PGTB
SEQUENCAGE | GENOTYPAGE

Newsletter

#9 - NOVEMBRE 2022

Le mot de la direction

Depuis plusieurs années Illumina domine le domaine du séquençage haut-débit court fragments. L'arrivée de nouveaux acteurs sur le marché force Illumina à réagir avec l'annonce de séquences plus longues et d'une nouvelle chimie dont bénéficiera le NextSeq2000 que nous avons récemment acquis. Les choses évoluent aussi du côté du séquençage longs fragments avec les mises à jour récentes des kits et *flow cell* chez Oxford Nanopore Technologies qui permettent d'améliorer significativement la qualité des séquences. Oxford Nanopore Technologies a par ailleurs annoncé un petit séquenceur, le P2, capable de séquencer deux *flow cells* PromethION en parallèle, rendant ainsi le très haut-débit plus accessible (jusqu'à 290Gb par flow cell). Plus de profondeur et de longueur de séquençage devraient affiner notre capacité à décrire le polymorphisme dans les (méta)génomes et à caractériser des régions génomiques qui se sont avérées difficiles à caractériser jusqu'alors.

Les articles de cette lettre illustreront les nouvelles possibilités offertes par ces évolutions techniques récentes et à venir prochainement.

Enfin, nous vous invitons à visiter notre nouveau site internet, où vous trouverez de nombreuses informations sur les services proposés et les technologies que nous développons et mettons en œuvre et leurs applications.

Un nouvel ingénieur dans l'équipe

Zachary Allouche a rejoint la PGTB pour deux ans en tant qu'ingénieur d'étude au sein de l'équipe depuis le 1er octobre. Il sera en charge des projets de séquençage Illumina et de la bioanalyse.

NextSeq 2000 : bilan après un an d'utilisation

En novembre 2021, nous avons mis en service notre NextSeq 2000, le dernier séquenceur haut-débit d'Illumina qui intègre la plateforme Bio-IT DRAGEN (Dynamic Read Analysis for GENomics), une solution ultra-rapide pour l'analyse secondaire des données produites par le séquenceur pour diverses applications (voir tous les détails dans notre Newsletter#8). Le bilan sur les runs réalisés depuis 12 mois est très satisfaisant : les *patterned flow cell* garantissent des densités de chargement optimales et donc des rendements élevés (en général 10 à 30% de reads obtenus en plus par rapport aux spécifications d'Illumina). Les trois *flow cell* disponibles (P1 : 100 millions | P2 : 400 millions | P3 : 1200 millions de paires de reads) et les différentes configurations de longueur de séquençage (1x50pb à 2x150pb) offrent une très grande flexibilité, qui permet de calibrer au mieux le séquençage en fonction du nombre d'échantillons et de l'application. Enfin, la grande facilité d'utilisation du NextSeq 2000 permet d'offrir des délais réduits pour l'obtention des résultats de séquençage.

A ce jour, nous avons déjà réalisé des runs de séquençage *shotgun*, RNAseq, *small* RNA, *single-cell*, CHIP-seq, Cut&Run, capture de gènes, capture bisulfite, et métatranscriptomique. La suite Bio-IT DRAGEN a été principalement utilisée pour des projets de RNAseq, sur des espèces modèles et non-modèles, et le sera prochainement sur des données single-cell. Les runs en *ready-to-load* (librairies préparées par l'utilisateur) représentent 75% des séquençages réalisés depuis un an.

Nouveaux kits et nouvelle chimie

Nous attendons avec impatience la chimie 2x300pb sur les *flow cell* P1 et P2, annoncée pour la fin de l'année 2022 par Illumina, qui permettra d'obtenir notamment des données de métagénomique plus précises et avec plus de profondeur que ce qui est aujourd'hui possible sur le séquenceur MiSeq, et par conséquent de réduire fortement le coût à l'échantillon. Illumina annonce également un kit 2x50pb sur la *flow cell* P1, qui n'offrait pour l'instant que du 2x150pb.

Enfin, nous savons désormais que le NextSeq 2000 bénéficiera début 2024 de nouvelles *flow cell* P4 générant plus de 500Gb et de la nouvelle chimie XLEAP-SBS présentée par Illumina et récemment déployée sur son nouveau séquenceur très haut-débit NovaSeq X. Avec cette chimie, Illumina annonce des runs deux fois plus rapides, des séquences deux fois plus longues et des données trois fois plus précises.

Oxford Nanopore : où en est la technologie aujourd'hui ?

