Avertissement, veuillez lire attentivement

Les opinions exprimées sont de bonne foi et, malgré le soin apporté à la préparation du présent document, la responsabilité de l'auteur ne peut être engagée en cas d'éventuelle utilisation abusive ou d'interprétation erronée des données et des opinions qu'il contient.

L'auteur déclare n'avoir aucun intérêt financier en concurrence et autofinance sa contribution à cette recherche.

L'auteur veut garder l'anonymat. Pour toute question ou information, telles que les fichiers de séquences ADN non identifiées, Alain Bonnet recueillera les coordonnées, les CV, les qualifications et une brève lettre de motivation (idéalement en anglais) des personnes intéressées. https://www.the-alien-project.com/contact/

De : « anonyme », PhD
À l'attention de : Thierry Jamin, Alain Bonnet

Objet : Présentation et discussion des analyses génomiques détaillées effectuées sur deux individus d'espèce non identifiée trouvées dans le désert de Nazca, au Pérou, en 2015

1. Introduction : caractéristiques phénotypiques des deux individus.

1.1 Maria :

L'individu nommé Maria pour plus de simplicité a été trouvée à Nazca, au Pérou, en 2015. Maria est une créature humanoïde de sexe non identifié, mesurant 165 à 170 cm, en position fœtale. Malgré des similitudes évidentes avec homo sapiens, l'individu présente des traits phénotypiques atypiques, parmi lesquels : un volume du crâne d'environ 25% supérieur à celui de l'homo sapiens (au niveau pariétal), trois doigts sur chaque main et chaque pied, de très longues phalanges. Deux laboratoires ont indépendamment estimé l'âge de l'individu par radiocarbone à la même datation de 1750 ± 30 BP.

1.2 Grande Main :

Sur le même site ont été trouvées plusieurs mains avec des caractéristiques similaires, trois doigts, six phalanges et de taille considérable. L'âge de l'une d'entre elles - qui sera nommée Grande Main pour plus de simplicité dans le présent rapport, a été estimé à 6420 ± 30 BP.

1.3 Questions posées - authenticité et nature des créatures.

Ces êtres sont-ils de véritables créatures biologiques ou seulement un arrangement à partir d'autres espèces déjà connues, animales et / ou humaines ? Telle est la question ayant été soulevée par les personnes ayant approché le matériel de près, et certainement par le public.

Avant de discuter de cette question, il convient de garder à l'esprit deux points :

(i) Le premier point est qu'en science, lors de la formulation d'une hypothèse, la première étape consiste à examiner ses présupposés et implications directes et à vérifier si elles sont cohérentes avec les données disponibles. Si cette étape préalable est réussie, une seconde étape consiste à tester les prédictions d'une telle hypothèse.

(ii) Ces corps ont respectivement 1700 et 6500 ans. Les tissus sont secs, durs et ont tendance à s'effriter. Par conséquent, d'éventuelles interventions chirurgicales ne peuvent pas avoir été effectuées récemment, mais plutôt au décès des sujets, à savoir il y a plus de 1700 ans pour Maria et plus de 6000 ans pour la Grande Main.

Ainsi, dans quelle mesure l'hypothèse d'un "arrangement à partir d'autres espèces déjà connues, animales et/ou humaines" est-elle vraisemblable ? Cette hypothèse implique l'existence de traces (lésions, cicatrices) qui devraient être visibles, car il n'y a pas de processus de cicatrisation après la mort. Elle présuppose également un certain niveau de technologie et de connaissances nécessaires pour produire de tels individus.

Les observations anatomiques minutieuses, dont des CT scans, ont révélé des détails particulièrement réalistes et précis (empreintes digitales, dents adultes, surfaces externe et interne du crâne dont les sutures, peau, vertèbres, côtes, articulations, organes internes). De plus, aucune lésion des os ou des tissus cutanés suggérant une intervention chirurgicale n'a pu être détectée.

