REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur de le Recherche Scientifique
Université SAAD DAHLEB – BLIDA 1

Faculté des Sciences de la Nature et de la Vie

Département De Biologie

Mémoire de fin d'études

En vue de l'obtention du diplôme de Master dans le domaine SNV

Filière Sciences Biologique

Option : Génétique

Thème :

PREDICTION FACIALE A PARTIR DE L'EXPRESSION DE CERTAINS GENES DANS LE CAS DE LA RECHERCHE D'ENQUETE DE CRIME OU DISPARITION

Présentée par :

BOUNAIM Khadidja Wanissa
SIDI-YAKHLEF Mahdjouba Nassima

Date de la soutenance :

17/07/2021

Devant le jury :

Nom	Grade	Lieu	Qualité
BENYAHIA N.	MAA	Blida	Président
BESSAAD M .E.A.	MCA	Blida	Examinateur
MOHAMED SAID R.	MCA	Blida	Promoteur

Promotion : 2020-2021

SOMMAIRE

Liste des figures

Liste des tableaux

Liste des abréviations

ADN : Acide désoxyribonucléique
AFR : Africaine
AIM's : Marqueur informatif sur l'ascendance
ARN : Acide ribonucléique
AMR : Amérique du nord
ARNm : Acide ribonucléique messager
ASC : Air sous la courbe
CE : Electrophorèse capillaire
CEPH : Centre d'etude du polymorphisme humain
CCA : Analyse des corrélations canoniques
CV : Vecteur de correspondance
CV : Cross validation
CPG : Cytosine-phosphate-guanine
CSA :Asie centrale et du sud
Dbsnp :Base de donnée SNP
EAS :Est-asiatique
EU :Etats-Unis
EUR :Européenne
EVC : Externally visible characteristics
FDP : Forensic dna phenotyping
FSCP : Paramètre de changement de forme du visage
GB :Grande Bretagne
GREAT :Genomic regions enrichment of annotations tools
GWAS :Genome wide association study
HapMap :Haplotype Map
HGDP : Projet de diversité du génome humain
HPS : HIrisPlex-S
IMC : Unité de masse corporelle
IUPUI US : Indiana university – Purdue university Indianapolis United States
MC : Centre médicale
ML : Machine-learning
MLR : Regression logistique multinomiale
Ng : Nanogramme
NGS : Next-generation sequencing
OBJ : Objet 3D
PC : Composantes principales
PCR :Réaction en chaine par polymérase
Pg : Picogramme
QPCR : Réaction en chaine par polymérase quantitative
RIP :Réponse de prédicteur imputées
Rsid ID : Référence des clusters de SNP
SBE : Extension de base unique
SNP : Single nucleotide polymorphisme
STRS : Short tandem repeat
WGDAM : Scientific working group on DNA analysis methods
TFS : Thermo fisher scientific
UV : Ultra-violet
Valeur P : Valeur prédictive
VPN :valeur prédictive négative
VPP :valeur prédictive positive
3D :3 dimension

Remerciment

En premier lieu, nous n'aurons pas pu faire ce travail sans la bénédiction et la puissance d'Allah soubhanaho wa taala..

Nos vifs remerciements sont d'abord adressés à monsieur MOHAMED SAID Ramdane qui a fait l'honneur de diriger ce travail, nous tenons à lui exprimer le profond respect

Nous tenons également à exprimer une reconnaissance au membre du jury

Monsieur BENYAHIA Nourredine de nous avoir honoré de présider le jury de la soutenance

Monsieur BESSAAD Mohamed El Amine, d'avoir bien accepté d'examiner les contenus du présent travail

Dédicace

Je dédie cette ouvrage

Aux êtres les plus chers de ma vie Maman

Et mon cher Papa

Qui m'ont soutenu et encouragé durant toutes mes années

qu'ils trouvent ici le témoignage de ma profonde reconnaissance

A mes frères Abida et Hamzouz qui ont partagé avec moi tous les moments d'émotion lors de la réalisation de ce travail

A mon époux d'amour qui m'a chaleureusement supporté, encouragé et qui m'a donné de la vivacité

A toutes mes amies que je considère comme des sœurs et à qui je souhaite plus de succès

Sans oublié mon binôme Nassima pour son soutienmoral, sa patience son bon humour et compréhension tout au long de ce projet.

Dieu puisse vous donner santé, bonheur et réussite.

Dédicace

Je dédie cette ouvrage

A l'être la plus cher de ma vie Maman,

Quoi que je fasse ou que je dise, je ne saurai point te remercier comme il se doit. Tes prières et ton affection me couvre, ta bienveillance me guide et ta présence à mes côtés a toujours été ma source de force pour affronter les différents obstacles

A mon cher Père, ma sœur chérie Zola et mon frère Midou,

Vous avez toujours été à mes côtés pour me soutenir et m'encourager

Que ce travail traduit ma gratitude et mon affection

A mes neveux et ma nièce chérie d'Amour

A ma très chère famille SIDI-YAKHLEF

Ma unique et chère tante Hayet,

Mes oncles et leurs épouses,

A toutes mes cousins et cousines,

A la famille Yahiaoui

Mon cher oncle Kader et sa femme chérie et ses belles petites filles

A ma chère tante Karima et ses enfants

A toutes mes amies que je considère comme des sœurs et à qui je souhaite plus de succès

Sans oublié mon binôme Wanissa pour m'avoir choisi comme binôme pour partager cette étude, son soutien moral, sa patience son bon humour et compréhension tout au long de ce projet

Puisse Dieu vous donne santé, bonheur et réussite.

Résumé

Il peut paraître impossible ou surréel, mais il est maintenant possible, de prédire les traits de pigmentation et faciaux de l'être humain. L'ADN a toujours été un mystère pour les scientifiques par rapport à sa complexité, ce dernier contient l'ensemble des caractères physique d'un individu produisant une large gamme d'apparence parmi les populations. Reste que la prédiction d'un portrait d'une personne inconnue est toujours un challenge pour les chercheurs aujourd'hui

La préoccupation la plus critique concernant la validité d'un modèle de prédiction est que les observations de chaque étape doivent être indépendantes.

Une analyse de prédiction systématique comprend généralement trois étapes.

La première étape est l'étape de découverte et qui est l'identification des relations entre génotypes et phénotypes réalisée par le GWAS. Ce système permet de reconnaitre des milliers de SNP qui peuvent être responsable des variétés faciales.

Ensuite l'étape de construction du modèle, cette étape se fait par 2 moyens : statistique qui englobe les différents modèles de régressions et les techniques du Machine Learning qui utilise multiple classificateurs.

La dernière étape est celle de la validité du modèle où les valeurs des variables obtenues dans les étapes précédentes sont comparées aux valeurs réellement observées pour estimer la précision de la prédiction et générer des paramètres d'applications.

Au final ces résultats sont projetés vers des ordinateurs qui pourront donner la forme réelle du portrait-robot. Bien que cette étude n'est qu'à ses début, plusieurs affaires criminelles datant des années 50 ont enfin était résolues. D'ici quelques années et avec le développement rapide de la science, la prédiction d'un profil inconnu pourra être un jeu d'enfant.

الملخص

قد يبدو الأمر مستحيلاً أو سريالياً ولكن من الممكن الآن التنبؤ بالصبغ وميزات الوجه للإنسان. لطالما كان الحمض النووي لغزا للعلماء بسبب تعقيده ، فهو يحتوي على جميع الخصائص الفيزيائية للفرد مما ينتج عنه مجموعة واسعة من المظهر بين السكان. ومع ذلك ، فإن توقع صورة شخص مجهول لا يزال يمثل تحديًا للباحثين حتى يومنا هذا ، وأهم مصدر قلق بشأن صحة نموذج التنبؤ هو أن ملاحظات كل خطوة يجب أن تكون مستقلة. يتكون تحليل التنبؤ المنهجي عادةً من ثلاث خطوات. تتمثل الخطوة الأولى في اكتشاف وتحديد العلاقات بين الأنماط الجينية والنمط الظاهري ، حيث يسمح هذا النظام بالتعرف على الآلاف من GWAS الذي تم تنفيذه بواسطة الأشكال المتعددة الأشكال التي قد تكون مسئولة عن أنواع الوجه. ثم خطوات بناء النموذج ، تتم هذه الخطوة من خلال مجالين: الإحصائيات التي تشمل نماذج الانحدار المختلفة وتقنيات التعلم الآلي التي تستخدم المصنفات المتعددة. الخطوة الأخيرة هي صحة النموذج حيث تتم مقارنة قيم المتغيرات التي تم الحصول عليها في الخطوات السابقة بالقيم التي تمت ملاحظتها بالفعل لتقدير دقة التنبؤ وإنشاء معلمات التطبيق. في النهاية ، يتم عرض هذه النتائج على أجهزة الكمبيوتر التي ستكون قادرة على إعطاء الشكل الحقيقي لصورة الروبوت. على الرغم من أن هذه الدراسة لا تزال في مراحلها الأولى ، فقد تم حل العديد من القضايا الجنائية التي يعود تاريخها إلى الخمسينيات من القرن الماضي. في غضون بضع سنوات ومع التطور السريع للعلم ، قد يكون التنبؤ بملف تعريف غير معروف أمرًا مفاجئ.

Abstract

It may seem impossible or surreal but it is now possible to predict the pigmentation and facial features of the human being. DNA has always been a mystery to scientists due to its complexity, it contains all the physical characteristics of an individual producing a wide range of appearance among populations. Still, predicting a portrait of an unknown person is still a challenge for researchers today. The most critical concern about the validity of a prediction model is that the observations of each step must be independent. A systematic prediction analysis typically consists of three steps. The first step is the discovery and identification of the relationships between genotypes and phenotype carried out by the GWAS. This system allows the recognition of thousands of SNPs that may be responsible for facial varieties. Then the model building steps, this step is done through 2 areas: statistics which include the different regression models and Machine Learning techniques which use multiple classifiers. The last step is that of the validity of the model where the values of the variables obtained in the previous steps are compared to the values actually observed to estimate the precision of the prediction and generate application parameters. In the end, these results are projected onto computers which will be able to give the real shape of the robot portrait. Although this study is only in its early stages, several criminal cases dating from the 1950s have finally been resolved. Within a few years and with the rapid development of science, predicting an unknown profile may be a snap.

INTRODUCTION :

« L'identité humaine est devenue très importante pour diverses raisons en premier lieu la capacité de reconnaître les personnes individuellement et en second lieu d'être en mesure de les identifier dans une grande foule d'individus. L'identité humaine, c'est simplement le fait que nous ne sommes pas identiques, nous sommes tous différents phénotypiquement dus au support génétique différents dans ses expressions appelé polymorphisme. En effet notre ADN détermine qui nous sommes et à quoi nous ressemblons. Donc de cause à effet comment alors établir l'identité de quelqu'un à partir de son ADN ? L'mage première que nous voyons et qui nous donne cette identité individuelle des personnes sont très exprimées dans notre visage qui est le panneau d'affichage biologique de notre identité dans ce monde. »

Incroyablement les humains partagent 99.9% de la constitution de leur génome qui veut dire 0.1% de leur ADN est responsable de la diversité entre les individus. Chacune de ces formes est appelée variante, (plusieurs formes) qu'on appelle Single Nucléotide Polymorphisme (SNP). Ces SNP sont essentiels pour comprendre les causes génétiques des traits humains, bien que certains traits soient entièrement environnementaux, mais d'autres, comme la couleur des yeux, sont extrêmement héritables. Les SNP peuvent nous aider à comprendre dans quelle mesure certains traits sont génétiques et quels mécanismes biologiques de notre corps peuvent affecter ces traits. (Claes, 2015)

L'identité faciale d'une personne peut être étudiée par le processus de caractérisation de l'identité (par exemple un visage d'une personne est défini par le fait qu'elle a un très petit nez retroussé, un menton et front proéminents, des yeux bleus et des cheveux marron. Donc ce que nous étudions ici c'est le rôle de certains de nos gènes contenus dans notre ADN et l'expression de notre visage. De ce fait si nous arrivons à établir la relation entre les deux (rôles des gènes et le faciès) cela signifierait que nous serons en mesure de prédire un visage à partir de l'ADN d'une personne inconnue. Ainsi, si nous trouvons de l'ADN sur un mégot de cigarette ou sur une scène de crime (salive poil, sperme etc..) nous pourrons ainsi identifier la personne à qui appartient cette ADN et ceci simplement par prédiction de son visage. De même on peut prédire un visage de personne disparu il y a longtemps et voir son facies maintenant.
Le GWAS l'un des principaux chevaux de bataille de toutes les recherches génétiques de nos jours qui se résume à une étude d'association à très grande échelle d'analyses d'association cas-témoin et de variantes génétiques analysés. Cette étude se fait en identifiant un ensemble informatif de plus d'un million de polymorphismes mononucléotidiques (dit SNP) à travers le génome. La disponibilité de quantités massives de données GWAS a nécessité le développement de nouvelles méthodes biostatistiques pour le contrôle de la qualité, l'imputation et les problèmes d'analyse, y compris les tests multiples. Les études d'association pangénomique ont identifié certains gènes clés et des sites au sein de ces gènes qui influencent la pigmentation des yeux et la couleur des cheveux, ainsi que les phénotypes de la couleur de la peau (Chaitanya et al, 2016).
Ces gènes ont été largement utilisés pour prédire la pigmentation à partir du génotype, principalement dans le contexte médico-légal (Forensic Dna Phenotyping (dit FDP). Par voie de conséquence des systèmes majeurs ont été développé et validé : système de test

d’ADN IrisPlex pour la prédiction de la couleur des yeux, système HIrisPlex pour prédiction combinée de la couleur des yeux et des cheveux et système de test d’ADN HIrisPlex-S (S pour la peau (skin) pour la prédiction simultanée de la couleur des yeux, des cheveux et de la peau.