Le séquençage de 3^{éme} génération par Oxford Nanopore Technologies (ONT) repose sur le passage d'un long fragment d'ADN à travers des pores permettant de générer un signal électrique pouvant être traduit en bases nucléotidiques. Cette technologie présente l'avantage de pouvoir séquencer des molécules d'ADN ou ARN natif de tailles très longues. Cependant depuis sa commercialisation, le séquençage par nanopore souffre d'une résolution trop imprécise générant entre 5 à 10 % d'erreurs de séquençage. Plusieurs facteurs technologiques (pores, enzymes, vitesse de passage de la molécule, algorithme de traitement du signal, ...) ont été optimisés au cours des années afin de réduire le taux d'erreur de séquençage. Récemment, ONT a développé un nouveau kit (appelé « Q20+ chemistry ») promettant une précision des lectures brutes supérieures au score Phred de 20 (i.e. ≤ 1 % d'erreur). De plus, la technologie offre dorénavant la capacité de lire certaines lectures (à un taux annoncé à environ 10 % des lectures) deux fois (sens et anti-sens) ce qui permet également de réduire le taux d'erreur de séquençage à un score Phred de 30 (i.e. 0,1 % d'erreur), une qualité s'approchant de celle des séquences produites en Illumina.

Depuis le début de l'été 2022, la PGTB a testé ces nouvelles améliorations. Afin de vérifier les promesses faites par ONT, nous avons testé ces avancées dans différents domaines (séquençage de génome entier et séquençage d'amplicons).

Grâce à un projet de séquençage de génome entier de mycoplasmes réalisé chaque année sur la plateforme, nous avons pu comparer sur un même échantillon les différents kits afin d'observer l'évolution de qualité des séquences. La Figure 1A représente la distribution de qualité des séquences en fonction des kits sur ce même échantillon. Dans un premier temps, on observe bien une évolution de la qualité moyenne des séquences en fonction des avancées technologiques. En effet, celle-ci passe d'un score Phred de 14,6 sur l'ancienne chimie (~3,5 % d'erreur) à 17,8 (~1,7 %) sur la chimie Q20+ et à 30 (0,1 % d'erreur) lorsqu'on s'intéresse aux lectures séquencées en duplex. Malgré une évolution de la qualité moyenne, on observe que la nouvelle chimie Q20+ (simplex) n'atteint pas le seuil promis des 1 % d'erreur.

En revanche, pour les lectures duplex le score moyen de Q30 est bien atteint, cependant le taux des 10 % des lectures du run pouvant être séquencé en duplex est loin d'être atteint puisque ces séquences ne représentent que 0,33 % de l'ensemble des séquences (54 lectures sur 16519). Ce faible taux de lectures en duplex peut s'expliquer par une taille de séquences longues rendant plus difficile le passage du fragment dans le pore à deux reprises.

