ABRAXAS
Biosystems

RAPPORT DES SERVICES D'EXPERTISE

Analyse génomique et bio-informatique de séquençage à haut rendement de l'ADN des échantillons extraits des corps desséchés retrouvés à Nazca.

Client:
GAIA INTERNATIONAL, INC.

JOSÉ JAIME MAUSSAN
FLOTA

9 Novembre 2018

ABRAXAS
Bringing knowledge to life

Table des matières

VUE D'ENSEMBLE

Le présent document fournit les détails de tous les travaux, tâches et procédures impliquées dans le service fourni par ABRAXAS BIOSYSTEMS S.A.P.I. DE C.V. pour GAIA INTERNATIONAL, INC. Et JOSE JAIME MAUSSAN FLOTA pour le projet «Génomique et bio-informatique analyse du séquençage à haut débit de l'ADN des échantillons extraits des corps desséchés trouvés à Nazca ". Nous présentons une description ordonnée des tâches principales et l'analyse développée pour ce projet.

Les échantillons de tissus extraits des corps desséchés trouvés à Nazca et utilisés pour les analyses présentées dans ce service ont été fournies, dirigées et gérées par JOSE JAIME MAUSSAN FLOTA et ses collègues scientifiques à toutes les étapes précédant l'extraction de l'ADN décrite dans ce rapport alors que les laboratoires CEN4GEN (6756 - 75 Street NW Edmonton, AB) Canada T6E 6T9) étaient chargés d'exécuter toutes les tâches sur les échantillons depuis l'extraction de l'ADN au séquençage à haut débit de l'ADN, également connu sous le nom de Séquençage de Prochaine Génération, et étapes de génération de données de séquençage propres.

ABRAXAS BIOSYSTEMS S.A.P.I. DE C.V. a effectué toute la génomique informatique et l'Analyse bio-informatique.

Pour ce projet, JOSE JAIME MAUSSAN FLOTA et ses collègues scientifiques ont assuré la livraison aux laboratoires du CEN4GEN, 7 échantillons, 3 échantillons de tissus et 4 échantillons d'ADN provenant des corps trouvés à Nazca, au Pérou. Après l'extraction de l'ADN, le contrôle de la qualité et les procédures d'amplification du MDA au laboratoire CEN4GEN, seuls 3 échantillons, sur les 7 originaux, ont passé les contrôles de NGS, les noms de ces échantillons, provenant des tubes originaux envoyés pour livraison, étaient comme suit :

Nom échantillon	Etiquette originale de l'échantillon	Identité
Ancient-0002	Os du cou entité assise	Victoria
Ancient-0003	1 Main 001	Main
Ancient-0004	Momia 5 - ADN	Victoria

Table 1: Le nom de l'échantillon indique le nom que CEN4GEN a attribué à l'échantillon, L'étiquette d'origine de l'échantillon est le nom dans le tube où l'échantillon a été initialement contenu lors de sa livraison à CEN4GEN, Identity est le nom du corps d'où provient l'échantillon.

C'est pourquoi toutes les tâches d'analyse mentionnées dans ce rapport après l'extraction de l'ADN, le contrôle de la qualité et les tâches d'amplification MDA n'ont été effectuées que pour ces 3 échantillons.

