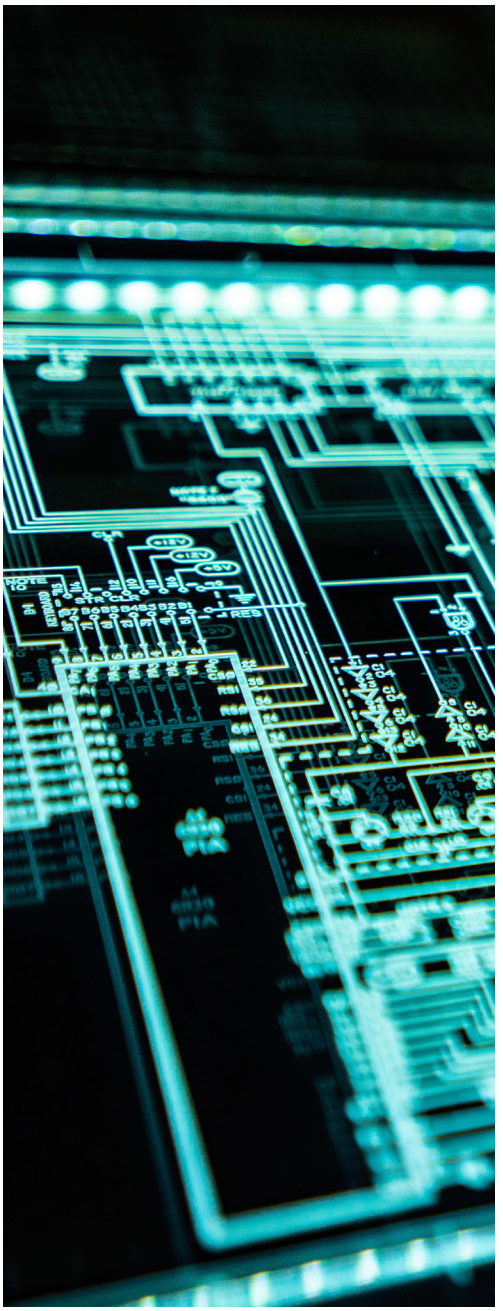
THE SIGNALS OF SUBTYPES: HOW AI CREATES PERSONALIZED CANCER TREATMENT

AN INTERVIEW WITH DR. HANG CHANG

BY TANYA SANGHAL, ALLISUN WILTSHIRE
ADDITIONAL CONTRIBUTOR AUTO YU



Hang Chang *is a computational biology research scientist at the Lawrence Berkeley National Laboratory (LBNL) in the Department of BioEngineering & BioMedical Sciences, within the Division of Biological Systems and Engineering. He holds a secondary affiliation at the LBNL in the Department of Cellular & Tissue Imaging, within the Division of Molecular Biophysics and Integrated Bioimaging. In 2014, Dr. Chang received the R&D 100 Award for the creation of a computational platform called BioSig3D. In 2017, he co-founded the Berkeley Biomedical Data Science Center and currently serves as Co-Director. Dr. Chang's lab focuses on the intersections between machine learning, computer vision, computational biology, informatics, and big data. In this interview, we discussed his work on AI-empowered multimodal biomarker discovery for precision cancer diagnosis and treatment.*

BSJ: Can you give a brief introduction to your lab's research?

HC: My research group works on questions at the interfaces between engineering, computation, and biomedical sciences. Our current research focuses on knowledge discovery at a large scale, looking at multimodal scientific data with applications to translational cancer research. I have been interested in the computational biosciences since I was a graduate student. Back then, I developed algorithms and software packages to quantify the plant cell wall. After that, I started focusing on the quantitative characterization of the heterogeneity of cancer. This heterogeneity results in large variations in clinical outcomes of patients after treatment, so the discovery of prognostic and predictive biomarkers remains a very urgent and unmet need. Since then, I have been focusing on the development and validation of multimodal biomarkers toward precision diagnosis and personalized treatment of cancer patients.

In the past few years, based on The Cancer Genome Atlas (TCGA), our group built an AI pipeline that extracts morphometric properties from whole slide images of tissue histology for cellular morphometric biomarker discovery. We also worked on plant cell wall characterization to try to optimize the energy generation from plant cell walls. The most exciting work that I have been doing during the past five years is on the association between cellular morphometric biomarkers, cancer heterogeneity, and clinical outcomes.

