Lung cancer biomark testing results from the MCM Team

Summary:

The MCM team's research into lung cancer biomarkers has identified 26 genes that are present with top scores across all the signature sizes considered. This update focuses on VAMP1, a gene linked to patient survival and differentially expressed in normal lung compared to lung cancer.

Terminology:

Gene signature: A set of genes shown to have a specific role in a disease is called gene signature. When such a signature can predict the presence of a disease, it is called a diagnostic gene signature. When signature relates to survival, it is called prognostic signature.

Matthews correlation coefficient: a statistical method used to evaluate the performance of a predictive model. It measures the differences between actual values and the predicted ones.

Probes: short DNA sequences targeting a small region of a transcript (gene). To make them more specific, probes are organized into probe sets, which are used to detect and quantify the presence of gene sequences through hybridization due to complementarity between the probe and the target.

Background

The Mapping Cancer Markers project aims to identify the markers associated with various types of cancer using a heuristic search algorithm. The project analyzes millions of data points collected from patient tissue datasets and identifies patterns that can detect cancer earlier, identify high-risk patients and customize treatment for individual patients. Initially focusing on lung cancer, we then investigated ovarian cancer, and currently analyzing sarcoma.

In November 2021, donated over 800 million work units for research into multiple types of cancer, with 193, 379 and 245 million work units crunched for lung, ovarian and sarcoma cancers respectively. To date, over 810,000 years of computational research has been donated to MCM, with close to 240 years generated every day. Thank you for helping us uncover insights into cancer signatures.

