NORME INTERNATIONALE

ISO 20397-2

Première édition
2021-02

Biotechnologie — Séquençage massivement parallèle —

Partie 2: Évaluation de la qualité des données de séquençage

Biotechnology — Massively parallel sequencing — Part 2: Quality evaluation of sequencing data

Numéro de référence
ISO 20397-2:2021(F)

ISO copyright office
Case postale 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Genève
Tél.: +41 22 749 01 11
E-mail: copyright@iso.org
Web: www.iso.org

Publié en Suisse

Sommaire

Page

Avant-propos

L'ISO (Organisation internationale de normalisation) est une fédération mondiale d'organismes nationaux de normalisation (comités membres de l'ISO). L'élaboration des Normes internationales est en général confiée aux comités techniques de l'ISO. Chaque comité membre intéressé par une étude a le droit de faire partie du comité technique créé à cet effet. Les organisations internationales, gouvernementales et non gouvernementales, en liaison avec l'ISO participent également aux travaux. L'ISO collabore étroitement avec la Commission électrotechnique internationale (IEC) en ce qui concerne la normalisation électrotechnique.

Les procédures utilisées pour élaborer le présent document et celles destinées à sa mise à jour sont décrites dans les Directives ISO/IEC, Partie 1. Il convient, en particulier, de prendre note des différents critères d'approbation requis pour les différents types de documents ISO. Le présent document a été rédigé conformément aux règles de rédaction données dans les Directives ISO/IEC, Partie 2 (voir www.iso.org/directives).

L'attention est attirée sur le fait que certains des éléments du présent document peuvent faire l'objet de droits de propriété intellectuelle ou de droits analogues. L'ISO ne saurait être tenue pour responsable de ne pas avoir identifié de tels droits de propriété et averti de leur existence. Les détails concernant les références aux droits de propriété intellectuelle ou autres droits analogues identifiés lors de l'élaboration du document sont indiqués dans l'Introduction et/ou dans la liste des déclarations de brevets reçues par l'ISO (voir www.iso.org/brevets).

Les appellations commerciales éventuellement mentionnées dans le présent document sont données pour information, par souci de commodité, à l'intention des utilisateurs et ne sauraient constituer un engagement.

Pour une explication de la nature volontaire des normes, la signification des termes et expressions spécifiques de l'ISO liés à l'évaluation de la conformité, ou pour toute information au sujet de l'adhésion de l'ISO aux principes de l'Organisation mondiale du commerce (OMC) concernant les obstacles techniques au commerce (OTC), voir www.iso.org/avant-propos.

Le présent document a été élaboré par le Comité technique ISO/TC 276, *Biotechnologie.*

Une liste de toutes les parties de la série ISO 20397 se trouve sur le site web de l'ISO.

Il convient que l'utilisateur adresse tout retour d'information ou toute question concernant le présent document à l'organisme national de normalisation de son pays. Une liste exhaustive desdits organismes se trouve à l'adresse www.iso.org/members.html.

Introduction

Le séquençage massivement parallèle (SMP) est une approche analytique de séquençage de l'acide nucléique à haut débit qui utilise un traitement massivement parallèle pour étudier des génomes entiers, des transcriptomes et des séquences ciblées d'acides nucléiques de différentes origines, en un laps de temps relativement court.

Le SMP est utilisé dans de nombreux domaines des sciences de la vie. Il permet une détermination et une analyse à haut débit de milliards de nucléotides. Du fait de la variabilité biologique des polymères d'acide désoxyribonucléique et d'acide ribonucléique à travers le vivant, la détermination précise de leurs séquences constitue un véritable défi. La qualité des séquences générées par SMP dépend de nombreux facteurs, notamment, entre autres, la qualité de l'échantillon, la préparation de la banque, le choix de la plateforme de lecture et la qualité des données de séquençage.

L'analyse des données de séquençage peut représenter de véritables défis bio-informatiques liés au stockage des données, au temps de calcul et à la précision de détection des variants. L'une des principales difficultés associées aux données de séquençage, trop souvent négligée, porte sur les mesures de contrôle qualité à tous les stades du pipeline de traitement des données, alors mêmes qu'ils sont essentiels à toute l'analyse en aval des données de séquences. Le contrôle qualité applicable au traitement et à l'analyse des données de séquençage de l'acide nucléique concerne trois niveaux distincts: données brutes, alignement et détection des variants. Le présent document fournit une liste d'éléments à prendre en compte lors de l'évaluation de la qualité des données de séquençage massivement parallèle, ainsi que les recommandations spécifiques à différentes plateformes SMP.