BSJ: Much of your research on breast cancer and lower-grade gliomas involves developing and validating cellular morphometric subtypes (CMS). Can you explain what these subtypes are and how they are useful?

HC: Tissue histology sections are the gold standard for the assessment of tissue neoplasm (abnormal growths that may either be benign or cancerous). They are rich in content, displaying the different cell types and overall tumor microenvironment of a given sample. All of this information can be read by pathologists and used to make accurate diagnoses; however, there are limits to pathologists' perceptual capabilities in discovering hidden information from large-scale complex data. In order to overcome these limits, we asked the following question: "Can we create a standardized quantitative pipeline that provides clinical value by characterizing the heterogeneity among these tissue sections that is beyond current pathological practice?" Similar to what was found with molecular subtyping, we hypothesized and demonstrated that the patient subtypes derived by machine learning techniques from cellular morphometric information, which is embedded in histology sections, provide significant and independent clinical value compared to well-known clinical and molecular factors in clinical practice. Thus, the introduction of CMS into cancer diagnosis and treatment outcome prediction can potentially provide a new avenue for the rapid, robust, and cost-effective precision diagnosis and personalized treatment of cancer patients.

For example, in our *Neuro-Oncology* paper, we identified two cellular morphometric subtypes. The patients with one of the subtypes demonstrate significantly shorter overall survival, which led us to ask what was behind this difference. Through quantitative analysis of the patient cohort, we show that the subtype with worse prognosis experiences some kind of immune suppression; although there is a high number of immune cells infiltrating the tumor tissue, they function deficiently. To validate this phenomenon , we performed immunohistochemistry (IHC) staining in hospital cohorts and confirmed that there indeed might be a mechanism of immune suppression in the subtype with worse prognosis. Why is this important? It means that our findings suggest potential benefit from immunotherapy for patients in this subtype.

BSJ: The algorithms you create also rely on identifying cellular morphometric biomarkers (CMBs) within those subtypes. How are CMBs identified, and how do they differ from cellular morphometric subtypes?

HC: Cellular morphometric biomarkers are a type of imaging biomarker that represents unique combinations of cellular morphometric properties, including the geometric properties of nuclear regions, cytoplasmic features, etc. I mentioned earlier that tissue histology sections contain a lot of cellular information, which can be characterized through their morphometric properties. These cellular morphometric biomarkers are identified through unsupervised feature learning algorithms to capture the heterogeneity embedded in tissue histology sections, such as tumor microenvironmental factors (which include the presence of tumor fibroblasts and infiltrated immune cells). Similar to molecular subtyping, which is typically built on gene expressions, the cellular morphometric subtype is constructed from the relative abundance of these cellular morphometric biomarkers.

There are many ways to validate these biomarkers. For example, in the *Neuro-Oncology* paper, using a public data set, we calculated the relative abundance of each individual biomarker by associating the biomarker abundance with the tumor microenvironmental factors that are already provided by these data sets. We also validate these biomarkers in terms of the patient's overall survival to ensure they are clinically meaningful. And, of course, we can improve our biological understanding of the biomarkers by evaluating them through their association with genomic data. In the hospital cohort, we further validated the relationship between CMSs and molecular factors with immunohistochemistry (IHC) staining.

BSJ: What was the purpose of using machine learning for cancer identification and treatment? What benefits did it pose?

HC: Cancer is a very complicated disease; there is no guaranteed way to treat it. Recent advances in