En tant que telle, l'hypothèse susmentionnée semble donc peu probable en raison de :

(i) l'absence de donnée suggérant une intervention chirurgicale ou une manipulation similaire

(ii) les détails anatomiques qui nécessiteraient, pour être imités, le déploiement de moyens biotechnologiques a priori non disponibles à cette époque et même de nos jours

(iii) la présence d'un autre individu, un nourrisson, ayant été trouvé sur le même site et présentant les mêmes caractéristiques atypiques que Maria. Des détails tels que les dents de lait et les proportions corps / tête confirment qu'il s'agit d'un bébé authentique et non d'un adulte de petite taille.

Une objection récurrente a été que les lésions ou les cicatrices pourraient être si subtiles que nous aurions pu les manquer. Effectivement, la résolution des scanners utilisés permettait de voir de nombreux détails précis, mais n'était pas la plus élevée disponible sur le marché. Cependant, cette objection présuppose toujours un niveau de technologie et de science incompatible avec le calendrier évoqué par les analyses au radiocarbone, et je voudrais attirer votre attention sur les implications de cette hypothèse, et ou cela nous mène. Quels appareils ou matériel chirurgical de ces époques anciennes auraient pu opérer de manière si raffinée que nos scanners modernes passeraient à côté des traces laissées sur les corps et les tissus ? Pouvons-nous raisonnablement supposer l'existence de laboratoires de biotechnologie de pointe ayant existé dans le désert de Nazca, entre 1700 et 6500 ans avant les temps présents ? Cette piste est sinon absurde, du moins incompatible avec les données disponibles.

Nous allons plutôt considérer les prédictions de cette hypothèse. En particulier, si ces créatures ont été construites avec des restes humains et / ou animaux, les analyses d'ADN devraient mettre en évidence, après retrait de l'ADN des contaminants, (i) soit des séquences ADN 100% homo sapiens moderne, (ii) soit des séquences ADN homo sapiens moderne et des séquences ADN animal en proportions significatives, probablement issues d'espèces animales locales.

Ce sont les analyses que nous avons effectuées et que nous allons présenter aujourd'hui.

1.3.2 Plan de la présentation

Après cette introduction longue mais nécessaire, nous procéderons comme suit.

1) Matériel et méthodes : brève présentation des sources des données sur lesquelles nous avons travaillé (ADN ancien, séquençage de l'ADN) et des méthodes (analyses fondées sur l'alignement de séquences).
2) Résultats du premier cycle d'analyse : identification des ADN contaminants, analyse de l'alignement avec l'Homo sapiens moderne. Isolement des séquences (indéterminées)
3) Résultats du deuxième cycle d'analyse, réalisée sur les séquences non classifiées : comparaison avec d'autres espèces
4) Conclusions et interprétations
5) Annexe

1. Sources de données et méthodes

L'extraction et le séquençage ont été effectués par un autre laboratoire basé au Mexique, BioTechMol http://biotecmol.mx/. Les analyses génomiques effectuées ne mentionnaient pas les méthodes utilisées et n'étaient pas exhaustives. Nous avons donc effectué les analyses des fichiers de données brutes afin de vérifier d'abord la qualité des données, et d'identifier l'ensemble de l'ADN séquencé - contaminants compris - sur des séquences de la plus haute qualité.

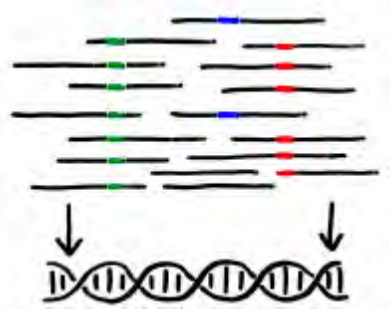
Extraction :

L'ADN est fragile et les anciens échantillons d'ADN sont généralement très endommagés et contaminés par des virus, des bactéries, des micro-organismes et l'ADN des personnes / animaux en contact direct avec l'échantillon. Pour les anciens échantillons, des procédures spécifiques doivent être suivies. Un échantillon d'os de 0,54 g prélevé sur le corps de Maria et un os / tissu inconnu de 2,38 g prélevé sur la grande main ont permis d'extraire suffisamment d'ADN (procédure de Shapiro B, Heslington M. 2012 +, kit de réparation également utilisé, Kit PreCR® Mélange de réparation de New England Biolabs M0309S).