Biotechnologie — Séquençage massivement parallèle —

Partie 2: Évaluation de la qualité des données de séquençage

1 Domaine d'application

Le présent document spécifie les exigences générales et les recommandations applicables à l'évaluation et au contrôle de la qualité des données de séquençage massivement parallèle (SMP). Il traite des modes opératoires faisant suite à la production des données brutes, en incluant la génération des alignements de séquences et la détection des variants.

Le présent document fournit également des lignes directrices générales applicables à la validation et à la documentation des données SMP.

Le présent document ne s'applique pas aux processus relatifs à l'assemblage de novo.

2 Références normatives

Le présent document ne contient aucune référence normative.

3 Termes et définitions

Pour les besoins du présent document, les termes et définitions suivants s'appliquent.

L'ISO et l'IEC tiennent à jour des bases de données terminologiques destinées à être utilisées en normalisation, consultables aux adresses suivantes:

— ISO Online browsing platform: disponible à l'adresse https://www.iso.org/obp

— IEC Electropedia: disponible à l'adresse http://www.electropedia.org/

3.1
séquence d'adaptateur
adaptateur
oligonucléotide artificiel d'une séquence connue qui peut être ajouté aux extrémités 3' ou 5' d'un fragment d'acide nucléique

Note 1 à l'article: Il fournit le site d'amorçage ainsi que les autres séquences nécessaires au séquençage de l'insert.

3.2
algorithme
séquence finie d'instructions complètement déterminée par laquelle les valeurs des variables de sortie peuvent être calculées à partir des valeurs des variables d'entrée

[SOURCE: IEC 60050-351:2013, 351-42-27, modifiée — Les notes ont été supprimées.]

3.3
détection de bases
lors du séquençage massivement parallèle, processus de calcul consistant à traduire les signaux électriques bruts en séquence nucléotidique

Note 1 à l'article: La performance de l'application de détection de bases et de l'algorithme est caractéristiquement définie par une précision de lecture et un consensus.

3.4
pipeline bio-informatique
suite de programmes individuels, scripts ou briques logicielles liés entre eux, dans lesquels les données brutes ou les résultats d'un programme sont utilisés comme données d'entrée dans l'étape suivante du traitement des données

EXEMPLE Les résultats d'un programme de découpage (trimming) peuvent être utilisés comme données d'entrée pour un assembleur *de novo*.

3.5
efficacité de capture
pourcentage de l'ensemble des lectures séquencées ou cartographiées qui chevauchent les régions ciblées

3.6
couverture
profondeur de couverture
nombre de fois qu'une position de base donnée est lue dans un cycle de séquençage

Note 1 à l'article: Nombre de lectures qui couvrent une position particulière.

3.7
largeur de couverture
fraction du génome cible qui est identifiée lors des cycles de séquençage

3.8
densité de clusters
nombre de clusters pour chaque flow cell de séquençage

Note 1 à l'article: La densité de clusters s'applique aux plateformes *SMP* (3.30) nécessitant une étape d'amplification.

Note 2 à l'article: La densité de clusters de séquence individuels, provenant chacun d'une seule molécule sur certaines plateformes de séquençage.

Note 3 à l'article: La densité de clusters est généralement exprimée en milliers par mm^2.

3.9
CCS
séquençage consensus sur séquences circulaires
mode de séquençage où la taille de l'insert est séquencée plusieurs fois lors d'une réaction de type amplification par cercle roulant, ce qui permet d'obtenir une haute précision.

Note 1 à l'article: Dans ce mode, plusieurs lectures de la même molécule peuvent être utilisées pour atteindre une précision moléculaire individuelle supérieure.

3.10
étendue de couverture
étendue de la profondeur de couverture d'un génome à l'issue des cycles de séquençage

3.11
CNV
variation du nombre de copies
variation du nombre de copies sur un segment d'ADN génomique d'un organisme

Note 1 à l'article: Les CNV sont des insertions, délétions, inversions et duplications contenant au moins 1 000 bases en longueur.

3.12
ADN
acide désoxyribonucléique
polymère de désoxyribonucléotides se présentant sous la forme de double brin (ADNdb) ou de brin simple (ADNsb)

[SOURCE: ISO 22174:2005, 3.1.2]

3.13
délétion
perte d'une (ou de plusieurs) paire(s) de bases nucléotidiques d'une séquence d'acide nucléique par rapport à sa séquence de référence

3.14
niveau de duplication
nombre de répétitions identiques pour chaque séquence d'une banque

Note 1 à l'article: Le niveau de duplication apparaît généralement sous la forme d'un graphique représentant le nombre relatif de séquences à différents degrés de duplication.