Le système FDP se compose de deux tests multiplex basés sur SNaPshot ciblant un total de 41 SNP via un nouveau test multiplex pour 17 SNP prédictifs de la couleur de la peau et le test HIrisPlex précédent et pour 24 SNP prédictifs de la couleur des yeux et des cheveux, dont 19 contribuent également à la prédiction de la couleur de la peau.
Le système HIrisPlex-S comprend en outre trois modèles de prédiction statistique, le modèle IrisPlex précédemment développé pour la prédiction de la couleur des yeux basé sur 6 SNP, l'ancien modèle HIrisPlex pour la prédiction de la couleur des cheveux basé sur 22 SNP et le modèle HIrisPlex-S récemment introduit pour la couleur de la peau, prédiction basée sur 36 SNP. Ils ont introduit ici des solutions de séquençage massivement parallèle (dit MPS) pour le système HIrisPlex-S (dit HPS) sur deux plates-formes MPS couramment utilisées en criminalistique, respectivement Ion Torrent et MiSeq (Illumina), qui couvrent les 41 variantes d'ADN en un seul test. En ce qui concerne notre étude, nous présenterons la validation du développement médico-légal des deux tests HPS-MPS.

1. Généralités

1.Généralités :

1.1 GWAS :

C’est une base de données appelée Genome-Wide Association Studies, qui permet aux généticiens d’identifier les gènes qui sont responsables des différences phénotypiques qui nous intéressent. Cette technique a permis aussi de fournir des informations auxquelles les entreprises pionnière de génomique privé telles que Parabon Nano-lab de développer une technologie pour prédire l'ascendance inconnue, la pigmentation et l'architecture du visage qui génère de nouvelles pistes d'enquête dans des cas de crimes.

Les outils et les pros logiciels introduits et utilisés pour les différentes taches du GWAS afin de réaliser une étude d'association pangénomique, où deux types de données sont disponibles pour tous les individus : l'étude de leur phénotype et de leur génotypequi peuvent être obtenu par séquençage du génome (génotypage).(Gumpinger et al ,2018).

I.2 Méta-analyse

Historiquement, la méta-analyse a été développée comme un outil permettant de combiner les résultats d'essais cliniques similaires. Après l'avènement des GWAS, la méta-analyse s'est avérée être une méthodologie robuste pour combiner les résultats obtenus à partir de différentes études. Comme chaque étude d'association pangénomique individuelle a normalement une taille d'échantillon modeste, une méta-analyse de plusieurs études d'association pangénomique a la capacité d'augmenter la puissance globale et de réduire les faux positifs. Dans le cadre d'une méta-analyse d'études d'association pangénomique, il est courant de mettre en commun les signaux d'association détectés dans différentes études sans utiliser explicitement les données génétiques sous-jacentes. Il s'agit d'un autre aspect qui fait de la méta-analyse une méthode attrayante, car l'accès aux données génotypiques est souvent réglementé par des règles strictes de protection de la vie privée. Les personnes qui participent à une étude peuvent consentir à ce que leurs données génétiques soient utilisées dans cette étude spécifique, mais ne permettent que la diffusion soit qu’auprès d'autres groupes de recherche. Il existe différentes façons d'intégrer les signaux provenant de différentes études. Les techniques les plus couramment utilisées sont la méthode de Fisher et la méthode pondérée de Stouffer, ainsi que les approches basées sur des modèles à effet fixe et aléatoire. La décision de la méthode la plus appropriée pour combiner les résultats de GWAS dépend fortement des hypothèses sous-jacentes des études en question. (Gumpinger et al ,2018).

I.3 Prédiction des caractéristiques visibles de l'extérieur à partir des SNP : couleur des yeux, des cheveux et de la peau

D’après Chaitanya, la pigmentation est le trait humain le plus visible, le plus variable et le plus perspicace, observé dans la coloration de l'œil (iris), des cheveux et de la peau. Cela a créé un penchant pour l'identification des gènes de pigmentation et de leurs variantes polymorphes qui représentent les différents traits phénotypiques existant entre les individus au sein et entre les populations.

La pigmentation humaine est le résultat de plusieurs voies génétiques et biochimiques complexes. La mélanine, pigment présent dans la couche basale de l'iris, du bulbe pileux et de l'épiderme de la peau, détermine la couleur des yeux, des cheveux et de la peau.

La mélanine est synthétisée dans les mélanosomes dans les mélanocytes. Deux principaux types de mélanine sont présents chez l'homme : l'eumélanine et laphéomélanine.
L'eumélanine, produite dans les eumélanosomes, est un pigment brun-noir responsable de la couleur sombre, et la phéomélanine, produite dans les phéomélanosomes, est un pigment rouge-jaune.
La pigmentation globale est régie par la quantité et type de mélanine (rapport entre l'eumélanine et la phéomélanine) la forme et la distribution des mélanosomes.
Les auteurs ont trouvé que les SNP dans un certain nombre de gènes de pigmentation étaient associés aux phénotypes de couleur des yeux, des cheveux et de la peau.
Le projet sur le génome humain et les projets HapMap ont fourni des informations scientifiques sur la séquence d'ADN humain et sa structure de bloc d'haplotype. Ces informations ont été utilisées pour développer des micros réseaux SNP commerciaux permettant l'analyse parallèle de dizaines (initialement) à des centaines à des milliers de SNP jusqu'à un million, qui sont utilisés dans les études d'association pangénomique (GWAS) pour identifier les SNP associés aux traits. Ces informations ont permis de sélectionner des SNP phénotypiques hautement associés et de tester leur effet prédictif de caractère dans des études de prédiction postérieures au GWAS.
Ces dernières années beaucoup de progrès dans la recherche pour mieux comprendre la base génétique de la couleur des yeux, des cheveux et de la peau via les GWAS, ainsi que la prévisibilité de l'ADN dans les trois traits de pigmentation via des études de prédiction dédiées.
En ce moment, les traits de pigmentation humains sont les EVC (Externally Visible Characteristics) qui sont le plus précisément prévisibles à partir de l'ADN, ce qui permet des applications FDP pratiques (Figure 1). Pour tous les autres EVC, les connaissances génétiques ne sont actuellement pas assez complètes pour considérer la FDP pratique, bien que des progrès dans une meilleure compréhension de la base génétique de plusieurs EVC soient constamment en cours. (Chaitanya, 2016).

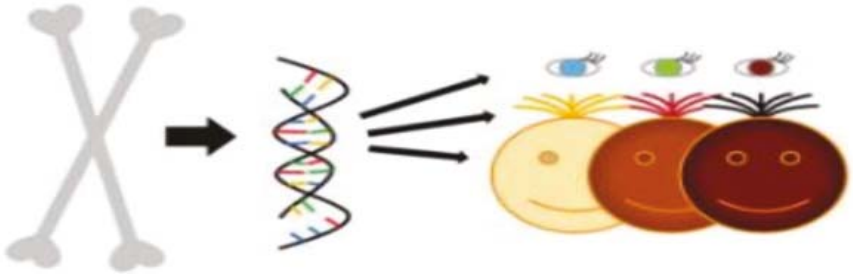


Fig.1 : l'ADN peut être extrait de restes squelettiques et de fluides corporels ; et les informations inhérentes peuvent être utilisées pour prédire la couleur des yeux, des cheveux et de la peau d'un individu. (Chaitanya, 2016).

I.3.1 Couleur des yeux

L'étude de Chaitanya a montré que la couleur de l'œil humain (iris) représente le premier EVC (après le sexe) qui pourrait être prédit à partir de l'ADN avec une précision significative.
La couleur des yeux humains peut être décrite sur la base d'un spectre continu des nuances de bleu les plus claires aux nuances les plus foncées de brun ou de noir avec des couleurs

intermédiaires telles que le vert, le jaune ou le noisette ainsi que des yeux de couleur mixte constitués de zones de couleur différente.

Dans les yeux, les mélanosomes ne se trouvent que dans les mélanocytes de l'iris. Le type et la quantité de mélanosomes et le type de mélanine dans les mélanocytes de l'iris déterminent la couleur des yeux. Un grand nombre de mélanosomes et des quantités plus élevées de mélanine sont des caractéristiques des iris brunes dominants ; en revanche, les iris bleus ont moins de mélanosomes et peu ou pas de mélanine.

Il faut noter que les humains ne portent pas de pigment bleu et les yeux bleus sont à cause de l'effet Tyndall, un phénomène qui est à l'origine du ciel bleu. L'effet Tyndall est la diffusion de la lumière par des particules dans une suspension liquide. La lumière pénétrant dans l'œil est diffusée et réfléchie dans l'espace et à la suite de l'effet Tyndall, une teinte bleue est générée.

Par conséquent, la couleur des yeux bleus dépend de la qualité et de la quantité disponible de la lumière dans l'environnement. Des quantités intermédiaires de mélanine et des proportions variables d'eumélanine par rapport à la phéomélanine sont responsables des autres couleurs des yeux telles que le vert, la noisette ou diverses nuances de couleur.

Des recherches intensives ont été menées pour comprendre le contrôle de la génétique dans la couleur des yeux humains via des études de gènes candidats, analyse de liaison et GWAS, suivis d'études prédictives.

Plusieurs SNP dans différents gènes ont été identifiés qui contribuent à la variation de la couleur des yeux humains. Un seul SNP, rs12913832 dans le gène HERC2 sur le chromosome 15 fournit une grande partie des informations prédictives de la couleur des yeux bleus et bruns Cependant, un SNP n'est pas suffisant pour fournir des informations précises sur les catégories de couleur des yeux, généralement classées comme bleu, marron et intermédiaire (y compris tout ce qui n'est pas classé comme bleu ou marron).

D'autres SNP dans plusieurs autres gènes, tels que SLC24A4, SLC45A2 (MATP), TYRP1, TYR et IRF4, ont été identifiés pour contribuer à la variation de la couleur des yeux. Sur la base de ces découvertes, plusieurs modèles de prédiction ont été proposés pour la prédiction de la couleur des yeux basée sur l'ADN. Cependant, le premier outil de prédiction de la couleur des yeux basé sur l'ADN pour les applications médico-légales était **l'IrisPlex**, très sensible et robuste.

L'Irisplex système, comprenant les six SNP les plus informatifs (voir tableau1) en un seul test de génotypages multiplex. La variation génétique de la couleur des yeux basée sur les génotypes de ces six SNP est décrite dans la Figure 2.