INDEX DES SERVICES

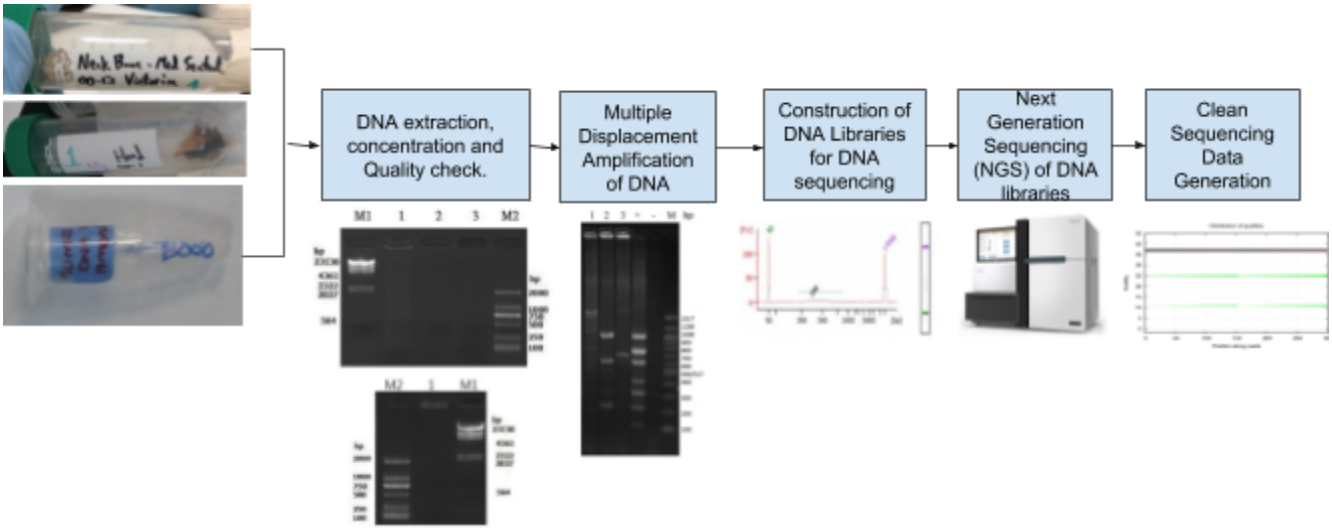
La solution complète proposée par Abraxas Biosystems comprend une large gamme de services allant de l'extraction d'ADN ancien au séquençage et à l'analyse de données (bio-informatique), permettant la génération de résultats précis à partir d'analyses d'échantillons anciens. Les tâches d'analyse réalisés dans le cadre de ce projet sont les suivants :

1. Extraction d'ADN.
2. Contrôle de la qualité de l'ADN.
3. Amplification de l'ADN par amplification par déplacement multiple.
4. Construction de banques d'ADN.
5. Séquençage de l'ADN de prochaine génération (NGS).
6. Génération de données de séquençage propre.
7. CQ des résultats de séquençage.
8. Analyse préliminaire par cartographie des lectures d'ADN sur la référence du génome humain.
9. Analyse de recoupement pour détecter de courts fragments communs à l'ADN ancien.
10. Cartographie des lectures d'ADN superposées de Ancient0003 à la version la plus récente du génome humain.
11. Analyse mitochondriale pour la détection de variants dans les boucles D et autres régions informatives pour déterminer les haplotypes mitochondriaux.
12. Détermination du sexe de l'échantillon Ancient0003.
13. Détection des éventuels organismes présents dans l'échantillon par la méthode d'esquisse d'ADN génomique (correspondances exactes de groupes de fragments courts, k-mers, avec des bases de données publiques) et filtrage itératif des lectures par correspondances exactes k-mer.
14. Assemblage de novo avec stratégies d'ADN mixte de lectures sans correspondance avec les organismes détectés dans la méthode d'esquisse.
15. Cartographie des lectures sans correspondance exacte dans le processus de filtrage itératif avec les séquences résultantes dans l'assemblage de novo.

16. Recherche dans les bases de données ADN de segments d'ADN assemblés de novo afin de détecter une similitude avec organismes connus.
17. Classification taxonomique des séquences non appariées dans les étapes précédentes par correspondance des recherches dans des bases de données génétiques complètes.