3.15
taux de GC
pourcentage de guanine et cytosine dans une ou plusieurs séquence(s) d'acide nucléique

Note 1 à l'article: La quantité de guanine et de cytosine dans un acide nucléique est généralement exprimée en fraction molaire (ou pourcentage) de bases azotées totales. Les bases azotées totales comprennent le nombre total de bases nucléotidiques lues après un ou plusieurs cycle(s) de SMP.

3.16
gène
séquence de nucléotides dans l'ADN ou l'ARN codant soit pour un ARN soit pour un produit protéique

Note 1 à l'article: Les gènes sont reconnus comme étant l'unité de base de l'hérédité.

Note 2 à l'article: Un gène peut comprendre des segments d'acide nucléique non contigus qui sont remaniés à l'occasion d'un processus nucléaire.

Note 3 à l'article: Un gène peut comprendre ou faire partie d'un opéron qui inclut des éléments de l'expression génétique.

3.17
indel
insertion (3.18) ou/et *délétion* (3.13) de nucléotides dans l'ADN génomique

Note 1 à l'article: Les indels ont des longueurs inférieures à 1 000 bases.

3.18
insertion
ajout d'une (ou de plusieurs) bases nucléotidiques dans une séquence d'acide nucléique

[SOURCE: ISO/TS 20428: 2017, 3.19, modifiée — Le terme «ADN» a été remplacé par «acide nucléique».]

3.19
séquençage
détermination de l'ordre et de la concentration des bases nucléotidiques (adénine, guanine, cytosine, thymine et uracile) d'une molécule d'acide nucléique

Note 1 à l'article: Une séquence est généralement décrite de l'extrémité 5' à l'extrémité 3'.

[SOURCE: ISO/TS 17822-1:2020, 3.19, modifiée — «ADN» a été supprimé dans le terme; «ADN» a été remplacé par acide nucléique, et «uracile» a été ajouté dans la définition.]

3.20
alignement de séquences
agencement de plusieurs séquences d'acides nucléiques en fonction de leurs régions de similarité

Note 1 à l'article: L'alignement de séquences ne nécessite pas forcément un génome de référence/une région d'acide nucléique cible de référence et son objectif n'est pas nécessairement de produire un assemblage.

3.21
données brutes
données de séquençage primaires produites par un séquenceur sans avoir recours à un pré-filtrage informatique pour l'analyse

3.22
ARN
acide ribonucléique
polymère de ribonucléotides se présentant sous la forme de double brin ou de brin simple

Note 1 à l'article: La synthèse des protéines dans les cellules est régie par les informations génétiques contenues dans la séquence de nucléotides d'une classe d'ARN connue sous le nom d'ARN messager (ARNm).

3.23
ribonucléotide
nucléotide contenant de la ribose comme composant pentosique formant le bloc élémentaire de construction de l'ARN

Note 1 à l'article: Les ribonucléotides comprennent l'adénylate (AMP), le guanylate (GMP), le cytidylate (CMP) ou l'uridylate (UMP).

3.24
lecture
séquence
séquence nucléotidique générée par un séquenceur

Note 1 à l'article: Une lecture (read) est une séquence déduite de paires de bases d'acide nucléique (ou de probabilités de paires de bases) correspondant à tout (ou partie) d'un fragment d'acide nucléique. La lecture peut être utilisée pour désigner les séquences obtenues par SMP.

3.25
type de lecture
catégorie de séquence qui dépend de la façon dont l'expérience de lecture des séquences est conçue et réalisée

EXEMPLE Le type de lecture peut être la lecture sur une seule extrémité (single-end), la lecture d'extrémités appariées (paired-end), la lecture d'extrémités appariées de plus grandes longueurs (mate-paired end), la lecture longue continue (long read) ou la lecture consensus sur séquences circulaires.

3.26
séquence de référence
séquence d'acide nucléique servant soit à aligner les lectures de séquençage, soit de référence pour des annotations telles que les gènes et les variations de séquence

3.27
démultiplexage
opération informatique inverse du processus de multiplexage, au cours de laquelle deux ou plusieurs échantillons sont mélangés afin de pouvoir les séquencer en une seule fois sur un instrument SMP

Note 1 à l'article: Les échantillons qui doivent être combinés doivent être munis d'un code-barres/indexés avant d'être mélangés.