En plus du test de génotypage multiplex, le système IrisPlex implique un modèle de prédiction statistique pour estimer les probabilités de couleur des yeux catégoriques des couleurs marron, bleu et intermédiaires initialement introduites via une feuille Excel interactive et conviviale.

Pour tester davantage la fiabilité du modèle de prédiction, le système IrisPlex a été évalué sur un vaste ensemble de données de plus de 3800 échantillons d'ADN provenant de sept sites dans sept pays à travers l'Europe, dans le cadre de l'étude European Eye (EUREYE).

Par conséquent, il a été prouvé que le système IrisPlex peut prédire avec précision les couleurs des yeux bleus et bruns avec une précision supérieure à 94%, alors que son potentiel à prédire la couleur des yeux non bleus et non bruns est beaucoup plus faible.

Notons qu'à l'heure actuelle, la compréhension génétique de la couleur des yeux et par conséquent la prédiction de la couleur des yeux basée sur l'ADN est principalement limitée

aux catégories de couleur des yeux, mais pas encore au spectre complet de la coloration des yeux humains.
Le système IrisPlex a été validé avec succès conformément aux directives strictes de SWGDAM. (Chaitanya, 2016).

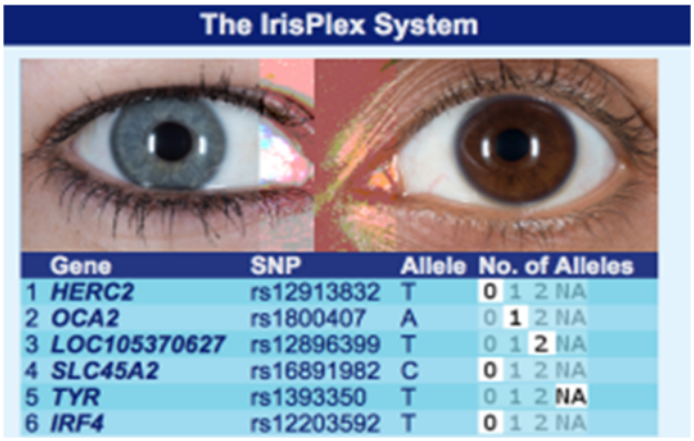


Tableau 01 : représentation des six SNP du système Irisplex.
(https://HIrisPlex.erasmusmc.nl/)

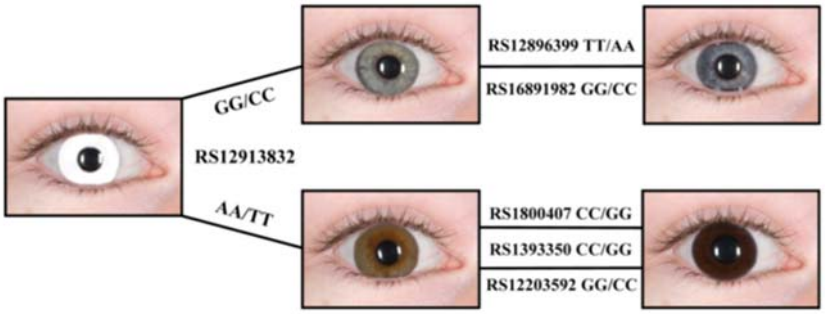


Fig.2:Détermination génétique des couleurs des yeux bruns et bleus montrant l'impact des génotypes SNP les plus influents du modèle 6-SNP. (Chaitanya, 2016)

I.3.2 Couleur des cheveux

La mélanine qui donne naissance à la couleur des cheveux humains est produite dans les mélanocytes folliculaires du bulbe pileux, d'où ils sont transférés aux kératinocytes environnants qui finalement se différencier et migrer en formant les tiges de cheveux pigmentées. Les différentes nuances de couleur des cheveux peuvent être décrites sur un spectre continu allant du rouge, du blond, du brun au noir.
Des niveaux élevés de phéomélanine caractérisent la couleur des cheveux roux, tandis que les rapports les plus élevés d'eumélanine à phéomélanine indiquent une couleur de cheveux noirs. Les couleurs blondes et brun clair sont le résultat de rapports intermédiaires d'eumélanine et de phéomélanine.
Les recherches de Chaitanya résultent que la variation de la couleur des cheveux humains, similaire à la couleur de l'iris, a été géographiquement limitée aux personnes d'origine européenne (au moins partielle) et aux régions environnantes du Moyen-Orient et de certaines parties de l'Asie occidentale.

Des études approfondies sur les gènes associés à la variation de la couleur des cheveux au cours des dernières années ont conclu que le gène MC1R sur le chromosome 16 est associé à la pigmentation des cheveux humains, représentant le premier gène identifié comme étant impliqué dans la coloration des cheveux humains.
Le gène MC1R a une association élevée avec la couleur des cheveux rouges et une association inférieure avec d'autres couleurs de cheveux non rouges.
Des recherches scientifiques récentes ont identifié plusieurs autres gènes (OCA2, HERC2, SLC45A2, SLC24A4, TYRP1, ASIP, TYR et KITLG) associés à des couleurs de cheveux non rouges. Il est actuellement suggéré que si les cheveux roux sont causés par un seul gène (MC1R), toutes les autres couleurs de cheveux sont des traits polygéniques déterminés par divers gènes dont tous ne sont pas connu pour le moment.
Les SNP dans le gène HERC2, comme pour la couleur des yeux, sont le marqueur ADN le plus contribuant à prédire la couleur des cheveux non rouges.
Les SNP dans les différents gènes ont été fortement associés aux différents phénotypes de couleur de cheveux
- SNP HERC2 avec phénotype de couleur de cheveux noir et brun foncé ;
-SNP OCA2 et TYR aux cheveux bruns.
-Il a été constaté que les SNP dans OCA2 sont les plus significativement associés aux cheveux bruns,
-SLC45A2 et IRF4SNP aux cheveux noirs,
-SLC24A4 SNP aux cheveux blonds et blonds foncés et
-ASIP SNP aux couleurs de cheveux roux, blond-rouge et blond foncé.
Sur la base des résultats précédents, et similaire à IrisPlex, un outil ADN a été développé pour prédire quatre catégories de couleur de cheveux (blond, marron, rouge et noir), les nuances de la couleur des cheveux (clair et foncé), ainsi que la couleur des yeux par catégorie.
Le système HIrisPlex comprend 24 SNP de 11 gènes (voir tableau 2),parmi ces 24 variantes, 6 de ces marqueurs IrisPlex, rs12913832, rs1800407, rs12896399, rs16891982, rs1393350 et rs12203592 sont utilisés pour la prédiction de la couleur des yeux, dont tous sauf deux (rs12896399 et rs1393350) sont utilisés pour la prédiction de la couleur des cheveux.
Un outil Excel interactif et convivial basé sur des macros est initialement introduit pour obtenir les probabilités de prédiction des quatre catégories de couleur de cheveux (blond, marron, rouge et noir), les nuances de couleur de cheveux (clair et foncé) et les trois catégories de couleur des yeux (marron, noir et intermédiaire).

La figure 3 illustre les différents SNP contribuant aux différentes catégories de couleur de cheveux. Pour faciliter les enquêtes policières visant à prédire la couleur des yeux et des cheveux des auteurs inconnus et des personnes disparues à partir de l'ADN, le système HIrisPlex doit être mis en laboratoires médico-légaux.
Le système HIrisPlex a été validé avec succès conformément aux directives strictes de SWGDAM. (Chaitanya, 2016).

The HIrisPlex System

	Gene	SNP	Allele	No. of Alleles
1	*MC1R*	rs312262906	A	**0** 1 2 NA
2	*MC1R*	rs11547464	A	0 1 2 **NA**
3	*MC1R*	rs885479	T	**0** 1 2 NA
4	*MC1R*	rs1805008	T	**0** 1 2 NA
5	*MC1R*	rs1805005	T	**0** 1 2 NA
6	*MC1R*	rs1805006	A	**0** 1 2 NA
7	*MC1R*	rs1805007	T	0 **1** 2 NA
8	*TUBB3*	rs1805009	C	**0** 1 2 NA
9	*MC1R*	rs201326893	A	0 **1** 2 NA
10	*MC1R*	rs2228479	A	**0** 1 2 NA
11	*MC1R*	rs1110400	C	**0** 1 2 NA
12	*SLC45A2*	rs28777	C	0 1 **2** NA
13	*SLC45A2*	rs16891982	C	**0** 1 2 NA
14	*KITLG*	rs12821256	G	**0** 1 2 NA
15	*LOC105374875*	rs4959270	A	0 **1** 2 NA
16	*IRF4*	rs12203592	T	**0** 1 2 NA
17	*TYR*	rs1042602	T	**0** 1 2 NA
18	*OCA2*	rs1800407	A	0 1 **2** NA
19	*SLC24A4*	rs2402130	G	0 **1** 2 NA
20	*HERC2*	rs12913832	T	**0** 1 2 NA
21	*PIGU*	rs2378249	C	**0** 1 2 NA
22	*LOC105370627*	rs12896399	T	0 **1** 2 NA
23	*TYR*	rs1393350	T	**0** 1 2 NA
24	*TYRP1*	rs683	G	**0** 1 2 NA

Tableau 02 : représente les 24 SNP de système HIrisPlex (https://HIrisPlex.erasmusmc.nl/)

Fig.3: L'effet de chaque SNP sur le modèle pour la prédiction de la couleur des cheveux dans le système HIrisPlex (Chaitanya, 2016).

I.3.3 Couleur de peau

La couleur de la peau humaine est un trait complexe et la couleur de la peau varie énormément sur une large gamme allant de très pâle à très sombre. La variation de la couleur de la peau est déterminée par la quantité et le type de mélanine produite dans les mélanocytes, ainsi que par la forme et la distribution des mélanocytes dans la couche basale de l'épiderme.

Les couleurs de peau plus foncées ont des niveaux plus élevés de mélanine enrichie en eumélanine et des unités uniques de plus grandes et plus pigmentées mélanosomes.

Une peau plus claire a des niveaux plus élevés d'eumélanine brun clair et de jaune / rouge phéomélanine avec des mélanosomes plus petits et moins pigmentés, conditionnés en groupes. La variation de la couleur de la peau est répartie globalement, contrairement à la variation de la couleur des yeux et des cheveux.

Il est fortement corrélé aux modèles géographiques, expliqués par la sélection naturelle via les adaptations aux changements environnementaux, tels que les niveaux variables de latitude de rayonnement ultraviolet (UV) à la suite de la migration humaine hors l'Afrique vers les régions du nord.

L'étude de Chaitanya a montrée que par rapport à la couleur des yeux et des cheveux, les connaissances sur la génétique sous-jacente à la variation de la couleur de la peau sont actuellement plus limitées. La couleur de la peau varie largement entre les populations et moins au sein des populations, ce qui pose des problèmes pour le GWAS qui ne peuvent être réalisés que dans des populations génétiquement homogènes avec des découvertes faussement positives limitées. Par conséquent, les GWAS sur la couleur de la peau ont été menées chez les Européens ou chez les Asiatiques, mais les deux groupes présentent respectivement moins de variation de couleur de peau que celle observée entre les groupes continentaux.

Par conséquent, nous comprenons actuellement moins la base génétique de la couleur de la peau que la couleur des yeux et des cheveux. Plusieurs SNP associés à la variation de la couleur des yeux et des cheveux semblent également être associés à la variation de la couleur de la peau. Un ensemble de 7-SNP a été décrit pour prédire la couleur de la peau non claire et non foncée. Le SNP rs6119471 (ASIP) était très variable dans les populations à peau foncée, alors que les six autres SNP rs1426654 (SLC24A5), rs12913832 (HERC2), rs16891982 (SLC45A2), rs12203592 (IRF4), rs1545397 (OCA2) et rs885479 (MC1R) ont été observés comme étant très variables dans les populations avec lumière couleur de la peau. Un ensemble complet avec un panel de 59 SNP précédemment associés à la couleur des yeux, des cheveux et de la peau a été publié en 2014. Sur ces 59 SNP, 29 ont été identifiés comme être fortement associés à la variation de la couleur de la peau et ont pu différencier les individus à peau blanche de ceux à peau intermédiaire / noire.

