Improving diagnosis and therapy of sarcomas and cancers of unknown primary using machine learning and RNA sequencing

Julien Vibert*[1], Gaëlle Pierron[1], Camille Benoist[1], Nadège Gruel[1], Delphine Guillemot[1], Anne Vincent-Salomon[1], Christophe Le Tourneau[1], Alain Livartowski[1], Odette Mariani[1], Sylvain Baulande[1], François-Clément Bidard[1], Olivier Delattre[1], Joshua Waterfall[1], and Sarah Watson[1]

[1]Institut curie – Institut Curie, PSL Research University – France

Abstract

Sarcomas represent about 15 % of all childhood cancers and are still lethal in a large proportion of cases. They comprise a heterogeneous group of bone and soft tissue tumors with more than 100 distinct histological entities. On the other hand, cancers of unknown primary (CUP) are metastatic cancers for which the primary tumor is not found despite thorough diagnostic investigations. Both entities thus represent a diagnostic challenge to pathologists.
To address this pathological complexity, RNA sequencing (RNA-seq) and machine learning techniques were used to improve the diagnosis and potentially guide therapy for these cancers. For sarcomas, RNA-seq was performed on more than 1600 patients, mainly children and young adults, addressed for sarcoma molecular diagnosis throughout France. For CUP, 20,918 public RNA-seq samples corresponding to 94 different categories, including 39 cancer types and 55 normal tissues, were used as training data, and RNA-seq was performed on a cohort of 48 patients with CUP.

A variational autoencoder (VAE) was used to perform unsupervised clustering and non-linear dimensional reduction of RNA-seq samples, allowing a finer and more insightful representation of the structure of the sarcoma and global cancer transcriptomic landscapes. A machine learning classifier tool based on the encodings learned by the VAE was developed to predict tissue of origin (TOO) for CUP.
Besides improving the diagnostic workflow for patients with sarcoma, the classifier exhibited an overall accuracy of 96% on reference data for TOO prediction in CUP. The TOO could be identified in 38 (79%) of 48 CUP patients. Eight of 11 prospective CUP patients (73%) could receive first-line therapy guided by the prediction, with responses observed in most patients. The variational autoencoder added further utility by enabling prediction interpretability. Thus, the classifier confidently predicted TOO for CUP and enabled tailored treatments leading to significant clinical responses.

*Speaker