interestingly, using pWGS samples that excluded variants not matching those found in tWGS, as the test data, the highest classification accuracy was 76% with the model trained with 25% of each patient's tWGS mutation profile.

Acknowledgement

We would like to thank Irina Kalatskaya, Quang Trinh, Jared Simpson, Katie Hoadley and David Louis for their helpful comments during preparation of this paper. We also gratefully acknowledge the assistance of Drs. Ludmil B. Alexandrov, Mi Ni Huang, Arnoud Boot, Steven Gallinger, Julie Wilson, Haiko J. Bloemendal, Laurens Beerepoot, Steven G. Rozen and Michael R. Stratton in providing independent WGS primary and metastatic tumour SNV profiles used for validation. We also thank W.J., L.S. and Q.M. supported by funding from the Province of Ontario, Canada. QM's research was supported by a gift from NVIDIA foundation, an advised fund of the Silicon Valley Community Foundation. RK was supported by the European Structural and Investment Funds grant for the Croatian National Centre of Research Excellence in Personalized Healthcare (contract #KK.01.1.1.01.0010), Croatian National Centre of Research Excellence for Data Science and Advanced Cooperative Systems (contract KK.01.1.1.01.0009), the European Commission Seventh Framework Program (Integra-Life; grant 315997) and Croatian Science Foundation (grant IP-2014-09-6400). J.d.R. is supported by a NWO-Vidi grant (016.Vidi.178.023). We acknowledge the contributions of the many clinical networks across ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonised variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects.

References

1. Greco, F. A. Molecular Diagnosis of the Tissue of Origin in Cancer of Unknown Primary Site: Useful in Patient Management. Current Treatment Options in Oncology. 14, 634–642 (2013).
2. Ocsenas, O., & Reimand, J. Chromatin accessibility of primary human cancers ties regional mutational processes and signatures with tissues of origin. PLoS Computational Biology.18, 8 (2022)
3. Jiao, W. et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nature Communications. 11, 1–12 (2020).
4. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. Nature. 578, 82–93 (2020).
5. Zhang, J. et al. The International Cancer Genome Consortium Data Portal. Nature Biotechnology. 37, 367–369 (2019).
6. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature. 575, 210–216 (2019).
7. Yuanchang Fang, Notta Lab
8. Bruhm, D.C. et al. Single-molecule genome-wide mutation profiles of cell-free DNA for non-invasive detection of cancer. Nature Genetics. 55, 1301-1310 (2023)
9. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. (2019)

Ontario Funding provided by the Government of Ontario.