Parmi les 29 SNP, les 10 meilleurs SNP prédictifs de la couleur de la peau (rs10777129, rs1426654, rs16891982, rs13289, rs3829241, rs6058017, rs6119471, rs2402130, rs1408799 et rs1448484) ont été identifiés pour la peau noire, 0,9803 pour la peau noire et 0,9803 pour le blanc.

Semblable à IrisPlex et HIrisPlex, un outil basé sur l'intelligence ADN validé par la médecine légale le système HIrisPlex-S (S-pour Skin), HIrisPlex-S comprend 24 SNP de système HIrisPlex plus les 17 SNP pour la prédiction de la couleur de peau.(Voir tableau03). (Chaitanya, 2016).

The HIrisPlex-S System

	Gene	SNP	Allele	No. of Alleles
1	*MC1R*	rs312262906	A	0 1 2 NA
2	*MC1R*	rs11547464	A	0 1 2 NA
3	*MC1R*	rs885479	T	0 1 2 NA
4	*MC1R*	rs1805008	T	0 1 2 NA
5	*MC1R*	rs1805005	T	0 1 2 NA
6	*MC1R*	rs1805006	A	0 1 2 NA
7	*MC1R*	rs1805007	T	0 1 2 NA
8	*TUBB3*	rs1805009	C	0 1 2 NA
9	*MC1R*	rs201326893	A	0 1 2 NA
10	*MC1R*	rs2228479	A	0 1 2 NA
11	*MC1R*	rs1110400	C	0 1 2 NA
12	*SLC45A2*	rs28777	C	0 1 2 NA
13	*SLC45A2*	rs16891982	C	0 1 2 NA
14	*KITLG*	rs12821256	G	0 1 2 NA
15	*LOC105374875*	rs4959270	A	0 1 2 NA
16	*IRF4*	rs12203592	T	0 1 2 NA
17	*TYR*	rs1042602	T	0 1 2 NA
18	*OCA2*	rs1800407	A	0 1 2 NA
19	*SLC24A4*	rs2402130	G	0 1 2 NA
20	*HERC2*	rs12913832	T	0 1 2 NA
21	*PIGU*	rs2378249	C	0 1 2 NA
22	*LOC105370627*	rs12896399	T	0 1 2 NA
23	*TYR*	rs1393350	T	0 1 2 NA
24	*TYRP1*	rs683	G	0 1 2 NA
25	*ANKRD11*	rs3114908	T	0 1 2 NA
26	*OCA2*	rs1800414	C	0 1 2 NA
27	*BNC2*	rs10756819	G	0 1 2 NA
28	*HERC2*	rs2238289	C	0 1 2 NA
29	*SLC24A4*	rs17128291	C	0 1 2 NA
30	*HERC2*	rs6497292	C	0 1 2 NA
31	*HERC2*	rs1129038	G	0 1 2 NA
32	*HERC2*	rs1667394	C	0 1 2 NA
33	*TYR*	rs1126809	A	0 1 2 NA
34	*OCA2*	rs1470608	A	0 1 2 NA
35	*SLC24A5*	rs1426654	G	0 1 2 NA
36	*ASIP*	rs6119471	C	0 1 2 NA
37	*OCA2*	rs1545397	T	0 1 2 NA
38	*RALY*	rs6059655	T	0 1 2 NA
39	*OCA2*	rs12441727	A	0 1 2 NA
40	*MC1R*	rs3212355	A	0 1 2 NA
41	*DEF8*	rs8051733	C	0 1 2 NA

Tableau 03 : représentation des 41 SNP du système HIrisPlex-S pour la prédiction de la couleur des yeux de cheveux et de la peau d'un individu. (https://HIrisPlex.erasmusmc.nl/)

I.4 Séquençage parallèle massif /Massively Parallel Sequencing (MPS) :

Au cours de la dernière décennie, la technologie de séquençage de nouvelle génération (NGS), alternativement le séquençage parallèle massif (MPS), a été appliqué à tous les domaines de la recherche biologique. Son introduction dans le domaine de la médecine légale a été plus lente, principalement en raison du manque de séquenceurs accrédités, de kits et de taux d'erreur de séquençage relativement plus élevés par rapport au séquençage standardisé de Sanger. Actuellement, la majorité des problèmes problématiques ont été résolus, ce qui est prouvé par l'ensemble des rapports de la littérature.

Les méthodes de séquençage de nouvelle génération (NGS) permettent efficacement de séquencer tous les types d'acides nucléiques, en utilisant un génome entier ou une approche ciblée, avec le séquençage de l'ADN, de l'ARNm et du petit ARN comme analyses standard. En outre, le séquençage à plus grande échelle de sous-types d'ARN spécifiques, tels que les ARN longs non codants et le snoARN, ainsi que l'ADN méthylé, est devenu possible avec l'introduction du NGS. La perspective d'analyser simultanément un grand nombre de marqueurs tels que les STR et les SNP en parallèle avec des analyses ciblées d'ARNm et de petits ARN fait du MPS un outil très puissant, relativement facilement applicable, dans les laboratoires médico-légaux.(Ballard et al, 2020).

I.5 Machine learning (apprentissage automatique)

Il s'agit d'une science moderne permettant de découvrir des patterns et d'effectuer des prédictions à partir de données en se basant sur des statistiques, sur du forage de données, sur la reconnaissance de patterns et sur les analyses prédictives. Le Machine Learning peut être défini comme une branche de l'intelligence artificielle englobant de nombreuses méthodes permettant de créer automatiquement des modèles à partir des données. Ces méthodes sont en fait des algorithmes. Un système Machine Learning ne suit pas d'instructions, mais apprend à partir de l'expérience. Par conséquent, ses performances s'améliorent au fil de son " entraînement " à mesure que l'algorithme est exposé à davantage de données.

Dans le cas de l'apprentissage supervisé, les données utilisées pour l'entraînement sont déjà" étiquetées ". Par conséquent, le modèle de Machine Learning sait déjà ce qu'elle doit chercher (motif, élément…) dans ces données. À la fin de l'apprentissage, le modèle ainsi entraîné sera capable de retrouver les mêmes éléments sur des données non étiquetées. (Https://www.lebigdata.fr/machine-learning-et-big-data)

2. Méthodologie

Pour pouvoir répondre à la problématique posée qui est : est-il possible de créer une image de profil inconnu à partir d'une petite quantité ADN. Nous avons rassemblé un ensemble d'information pour parvenir à la réponse.

Notre étude s'agit de prédire un portait d'un profil inconnu des différents traits de pigment au traits du visage et cela à partir de traces biologique trouvés dans des scènes de crime.

Nous avons en premier contacté une société privé appelé Parabon nanolab qui nous a envoyé quelques articles qu'on a analysé puis nous a permis de s'approfondir plus à propos de ce travail. Nous avons subdivisé ce dernier en 3 études essentielles. La première étude présente un système HIrisPlex (dit HPS) pour la prédiction de 3 traits de pigmentations couleurs de cheveux, des yeux et de la peau.

La deuxième étude est basée sur l'architecture génétique du visage et les différents SNP responsables des variations faciales.

La troisième et dernière étude est celle du résultat final et global. Elle a été fait en fusionnant les résultats des 2 études précédentes et qui est la prédiction d'un profil d'un être connu et inconnu et cela se fais grâce à un ensemble d'algorithme appelé Machine Learning.

- La première étude présente la possibilité de prédire les caractéristiques visibles de l'extérieur offert par le phénotypage ADN médico-légal (FDP) à partir de quantités infimes d'ADN sur les lieux du crime, ce qui peut aider à trouver des auteurs inconnus qui sont généralement non identifiables via le profilage ADN médico-légal conventionnel. La recherche fondamentale en génétique humaine a permis de mieux comprendre les variantes spécifiques de l'ADN responsables des caractéristiques de **l'apparence physique**, en particulier **la couleur des yeux, des cheveux et de la peau**.

Récemment, Walsh et al ont introduit le système **HIrisPlex-S** pour la prédiction simultanée de la couleur des yeux, des cheveux et de la peau sur la base de **41 variantes d'ADN** générées à partir de deux tests multiplex SNaPshot validés de manière médico-légale utilisant l'électrophorèse capillaire (CE).

Ici, nous présentons des solutions de **séquençage massivement parallèle (MPS)** pour le système HIrisPlex-S (HPS) sur deux plates-formes MPS couramment utilisées en médecine légale, **Ion Torrent et MiSeq**, qui couvrent les 41 variantes d'ADN dans un dosage unique, respectivement. De plus, nous présentons la validation médico-légale du développement des deux dosages HPS MPS.

Le test **Ion Torrent MPS**, basé sur la technologie **Ion AmpliSeq**, a illustré la génération réussie de profils génotypiques HIrisPlex-S complets à partir de **100 pg d'ADN** de contrôle d'entrée, tandis que le test **MiSeq MPS** basé sur une conception interne un produit des profils complets à **partir de 250 pg de ADN d'entrée**.

Ensuite on a évalué **des dommages de cas médico-légaux simulés** tels que la salive, les cheveux (bulbe), le sang, le sperme et l'ADN tactile en faible quantité, ainsi que les artificiels des échantillons d'ADN, **des tests de concordance et des échantillons de**

nombreuses espèces (espèce animal, des échantillons non humain), ont tous illustré la capacité des deux versions du test HIrisPlex-S MPS à produire des résultats qui motivent les applications médico-légales.

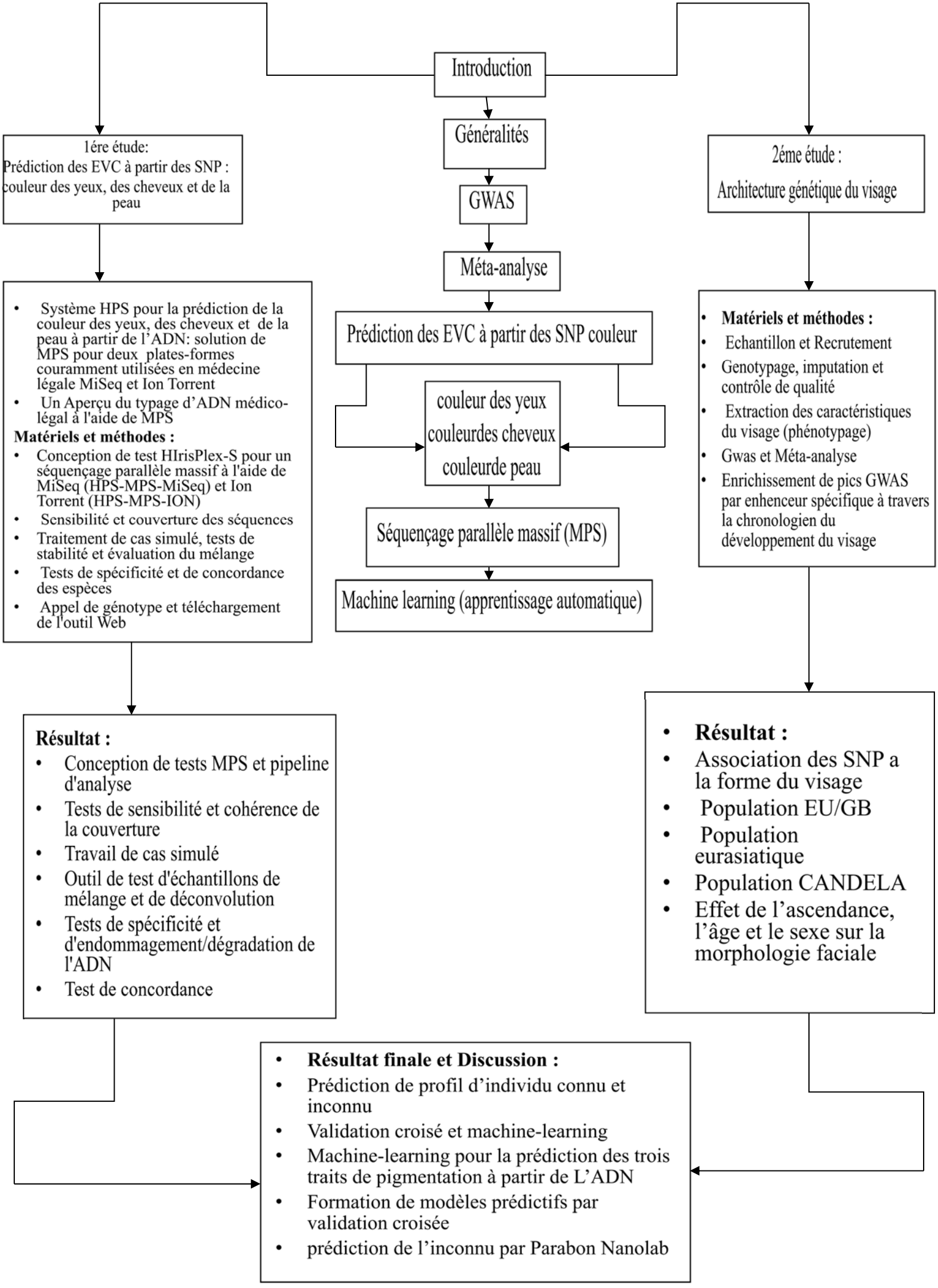
En fournissant également un **pipeline d'analyse bioinformatique** intégré, les données MPS peuvent désormais être analysées et un fichier généré pour être téléchargé sur l'outil Web en ligne HIrisPlex accessible au public **(https://hirisplex.erasmusmc.nl).**

