Development of *in silico* samples and benchmarking of splicing and fusion detection tools applied in a clinical setting

Context

Tumours consist of various cells with distinct genetic variants (DNA) and abnormal transcript expression patterns (RNA), which can vary widely even with a similar clinical presentation. To face this heterogeneity, treatments are now prescribed to patients according to the genetic characteristics of their own tumours, in a so-called "personalized medicine".
RNA sequencing can be used to detect abnormal transcript expression, including splicing (alternative transcripts originating from the same gene) and fusion events (two different transcripts merged into a single chimeric transcript). However, its application in a clinical context remains challenging, especially with Formalin-Fixed Paraffin-Embedded (FFPE) tumours in which DNA and RNA are scarcer and of lower quality. To assess the limits of bioinformatics tools developed in-house or published by others in such a challenging context, a thorough benchmark is required. As only very few control samples with known events are available, synthetic datasets could provide the statistical robustness needed for this benchmark.

Mission

The M2 student will develop a protocol for generating synthetic sequencing data (FASTQ) as similar as possible to real data produced in our facility. This protocol should enable the following:

1. Adjusting parameters, such as the number of total and altered reads, insert lengths, or read quality.
2. Introducing known splicing and fusion events with a specific number of reads.

The student will also test splicing detection tools on the generated in silico samples.
She/He will be supervised by a bioinformatician and a clinician.

Location

The student will be located at the Hospices Civils de Lyon (HCL), Groupement Hospitalier Est (59 Boulevard Pinel; 69677 Bron; France), specifically in the bioinformatic group of the Centre de Biologie Pathologie Est (CBPE).
The team develops and implements bioinformatics tools for DNA and RNA analysis in a clinical context, later used by molecular pathologists to identify relevant genetic variants for diagnosis and to guide specific treatment decisions.

Profile

We are seeking a highly motivated and detail-oriented M2 student. The ideal candidate should possess the following qualifications and skills:

- Educational Background: Currently enrolled in or recently graduated from an M2 program in bioinformatics, computer science, or a related discipline.
- Programming Skills: Programming skills in scripting languages such as bash, Python or R are critical.
- Knowledge: Proficiency in RNA-seq data analysis and bioinformatic tool, knowledge of molecular biology, especially in the context of cancer biology, will be appreciated.

Contact

Marc Barritault: marc.barritault@chu-lyon.fr
Valentin Wucher: valentin.wucher@univ-lyon1.fr