Séquençage et analyse :

Quelques mots peut-être pour un public non spécialiste. L'ADN est constitué principalement de quatre nucléotides, ou bases. Combinées ensemble, ces nucléotides forment (4 * 4 * 4) 64 triplets nommés « codons » qui constituent un alphabet. Nous sommes loin de maîtriser la syntaxe de l'ADN, mais certaines règles sont connues, par exemple, certains codons indiquent le début d'une séquence codante, tandis que d'autres indiquent la fin.
Séquencer l'ADN signifie déterminer l'ordre des nucléotides composant l'ADN. C'est une technique moléculaire maintenant réalisée par des machines et des ordinateurs, et seulement supervisée par l'humain. Une difficulté du séquençage est que la molécule d'ADN est très longue (3 milliards de nucléotides pour l'ADN humain), de sorte que le séquençage ne peut pas être effectué en une seule fois. Le séquençage de l'ADN est plutôt effectué sur de « petits morceaux » d'ADN » ou « read » d'une longueur comprise entre 50 et 150 nucléotides, ou bases. Ensuite, le génome est reconstruit en utilisant une procédure appelée *assemblage du génome.*
Pour faciliter la compréhension, on peut imaginer un texte qui serait coupé de manière aléatoire tous les 5 à 10 mots. Ensuite, les morceaux seraient copiés et enfin, le texte devrait être reconstruit de manière cohérente et sémantiquement correcte, tout en repérant les éventuelles erreurs de copie.

Pour permettre la reconstruction, les « petits morceaux » sont de longueurs différentes, séquencés et superposés plusieurs fois. Les superpositions et les règles syntaxiques permettent de reconstruire le génome une fois séquencé. Les superpositions sont également utilisées pour identifier les éventuelles erreurs de séquençage (par exemple, pour simplifier, si une région est séquencée 10 fois et que seulement 5 lectures ou « read » sont identiques, on peut considérer qu'il existe un taux d'erreur élevé dans ces régions et éventuellement l'éliminer de la prochaine analyse).



Le séquençage a été effectué sur la plateforme Myseq Illumina (les plateformes Illumina sont parmi les plateformes les plus utilisées)

Analyses :

Le but de ces analyses était de définir quantitativement et qualitativement l'ADN de ces échantillons, y compris les contaminants.

Notez que cette procédure nécessite plusieurs étapes après le séquençage, en particulier l'assemblage du génome comme mentionné précédemment et l'alignement. L'alignement est utilisé pour la comparaison du génome lorsqu'il est nécessaire d'identifier une espèce, par exemple. Il consiste à positionner l'ADN séquencé sur l'ADN de tous les génomes disponibles pour une espèce donnée.

Les données résultantes (fichier Illumina fastq) ont donc été contrôlées et analysées (voir Annexe pour plus de détails) afin de déterminer qualitativement et quantitativement l'ADN, en référence à la base de données RefSeq Complete Genomes https:https://www.ncbi.nlm.nih.gov/refseq/.

2.Identification des contaminants / virus, comparaison avec Homo sapiens moderne, Isolement de séquences non classifiées (indéterminées)

2.1 Maria

Les résultats sont affichés en termes de pourcentage de cartographie absolue sur les régions du génome humain et bactérien / contaminant, le reste étant considéré comme non mappé / non classifié.

33,7% des lectures furent alignés avec homo sapiens moderne
18,4% des lectures furent alignés avec des contaminants
47,9% des lectures ne furent alignés avec rien, autrement dit non classifiés.