En outre, nous présentons un outil **de séparation de mélanges** à 2 personnes qui évalue non seulement la fiabilité du génotype en ce qui concerne la confiance du génotypage, mais fournit également le scénario de mélange le plus approprié pour les contributeurs mineurs et majeurs, y compris la séparation des profils. Nous envisageons cette mise en œuvre MPS du système HIrisPlex-S pour la prédiction de la couleur des yeux, des cheveux et de la peau à partir de l'ADN comme point de départ pour étendre davantage l'ADN médico-légal basé sur MPS

- Dans cette seconde étude, le but est de parvenir a lier un SNP a un trait facial, un travail qui est très compliqué, c'est pour cela que les auteurs de cette recherche ont pu créer un système de corrélation. Dans ce travail nous avons combiné 3 différentes populations et cela pour la diversité des traits faciaux de populations Eurasiatiques, population candela et population Américaine et anglaise (combinées). L'ADN pris des participants est génotypé puis imputé et finalement filtré. Le phénotype est ensuite extrait à partir de leurs visages en utilisant des appareils de photographie spécialisés. Par la suite, une cartographie de maillage est posée sur les images pour créer des repères facilitant l'étude de la morphologie du visage. Ces derniers sont analysées par plusieurs systèmes et logiciels pour obtenir des scores appelés valeur P. Ces valeurs sont corrélées à un SNP et dit hautement significatif à l'échelle du génome si la valeur P obtenue est supérieur au seuil de 10^{-8}. A la fin nous avons obtenu un ensemble de SNP qui sont fortement lié à un trait facial et qui diffère d'une population a une autre.

- La troisième et dernière étude est le résultat global. Après avoir classé les différents SNP pour les 3 traits de pigments et les traits du visage, les données sont intégrées dans des algorithmes du « machine Learning » pour pouvoir lier chaque SNP a son propre caractère physique. Un ensemble de test mis en œuvre pour évaluer le rendement du processus et juger sa fiabilité. A fur et à mesure que le système acquis de l'expérience et obtient des données génomiques provenant de différente population issue du GWAS et des multiples méta-analyses, la prédiction de profil inconnu devient de plus en plus facile à gérer. Cette technique a permis à plusieurs enquêteurs de trouvé des pistes pour leurs enquêtes de crime et de disparition.

Un organigramme global récapitulatif de notre étude (voir ci-dessous) :

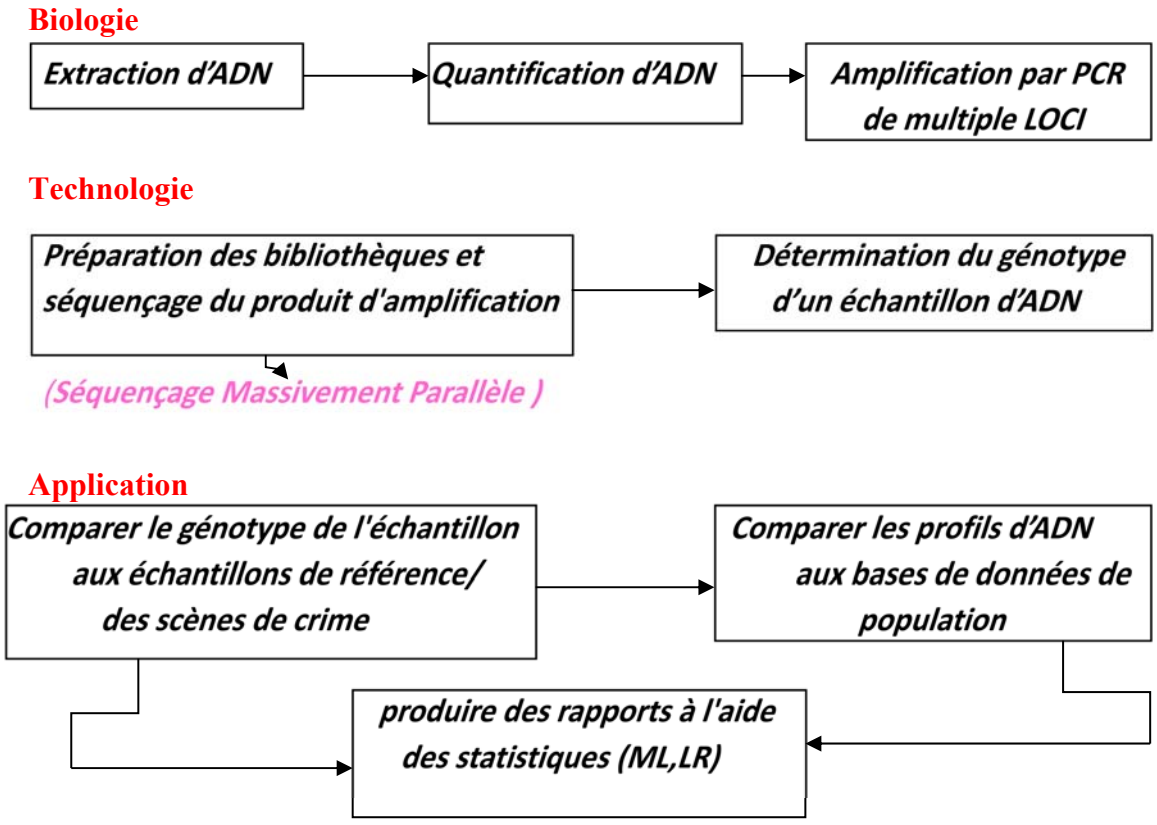

Introduction
Généralités
GWAS
Méta-analyse
Prédiction des EVC à partir des SNP couleur
couleur des yeux
couleurdes cheveux
couleurde peau
Séquençage parallèle massif (MPS)
Machine learning (apprentissage automatique)
1ére étude:
Prédiction des EVC à partir des SNP : couleur des yeux, des cheveux et de la peau
Système HPS pour la prédiction de la couleur des yeux, des cheveux et de la peau à partir de l'ADN: solution de MPS pour deux plates-formes couramment utilisées en médecine légale MiSeq et Ion Torrent
Un Aperçu du typage d'ADN médico-légal à l'aide de MPS
Matériels et méthodes :
Conception de test HIrisPlex-S pour un séquençage parallèle massif à l'aide de MiSeq (HPS-MPS-MiSeq) et Ion Torrent (HPS-MPS-ION)
Sensibilité et couverture des séquences
Traitement de cas simulé, tests de stabilité et évaluation du mélange
Tests de spécificité et de concordance des espèces
Appel de génotype et téléchargement de l'outil Web
Résultat :
Conception de tests MPS et pipeline d'analyse
Tests de sensibilité et cohérence de la couverture
Travail de cas simulé
Outil de test d'échantillons de mélange et de déconvolution
Tests de spécificité et d'endommagement/dégradation de l'ADN
Test de concordance
2éme étude :
Architecture génétique du visage
Matériels et méthodes :
Echantillon et Recrutement
Genotypage, imputation et contrôle de qualité
Extraction des caractéristiques du visage (phénotypage)
Gwas et Méta-analyse
Enrichissement de pics GWAS par enhenceur spécifique à travers la chronologien du développement du visage
Résultat :
Association des SNP a la forme du visage
Population EU/GB
Population eurasiatique
Population CANDELA
Effet de l'ascendance, l'âge et le sexe sur la morphologie faciale
Résultat finale et Discussion :
Prédiction de profil d'individu connu et inconnu
Validation croisé et machine-learning
Machine-learning pour la prédiction des trois traits de pigmentation à partir de L'ADN
Formation de modèles prédictifs par validation croisée
prédiction de l'inconnu par Parabon Nanolab

3. Première étude : Prédiction des caractères visibles de l'extérieur à partir des SNP couleur des yeux, des cheveux et de la peau

3.1 Système HIrisPlex-S pour la prédiction de la couleur des yeux, des cheveux et de la peau à partir de l'ADN: solution de séquençages massivement parallèles pour deux plates-formes couramment utilisées en médecine légale

Le système FDP se compose de deux tests multiplex basés surSNaPshot ciblant un totale de **41 SNP** via un nouveau test multiplex pour 17 SNP prédictifs de la couleur de la peau est le systeme **HIrisPlex-S**. (l'ancienne étude voir annexe3).
Le premier systéme de test d'ADN est IrisPlex pour la prédiction de la couleur des yeux et le système HIrisPlex pour prédiction combinée de la couleur des yeux et des cheveux à partir des traces d'ADN (voir annexe 1et 2).
Cette étude a été mise à jour pour introduire des solutions de séquençage massivement parallèle ou (MPS) en anglais, pour le système HIrisPlex-S (HPS) sur deux plates-formes MPS couramment utilisées en criminalistique, Ion Torrent et Illumina MiSeq qui couvrent les 41 variantes d'ADN en un seul test.

3.1.1Un Aperçu du typage d'ADN médico-légal à l'aide de MPS:



-Regardant l'image globale du typage médico-légal de l'ADN, il ya trois composants principaux :
Biologie, technologie et application
-Dans la biologie nous avons cette preuve de scène de crime qui entre dans le laboratoire, on effectue une extraction d'ADN afin d'obtenir l'ADN à partir de la preuve biologique à partir de là, la quantification de l'ADN aura lieu, elle est utilisée pour générer ou estimer la quantité totale d'ADN dans notre échantillon, puis amplification par PCR.

Nous sachions que nous recherchons des régions particulières dans l'ADN on utilise des amorces afin de nous concentrer sur ces régions.
-utilisant la technologie, on utilise l'électrophorèse capillaire, qui prend le produit PCR et le sépare
Par taille à travers un capillaire afin de préparer des bibliothèques et les séquencé par l'MPS.
Après séquençage le résultat est notre profil d'ADN d'intérêt qui va être comparéà d'autres profils
D'ADN, soit dans la base de données soit dans le dossier de cas.
Au final un rapport à l'aide des statistiques (regression logistique) est généré.

3.2Matériels et méthodes :

Matériels :

-SNaPshot : test multiplex utilisé pour analyser l'identité et les SNP associé à la FDP.

-Illumina MiSeq : Le MiSeq est un instrument intégré qui effectue l'amplification clonale, le séquençage de l'ADN génomique et l'analyse des données avec appel de base, alignement, appel de variante et rapport en une seule analyse. L'instrument de paillasse MiSeq utilise une Flow Cell double face et une seule voie et une cartouche de réactifs fournis sous forme de kit. Le séquençage est effectué en enregistrant la synthèse de brins d'ADN dans des grappes de modèles d'échantillons attachés à la Flow Cell. Chaque base nouvellement attachée libère un colorant fluorescent qui est excité par des diodes laser (530 et 660 nm) et imagé à l'aide de deux caméras numériques. L'interrogation séquentielle des bases permet un réglage flexible de la longueur de lecture au cours d'une analyse. Jusqu'à 96 échantillons peuvent être séquencés en une seule analyse avec des bibliothèques d'ADN préparées avec des adaptateurs indexés ou à code-barres.

