Éléments de validation de la qualité de séquençage, d'alignement, de la couverture des designs et panels de gènes

Définitions

- Un **design** est la liste des régions séquencées. Généralement un fichier fourni par le fournisseur du kit (Manifest/BED), et traduit au format BED.

- Un **panel de gènes** est une liste de régions regroupées par gènes. Un gènes est donc une liste de régions (généralement les exons). Il s'agit d'un fichier au format BED. Un Panel peut être regroupé autrement que par gènes : par exons, par groupe de gènes (le groupe des gènes BBS, BRCA...).

- Un **fichier fastq** est la liste des reads séquencés, dont chaque base est associée à une valeur de qualité. Pour la technologie Paired-End, 2 fichiers fastq sont générés, chacun ayant exactement le même nombre de reads.

- Un **fichier BAM** est une liste de reads, et un ensemble d'information associées : les qualités de chaque bases, un tag (duplicat, short-read, alignement secondaire, forward ou reverse...), les coordonnées de l'alignement (si alignement), le code CIGAR (si alignement)... Un fichier BAM non-détruit est un fichier BAM qui contient au moins toutes les informations du/des fichier-s fastq.

- Un **fichier BAM de validation** est une liste de reads « validés » (qualité >10, non-dupliqués, non alignement secondaire...), représentant les reads utilisés pour le calling. Ce BAM est utilisé pour évaluer la qualité de la couverture, du design, des panels de gènes.

- La **profondeur** est le nombre de base séquencé à une position sur le génome. e.g. une profondeur de 42X à la position chr1:123456 signifie que 42 bases (A, T, G ou C) ont été séquencées à la position chr1:123456.

- La **couverture** est le pourcentage de base dont la profondeur est supérieure à un seuil, pour une région données. e.g. 98% des bases ont une profondeur supérieure à 30X sur l'ensemble des exons du gènes X.

Validation de la qualité de séquençage et d'alignement

Sur le BAM non-détruit, certains indices de qualité sont calculés.
Exemple :
- nombre de reads total,
- Q30 (pourcentage de bases de qualité >30, moyenne sur l'ensemble de la longeur du read),
- les qualités par bases (pour chaque position des reads)
- nombre de reads alignés (sur l'ensemble du génome),
- reads non-dupliqués (ou reads uniques)

Validation de la couverture

Les calculs de profondeur et de couverture sont réalisés sur le BAM de validation.
Ces calculs peuvent être utilisé sur le design et sur les panels de gènes, pour une validation initiale et pour les validations continues (pour chaque run).
Les indices de qualités calculables peuvent être réalisés sur l'ensemble des régions d'un BED (globale), et par régions/gènes.

Les **paramètres** utilisés sont :

- le **seuil de profondeur minimum** (e.g. 30X). Sous ce seuil, la position est considérée comme insuffisamment séquencée (FAIL)

- le **seuil de profondeur attendue** (e.g. 100X). Sous ce seuil, la position est considérée comme correctement séquencée mais en alerte (WARN)

- la **couverture minimum** (e.g.). Sous ce seuil, la région est considérée comme insuffisamment couverte (dépend de la profondeur considérée)

Les **indices** calculables :

- le nombre de reads **ON target**, c'est-à-dire le nombre de reads « validés » sur les régions/gènes considérés

- la **couverture globale de séquençage**, c'est à dire le pourcentage de base dont la profondeur est nulle. Un pourcentage inférieur à 100% indiquerait un problème dans le design

- la **couverture globale au seuil minimum** (e.g. 98% à 30X). Une couverture inférieur à la couverture minimum (e.g. 95%) indiquerait une erreur de séquençage, comme problème technique ou particularité biologique comme la délétion d'une région (e.g. couverture PASS)

- la **couverture globale au seuil attendue** (e.g. 94% à 100X). Une couverture inférieur à la couverture minimum (e.g. 95%) indiquerait un séquençage à surveiller (e.g. couverture WARN)

- les **couvertures par régions/gènes**, identifiant les régions/gènes :
    - non séquencés (e.g. couverture <100% à 1X),
    - insuffisamment séquencés (e.g. couverture <95% à 30X), ou
    - en alerte (e.g. couverture <95% à 100X).

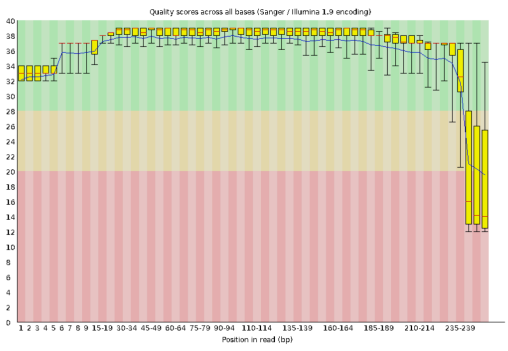
Capture d'écran des indices disponibles dans le future rapport de STARK

Ces captures d'écran montrent l'état d'avancement de la construction du prototype du future rapport (HTML, page web, PDF) de la version suivante de STARK (0.9.18b), pour la partie validation du séquençage, de l'alignement et des couvertures des panels de gènes.
S'agissant d'un prototype, certains points ne sont pas encore définitif, d'autres à titre d'exemple, d'autres manques (e.g. couverture globale de séquençage)

Sequencing & Mapping

Sequenced reads were mapped to the hg19 assembly version using bwamem aligner(s).
Statistics were determined with FastQC on fastq files, and Samtools flagstat on BAM file(s).
On target reads are uniq (no duplicates), with quality higer than 10, and aligned on the defined design.

Per base quality



Reads metrics

Total number of reads - 230133	100 %
Aligned reads (bwamem) - 220453	90 %
Uniq reads (bwamem) - 210836	80 %
On target reads (bwamem) - 200743	70 %

Depth & Coverage

Depth and coverage generated with Samtools mpileup and Picard CollectHsMetrics
on reads with quality higher than 10, without duplicated read, and with clipped overlapped bases

IFU176 EGFR18-21MASTR Manifest v130729.AmpliconManifest.by_exon.genes
X Gene Panel on **bwamem** aligner reads
58 Genes are fully covered

96	NaN	2	4
100X	**30X**	**Warning coverage**	**Failed coverage**
Percentage of bases with depth higher than 100	Percentage of bases with depth higher than 30	Number of genes with a warning coverage <100% bases >100X	Number of genes with a failed coverage <100% of bases >30X