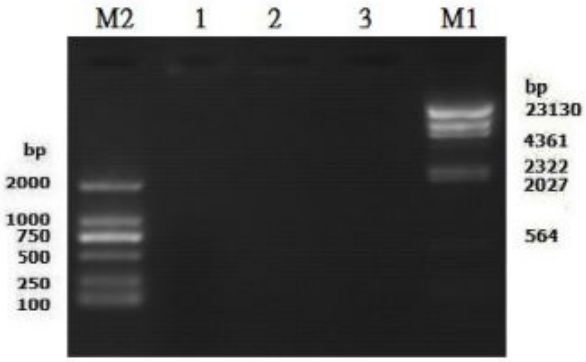
DEROULEMENT DES OPERATIONS DES SERVICES DU LABORATOIRE

Pour les processus d'analyse de laboratoire, tâches 1 à 6, les opérations générales étaient les suivantes.

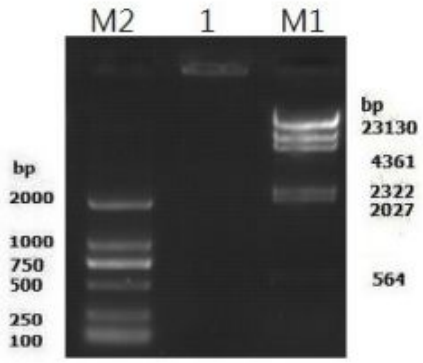


EXTRACTION DE L'ADN

Les 3 échantillons de tissus ont été analysés selon un protocole d'extraction d'ADN spécifique pour échantillons anciens et développés dans les laboratoires CEN4GEN, basés sur les protocoles décrits à l'article suivant (Gamba et al., 2016). Après le processus d'extraction de l'ADN, l'ADN a été passé sur des gels d'agarose pour vérifier la présence de bandes indiquant la présence de quantités adéquates d'ADN (indiquées par des bandes horizontales lumineuses visibles sur chaque piste correspondant à chaque échantillon). De plus, l'ADN déjà extrait de l'Ancient-004 était vérifié par cette méthode car il contenait l'ADN d'un échantillon de tissu qui n'était plus disponible sur le corps de Victoria. Les résultats sont présentés dans la figure suivante :



Lane No.	Sample Name	Dilution Ratio(×)	Test Volume(μL)	Sample Integrity
M1	λ-Hind III digest(Takara)	1	3	
1	CEN4GEN-Ancient0001	1	3	Degraded completely
2	CEN4GEN-Ancient0002	1	3	Degraded completely
3	CEN4GEN-Ancient0003	1	3	Degraded completely
M2	D2000 (Tiangen)	1	6	



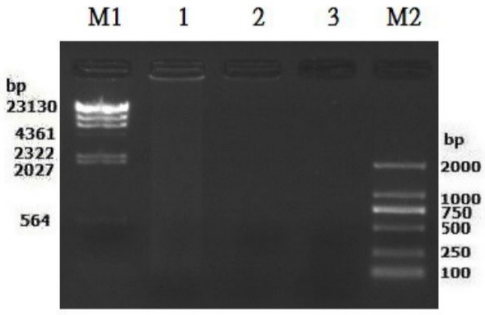
Lane No.	Sample Name	Dilution Ratio(×)	Test Volume(μL)	Sample Integrity
M1	λ-Hind III digest(Takara)	1	3	
1	CEN4GEN-Ancient0004	1	3	
M2	D2000 (Tiangen)	1	6	

Figure 1 : Résultats de l'extraction de l'ADN (en haut) et contrôle de la qualité de l'ADN déjà extrait (en bas). M2 et les lignes M1 sur chaque gel sont les marqueurs moléculaires utilisés pour mesurer la taille des fragments d'ADN.

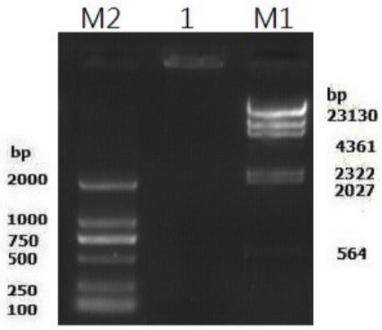
La figure 1 montre que les bandes au niveau des échantillons d'ADN extraits présentaient une visibilité insuffisantes, donc pas assez d'ADN pour des résultats NGS corrects. Cela a obligé les laboratoires à évaluer les 4 échantillons d'ADN pour voir si ceux-ci avaient assez d'ADN.