-Ion Torrent : est un séquenceur à semi-conducteur qui mesure les changements de pH, conséquence de la libération d'hydrogène. Ions lors de la synthèse de l'ADN.

-Applications, différents kits et logiciels utilisées par les deux plateformes (Voir annexe 4)

-Différents formats bio-informatique pour l'appel de génotype et téléchargement de l'outil Web(Voir annexe 4)

Les échantillons d'étude ont été collectés conformément à l'Université de l'Indiana

Pour la prédiction de la couleur des yeux, la base de données du modèle comprend des échantillons sur 9 188 individus de huit régions d'Europe (Pays-Bas, Norvège, Estonie, Royaume-Uni, France, l'Italie, l'Espagne et la Grèce) (Liu et al, 2009 ; Walsh et al ,2012), ainsi que 278 personnes supplémentaires d'une collection utilisée pour la prédiction de la couleur de la peau.(Walsh et al ,2017).

Nombre total d'individus pour le modèle catégorique de couleur des yeux = 9466

Pour la prédiction de la couleur des cheveux, la base de données du modèle comprend 1601 individus, d'Irlande, de Grèce et de Pologne et 50 Individus d'un nouvel ensemble japonais pour lequel un phénotype de cheveux noirs est enregistré(Walsh et al, 2013), ainsi que 277 individus supplémentaires d'une collection basée aux États-Unis utilisée pour la prédiction de la couleur de la peau.(Walsh et al ,2017).

Nombre total d'individus pour le modèle de couleur de cheveux catégorique = 1878

Pour la prédiction de la couleur de la peau, la base de données du modèle comprend 1423 individus, d'Irlande, de Grèce et de Pologne, basée aux États-Unis (y compris les lieux de naissance des parents au Nigeria, Mexique, Colombie, Inde, Bangladesh, Palestine, Canada, Chine, Honduras, Allemagne, Philippines, Russie, Soudan, Japon, Arabie saoudite, Pakistan, El Salvador, Espagne, Haïti, Corée du Sud, Vietnam) ainsi que des personnes du CEPH HGDP (Centre d'Etude du Polymorphisme Humain /Human Genome Diversity Project) du Sénégal, du Nigéria, du Kenya et de Papouasie-Nouvelle-Guinée).(Walsh et al ,2017).

Nombre total d'individus pour le modèle catégorique de couleur de peau = 1423

Méthodes :

3.2.1Conception de test HIrisPlex-S pour un séquençage parallèle massif à l'aide de MiSeq (HPS-MPS-MiSeq) et Ion Torrent (HPS-MPS-ION) :

Un protocole MPS personnalisé a été utilisé pour développer les deux tests l'Illumina MiSeq etIon Torrent (voir les deux protocoles détaillésen annexe 5).

Les échantillons de test composés d'échantillons de source unique et multiple, y compris des cas simulés (salive, sang, sperme, cheveux (y compris le bulbe), écouvillons vaginaux et objets touchés) et non humains.

Les phénotypes ont été enregistrés à l'aide des données du questionnaire et vérifiés à l'aide d'images.

L'ADN a été extrait en utilisant un protocole de relargage interne

Les concentrations d'ADN des échantillons ont été déterminées à la fois par qPCR via InnoQuant Kit d'évaluation de la quantification et de la dégradation de l'ADN humain et par le kit Quantifiler Trio DNA Quantification (ThermoFisherScientific, Waltham, MA, USA). Et plaquées sur le site IUPUI US (Indiana University – Purdue University Indianapolis).

Ces échantillons ont été utilisés pour générer des données et tester les performances de deux tests MPS, l'un conçu en interne pour la plate-forme Illumina MiSeq par le site américain IUPUI, et un test Ion AmpliSeq distinct conçu par Erasmus MC Rotterdam (Centre médical universitaire Erasmus) en collaboration avec l'Université Jagiellonion et ThermoFisher Scientific (TFS) pour une utilisation sur la plate-forme Ion Torrent.

De plus, tous les génotypes générés par séquençage d'amplicon ont également été confirmés par électrophorèse capillaire génotypage par extension de base unique (CE-SBE) à l'aide de tests HIrisPlex et HIrisPlex-S SNaPshot.

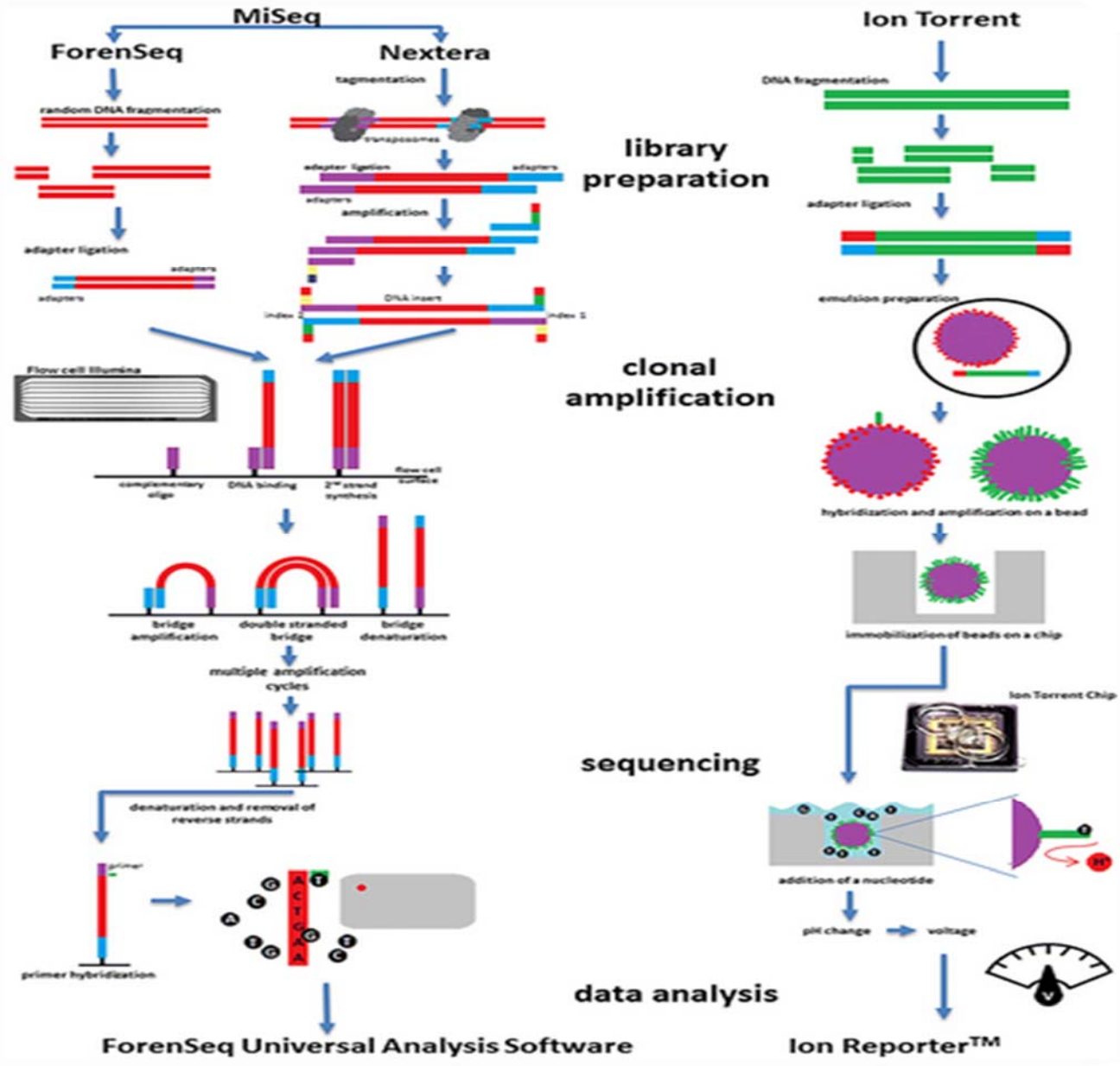


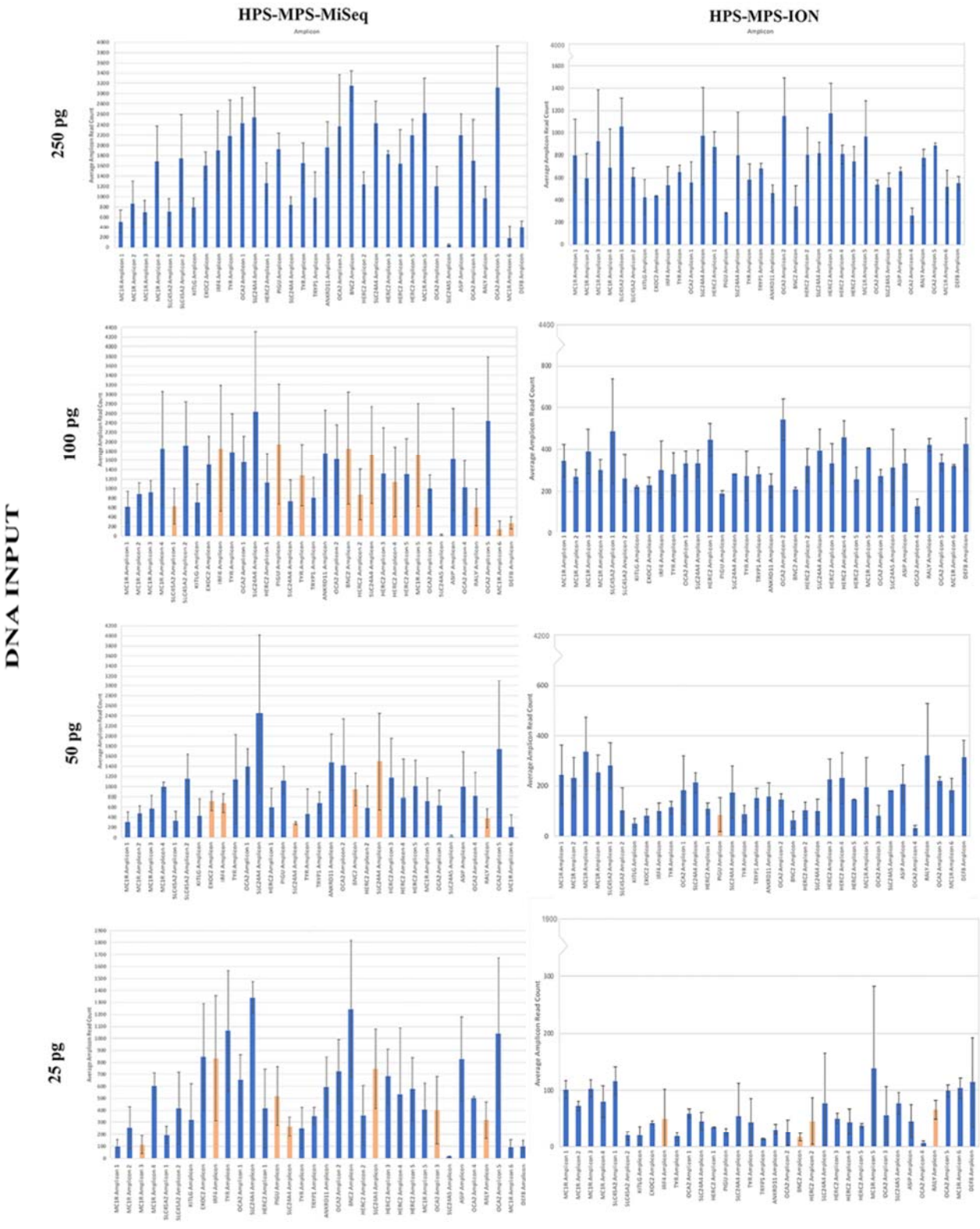
Fig 4 : Illustration schématique du séquençage avec **MiSeq** et **Ion Torrent** (Ballard et al, 2020).

