GenomEast Platform
CERBM GIE – IGBMC
1 rue Laurent Fries
67404 Illkirch Cedex
France

+33 3 88 65 34 26
thibault@igbmc.fr
jost@igbmc.fr
genomeast.igbmc.fr

FICHE PRODUIT : SEQUENÇAGE DES LIBRAIRIES DES UTILISATEURS

La plateforme propose une prestation de séquençage de librairies préparées par les utilisateurs.

1 Services proposés

- La vérification des librairies préparées par l'utilisateur :
  - Quantification et vérification de la qualité des librairies par électrophorèse capillaire (Bioanalyzer d' Agilent ou Fragment Analyzer d'AATI).
- Le séquençage avec le séquenceur HiSeq 4000 d'Illumina :
  - Chargement des librairies sur la flow cell et génération des clusters avec la Cbot (Illumina).
  - Séquençage simple ou pairé avec des tailles de lecture de 50 ou 100 pb selon les options choisies par le porteur de projet. Nous pouvons réaliser d'autres longueurs de lecture, mais uniquement pour des flow cells complètes (8 pistes).
  - Nous ne séquençons les librairies fournies par les utilisateurs que sur des pistes entières (c.-à-d. nous ne multiplexons pas sur la même piste des librairies préparées par un utilisateur avec des librairies préparées par notre plateforme ou par un autre utilisateur). Nous ne commençons un run de séquençage que lorsque la flow cell (8 pistes) est complète avec des échantillons dont la longueur de lecture est similaire.
- L'analyse primaire des données
  - Démultiplexage et création des fichiers FASTQ.
  - Suppression des dimères d'adaptateur.
  - Contrôle de la qualité du séquençage.
  - Détection d'éventuelles contaminations.
  - Création d'un rapport synthétisant les méthodes utilisées pour l'analyse primaire et les résultats obtenus (un rapport détaillé pour chaque échantillon et un rapport global plus synthétique pour chaque projet).
- L'analyse ultérieure des données (optionnelle, voir section 5 pour plus d'information)

2 Informations à fournir par les utilisateurs

Afin de vérifier leur compatibilité avec la technologie HiSeq 4000 d'Illumina et pour optimiser les conditions de séquençage, il est important de transmettre à la plateforme le détail de la construction des librairies générées. Le tableau ci-dessous liste les documents à fournir avec les informations recherchées et leur finalité.

Documents à fournir	Informations recherchées	Finalité
Protocole de préparation des librairies (référence du kit, publication …)	Schéma de la construction des librairies (indexes simples ou doubles, barcodes, UMI …) Méthode d'ajout des adaptateurs (PCR ou ligation)	Déterminer la compatibilité avec la technologie Illumina et définir les conditions de séquençage Limiter les problèmes d'index hopping liés à un reliquat d'adaptateurs dans les librairies

GenomEast Platform
CERBM GIE – IGBMC
1 rue Laurent Fries
67404 Illkirch Cedex
France

+33 3 88 65 34 26
thibault@igbmc.fr
jost@igbmc.fr
genomeast.igbmc.fr

Documents à fournir	Informations recherchées	Finalité
	Diversité en ATGC des 25 premières bases de la librairie (***Ref 2***)	quand ils sont ajoutés par PCR (***Ref 1***) Déterminer la nécessité de multiplexer les librairies fournies avec la librairie contrôle PhIX d'Illumina ou avec une autre librairie équilibrée pour optimiser le nombre de clusters sur la piste
	Nécessité d'utiliser un primer de séquençage custom (à fournir : ≥20µl à 100 µM dans de l'eau MilliQ dans des tubes LowBind)	Séquencer les librairies avec des adaptateurs custom
Listing des échantillons et des indexes utilisés (à renseigner sur le LIMS de la plateforme ou sur unfichier Excel : *voir modèle page 4*)	Nom, concentration et volume des échantillons (**si les échantillons sont soumis en un pool, le détail de chaque échantillon du pool est à fournir**) Nom, séquence et fournisseur des indexes utilisés.	Démultiplexer les échantillons sur les données brutes Plusieurs sets d'indexes commerciaux sont déjà pré-enregistrés dans notre LIMS
Information sur l'utilisation de spikes lors de la génération des librairies (par exemple chromatine de drosophile dans les librairies ChIPseq)	Nature et proportion	Interpréter les résultats d'analyse des contaminants dans le FastQscreen
Si disponible, profils Bioanalyzer (Agilent) ou Fragment Analyzer (AATI)	Taille de l'insert (= taille de la librairie - taille des adaptateurs) Présence de dimères d'adaptateurs ou de reliquats de primers Quantité totale des librairies	Déterminer la proportion de fragments « séquençables » (fragments <600 pb sur le HiSeq 4000) Si un séquençage en paired-end (PE) est demandé, vérifier la taille minimale des librairies pour des séquences non chevauchantes (insert >320 bp pour un séquençage 2x100 PE et insert >220 bp pour un séquençage 2x50 PE) Optimiser la génération des clusters sur la flow cell et limiter le nombre de séquences non informatives Quantité optimale : 20µl à 5nM

GenomEast Platform
CERBM GIE – IGBMC
1 rue Laurent Fries
67404 Illkirch Cedex
France

+33 3 88 65 34 26
thibault@igbmc.fr
jost@igbmc.fr
genomeast.igbmc.fr

Documents à fournir	Informations recherchées	Finalité
		Quantité minimale : 10µl à 3nM (suffisant pour 2 séquençages)

3 Contrôles qualité réalisés par la plateforme

Les librairies sont contrôlées selon des critères qualité inhérents à la technologie de séquençage (voir tableau ci-dessous). Les résultats sont envoyés par e-mail au porteur de projet et/ou accessibles sur le LIMS de la plateforme (http://ngs-lims.igbmc.fr).