ANALYSE DE LA QUALITÉ D'ADN

Après extraction de l'ADN des 3 échantillons de tissus, les 4 échantillons d'ADN ont été analysés par un processus de contrôle de la qualité permettant d'évaluer la présence de bonnes quantités et de tailles d'ADN afin de voir s'ils pouvaient sauver l'ADN nécessaire pour le NGS. Le contrôle de qualité a également été effectué dans des gels d'agarose comme dans l'étape précédente. Les résultats sont montrés plus bas :



Lane No.	Sample Name	Dilution Ratio(×)	Test Volume(μL)	Sample Integrity
M1	λ-Hind III digest(Takara)	1	3	
1	CEN4GEN-Momia1	1	3	Degraded completely
2	CEN4GEN-Momia3	1	3	N/A
3	CEN4GEN-Momia4	1	3	N/A
M2	D2000 (Tiangen)	1	6	



Lane No.	Sample Name	Dilution Ratio(×)	Test Volume(μL)	Sample Integrity
M1	λ-Hind III digest(Takara)	1	3	
1	CEN4GEN-Ancient0004	1	3	Degraded completely
M2	D2000 (Tiangen)	1	6	

Figure 2 : Contrôle de la qualité de l'ADN d'échantillons d'ADN déjà extraits non analysés lors du contrôle de qualité précédent (en haut) et contrôle de la qualité de l'ADN déjà extrait préalablement analysé avec les échantillons de tissus de l'extraction d'ADN (en bas).

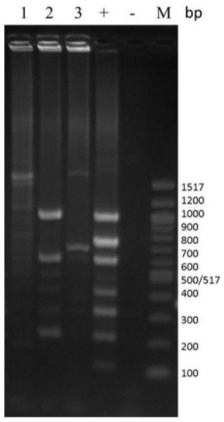
Les résultats sont les mêmes que pour l'ADN extrait des tissus par les laboratoires CEN4GEN montrant très faible présence de DNAe. Cette absence de quantités élevées d'ADN et l'avantage de NGS pour récupérer des données de faibles entrées d'ADN avec quelques efforts d'amplification a conduit l'équipe du laboratoire CEN4GEN à exécuter un processus appelé Amplification à Déplacement Multiple qui avait montré un bon résultat dans leurs installations avec des échantillons anciens pour amplifier les niveaux d'ADN disponible nécessaires pour le séquençage NGS.

AMPLIFICATION DE L'ADN

Par Amplification à Déplacement Multiple

Après avoir trouvé de très faibles quantités résultant de l'état dégradé des échantillons de tissus, les laboratoires se sont tournés vers le processus de la MDA pour amplifier les quantités de fragments d'ADN. Ce processus a été personnalisé pour les caractéristiques de l'ADN extrait à l'aide des méthodes exclusives des Laboratoires CEN4GEN. Les résultats de la MDA étaient acceptables pour pré-alimenter avec NGS pour 2 des échantillons de l'ADN extraits par CEN4GEN (Ancient0002 et Ancient0003) et pour 1 des échantillons déjà livré extrait (Ancient0004) comme indiqué ci-dessous :

Electrophoretogram:



Lane No.	Sample Name	Dilution Ratio(×)	Test Volume(μL)	Number of housekeeping genes detected
1	CEN4GEN-Ancient0002	1	10	0
2	CEN4GEN-Ancient0003	1	10	3
3	CEN4GEN-Ancient0004	1	10	0
+	Positive control	1	10	7
-	Negative control	1	10	0
M	100bp DNA ladder (NEB)		6	/

Figure 3 : Résultats de l'amplification MDA, les 3 premières voies correspondent aux valeurs amplifiées avec succès des échantillons et les quatrième et cinquième dernières voies sont destinés respectivement au contrôle négatif et aux marqueurs moléculaires.