Un MiSeq : Forenseq est une préparation de bibliothèque utilisée pour le séquençage STR et SNP (STR autosomiques, sexe, ascendance géographique et SNP phénotypiques) ; alternativement, Nextera est utilisé pour le séquençage de l'ADNmt.
Dans le processus de préparation des échantillons, des adaptateurs sont ajoutés aux fragments d'ADN dans une réaction PCR en deux étapes afin de permettre la liaison de l'ADN à une lame de verre.
Dans l'étape suivante, les fragments sont amplifiés par clonage sur la lame et séquencés. Le brin matrice est étendu avec un nucléotide à la fois. La réaction de polymérisation est stoppée grâce à l'utilisation de 3'- *O* -azidométhyl-dNTP marqués par fluorescence. L'incorporation de la base est suivie de l'élimination des bases non incorporées et de l'imagerie à l'aide d'une caméra CCD. Par la suite, le bloc 3' et l'étiquette fluorescente sur le nucléotide incorporé sont retirés et la réaction passe au cycle suivant.

Ion Torrent : la préparation des échantillons de fragments d'ADN pour le séquençage sur Ion Torrent est complètement différente, suivi par l'amplification de l'ADN ligaturé par adaptateur hybridé à des billes par PCR en émulsion. Les billes sont distribuées dans des micropuits, où se produit le séquençage par synthèse. Le capteur situé au fond du puits convertit les changements de pH en un signal de tension proportionnel au nombre de bases incorporées.

3.2.2 Sensibilité et couverture des séquences

La sensibilité des deux dosages MPS a été évaluée pour déterminer l'entrée minimale nécessaire pour obtenir un profil HPS 41-SNP complet. Deux échantillons d'ADN de contrôle commercial, 9947A (femelle) (OriGene, Rockville, MD, USA) et 9948(mâle) (OriGene, Rockville, MD, USA), ont été utilisés pour préparer des dilutions en série à des concentrations de 5 pg, 10 pg, 25 pg, 50 pg, 100 pg, 250 pg, 500 pg et 1 ng. Ces échantillons de contrôle de haute qualité ont été utilisés pour évaluer la précision et la couverture de séquençage de chaque amplicons HPS à des concentrations différentes pour chaque conception de test et ont été utilisés pour définir des seuils pour l'appel du génotype utilisé dans l'outil de seuil et de mélange.Pour les appels HPS-MPSMiSeq, chaque concentration a été réalisée en double pour les deux contrôles, ces valeurs de seuil ont été calculées à partir de deux échantillons de contrôle exécutés en double à 100 pg et 50 pg pour un total de 4 échantillons à chaque concentration.Pour HPS-MPS-ION, ces valeurs de seuils ont été calculées à partir de deux échantillons témoins analysés à 100 pg et 50 pg pour un total de 2 échantillons à chaque concentration. Le pourcentage d'erreur de séquençage des contrôles a été calculé comme le nombre d'appels incorrects à ce site variant dans l'amplicon.(barres orangedans la figure05)En outre, une évaluation des appels génotypiques (homozygote et hétérozygote) et la couverture de chaque site de variante HPS avec une entrée d'ADN de 500 pg provenant de plusieurs individus (n = 8), générée par le pipeline HPS-MPS, a également été évaluée, y compris un minimum et le nombre maximum de lectures.



Control	9948	9947A	2800M
rs796296176	CC	CC	CC
rs11547464	GG	GG	GG
rs885479	GG	GG	GG
rs1805008	CC	CC	CC
rs1805005	GG	GG	GG
rs1805006	CC	CC	CC
rs1805007	CC	CC	CC
rs1805009	GG	GG	GG
rs201326893	CC	CC	CC
rs2228479	GG	GG	GG
rs1110400	TT	TT	TT
rs28777	AA	AA	AA
rs16891982	GG	GG	GG
rs12821256	CT	TT	TC
rs4959270	CA	CA	CA
rs12203592	CT	CT	CC
rs1042602	CC	CC	CA
rs1800407	CC	CC	CC
rs2402130	AA	AA	AA
rs12913832	GG	GG	AG
rs2378249	AA	GA	AA
rs12896399	GT	TT	GG
rs1393350	GG	AA	GG
rs683	AA	AA	CA
rs3114908	CC	CC	CT
rs1800414	TT	TT	TT
rs10756819	GA	AA	AA
rs2238289	AA	AA	AA
rs17128291	AG	AG	AA
rs6497292	AA	AA	AA
rs1129038	TT	TT	CT
rs1667394	TT	TT	CT
rs1126809	GG	AA	GG
rs1470608	GT	GG	GG
rs1426654	AA	AA	AA
rs6119471	CC	CC	CC
rs1545397	AA	AA	AA
rs6059655	GG	AG	GG
rs12441727	GA	GG	GG
rs3212355	CC	CC	CC
rs8051733	AA	AG	AG

Fig.5: Test de sensibilité des dosages HPS-MPS-MiSeq et HPS-MPS-ION en utilisant les échantillons d'ADN de contrôle 9947A et 9948 présentés pour les 34 amplicons utilisés pour séquencer les 41 variantes d'ADN HIrisPlex-S. Les concentrations d'ADN de 250 pg, 100 pg, 50 pg et 25 pg sont indiquées sur la base d'une évaluation à partir d'échantillons en double. Barres bleues indiquent des appels corrects dans tous les échantillons analysés à cette concentration d'ADN, les barres orange indiquent qu'un appel incorrect a été effectué dans un échantillon à cette concentration. Le génotype d'un contrôle Illumina 2800M a été fourni dans cette figure 05 aux côtés des génotypes

de 9947A et 9948 pour que les utilisateurs puissent les comparer lorsqu'ils exécutent leur propre échantillon de contrôle, mais l'échantillon 2800M n'a pas été utilisé dans cette évaluation de sensibilité. Le graphique Box et Whisker représente la moyenne, y compris la distribution du nombre de lectures minimum et maximum pour cette variante, respectivement.(Breslin et al, 2019).

3.2.3Traitement de cas simulé, tests de stabilité et évaluation du mélange

Pour les échantillons de cas simulés, des échantillons ont été fabriqués avec du sang séché et endommagé par les UV, de la salive séchée et endommagée par les UV, salive humide, l'ADN tactile, les cheveux, l'écouvillon vaginal et le mélange d'écouvillon vaginal avec sperme.

Ces échantillons ont été extraits avec la méthode de relargage et quantifiés à l'aide du kit Quantifiler Trio DNA Quantification (TFS) pour évaluer la quantité et la qualité des échantillons avant la préparation de la bibliothèque.

L'ADN d'un individu mesuré à 500 pg d'ADN a ensuite été exposé à la lumière UV pour des intervalles de temps de 0, 5, 10 et 20 min en utilisant le CL- 1000 Ultraviolet Crosslinker (Ultra-Violet Products Ltd, Upland, CA, USA) à une force de 50 J / cm2 afin de tester la robustesse de chaque test pour analyser l'ADN endommagé.

Des mélanges d'ADN pour deux personnes ont été simulés et testés dans des rapports de 1:1, 1:2, 1:5 et 1:10 en double.

Pour s'assurer qu'un mélange de variantes d'ADN était présent dans l'échantillon, deux ensembles de mélanges pour 2 personnes (nombre d'individus = 4) ont été mis en place pour contribuer aux mélanges d'échantillons qui avaient des couleurs différentes des yeux, des cheveux et de la peau.

L'outil de déconvolution de mélange pour 2 personnes a été conçu en utilisant un calcul de ratio pour 2 personnes (ratio mineur : majeur sur 1, par exemple un ratio 1: 1 est ½ et a été saisi comme 0,5) en plus Le nombre de lectures d'allèles hétérozygotes varie comme observé à partir des échantillons de couverture de variantes de 500 pg de 8 individus (réalisés en double).(Pour les résultats voir annexe 7).

3.2.4Tests de spécificité et de concordance des espèces

Chaque essai a été testé pour la spécificité humaine contre le chat, le chimpanzé, le chien,ADN de souris et de porc à 1 ng d'entrée. Des échantillons non humains ont été extraits par la méthode d'extraction interne (à l'exception de l'échantillon de chimpanzé obtenu du Dr Brenda Bradley, Université George Washington).

Pour terminer les tests de concordance de génotypage, une plaque de concordance (échantillon n = 16 sous-ensemble) a été générée à partir de l'ensemble de 96 échantillons utilisé par les laboratoires IUPUI US (plateforme Illumina MiSeq) et Erasmus MC Rotterdam (plateforme Ion Torrent) et envoyée à cinq laboratoires européens externes comme collaborateurs.

On a demandé aux utilisateurs d'indiquer si l'échantillon était un mélange ou une source unique. Si une seule source était indiquée, les utilisateurs étaient également invités à fournir un profil prévu final.

3.2.5 Appel de génotype et téléchargement de l'outil Web

Par souci de cohérence, un pipeline a été conçu pour que les deux plates-formes soient évalués à l'aide des mêmes algorithmes pour générer les 41 appels de génotypes nécessaires à l'entrée du modèle de prédiction dans l'outil de prédiction HIrisPlex-S basé sur le Webdisponible à l'adresse https://hirisplex.erasmusmc.nl/hps/hps.(Voir annexe 6).

3.3Résultats :

3.3.1Conception de tests MPS et pipeline d'analyse :

Les deux versions de test HPS-MPS appelées : HPS-MPS-MiSeq et HPSMPS-ION ciblaient chacune 41 variantes d'ADN dans 34 amplicons.
Ces amplicons ont été conçus pour être aussi courts que possible: les tailles de fragments amplifiés variaient entre 130 et 225 pb (longueur moyenne 124 pb) dans HPS-MPS-MiSeq; les tailles des plaquettes variaient entre 44 et 113 pb (taille moyenne des plaquettes 71 pb) dans HPS-MPS-ION.
Le pipeline d'analyse HPS-MPS a été conçu pour être convivial et semi-automatisé pour faciliter l'ensemble du processus, de l'échantillon d'ADN aux probabilités de prédiction de la couleur des yeux, des cheveux et de la peau des donneurs d'échantillons, estimées via l'outil Web HIrisPlex (https://HIrisPlex.erasmusmc.nl/).

3.3.2Tests de sensibilité et cohérence de la couverture

Analyse des concentrations d'entrée d'ADN de 5 pg, 10 pg, 25 pg, 50 pg, 100 pg, 250 pg, 500 pg et 1 ng ont produit des profils HPS complets jusqu'à 100 pg dans HPS-MPS-MiSeq (Fig. 5), à l'exception d'un seul résultat (9947A) avec un abandon significatif pour 12 variantes.
À 50 pg d'entrée, une chute a été observée à moins de 7 loci pour le contrôle 9947A. Sur la base de ces résultats, le seuil de sensibilité du test HPSMPS-MiSeq a été fixé à 250 pg.

Les mêmes échantillons d'ADN dans les mêmes dilutions (mais pas en double) ont été testés avec le test HPS-MPS-ION. Comme le montre la figure05, des profils HPS complets ont été observés à 100 pg d'ADN d'entrée dans tous les échantillons testés.
L'abandon a commencé à se produire à une entrée d'ADN de 50 pg, ce qui a affecté un amplicon avec un variant d'ADN HPS (rs683 dans l'amplicon TYRP1).
À 25 pg d'ADN d'entrée, davantage de pertes se sont produites à rs12203592 (amplicon IRF4)
Et rs2238289 (amplicon HERC2 2) pour l'échantillon 9948,et rs2238289 (amplicon HERC2 2) et rs6059655 (amplicon RALY) pour l'échantillon 9947A.