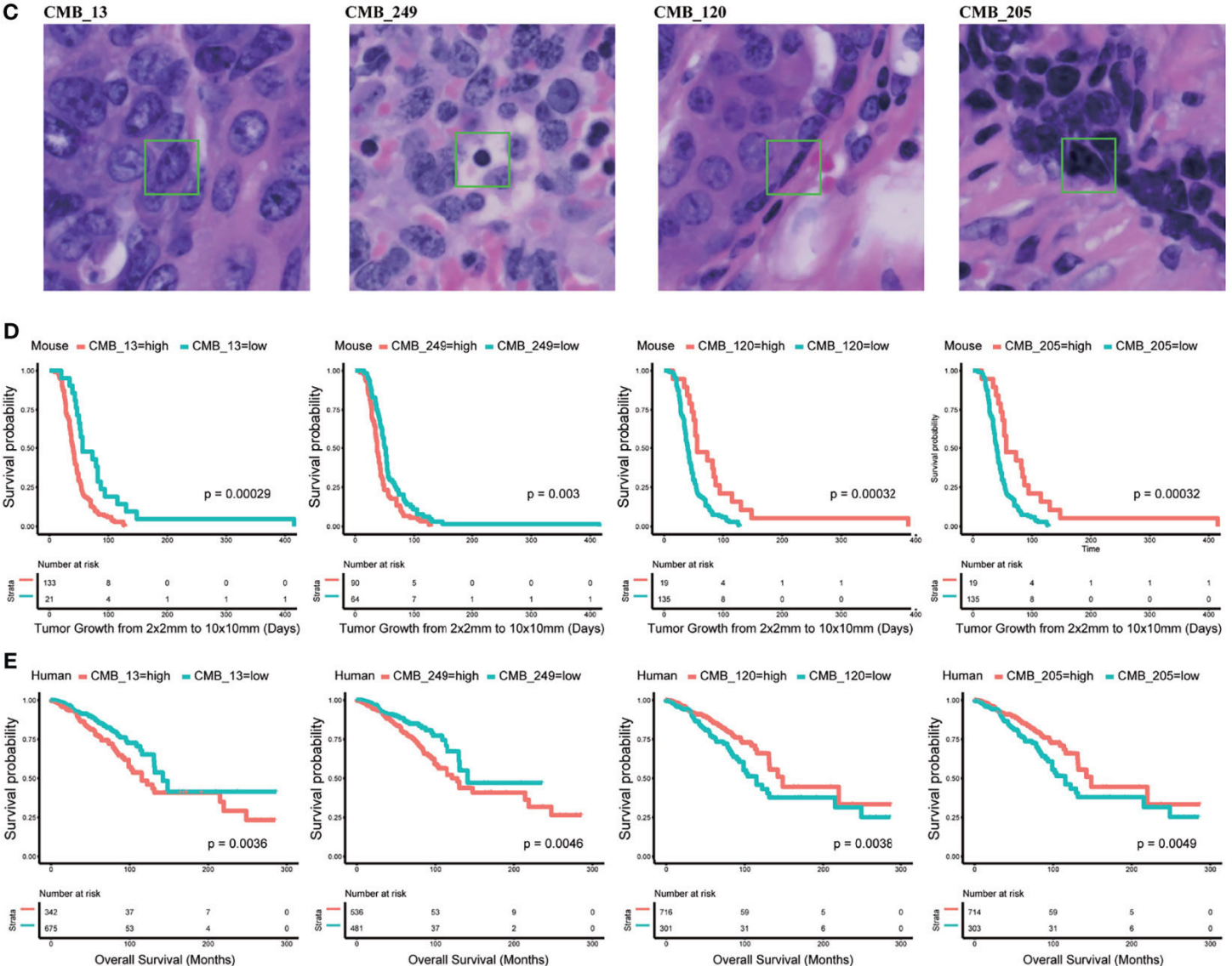


Figure 1: The images in Line C show examples of tissue histology slides used to create different cellular morphometric biomarkers (CMBs) under different artificial intelligence algorithms. *The graphs in Lines D and E demonstrate the overall survival probability for different cellular morphometric subtypes, the red line representing high probability of survival and the blue representing a low probability of survival.*[1]

biotechnologies, such as next generation sequencing (NGS), enable the generation of a massive amount of multimodal data. The problem for physicians is effectively using data that is beyond their physical capability to integrate and interpret. One of the cancer research focuses is to find a way to integrate this data to improve our characterization and understanding of cancer. The development and deployment of advanced machine learning techniques facilitate efficient and effective data integration toward the discovery of prognostic and predictive biomarkers for cancer patient management.

BSJ: To identify CMS' from mouse mammary tumors, you created an AI pipeline based on stacked predictive sparse decomposition (SPSD).[1] Could you briefly walk us through SPSD and why you chose it for your studies?