Le reste des échantillons n'a pas montré les résultats de l'amplification, alors le reste des tâches d'analyse ont été effectuées uniquement pour ces 3 échantillons.

ANALYSE DE SIMILARITÉ À DES ORGANISMES CONNUS

Les contigs résultants ont été analysés par rapport à la base de données NT à l'aide de blastn v 2.7.1 (à l'aide d'une Valeur E de 10, à une taille de mot de 20 et un pourcentage d'identité de 30) pour rechercher des correspondances possibles avec les organismes connus dans la base de données NT et le nombre de résultats a été compté pour déterminer si les fragments assemblés avaient de meilleurs résultats en obtenant une correspondance avec des organismes connus. Au total, pour l'échantillon Ancient0002, seuls 1 256 (sur 60 852) contigs n'ont eu aucune correspondance et pour Ancient0004, seulement 1 768 (sur 54 273) contigs n'ont obtenu aucune correspondance.

Également un autre assemblage de novo qui a utilisé comme entrée les deux ensembles de lectures uniques non appariées de Ancient0002 et Ancient0004 ensemble, mais les résultats de montage ont été beaucoup moins assemblé et avait des résultats beaucoup plus fragmentés donc cet assemblage supplémentaire était mis au rebut.

CLASSIFICATION TAXONOMIQUE ULTRA COMPLETE

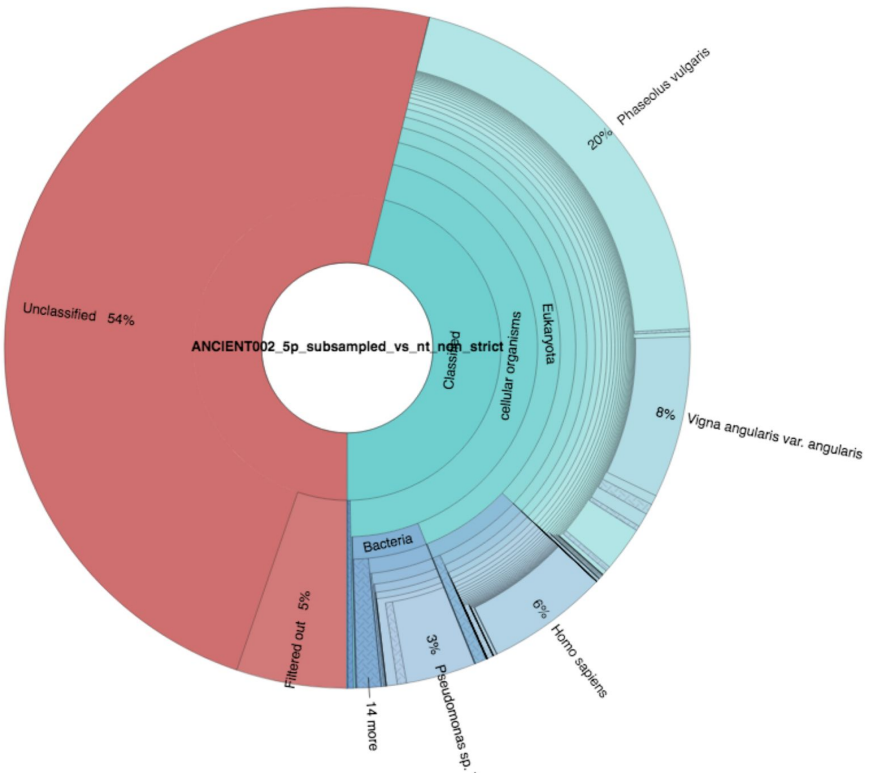
Pour des lectures des sous-échantillons uniques et brutes

