Histoire évolutive d'une tumeur

Projet. Au cours d'un cancer, différentes mutations s'accumulent sur les cellules cancéreuses (Fig. 1-a), générant plusieurs lignées cellulaires qui co-existent dans une tumeur donnée. L'objectif de ce projet est d'étudier l'histoire évolutive d'une tumeur à partir de données de séquençage haut débit dites "*bulk*", c'est-à-dire mélangeant différentes cellules de la tumeur (Fig. 1-c et d).

Ces données sont complexes à la fois pour des raisons biologiques et techniques. L'évolution d'un cancer dépend en effet de nombreux processus biologiques, qui induisent notamment des mutations, des altérations structurelles de certaines régions du génome dans certaines cellules, ainsi que des variations de taille de la tumeur. D'un point de vue technique, le séquençage haut-débit ne fournit pas des séquences entières de génomes, mais renvoie un très grand nombre de petits fragments, appelés "*reads*", que l'on place sur une séquence de référence pour pouvoir les exploiter. Dans le cas de données *bulk* où l'on séquence plusieurs cellules, il n'est de plus pas possible de déterminer directement à quelle cellule appartient tel ou tel *read* (Fig. 1-d).

Le stage proposé s'inscrit dans un projet plus général dont l'objectif est de reconstruire l'histoire de la composition cellulaire de la tumeur d'un patient à partir de biopsies de suivi réalisées à plusieurs temps différents et séquencées. L'approche envisagée repose sur la mise au point d'un modèle stochastique des données de séquençage *bulk* d'une tumeur. Un tel modèle se décompose naturellement deux parties principales. Un processus de naissance et mort (pour la division et la mort cellulaire), couplé à un processus de Poisson (pour les mutations), peut en premier lieu être utilisé pour modéliser l'évolution du nombre de cellules de chaque lignée et l'apparition de nouvelles lignées (Fig. 1-a). Conditionnellement à cet effectif des lignées cellulaires, la seconde partie modélise le prélèvement des cellules tumorales et leur séquençage haut débit, qui produit l'ensemble des *reads* observés (Fig. 1-b, c et d).

Ce modèle pourra être utilisé pour simuler des données de séquençage sous diverses hypothèses biologiques, afin de tester la robustesse et la précision des méthodes de reconstructions déjà existantes, telles que Pairtree (Wintersinger et al., 2022) ou CALDER (Myers et al., 2019). Un second objectif, plus ambitieux, sera de calculer la vraisemblance de données de séquençage *bulk* sous ce modèle afin de proposer une nouvelle méthode d'inférence statistique, en adaptant par exemple l'approche de Didier and Laurin (2020) pour la première partie du modèle.

Environnement et compétences attendues. Le stage se déroulera entre l'IMAG à Montpellier et le MAP5 à Paris, le ou la stagiaire étant libre de choisir sa localisation principale. Il sera encadré par Gilles Didier et Paul Bastide, en collaboration avec Alice Cleynen et Sophie Lèbre. Le projet s'inscrit dans l'**ANR IdenTHiC** (*Identification of Tumor HIstory at the Clone level*), qui porte sur l'étude de données cliniques de patients atteints de cancers pour l'aide au diagnostic. Une bourse de thèse sur le sujet pourra être disponible (à l'IMAG). Le stage, d'une durée de 4 à 6 mois, pourra commencer dès **février ou mars 2025**, et sera rémunéré en fonction du taux légal d'indemnité de stage en vigueur. Le développement de l'outil de simulation nécessite un goût pour la programmation, notamment en R ou python. L'étude du modèle met en œuvre des compétences en probabilité et statistiques et bénéficierait d'un intérêt pour les applications biologiques.

Contact. CV et lettre de motivation sont à adresser à Gilles Didier et Paul Bastide (`gilles.didier@umontpellier.fr`, `paul.bastide@u-paris.fr`).

Références

Didier, Laurin. 2020. *Systematic Biology*. 69:1068–1087.
Myers, Satas, Raphael. 2019. *Cell systems*. 8:514–522.
Wintersinger, Dobson, Kulman, et al. 2022. *Blood Cancer Discovery*. 3:208–219.

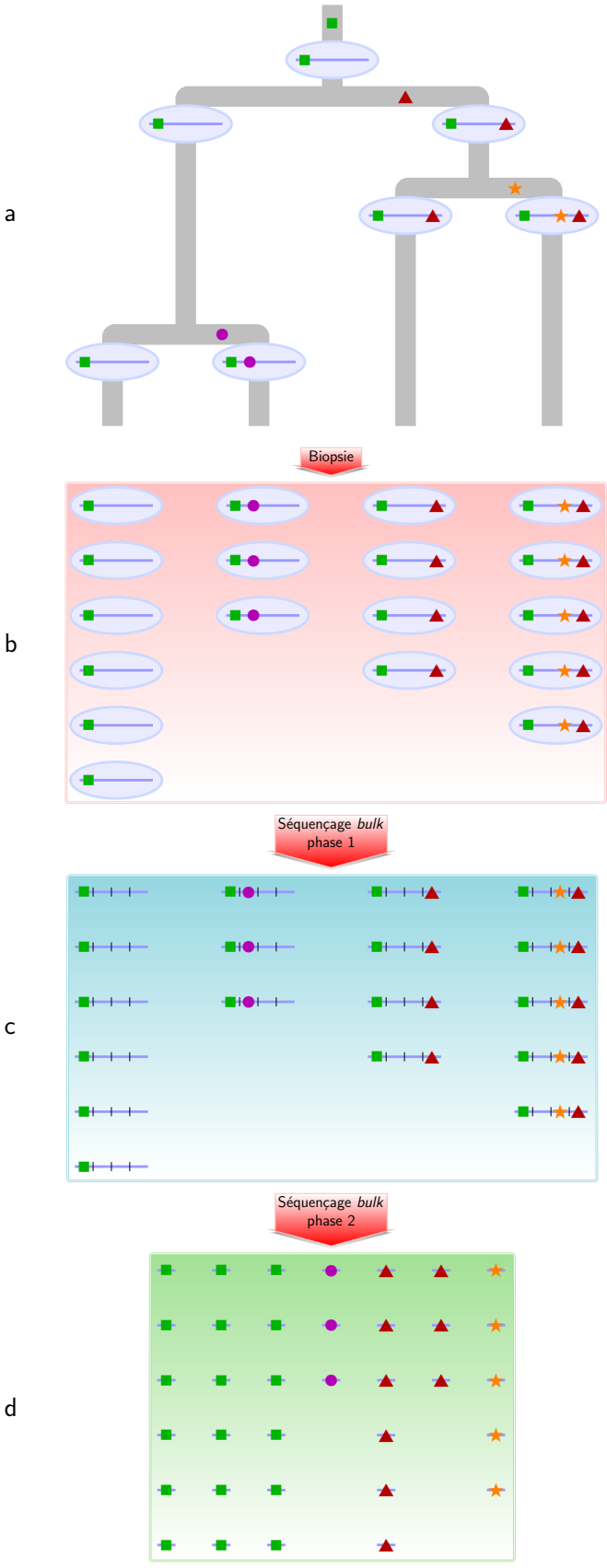

Figure 1: a) Évolution des cellules tumorales. Les mutations sont représentées par des pictogrammes colorés. b) Ensemble des cellules tumorales prélevées lors d'une biopsie. c) Phase 1 du séquençage *bulk*. Les séquences génétiques des cellules prélevées sont extraites, mélangées et découpées en *reads*. d) Phase 2 du séquençage *bulk*. Les *reads* sont séquencés et les mutations sont identifiées et comptées.