Vérification des librairies	
Profil des librairies (électrophorèse capillaire)	Taille moyenne comprise entre 200 et 600 pb.
Pureté des librairies (électrophorèse capillaire)	Présence minoritaire de reliquats de primers et de dimères d'adaptateur (bande à 120-130 pb), si applicable.
Quantité totale des librairies	>10µl à 3nM

Après validation des librairies, la plateforme s'engage à utiliser la technologie de séquençage Illumina selon les recommandations du fournisseur. N'ayant aucun contrôle sur la génération des librairies, la plateforme déclinera toute responsabilité sur la qualité des résultats de séquençage finaux. Nous vérifions cependant la qualité des données générées selon nos critères habituels suivants.

Vérification des données de séquençage	
Nombre attendu de séquences par piste	>250 millions en single-end >500 millions en paired-end
Scores de qualité attendus (Phred score)>30	≥ 85% des bases pour des lectures de 50 bases ≥ 75% des bases pour des lectures de 100 bases

4 Livraison des résultats

Pour chaque échantillon, les résultats suivants sont mis à la disposition du porteur du projet :

- Les données brutes de séquençage (séquences nucléotidiques au format FASTQ ayant passé le filtre qualité et ne correspondant pas à des dimères d'adaptateur).
- Un rapport présentant les contrôles qualité de séquençage (fichier PDF).

Deux fichiers additionnels sont fournis par projet :

- Un rapport (fichier PDF) présentant le nombre de lectures brutes, le pourcentage de bases ayant un score de qualité Phred supérieur à 30 et la taille de tous les fichiers bruts (FASTQ) à télécharger.
- Un fichier texte fournissant les chaines de caractères MD5 associées à chaque fichier FASTQ à télécharger. Le porteur de projet peut utiliser ces informations pour vérifier l'intégrité des fichiers après leur téléchargement.

A la place ou en plus de l'ensemble des fichiers précédemment cités, le porteur de projet peut demander à la plateforme de lui fournir des fichiers au format BCL (binary base call), non démultiplexés. Dans ce cas, les fichiers suivants sont mis à disposition du porteur du projet :

- Une archive au format tar.gz contenant les fichiers BCL, xml et RTA*.txt issus du dossier de run de séquençage, en respectant les noms et l'arborescence des répertoires de ce dernier, pour les lignes sur lesquelles des échantillons du projet ont été séquencés.

GenomEast Platform
CERBM GIE – IGBMC
1 rue Laurent Fries
67404 Illkirch Cedex
France

+33 3 88 65 34 26
thibault@igbmc.fr
jost@igbmc.fr
genomeast.igbmc.fr

- Un fichier SampleSheet.csv (fichier csv contenant des informations relatives aux échantillons séquencés, nécessaire pour démultiplexer les fichiers BCL à l'aide de Cell Ranger mkfastq ou Illumina bcl2fastq).
- Un fichier texte fournissant les chaines de caractères MD5 associées à chaque fichier à télécharger. Le porteur de projet peut utiliser ces informations pour vérifier l'intégrité des fichiers après leur téléchargement.

Un e-mail de livraison des données informe le porteur de projet qu'il peut télécharger ses données de séquence en utilisant un login et un mot de passe sur le serveur FTP de la plateforme.

Conformément aux « Conditions Générales de la Plateforme GenomEast », il est rappelé que le porteur du projet est responsable de la sauvegarde et de l'archivage de ses données. La plateforme ne s'engage à les conserver que pour une durée limitée de six mois après leur mise à disposition.

5 Analyse ultérieure (optionnelle)

L'analyse ultérieure des données n'est pas prise en charge dans la prestation standard, mais elle peut être réalisée sous la forme d'une collaboration avec des membres de la plateforme. Le type d'analyses réalisées dépend de la nature des librairies séquencées. Nous recommandons aux porteurs de projet qui souhaiteraient initier une collaboration avec la plateforme de nous contacter avant de commencer leur projet pour que nous puissions les aider à définir les analyses qui pourraient répondre au mieux aux questions biologiques posées.

6 Références

(1) Effects of Index Misassignment on Multiplexing and Downstream Analysis. White paper Illumina. Pub. No. 770-2017-004-D.
(2) Low-Plex Pooling Guidelines for Enrichment Protocols. Technical note Illumina. Pub. No. 770-2013-060, 23 September 2015.

* Modèle de fichier Excel pour la description des librairies

Sample name	Concentration (ng/µl)	Quantification method	Volume (µl)	Experimental condition	Remarks		Index 1 (i7) Supplier	Index 1 (i7) Code	Sequence Index 1 (i7)	Index 2 (i5) Supplier 2	Index 2 (i5) Code 2	Sequence Index 2 (i5)
WT1	*10*	*Qubit*	*10*	*Controle*			*Illumina:Nextera*	*N704*	*TCCTGAGC*	*Illumina:Nextera*	*S502*	*CTCTCTAT*