HC: SPSD is an unsupervised machine learning pipeline that we created a long time ago for the mining of robust cellular morphometric biomarkers with speed and accuracy. It finds the underlying correlation or relationship within the higher dimensional feature space. One of the reasons that we use SPSD is that it is fast; this is very critical in clinical practice because sometimes, the turnaround time needs to be very short when deciding a treatment plan. Through our studies, we have been demonstrating that this algorithm can be used for different fields. For example, we used it for both white blood cell subtype classification and for environmental risk exposure studies. All of these studies give us confidence that we can learn, beyond the statistical noise, the real characteristics of the data.

BSJ: In a following study, to improve risk stratification for breast cancer patients you created an algorithm called iCEMIGE, a novel approach to the difficulties of integrating multimodal data.[3] Could you explain iCEMIGE and any challenges you encountered during its creation?

HC: We have previously identified cellular morphometric biomarkers, gene biomarkers and tumor microbiome biomarkers for precision prognosis of breast cancer patients. The goal of iCEMIGE is to investigate whether the multi-modal integration of these biomarkers improves risk stratification of breast cancer patients. During the creation of iCEMIGE, we were exploring two different ways of multi-modal integration: one was an integration from raw data using a black-box-like framework and the other was to perform the integration at biomarker level. In machine learning, one of the major ways to perform this integration is at the raw data level. We throw the high dimensional vectors of gene expression, the high dimensional vectors of the microbiome, and the cell morphometric data into a black box, and the algorithms give back an output. One of the challenges with this black-box model is the lack of explainability (e.g. we do not know why certain biomarkers are picked up out of the high dimensional space). Even small amounts of noise in independent data can disturb the whole performance curve. To overcome these challenges, we asked the question, "Can we utilize human knowledge and/or pre-identified biomarkers to create a hybrid system?" We used this information in the first step instead of the raw data and the pipeline optimized their combinations. The final iCEMIGE framework was built upon the biomarker-level-integration strategy due to its improved robustness and expandability compared to the other strategy. We have shown that the iCEMIGE model exceeds the performance of individual biomarkers, as well as WHO standards.

BSJ: What is the process like of translating subtypes/biomarkers identified in mouse models to human patients? How do you ensure that an algorithm or model applies to an entire population set?

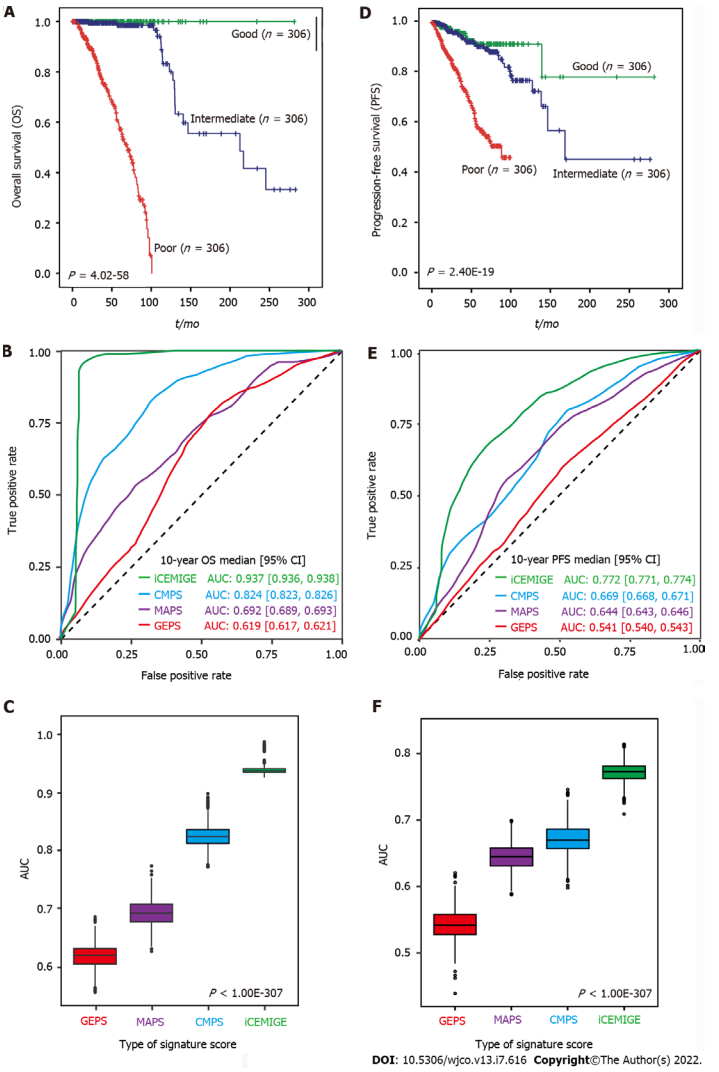
HC: The only way is through biomedical validation on independent cohorts. The translation from mouse model to human patients consists of four steps: 1) biomarker detection from mouse data; 2) subtype construction from mouse data; 3) biomarker extraction and subtype prediction on human patients using pre-built models from mouse data; and most importantly, 4) during the last step, we need to validate the consistency of biomarkers and subtypes (in terms of their clinical value) before and after translation.