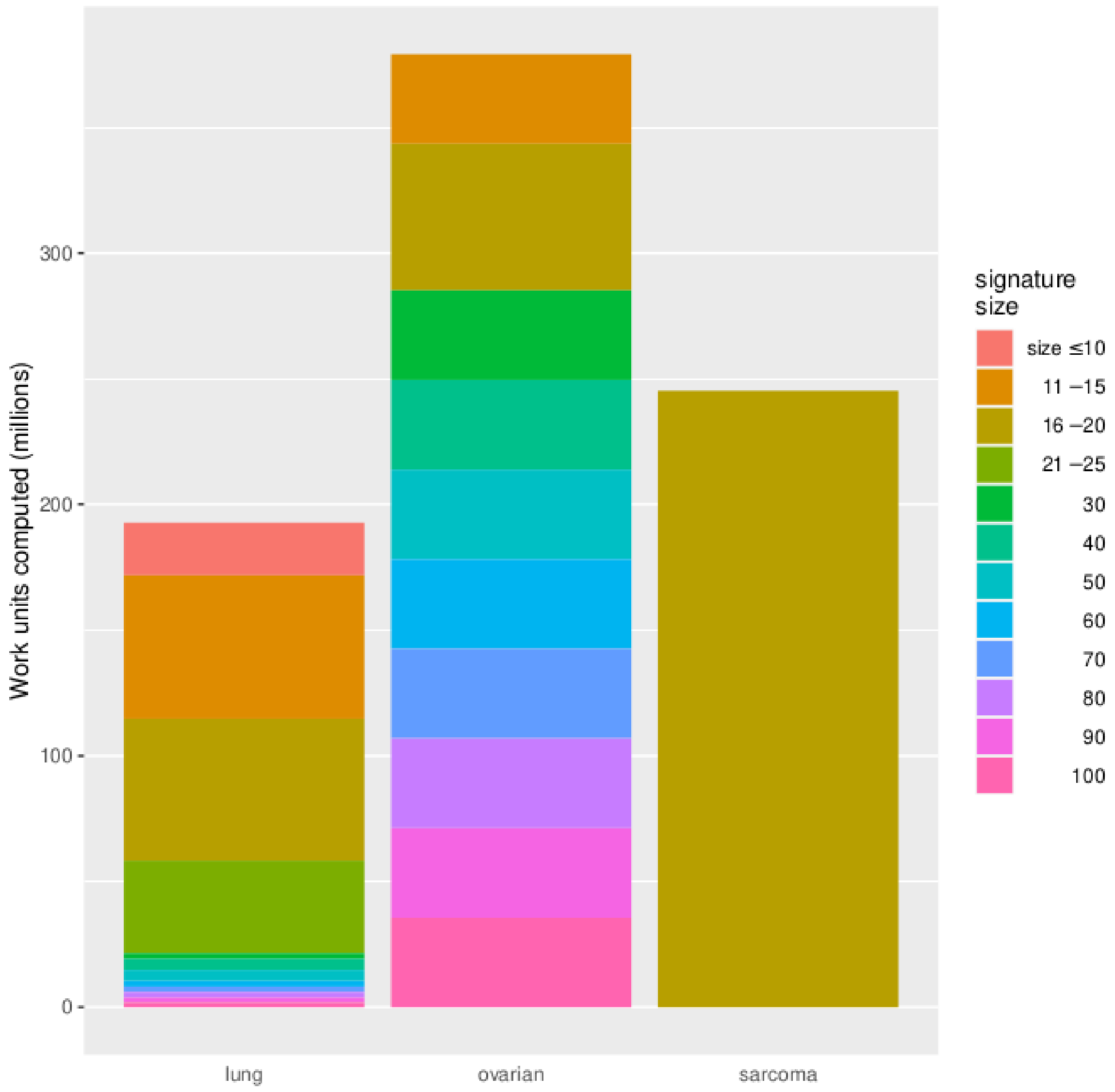

Figure 1: Number of completed work units per cancer type and signature size.

Lung cancer analysis

Several methods are available for lung cancer diagnosis but transthoracic needle aspiration and thoracoscopic biopsy are the methods with the highest sensitivity. Despite being highly accurate, these methods are invasive and scientists have searched for alternative screening methods or biomarkers to identify patients with cancer, especially in early stages. To identify new potential biomarkers, we tested multiple signatures in a dataset of tissues belonging to patients who have a history of lung cancer to find any groups of probes that could indicate the patient has early-stage lung cancer.

The dataset we chose to run on WCG comprises 192 histologically normal bronchial epithelium of smokers obtained at the time of clinical bronchoscopy. This procedure is routinely done, and thus being able to identify cancer markers expressed in the normal tissue would be an advantage. Of the 192 patient samples, 97 had lung cancer, 92 did not have lung cancer and 5 were suspected to have lung cancer. Our analyses focus on differentiating lung cancer from 92 non-cancerous samples.

WCG volunteers tested 4.5 trillion (4.5×10^{12}) candidate lung cancer signatures divided into 26 different diagnostic signature sizes. We then considered the signatures with Matthews correlation coefficient in the 99.999 percentile among all signatures of that same size. Figure 2 shows the distribution of biomarkers in the signatures. Count is the number of times a probe is present in the top signatures for its size.