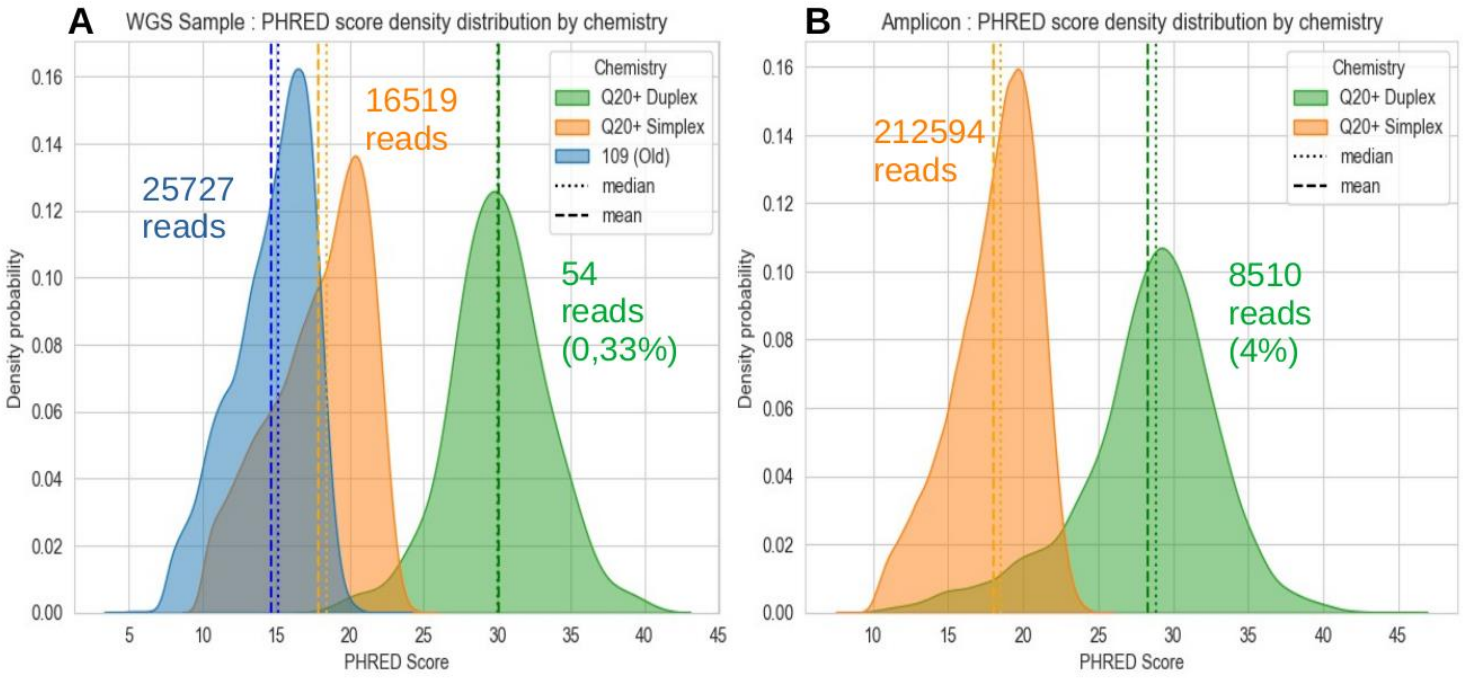

Figure 1 : Distribution des scores Phred de qualité des séquences issues des kits LSK109 (bleu) et LSK112 (Q20+ simplex et duplex reads) pour plusieurs applications : séquençage de génome complet (A), séquençage d'amplicon HLA-G (B)

Par ailleurs, d'un point de vue de la qualité d'assemblage, les données obtenues à partir de la nouvelle chimie Q20+ donne une amélioration significative. En effet, lorsqu'on regarde l'intégrité du core-genome des mycoplasmes, on retrouve 139 gènes BUSCO sur les 151 de la base de données avec la chimie Q20+, alors que l'ancienne chimie (LSK109) ne permettait d'en identifier que 79.

Nous avons également pu tester la nouvelle chimie Q20+ et le mode duplex dans le domaine du séquençage de fragments d'ADN d'intérêt amplifiés par PCR. Ce projet par séquençage d'amplicons consistait à séquencer un amplicon ciblant le gène codant pour l'antigène du complexe majeur d'histocompatibilité (HLA-G) d'une longueur de 5 kb chez 93 patients dans le but de déterminer les haplotypes. Un total de 212 594 lectures avec une qualité moyenne de 18,1 (~ 1,55 % d'erreur) a été obtenue en utilisant la chimie Q20+ après démultiplexage et trimming sur la taille de l'amplicon (Figure 1B). Le mode duplex a quant à lui pu être appliqué à 8510 lectures (soit 4% du total) ce qui représente une couverture moyenne de 91x par échantillon.

D'autre part, ce séquençage en mode duplex permet d'obtenir une qualité moyenne de 28,3 (~ 0,15 % d'erreur). Ainsi, même si le taux de séquences pouvant être séquencé en mode duplex est encore bien en dessous des 10 % et que la qualité moyenne est plus faible que les Q30 promis, la production de ces données a largement permis d'obtenir la phase pour cette région cible de 5kb.

Ainsi, grâce à la poursuite des améliorations, le séquençage ONT nous paraît prometteur pour ce type d'application, avec sûrement la possibilité de cibler plusieurs régions de plus grandes longueurs et d'obtenir l'information de la phase sur un plus grand nombre d'échantillons. Ce type de données se prêterait par exemple bien aux études de phylogénie ou de phylogéographie.

Selon nous, il serait également possible d'extrapoler le séquençage d'amplicon à des analyses de metabarcoding par exemple sur le 16S entier, ce qui devrait permettre d'assigner les communautés taxonomiques jusqu'au rang de l'espèce.

En conclusion, on remarque que les chiffres donnés par ONT sont pour le moment difficile à atteindre (taux de lectures duplex et qualité moyenne) chez les espèces non-modèles. Cependant, les progrès réalisés par ONT sur la qualité des séquence produites ces derniers mois sont importants et ouvrent de belles perspectives d'évolutions pour différentes applications, notamment si on prend en compte que la majorité des erreurs de séquençage produites par cette technologie concerne les homopolymères (motifs ADN de répétitions).

Ne pas jeter sur la voie publique
Crédits photos : INRAE, Pixabay, ONT et PGTB
Textes et conception graphique de la PGTB