Algorithms for cancer genome data analysis - Learning techniques for ITH modeling and gene fusion classification

The introduction of next-generation sequencing (NGS) technology resulted in an explosion of genomic sequencing data. To extract new and useful knowledge, new computational strategies for managing and investigating such data are required.

The first part of this thesis is dedicated to computational methods developed to model intra-tumor heterogeneity. Cancer is an evolving entity, and the evolutionary properties of each tumor are likely to play a critical role in shaping its natural behavior and how it responds to therapy. In fact, during the evolution of the disease, cancer cells differentiate, giving birth to subpopulations (subclones) characterized by a distinguishable set of mutations. This phenomenon, known as intra-tumor heterogeneity (ITH), may be studied using Copy Number Aberrations (CNAs). Nowadays, ITH can be assessed at the highest possible resolution using single-cell DNA (scDNA) sequencing technology. However, since the technology required to generate large scDNA sequencing datasets is relatively recent, dedicated analytical approaches are still lacking.

The first part of this Ph.D. thesis has been dedicated to designing new computational methods based on statistical and machine learning techniques to manage scDNA data and unveil spatial ITH.

- In this context, a tool capable of producing multi-sample CNA analysis on large-scale scDNA sequencing data and investigating spatial and temporal tumor heterogeneity has been developed. The main methodological contribution has been leveraging the advantages the of existing approaches, through a different, completely open, pipeline which, for the first time, integrates scCNA data from multiple samples to start investigating ITH from a qualitative point of view.
- Secondly, a study on clustering methods applied to scCNA data is presented. Clustering methods are increasingly applied to scDNA sequencing data to infer the subclonal structure of cancer. However, the complexity of these data exacerbates some data-science issues and affects clustering results. Additionally, determining whether such inferences are accurate and clusters recapitulate the actual cell phylogeny is not trivial, mainly because ground truth information is unavailable for most experimental settings. Here, by exploiting simulated sequencing data representing known phylogenies of cancer cells, a formal and systematic assessment of well-known clustering methods is presented to study their performance and identify the approach providing the most accurate reconstruction of phylogenetic relationships.
- Finally, a tool to explore the extent of spatial heterogeneity in multi-regional tumor sampling is proposed. The spatial distribution of subclones within a tumor mass can, in principle, be studied using scCNA profiles from multiple samples of the same tumor. However, the existing methods for scCNA analysis are still limited. Many of them only identify the total copy-number, while a few infer the tumor phylogeny using the computed CNAs. An instrument capable of exploiting both the granularity of single-cell DNA data and multi-sample analysis to quantify ITH still does not exist. For this reason, PhyliCS has been developed. PhyliCS is the first tool that exploits scCNA data from multiple samples from the same tumor to estimate whether the

different clones of a tumor are well mixed or spatially separated. In this regard, the SHscore (Spatial Heterogeneity score) is the key methodological contribution. It is a novel metric that allows to quantify how far cells from various samples from the same patient have diverged in their CN landscapes. The SHscore has been evaluated in a variety of simulation settings. Results show that the proposed score accurately represents heterogeneity in the clonal structure of multiple samples and indirectly reflects the evolutionary history of tumor subsamples.

Given the significant contribution of AI techniques in the study of complex biological phenomena characterized by a lack of domain understanding, they were adopted to investigate the oncogenic potential of gene fusions. Gene fusions are one of the most common somatic mutations and are considered to be responsible for 20% of global human cancer morbidity. However, not all gene fusions are oncogenic. Indeed, some are genuinely expressed in normal human cells or constitute passenger events. Nevertheless, the biological mechanisms which lead from gene fusions to tumorigenesis are not fully understood, and theoretical formulations of this complex phenomenon are still lacking. Therefore, AI algorithms represent an opportunity to infer the causal links between gene fusions and carcinogenesis directly from data.

The second part of this thesis has been devoted to the application of deep-learning techniques to the complex task of classifying gene fusions as oncogenic or not oncogenic.

- In this context, a tool based on a specifically designed neural network has been proposed to classify gene fusions as oncogenic or not oncogenic. Identifying potentially oncogenic gene fusions may improve affected patients' diagnosis and treatment. Previous approaches to this issue exploited protein domains, specific gene-related information, to predict the oncogenic potential of the gene functions. The proposed model profits from the earlier findings and includes the microRNAs in the oncogenic assessment. Specifically, the designed neural network integrates information related to transcription factors, gene ontologies, microRNAs, and other detailed information related to the functions of the genes involved in the fusion and the gene fusion structure. The designed neural network outperformed state-of-the-art tools.