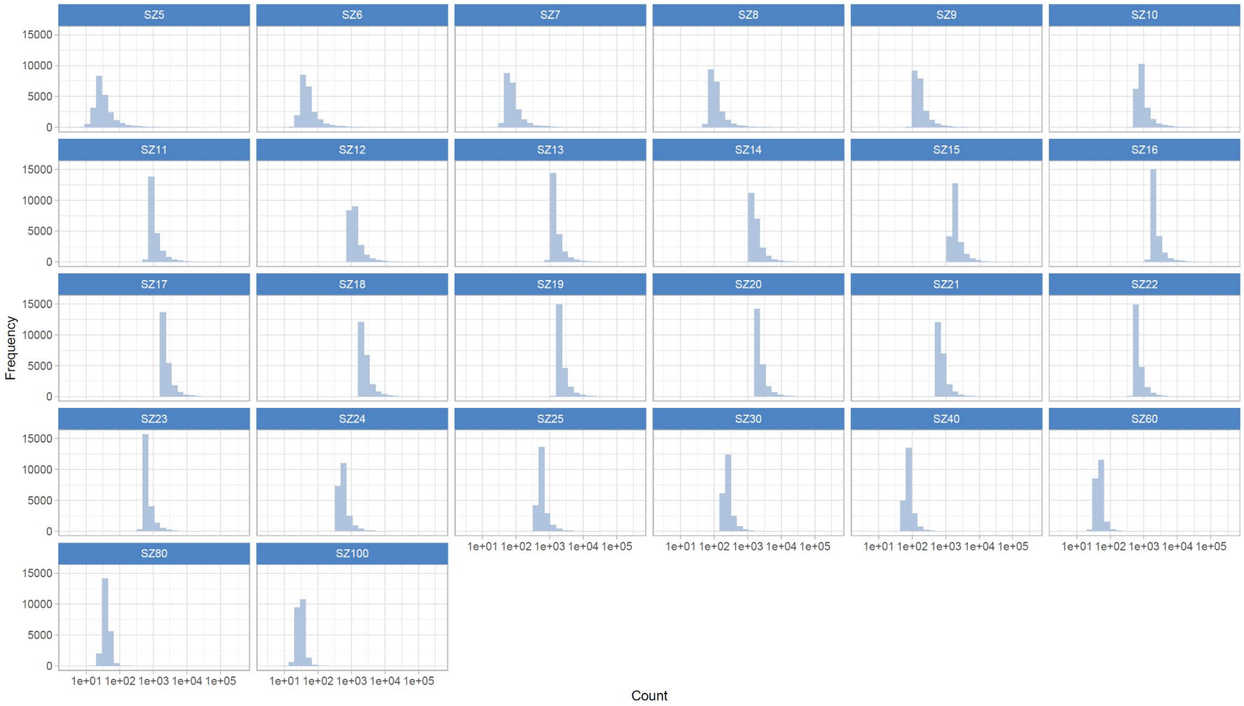

Figure 2: Frequency of each count (i.e., number of times a probe is present) per signature size.

As expected, smaller signatures have frequently lower counts, likely because they are not large enough to pick up the signal. Biological mechanisms are executed by multiple genes, creating a redundancy in the signature signal that can pick any gene from the same mechanism. Different genes could then be linked to the same mechanism, but at first sight (and without exploring the mechanism they are linked to), the genes in smaller signatures can be different.
As sizes increase, we see that the counts stabilize at around 1000, before reducing again when the signature sizes increase, likely due to the fact that it becomes too big to extrapolate the signal avoiding redundancy.

When considering the probesets in the 99th count percentile and present in at least 20 signature sizes, we can see that there is higher overlap in the signatures of sizes from 10 to 23, providing an indication that the best signature size window (for lung cancer at least) is in this range (Figure 3). This is going to be important for future studies on diagnostic signatures. Our aim is to release all data calculated with the help of WCG volunteers in a freely accessible database, where other researchers will be able to compare their results to the “lung molecular landscape” we obtained.