BSJ: In what ways can the artificial intelligence you have developed for cancer detection be applied to other fields of research?

HC: We have shown that the algorithms we developed for cancer research can be used for white blood cell classification, environmental risk assessment, etc. Most importantly, the system-biology-driven multi-scale and multi-modal integration strategy can be widely used for various biomedical studies beyond cancer. In our research, we aim to integrate meaningful information into robust algorithms that will output explainable findings with high confidence.

BSJ: What is your intended research direction for the future?

HC: The major focus of my future research is to validate the cellular morphometric biomarkers in multicenter prospective clinical studies toward approval for clinical practice. My goal with computational biomedical science is to maximize the translational impact of our techniques and discoveries via clinical



DOI: 10.5306/wjco.v13.i7.616

Figure 2: Panels A and D depict the survival rates, overall and progress-free, of breast cancer patients from the TCGA cohort. *The patients have been divided into three groups by the prognoses predicted by the iCEMIGE algorithm.*[3]

"In our research, we aim to integrate meaningful information into robust algorithms that will output explainable findings with high confidence."

validation toward FDA approval for clinical application. In order to do that, besides independent validation, we need to put these biomarkers through preclinical studies and clinical trials before we can use the technology to help guide the treatment choices for cancer patients. For this, we have been working with international collaborators on skin research, including melanoma, squamous cell carcinoma, and other types of human cancers. Currently, we are in a stage of retrospective study with archived patient cohorts to make sure that we are fully confident about the robustness of the biomarkers we developed. So far, the iCEMIGE integration of different biomarkers is very promising for improving the assessment of prognostic risk.

We are also working on the application of AI techniques, which were developed in-house, to define the best combination of drug therapy based on the integrated molecular, clinical, and cellular morphometric profiles of cancer patients. It is about precision therapy– due to the heterogeneity of cancer patients, many respond differently to treatments. An important goal of mine, besides clarifying a patient's prognosis, is to find the best possible combination of treatments for each patient.

At the moment, we have exciting results on the prediction of chemotherapy with glioblastoma patients. These results arose through an international collaboration, and the publication is in preparation.

REFERENCES

1. Chang, H., Yang, X., Moore, J., Liu X.P., Jen, K.Y., Snijders, A.M., Ma, L., Chou, W., Corchado-Cobos, R., García-Sancha, N., Mendiburu-Eliçabe, M., Pérez-Losada, J., Barcellos-Hoff, M.H., Mao, J.H. (2022). From Mouse to Human: Cellular Morphometric Subtype Learned From Mouse Mammary Tumors Provides Prognostic Value in Human Breast Cancer. *Frontiers in Oncology*, 11. https://doi.org/10.3389/fonc.2021.819565
2. Liu, X.P., Jin X., Ahmadian, S.S., Yang X., Tian, S.F., Cai, Y.X., Chawla, K., Snijders, A.M., Xia, Y., van Diest, P.J., Weiss, W.A., Mao, J.H., Li, Z.Q., Vogel, H., & Chang, H. (2022). Clinical significance and molecular annotation of cellular morphometric subtypes in lower-grade gliomas discovered by machine learning. *Neuro-Oncology*, noac154. https://doi.org/10.1093/neuonc/noac154
3. Mao, X. Y., Perez-Losada, J., Abad, M., Rodríguez-González, M., Rodríguez, C. A., Mao, J. H., & Chang, H. (2022). iCEMIGE: Integration of CEll-morphometrics, MIcrobiome, and GEne biomarker signatures for risk stratification in breast cancers. *World journal of clinical oncology*, 13(7), 616–629. https://doi.org/10.5306/wjco.v13.i7.616