Les tâches précédentes visaient à trouver la quantité de lectures d'ADN classifiables à partir des Échantillons Ancient0002 et Ancient0004 afin de comprendre dans quelle mesure les échantillons avec une faible cartographie sur le génome humain ressemblent à des organismes connus au niveau de l'ADN. Même si notre schéma génomique et notre approche de filtrage itératif étaient complet à de nombreuses gammes taxonomiques et types d'organismes, dont nous avions besoin de produire une classification encore plus détaillée des lectures du séquençage de l'ADN à tous les niveaux taxonomiques possibles. et avec une base de données encore plus grande avec un algorithme de correspondance plus flexible pour élargir, dans la mesure du possible, les spectres et la puissance de détection de nos méthodes d'appariement de l'ADN à des organismes connus et séquencés de la Terre. Pour parvenir à ce spectre accru de détection et de classification détaillée à tous les rangs taxonomiques possibles nous mettons en œuvre une stratégie ultra complète et très sensible basée sur la création d'une nouvelle base de données avec encore plus d'entités, avec l'un des plus jeux complets de données de séquence connus en bio-informatique sous le nom de base de données NCBI, construits à l'aide d'algorithmes de compression et de redondance et reposant également sur la mise en oeuvre d'une recherche de correspondance non exacte optimale comparable en sensibilité au très

sensible BLAST mais, dans la pratique, cette recherche utilisant une recherche BLAST aurait pris des mois en élargissant ainsi la puissance de recherche de l'esquisse et le filtrage itératif bien que cette stratégie soit limitée à l'exactitude des correspondances. Cette stratégie a été mise en œuvre à l'aide du logiciel taxmaps v 0.2.1 (Corvelo, Clarke, Robine, & Zody, 2018) et appliqué à un sous-ensemble de 5% de toutes les lectures brutes non filtrées pour la Échantillons Ancient0002 et Ancient0003.

La même analyse a été répétée pour un sous-ensemble de 25% de l'échantillon Ancient0004, trop juste pour confirmer ce que nos méthodes prédisaient correctement pour les proportions des lectures classifiés et non classées au fur et à mesure que les échantillons se rapprochent de l'ensemble des lectures (ce qui a nécessité des dizaines de jours supplémentaires).

Cette stratégie nous a également permis de comparer si le comportement des processus de chevauchement et de filtrage était différent par rapport aux lectures non filtrées originellement séquencées dans leurs correspondances avec des organismes connus ».
ADN Les résultats sont indiqués ci-dessous :



