Comment identifier et sélectionner les mutations « drivers » oncogéniques ?

Un nouvel outil de détection et d'annotation de variants obtenus après NGS en absence de matériel contrôle non-tumoral

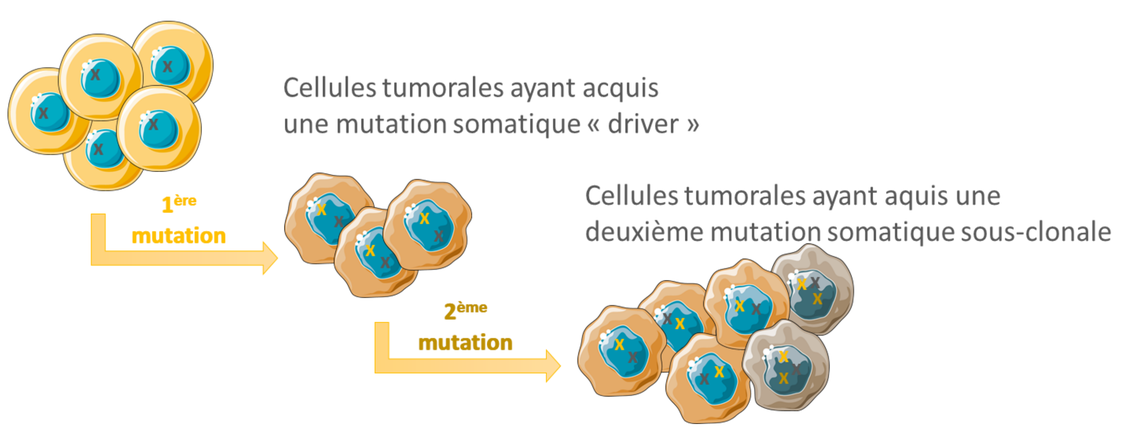
V. CHESNAIS[1], S. DIRY[1], S. HENNE[1] and E. GINOUX[1]

[1] Life & Soft, Plessis-Robinson, France

INTRODUCTION

Le récent développement des technologies de séquençage nouvelle génération (NGS) entraîne l'accumulation de données, riches d'informations mais complexes à analyser. En oncologie, la classification des tumeurs repose de plus en plus sur des données de génomique, et notamment sur des profils mutationnels particuliers.

L'une des principales difficultés rencontrées est **l'absence de matériel contrôle** associé aux échantillons tumoraux rendant difficile l'interprétation des variants détectés. Le défi est donc de développer un outil capable de discriminer les mutations dites « **drivers** » responsables de l'émergence d'une tumeur, des variants dits « **passengers** » acquis dans les cellules au cours de la vie et germinaux. De plus, le séquençage à haut-débit entraîne un **taux d'erreur** de l'ordre de 0.1% augmentant le bruit de fond lors de la détection de variants.



METHODE

Le pipeline d'analyse que nous présentons a deux objectifs principaux répondant aux besoins d'analyses de variants après NGS :

1. **Discriminer les erreurs de séquençage des vrais variants**. Il associe pour cela plusieurs outils d'appel de variants afin d'augmenter à la fois la spécificité et la sensibilité de détection des variants notamment les insertions/délétions ou ceux de faibles fréquences alléliques (VAF). Il fournit également de nombreuses annotations qualitatives, ainsi qu'une analyse de la récurrence de chaque variant dans toutes les analyses effectuées.
2. **Différencier les mutations « drivers » des mutations « passengers »**. Des annotations fonctionnelles intègrent des bases de variants telles que 1000 Genomes, ExAC ou ClinVar ainsi que des prédictions d'impact fonctionnel afin de faciliter l'interprétation des résultats.

Une première validation de notre pipeline a été réalisée grâce à des données *in silico* générées avec le simulateur wgsim. Un panel de 11kb ciblant les 5 gènes les plus fréquemment mutés dans les SMD (source Cosmic) a été généré *in silico*: 15 mutations ont été introduites à des VAF variables (5%, 25% et 45%). Une deuxième validation a été effectuée sur 5 échantillons Horizon séquencés sur MiSeq avec le panel Cancer Hotspot v2 (22kb) et disponibles sur la plateforme de partage Illumina. Ces échantillons sont porteurs de 12 à 44 variants de VAF médiane de 8% (range: 1% - 30%).

RESULTATS

Les données *in silico* ont été générées afin d'avoir une profondeur moyenne du panel de 1200X. Nous avons ainsi pu observer que les frameshift sont détectés principalement par 1 ou 2 variant callers alors que les SNPs sont détectés le plus souvent par les 3 variant callers (**figure A**). De la même façon, plus la VAF des mutation est faible, plus le nombre d'outils capables de les détecter est faible (**figure B**). De manière intéressante, les combinaisons d'outils sont différentes selon le type et la VAF des variants à détecter. Dans tous les cas, les VAF estimées par les différents variants callers sont similaires aux VAF théoriques avec un écart maximum de 3.9% pour une VAF théorique de 35%. **Ces résultats tendent à montrer l'importance de combiner différents outils d'appels de variants afin d'optimiser la détection de tous les types de variants en choisissant des outils fonctionnellement complémentaires.**

Nous avons également regardé l'impact de la profondeur de séquençage et du taux d'erreur sur la détection des variants. Comme attendu, la diminution de la profondeur de séquençage entraîne une diminution de la sensibilité de détection des variants pour 2 outils (**figure C**). De plus, le nombre de variants détectés par les 3 outils diminue, au profit d'une détection par seulement 2 outils. L'augmentation du taux d'erreur de séquençage a les mêmes effets (**figure D**). **Ces résultats tendent à montrer l'importance des conditions de séquençage sur la détection. Selon l'outil d'appel de variant, ces conditions peuvent avoir plus ou moins d'effet sur les résultats.**

Les données Illumina ont été séquencées à une profondeur moyenne de 3000X. Sur les 91 mutations recherchées, notre pipeline en détecte 79 (sensibilité de 87%). La sensibilité de chacun des outils est très variable et s'explique par la VAF des mutations recherchées (**figure E**). Enfin, le taux de prédictions positives obtenu pour chacun des outils de détection de variants est compris entre 48% et 69% (**figure F**). **Ces résultats confirment la robustesse de notre pipeline pour la détection de variants somatiques et notamment de variants sous-clonaux jusqu'à une VAF minimale de 1%.**

CONCLUSION

Notre pipeline permet de détecter tous les types de variants jusqu'à une VAF de 1%. Chacun des outils ayant ses spécificités, il semble important de choisir les outils à considérer en fonction de ce que l'on recherche: types de mutations, contexte génomique, VAF… En effet, en absence de matériel contrôle, l'interprétation des résultats repose essentiellement sur la sensibilité et la spécificité des outils d'appels de variants, ainsi que sur l'annotation qualitative et fonctionnelle des variants détectés.

Enfin, les CNVs peuvent avoir un impact sur la détection des variants et sur leur VAF. L'intégration d'une analyse du nombre de copies des régions séquencées permettra, par la suite, d'augmenter la qualité des résultats rendus par notre pipeline.

Life &Soft