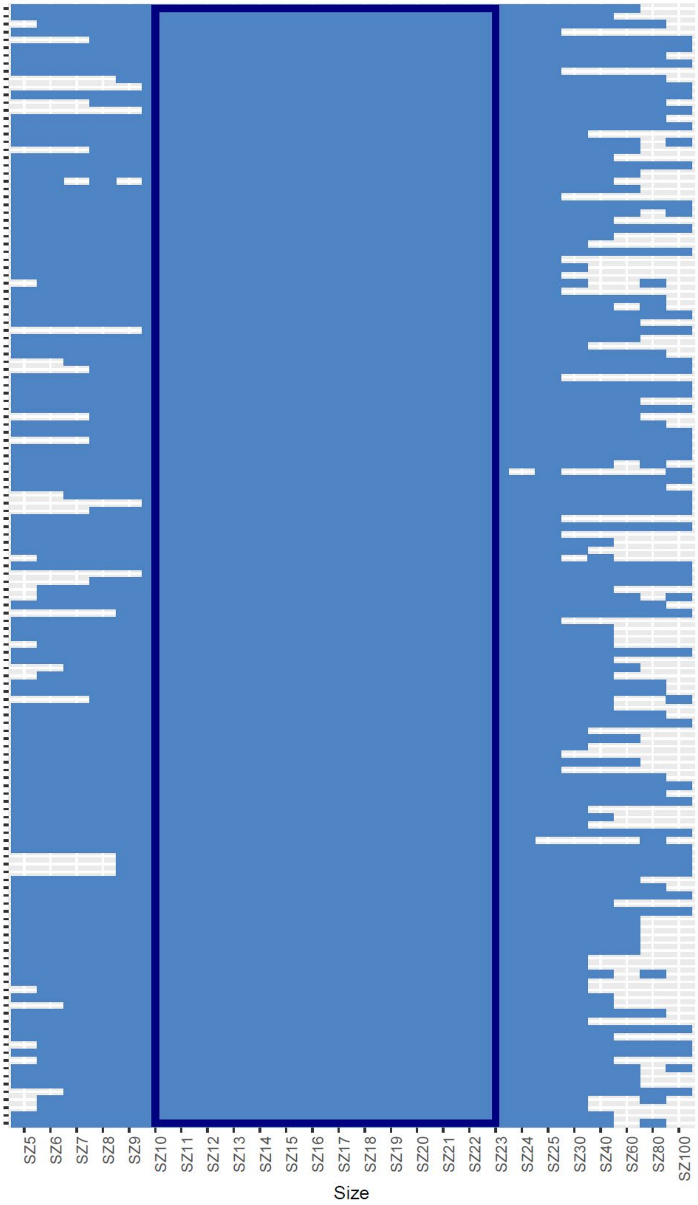

Figure 3: Heatmap of the 143 probes present in more than 20 signature sizes, and the sizes where they are present.

Exploring the results further, we found that 28 probesets (mapped to 26 genes), are present in all 26 signature sizes. One of these genes, VAMP1, is particularly promising as its protein is involved in synaptic vesicles fusion to the presynaptic membrane. Below are the cellular localizations for VAMP1 as per UniProt, and its 3D structure predicted by AlphaFold.

GO ID	Qualified GO term

GO:0005739	mitochondrion
GO:0005741	located_in mitochondrial outer membrane
GO:0005829	located_in cytosol
GO:0005886[5]	is_active_in plasma membrane
GO:0005887	plasma membrane

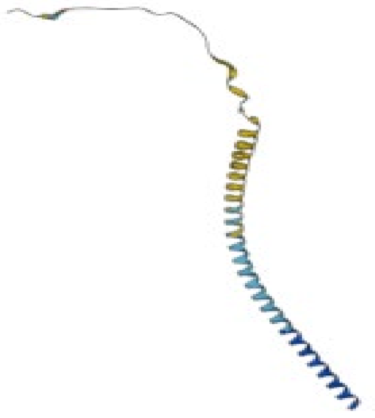

Figure 4. Cellular component annotations and protein structure for VAMP1 (as predicted by AlphaFold).

When investigating VAMP1 in external datasets to validate its importance, we found that it is significantly linked to the smoking status in lung cancer patients (through the cBioPortal, www.cbioportal.org). From our earlier study (https://pubmed.ncbi.nlm.nih.gov/25342220), we considered smoking-specific microRNAs that may regulate VAMP1 (as determined by our mirDIP data portal; http://ophid.utoronto.ca/mirDIP). Interestingly, hsa-miR-1262, a microRNA specific for former smokers, regulates VAMP1. None of the microRNAs specific to current smokers or never smokers appear to be relevant. Importantly, VAMP1 is protective for overall survival in lung cancer (high expression results in longer survival), making it a potential prognostic marker.

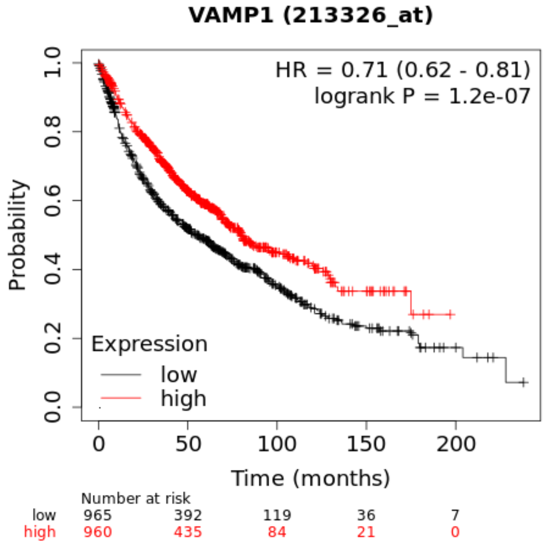

Figure 5. Kaplan-Meier plot shows that individuals with high expression of VAMP1 (in red) survive significantly longer compared to individuals with low expression of VAMP1 (in black). This shows that VAMP1 is a prognostic marker.