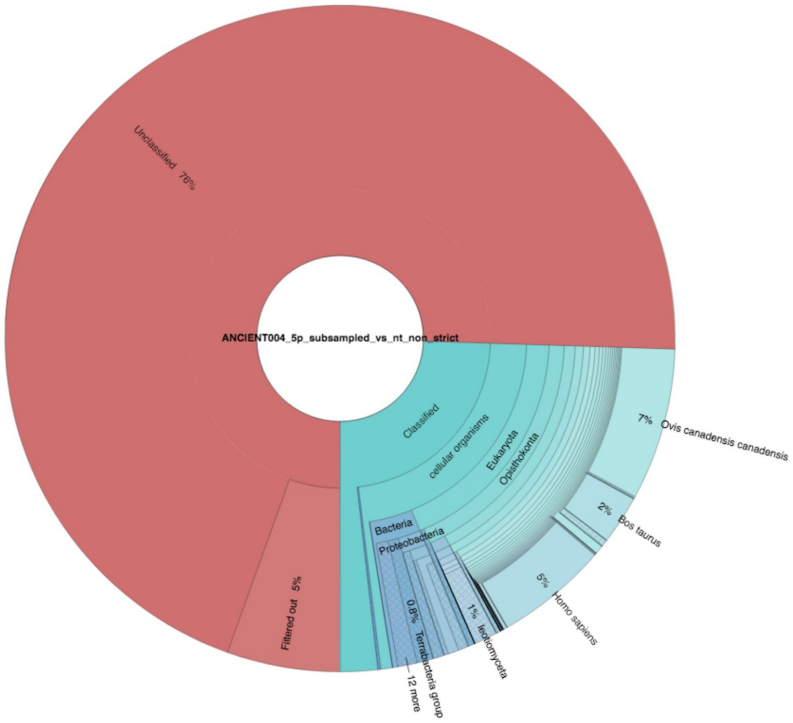


Figure 8: Proportion de lectures classifiées et non classées d'un sous-échantillon de 5% (28 073 655 lectures pour Ancient0002 et 25 084 962 pour Ancient0004) des lectures complètes de séquençage brut pour Ancient 0002 (en haut) et Ancient0004 (en bas) par rapport à la base de données NCBI telle que mise en œuvre dans les cartes de taxe 0.2.1 qui comprend 34 904 805 séquences d'ADN représentant 1 109 518 de Taxa.

Cette approche a confirmé la présence de taux très élevés d'ADN non correspondant et non classifiés contenu dans les échantillons séquencés par rapport à l'une des plus complètes bases de données compilées publiquement pour les informations génomiques dans les paramètres considérés (un distance de modification autorisée de maximum 0,2 entre les kmers recherchés par taxmaps contre la base de données non redondante mise en œuvre pour la base de données nt).

CONCLUSIONS

Abraxas Biosystems a effectué une vaste gamme d'analyses bio-informatiques et génomiques afin d'identifier l'origine biologique possible et l'ascendance des échantillons fournis par Jaime Maussan et ses collègues scientifiques et extraits / séquencés dans les laboratoires CEN4GEN. Après la conception d'un protocole méticuleusement personnalisé pour maximiser le taux de réussite de l'extraction de l'ADN ancien, le séquençage (avec CEN4GEN Labs) et l'analyse bio-informatique des échantillons, les résultats montrent une correspondance

très faible avec les données du génome humain pour les échantillons Ancient0002 et Ancient0004 contrairement à l'échantillon Ancient0003 qui montrait une cartographie très élevée correspondant au génome humain. Il faut aussi noter que les échantillons Ancient0002 et Ancient0004 ne montrent que très peu de correspondances avec l'une des bases de données les plus fiables et les plus précises (à partir de NCBI). Cependant, les bases de données NCBI ne contiennent pas tous les organismes connus existant dans le monde, donc il pourrait y avoir beaucoup d'organismes possibles qui pourraient correspondre à cet ADN ou certaines régions qui pourraient être exclues ou difficiles à séquencer, communes à de nombreux organismes dans ces échantillons, et dans les protocoles appliqués aux génomes déclarés au NCBI.

Les protocoles de laboratoire et de calcul pour l'analyse de l'ADN ancien, compte tenu de la nature des échantillons, comprennent plusieurs étapes susceptibles de générer des interférences dans les données et d'avoir un impact direct sur les résultats. L'un des exemples les plus courants est la manipulation de tissus par plusieurs individus et à l'environnement ouvert antérieur à son isolement, compliquant les possibilités que tout l'ADN séquencé provient de l'ADN endogène des corps individuels échantillonnés. Un moyen d'éviter ce type de bruit et d'obtenir de meilleurs résultats consiste à séquencer des échantillons internes d'os et non des tissus exposés.

Enfin, les bases de données actuelles du NCBI se développent constamment, il est donc possible qu'une meilleure bases de données plus complète pourraient bientôt être développée qui comprendrait plus de génomes microbiens et / ou eucaryotes disponibles pouvant éclairer la nature des échantillons d'ADN non correspondant. De plus, une analyse ciblée des segments d'ADN non correspondants pourrait être développée pour confirmer que ce ne sont pas des restes du séquençage ou des protocoles de l'amplification. Les anciens protocoles d'ADN font l'objet d'une amélioration continue compte tenu de sa sensibilité caractéristiques de dégradation de ce type d'échantillons. Nous recommandons des études supplémentaires pour accepter ou rejeter toute autre conclusion.