MARIA - Sommaire Général

Nombre total de Lectures (READS) (329037x2)	658074	
Nombre de bases - Lectures1	46674295	4.6Mbp
Nombre de bases - Lectures2	46928318	4.6Mbp
Nombre total de bases	93602613	9.36Mbp

Génome Humain	
Lectures alignées sur le génome humain GrCh38	221623
Nombre de bases alignées (alignement unique)	5782559
Pourcentage de lectures cartographiées sur le génome humain	33.6775

Génomes bactériens et autres	
Lectures alignées sur génomes bactériens	121186
Nombre de bases alignées (alignement unique)	17183634
Pourcentage de lectures cartographiées sur génomes bactériens et autres	18.415

Inconnus	
Nombre de lectures non alignées/non cartographiées/non classifiées	315265
Nombre total de bases non classifiées	44756142
Pourcentage du total des lectures non cartographiées/non classifiées	47.907

Maria - Comparaison avec Homo sapiens

CATEGORY	FIRST_OF_PAIR	SECOND_OF_PAIR	PAIR
TOTAL_READS	329037	329037	658074
FilterPassed_READS	329037	329037	658074
Percent_FilterPassed_READS	1	1	1
FilterPassed_NOISE_READS	0	0	0
FilterPassed_READS_ALIGNED	110328	111295	221623
Percent_FilterPassed_READS_ALIGNED	0.335306	0.338245	0.336775
FilterPassed_ALIGNED_BASES	2873043	2909516	5782559
FilterPassed_HQ_ALIGNED_READS	64305	64227	128532
FilterPassed_HQ_ALIGNED_BASES	1894551	1874692	3769243
FilterPassed_HQ_ALIGNED_Q20_BASES	1835735	1781625	3617360
FilterPassed_HQ_MEDIAN_MISMATCHES	0	0	0
FilterPassed_MISMATCH_RATE	0.002944	0.003927	0.003439
FilterPassed_HQ_ERROR_RATE	0.002706	0.003187	0.002945
FilterPassed_INDEL_RATE	0.00017	0.000299	0.000235
MEAN_READ_LENGTH	140.851205	141.623225	141.237215
READS_ALIGNED_IN_PAIRS	109707	109707	219414
Percent_READS_ALIGNED_IN_PAIRS	0.994371	0.985732	0.990033
BAD_CYCLES	0	0	0
STRAND_BALANCE	0.502737	0.498495	0.500607
Percent_CHIMERAS	0.003847	0.003948	0.003897
Percent_ADAPTER	0.061972	0.000243	0.031107

Maria-Cartographie avec des bactéries et d'autres génomes contaminants

Nombre total de lectures cartographiées	121186
Porcentage de lectures alignées	18.415

Bactéries	N^{bre} de lectures
Alteromonas_macleodii_str._'Ionian_Sea_U8'	21355
Caulobacter_sp._K31	8024
Phenylobacterium_zucineum_HLK1	5099
Delftia_acidovorans_SPH-1	4913
Delftia_sp._Cs1-4	4869
Caulobacter_segnis_ATCC_21756	4504
Caulobacter_crescentus_CB15	3769
Delftia	3762
Bradyrhizobium_sp._BTAi1	2997
Propionibacterium_acnes	2095
Caulobacter	1871
Brevundimonas_subvibrioides_ATCC_15264	1690
Ralstonia_pickettii	1503
Rhodopseudomonas_palustris_CGA009	1279
Ralstonia_pickettii_12J	1188
Alphaproteobacteria	1092
Proteobacteria	1026
Ralstonia_pickettii_12D	1012
Escherichia_coli	997
Enterobacteriaceae	922
Ralstonia_solanacearum	903
Propionibacterium_acnes_ATCC_11828	825
Bacteria_2	790
Caulobacteraceae	643
Autres	34674
Virus et autres	9384

2.2 Grande Main

Comme précédemment, les résultats sont affichés en termes de pourcentage d'alignement parfait avec les génomes Humain et Bactérien / Contaminant, le reste étant considéré comme non classifié.

0,37% des lectures furent alignés avec homo sapiens moderne
26,7% des lectures furent alignés avec des contaminants
72,9% ne furent pas classifiés.