The protective value of VAMP1 is further supported by the gene expression in the two main lung cancer histologies, as shown in Figure 6. Normal samples express a much higher level of VAMP1 compared to cancer samples. This highlights that VAMP1 is also a potential diagnostic marker.

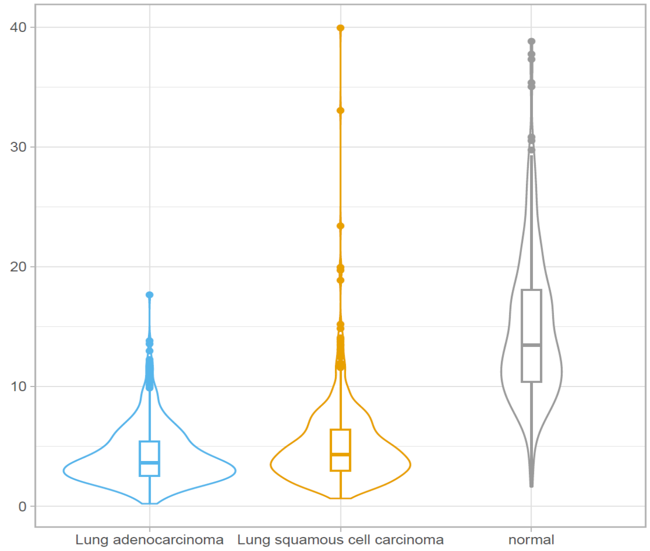

Figure 6. Expression of VAMP1 in normal and lung cancer samples.

Extending our observation beyond lung cancer, as described in our initial WCG project description, we find that VAMP1 behaves in a similar way in the majority of tested cancer as shown in Figure 7. This suggests that VAMP1 is a key player in carcinogenesis, and may be involved in a mechanism linked to the hallmarks of cancer.

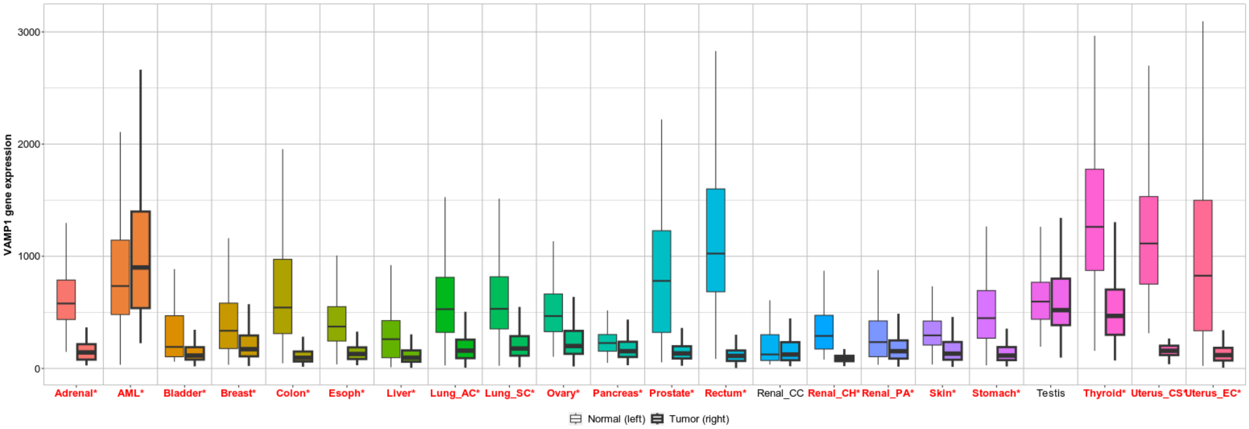

Figure 7. VAMP1 expression in normal and cancer for 22 different cancer types or subtypes. (Significant differences by Mann-Whitney U test are marked with red font).

We are excited about these results and will share more in the future as we continue with the validation of our results from lung cancer analyses.

WCG Team