Note 2 à l'article: Le démultiplexage est un algorithme de calcul qui sépare un groupe de lectures en fonction de leur échantillon d'origine d'après le code-barres.

3.28
cartographie
assemblage de séquences d'acides nucléiques en fonction d'une séquence de référence existante, servant à construire une séquence consensus

3.29
lecture d'extrémités appariées de fragments de grandes longueurs
lecture des deux extrémités appariées situées aux deux extrémités d'un fragment de séquence d'acide nucléique long de plusieurs kilobases (plus de 2 kb ou au moins 2 kb)

3.30
SMP
séquençage massivement parallèle
technique de séquençage permettant la détermination simultanée de la séquence de multiples molécules d'acides nucléiques indépendantes basée sur le modèle incrémentiel

Note 1 à l'article: La technologie de séquençage massivement parallèle permet d'obtenir plusieurs millions ou milliards de lectures courtes par cycle.

3.31
lecture d'extrémités appariées
lecture par SMP des deux extrémités appariées situées aux deux extrémités d'un fragment d'ADN

Note 1 à l'article: Dans le séquençage d'extrémités appariées, l'instrument séquence les deux extrémités de courts fragments d'une longueur généralement comprise entre 200 pb et 800 pb.

3.32
score de qualité
score *Q*
score de qualité Phred
mesure de la qualité de séquençage d'une base nucléotidique donnée

Note 1 à l'article: Q est défini par la formule suivante:

$$Q=-10\log10(p)$$

où p est la probabilité estimée pour que la détection de bases soit erronée.

Note 2 à l'article: Un score de qualité de 20 représente un rapport d'erreur de 1 sur 100, avec une précision de détection correspondante de 99 %.

Note 3 à l'article: Des scores de qualité élevés indiquent une plus faible probabilité d'erreur de séquençage. Des scores de qualité faibles peuvent rendre inutilisables les lectures correspondantes. Des scores de qualité faibles peuvent également conduire à des appels de variants faux-positifs, aboutissant à des conclusions inexactes.

3.33
run
processus complet de réalisation de l'opération de séquençage, de la charge des échantillons jusqu'à l'obtention des données brutes

3.34
annotation de séquences
processus consistant à ajouter une note d'explication, un commentaire ou une référence sur les caractéristiques spécifiques présentes dans une séquence d'ADN, d'ARN ou de protéines, grâce à des informations descriptives sur la structure ou la fonction

Note 1 à l'article: Le processus d'annotation de séquences peut être considéré comme une assignation de données à la séquence.

3.35
lecture d'extrémité simple
lecture de séquence obtenue en lisant un seul des deux brins d'un fragment d'ADN à partir d'une des deux extrémités

3.36
SNV
variant nucléotidique unique
variation sur un seul nucléotide d'une molécule d'acide nucléique

3.37
SV
variation structurale
région d'ADN d'environ 1 000 bases ou de taille supérieure, pouvant inclure des inversions et des translocations équilibrées ou des déséquilibres génomiques

Note 1 à l'article: Il existe plusieurs types fréquents de variants structurels: variants du nombre de copies (délétions, insertions, amplifications, duplications), délétions neutres du nombre de copies (perte d'hétérozygosité), inversions, duplications segmentaires et translocations (équilibrées ou déséquilibrées).

3.38
sous-séquence
fraction d'une séquence présente entre des adaptateurs en épingle à cheveux

3.39
découpage (trimming) des lectures brutes
opération visant à supprimer les parties de faible qualité ou les séquences contaminantes tout en préservant la partie de haute qualité d'une lecture SMP la plus longue possible

3.40
variation
différences d'une ou de plusieurs bases d'acides nucléiques dans une séquence par rapport à la base/aux bases prévue(s)

3.41
détection de variants
processus d'identification précise des variations des données de séquence par rapport à une séquence de référence

3.42
ZMW
guide d'onde mode zéro
guide d'onde optique qui guide l'énergie lumineuse dans un volume de petite dimension relativement à la longueur d'onde de la lumière

Note 1 à l'article: Une polymérase est ancrée au fond de ce ZMW et l'incorporation de nucléotides est mesurée par une hausse de fluorescence pendant la fixation puis par une réduction ultérieure après incorporation.

4 Données brutes

4.1 Généralités

Il convient d'attribuer à chaque nucléotide d'une séquence une valeur numérique (score de qualité de base) correspondant à la précision présumée du processus de détection de bases, le cas échéant.