Grande Main - Sommaire général

Nombre total de Lectures (READS) (341311x2)	682,622	
Nombre de bases - Lectures1	51,160,199	5.1Mbp
Nombre de bases - Lectures2	51,205,429	5.1Mbp
Nombre total de bases	102,365,628	10.2Mbp

Génome Humain	
Lectures alignées sur le génome humain GrCh38	2,518
Nombre de bases alignées (alignement unique)	366,819
Pourcentage de lectures cartographiées sur le génome humain	0.3689

Génomes bactériens et autres	
Lectures alignées sur les génomes bactériens, viraux et autres	182,243
Nombre de bases alignées (alignement unique)	27,331,623
Pourcentage de lectures cartographiées sur génomes bactériens et autres	26.7

Inconnus	
Nombre de lectures non alignées/non cartographiées/non classifiées	497,836
Nombre total de bases non classifiées	74,655,252
Pourcentage du total des lectures non cartographiées/non classifiées	72.93

Grande Main – Comparaison avec Homo Sapiens

CATEGORY	FIRST_OF_PAIR	SECOND_OF_PAIR	PAIR
TOTAL_READS	341311	341311	682622
FilterPassed_READS	341311	341311	682622
Percent_FilterPassed_READS	1	1	1
FilterPassed_NOISE_READS	0	0	0
FilterPassed_READS_ALIGNED	1274	1244	2518
FilterPassed_ALIGNED_BASES	186810	180009	366819
Percent_FilterPassed_READS_ALIGNED	0.003733	0.003645	0.003689
FilterPassed_HQ_ALIGNED_READS	1184	1152	2336
FilterPassed_HQ_ALIGNED_BASES	174577	167615	342192
FilterPassed_HQ_ALIGNED_Q20_BASES	170494	159118	329612
FilterPassed_HQ_MEDIAN_MISMATCHES	0	0	0
FilterPassed_MISMATCH_RATE	0.003014	0.004873	0.003927
FilterPassed_HQ_ERROR_RATE	0.002326	0.004135	0.003212
FilterPassed_INDEL_RATE	0.000252	0.000361	0.000305
MEAN_READ_LENGTH	148.893203	149.025721	148.959462
READS_ALIGNED_IN_PAIRS	1214	1214	2428
Percent_READS_ALIGNED_IN_PAIRS	0.952904	0.975884	0.964257
BAD_CYCLES	0	0	0
STRAND_BALANCE	0.515699	0.47508	0.495631
Percent_CHIMERAS	0.007143	0.007347	0.007243
Percent_ADAPTER	0.102754	0.000319	0.051537

Grande Main - Cartographie avec les bactéries et autres génomes contaminants

Bactéries	N^{bre} de lectures
Acinetobacter baumannii	6993
Ralstonia sp. MD27	831
Franconibacter helveticus	784
Pseudomonas sp. UBA6753	438
Acinetobacter pittii	413
Acinetobacter sp. 1542444	389
Acinetobacter sp. UNC434CL69Tsu2S25	382
Acinetobacter sp. 826659	370
Acinetobacter sp. 742879	331
Delftia sp. 67-8	272
Acinetobacter sp. LMB-5	267
Achromobacter denitrificans	213
Acinetobacter sp. UBA1297	201
Clostridium cochlearium	192
Clostridium novyi	184
Clostridium botulinum	175
Caulobacter mirabilis	173
Bradyrhizobium sp. BTAi1	171
Acinetobacter sp. UBA4567	170
Acinetobacter nosocomialis	162
Caulobacter henricii	160
Acinetobacter lactucae	127
Acinetobacter sp. UBA3098	119
Acinetobacter sp. WC-141	111
Pseudomonas sp. Irchel 3E13	100
Autres Bactéries	167154
Virus et Plasmides	1361

3. Séquences non classifiées : comparaison avec d'autres espèces

Nous avons effectué un second tour afin d'identifier les séquences jusqu'alors non classifiées. Plusieurs espèces furent considérées pour la comparaison avec les génomes de Maria et de la Grande Main, notamment : alpaga, babouin, chien, chat, cheval, chimpanzé, macaque rhésus. Les résultats ont été négatifs pour les deux sujets.

Organismes	Total des lectures	Lectures alignées	% Alignements
Maria non cartographié			
Alpaga	315265	1	0.00032
Babouin	315265	0	0.00000
Chien	315265	366	0.11609
Chat	315265	31	0.00983
Cheval	315265	0	0.00000
Chimpanzé	315265	1	0.00032
Macaque Rhésus	315265	1	0.00032
Grande Main non cartographié			
Alpaga	497861	11	0.00221
Babouin	497861	0	0.00000
Chien	497861	240	0.04821
Chat	497861	38	0.00763
Cheval	497861	2	0.00040
Chimpanzé	497861	25	0.00502
Macaque Rhésus	497861	3	0.00060

NB : Notez que le nombre de read alignés avec le genome du chien n'est pas significatif pour la comparaison en vue d'identifier l'espèce, mais est significativement supérieur à celui des autres espèces de référence. Nous avons supposé que les huaqueros qui ont trouvé les corps avaient des chiens.

Des analyses d'alignement sont encore en cours afin d'identifier la nature de ces séquences non identifiées - près de la moitié de l'échantillon pour Maria, écrasante majorité pour la Grande Main. Pour le moment, ces séquences sont considérées comme étant non identifiées.

4. Conclusions générales

Plusieurs questions ont été soulevées au cours de ces enquêtes. Les plus fréquentes étaient les suivantes : (a) l'authenticité des individus, (b) leurs éventuels liens ou similitudes avec l'homo sapiens, (c) leurs origines.

(a) l'authenticité des individus

Les points suivants :

(i) Les détails anatomiques subtils (jointures et articulations, tissus cutanés, surface interne du crâne, sutures crâniennes, gradient de densité osseuse, organes internes, empreintes digitales, etc.)

(ii) L'absence de cicatrices ou de lésions mécaniques ou chirurgicales détectées sur les tissus

(iii) Le fait qu'aucun ADN animal n'a été trouvé (plusieurs espèces locales testées)

(iv) Le fait que des séquences d'ADN d'humain moderne étaient présentes chez un individu en pourcentage mineur

(v) la présence d'un nourrisson trouvé sur le même site, présentant les mêmes caractéristiques atypiques suggèrent que, bien que nous ne puissions pas formellement confirmer l'authenticité, aucune donnée matérielle n'accrédite l' « hypothèse de la contrefaçon».

Au contraire, cet ensemble croissant de données suggère davantage **que nous pourrions être en présence d'espèces biologiquement non identifiées, qui méritent des recherches supplémentaires.**

(b) Leurs éventuels liens ou similitudes avec l'homo sapiens

Ici, les définitions doivent être posées. Les espèces ne sont pas définies sur la base de leur apparence ou morphologie similaire. Deux individus appartiennent à la même espèce si, et seulement si leur croisement donne lieu à une progéniture viable et non stérile.
Dans le cadre de la théorie de l'évolution, on peut dire qu'il existe un continuum entre les espèces. Deux espèces ayant un ancêtre commun sont considérées comme totalement différenciées si elles ne peuvent pas se croiser - ce qui inclut l'accouplement et la descendance fertile. Par exemple, les lions et les tigres peuvent s'accoupler, mais leur progéniture est généralement stérile. Même chose pour les chevaux et les ânes. Ils ont tous deux un ancêtre commun à partir duquel ils ont évolué différemment (le processus s'appelle la spéciation), mais ils ne sont pas encore totalement différenciés, dans la mesure où ils peuvent encore s'accoupler. En revanche, l'Homo sapiens et les singes (les chimpanzés, par exemple), qui sont supposés avoir un ancêtre commun, ne peuvent pas se reproduire ni avoir de progéniture. Ils sont considérés comme deux espèces différentes.
Néanmoins, les espèces connues, notamment les mammifères, présentent un très haut pourcentage de séquences en commun – plus de 95% selon les cas. Nous ne pouvons donc pas répondre à cette question sur la seule base d'analyses ADN.

(c) Leurs origines E.T

« Non identifié » ne signifie pas « Extraterrestre ». Notez que Maria semble être entièrement équipée pour survivre et se déplacer dans la biosphère terrestre. Par conséquent, d'un point de vue biologique, rien ne suggère a priori qu'elle viendrait d'une autre planète.

De plus, il n'existe pas de base de données sur les génomes d'E.T ou l'exobiologie en général. Nous ne pouvons donc pas comparer son génome avec quoi que ce soit considéré comme E.T, et par conséquent nous sommes incapables de parler de ses éventuelles origines E.T. Aujourd'hui, l'origine E.T n'est PAS (ou pas encore) une information génomique.

Les séquences non classifiées sont disponibles sur demande - les coordonnées, les preuves de qualifications et le type de ressources informatiques (le Cloud convient) peuvent être transmises à Alain Bonnet.

Annexe

Matériaux et méthodes

Les données

Les fichiers Illumina en paires (paired-end) étaient disponibles au format fastq. Chaque fichier (forward et reverse) contenait 329037 reads chacun pour Maria, et 341311 pour la Grande Main. Les fichiers fastq ont été convertis au format fasta à l'aide de scripts personnalisés permettant de compter le nombre de bases.

Contrôle Qualité Données brutes QC

Les fichiers de séquences en paires (paired-end séquences) ont été soumis à un contrôle en vue de déterminer la qualité par séquence et par base, le contenu de k-mer, l'adaptateur et toute autre contamination. L'outil de contrôle de qualité NGS Fastqc a été utilisé à cette fin. Les adaptateurs Illumina et la liste des contaminants standard ont été utilisés pour filtrer d'éventuelles source de bruit connues (le *bruit* en traitement du signal s'oppose au *signal* la qualité d'un signal se définit selon un rapport *signal/bruit*). Les reads étaient de longueurs variables allant de 35 à 151 bases.
Plus de 90% des reads ont obtenu le score exprimé de Q30 et un nombre minimal inférieur à Q20.
Globalement, les données brutes étaient de bonne qualité et ne présenterent aucun signe d'épuisement de réactifs ou d'artefacts de séquençage.

Alignement sur le génome humain (GRCh38)

Après avoir passé le contrôle de qualité, les reads ont été alignées sur le dernier et plus complet assemblage plus complet du génome humain, GRCh38. L'alignement a été effectué à l'aide de l'algorithme bwa-mem avec des options strictes. Bwa-mem est l'algorithme d'alignement de choix pour les reads de plus de 100 pb de longueur en raison de sa vitesse et de sa précision, en particulier avec les génomes de mammifères. Environ 33% et moins de 1% des reads furent alignés sur le génome humain GRCh38 et furent ensuite stockés au format *binary align map* (bam).

Bamtools a été utilisé pour filtrer les reads alignés et non- alignés. Les reads non alignés extraits de ce fichier d'alignement furent convertis au format fastq pour des analyses ultérieures.

Classification des reads non classifiés par alignement exact de k-mers

Les reads non classifiées à l'étape précédente ont été soumis à une analyse de classification à l'aide de Kraken. Kraken attribue des étiquettes taxonomiques aux lectures de séquençage en fonction de l'alignement exact des k-mères sur des groupes de génomes (bactérien, plasmide, virus, etc.). Une base de données de référence a d'abord été construite à partir de génomes bactériens, archéens et viraux complets dans RefSeq. Cette base de données était considérablement volumineuse et mesurait environ 8 gigaoctets. Pour éliminer la source principale de faux résultats positifs, tels que les séquences de faible complexité dans les génomes eux-mêmes (par exemple, pour une chaîne de 31 A ou plus consécutifs) nous avons exécuté le programme "dust" sur tous les génomes, puis construit la base de données à partir de ces génomes "nettoyés". L'analyse kraken a duré 6 heures (plateforme t2.xlarge EC2 hébergée sur Amazon AWS avec 4 unités de calcul haute puissance et 16 Go de RAM). À la fin, environ 27,7% des reads non classifiés ont été identifiés. Cela a contribué à 18,5% du total des lectures de l'échantillon. Diverses étiquettes taxonomiques bactériennes et virales ont été attribuées et les read non classifiés ont été rassemblés dans un fichier fastq.

Globalement, environ 33,7% des reads furent alignés sur le génome humain, 18,4% correspondaient aux génomes bactériens et les 47,9% restants ne furent pas identifiés. La même procédure fut suivie pour la grande main, avec des pourcentages différents.