L'erreur de séquençage par variante HPS SNP par test et plate-forme est basé sur les deux échantillons d'ADN de contrôle 9948 et 9947A à des concentrations allant de 250 pg à 25 pg d'entrée d'ADN. À titre d'exemple, environ. Une erreur de 50 % indique qu'au moins un échantillon a subi un abandon complet pour cette variante d'ADN, plus proche de 100 % indique un abandon pour tous les échantillons sur ce site, et enfin environ. 25 % indiqueraient qu'au moins un allèle de cette variante a abandonné pour cet échantillon. Outre l'abandon de certains allèles à des niveaux d'entrée d'ADN inférieurs au seuil de

sensibilité identifié, le pourcentage d'erreur était globalement comparable entre les tests et les plates-formes. Cependant, le test HPS-MPSION présentait une erreur de séquençage par variant d'ADN inférieure à celle du test HPSMPS-MiSeq, par ex. 0,07 % HPS-MPS-ION contre 0,32 % HPS-MPSMiSeq à 250 pg d'ADN d'entrée après application.

Dans l'ensemble, comme le montre la Fig05, le HPS-MPS-ION a atteint plus uniformément couverture de séquençage distribuée à travers les amplicons et les concentrations d'entrée d'ADN par rapport aux analyses HPS-MPS-MiSeq.

Cependant, le test HPS-MPS-MiSeq a affiché des couvertures de lecture considérablement plus élevées (jusqu'à 3 fois les lectures sur certains amplicons) que le test HPS-MPS-ION, où certains amplicons avaient moins de 100 lectures, même à 250 pg d'entrée d'ADN.
La figure 5 comprend également les profils de génotype générés de manière cohérente par les deux dosages HPS-MPS (ainsi que les dosages HPS SBE-CE) des échantillons d'ADN témoins 9947A, 9948 et 2800M (Promega, Madison, WI). 2800M n'a pas été évalué dans cette étude de sensibilité.

L'une des raisons des différences de performances observées avec les deux tests HPS-MPS peut être due au nombre inégal d'échantillons d'ADN inclus dans les séquences de séquençage singulières respectives. Pour ce test de validation, 96 échantillons ont été séquencés à partir d'une cartouche pour HPS-MPS-MiSeq, tandis que pour HPS-MPS-ION, ils ont été séquencés avec deux puces exécutant chacune jusqu'à 48 échantillons en parallèle.
Réduire le nombre d'échantillons dans l'analyse MiSeq peut augmenter la sensibilité et donc la probabilité de récupérer un profil génotypique complet à une quantité d'ADN d'entrée inférieure à celle indiquée par le seuil de sensibilité obtenu ici.

Pour évaluer la cohérence de la couverture des comptes de lecture pour les homozygotes et allèles hétérozygotes pour les deux tests HPS-MPS, plusieurs individus présélectionnés (N = 8) avec des profils de phénotype et de génotype variables, ont été analysés en double pour un apport total d'ADN de 500 pg par échantillon. (Figure 6).
Les lectures en moyennes par allèle ont été évaluées pour les appels de génotypes homozygotes et hétérozygotes à l'aide du pipeline d'analyse HPS-MPS.
Les génotypes hétérozygotes par variant par rapport à HPS-MPS-ION (HPS-MPS-MiSeq : en 1039 lectures moyenne homozygotes, 570 lectures moyenne d’appels hétérozygotes,
HPS-MPS-ION : (750 lectures moyenne homozygotes, 444 recomptes moyenne d’appels hétérozygotes par variant).
Cependant, HPS-MPS-ION a montré un profil plus équilibré avec des comptages de lecture répartis plus uniformément à travers les différents amplicons. Notamment, pour HPS-MPS-MiSeq, DNA les variantes rs1426654 (amplicon SLC24A5) et rs1545397 (amplicon OCA2) ont affiché un nombre de lectures beaucoup plus faible par rapport aux autres variantes d'ADN avec moins de 100 lectures en moyenne à 500 pg d'entrée d'ADN.
La plage de lectures à 500 pg d'ADN d'entrée pour HPS-MPS-MiSeq était de 14 à 4490 lectures homozygotes et de 2 à 1771 lectures hétérozygotes utilisant des génotypes sur un total de 16 profils (8 échantillons en double).Pour le HPS-MPS-ION, il s'agissait de 199-1590 lectures homozygotes et de 176-1208 lectures hétérozygotes.

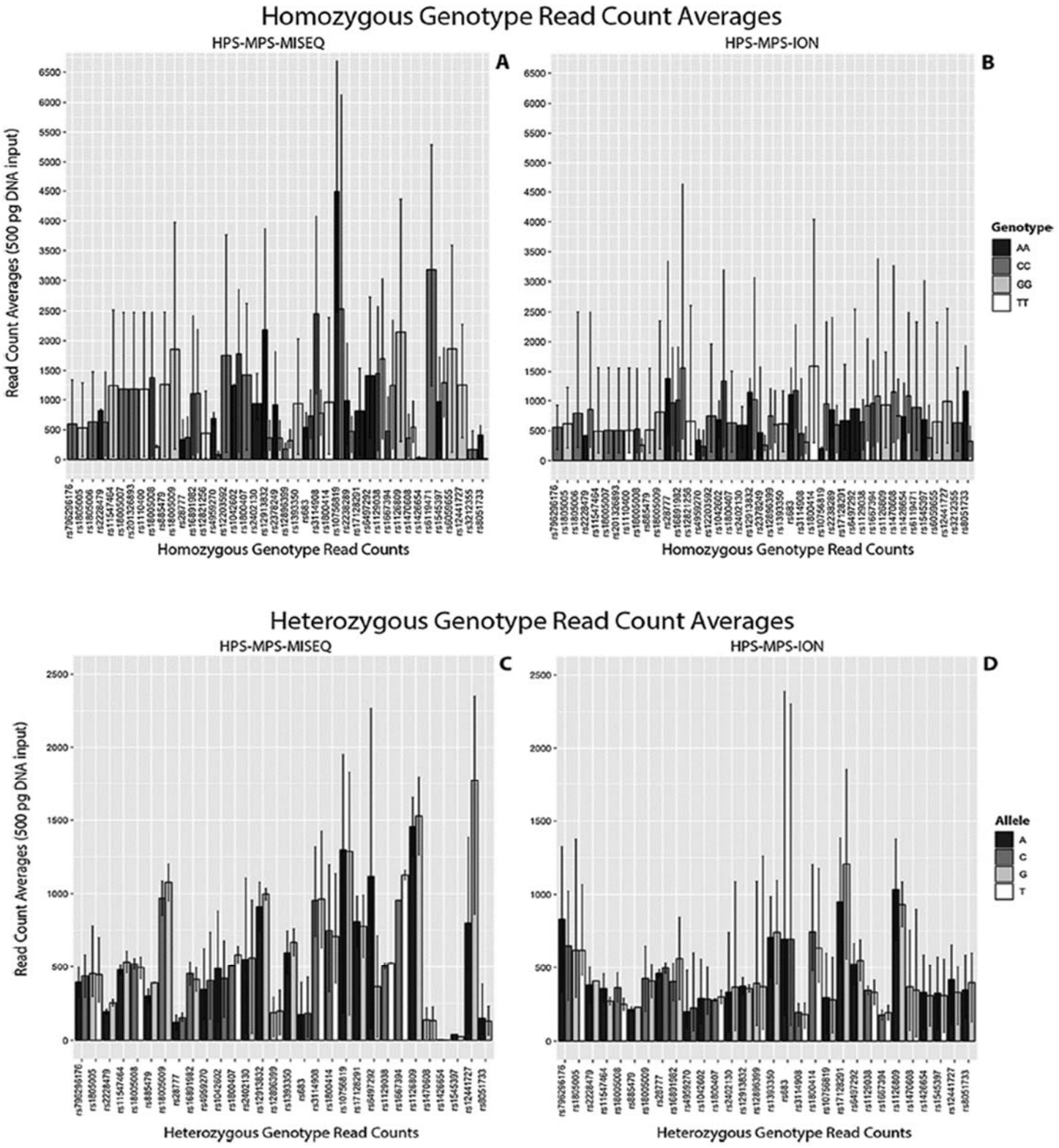


Fig.6 : Nombres de lectures moyens d'homozygotes et d'hétérozygotes, y compris les nombres de lectures de plage minimum et maximum par génotype (homozygote) et allèle (Hétérozygote) pour l'analyse HIrisPlex-S MPS avec les tests HPS-MPS-MiSeq et HPS-MPS ION. Les analyses sont basées sur 8 individus (effectués en double, donc n = 16 échantillons ont été utilisés pour cette évaluation) à 500 pg d'ADN d'entrée pour les 41 variants d'ADN HIrisPlex-S observés. Il convient de noter que tous les allèles (sous forme hétérozygote ou homozygote) ne peuvent pas être pris en compte dans cette évaluation

moyenne du nombre de lectures en raison de l'absence de l'allèle chez ces 8 individus. À des fins de visualisation, un seul échantillon peu performant a été supprimé de l'analyse HPS-MPS-MiSeq en raisond'une
amplification infructueuse, par conséquent 15 échantillons ont été utilisés pour l'évaluation HPS-MPS-MiSeq.(Breslin et al,2019).

3.3.3 Travail de cas simulé

Neuf échantillons de cas simulés (mock) provenant de six individus différents ont été réalisés en double à partir de sources d'ADN de sang, de sperme, de salive, de cheveux et de toucher et analysés avec les tests HPS-MPS-MiSeq et HPSMPS-ION.

Les deux tests ont bien fonctionné, analysant des échantillons avec des entrées d'ADN supérieures à 100 pg ; en accord avec les résultats des tests de sensibilité.

Les échantillons de cas fictifs provenant de salive, l'ADN de salive séchée et endommagée, de sperme, d'écouvillons vaginaux, de cheveux et d'échantillons de sang séché, avec une entrée d'ADN allant de 121 à 6890 pg, ont généré des profils 41-SNP HPS complets et corrects (par rapport aux profils de référence généré par le typage CE) avec les deux tests MPS.

Des résultats incorrects dus à l'abandon/à l'entrée d'allèles n'ont été observés que pour les deux échantillons d'ADN tactile analysés, probablement en raison d'un faible niveau d'ADN d'entrée (˜1 pg et 18 pg, respectivement).

Les échantillons d'ADN tactile préparés pour cet ensemble de validation ont été obtenus à partir d'empreintes digitales physiques prélevées sur une lame de verre immédiatement (fraîche) et après 24 h d'exposition sur une paillasse. L'échantillon d'ADN tactile fraîchement préparé (1 pg d'ADN) a montré des résultats incorrects à 16 (39%) des 41 SNP avec le dosage HPS-MPS- MiSeq ;et 19 (46 %) SNP avec le dosage HPS-MPS-ION,

Les deux dosages ont donc eu des problèmes similaires avec cet échantillon. L'échantillon d'ADN tactile vieilli (˜18 pg d'ADN) a révélé des résultats incorrects pour 6 (15%) des 41 variantes HPS avec le HPS-MPS-MiSeq et une variante (2%) avec le HPS-MPS-ION.

La différence de performance des deux tests entre l'ADN tactile frais et vieilli s'explique probablement par des différences de quantité d'ADN d'entrée collectée à partir de l'écouvillon d'empreintes digitales, plutôt que le temps entre le toucher et la collecte de traces.

Notamment, les deux échantillons d'ADN tactile avaient des quantités d'entrée bien inférieures au seuil de sensibilité établi pour les deux tests, contrairement à tous les autres échantillons de cas fictifs utilisés qui étaient proches ou supérieurs au seuil de sensibilité.

Le test HPS-MPS-MiSeq a affiché des génotypes précis (basés sur le typage HP et HPS CE comparaison) dans la plage entre 100 pg et 250 pg d'ADN d'entrée pour ces exemples de cas fictifs.

Cette constatation appuie l'idée que l'échantillon en double unique dans le test de sensibilité qui a montré des abandons à 100 pg d'ADN d'entrée peut représenter une valeur aberrante, et que la vraie sensibilité du test HPS-MPS-MiSeq peut être plus proche de 100 pg (correspondant à la sensibilité de HPS-MPS-ION) plutôt que les 250 pg indiqués par le test de sensibilité.

Dans l'ensemble, les deux tests HPS MPS ont été en mesure de générer des résultats HIrisPlex-S complets et précis à partir de tous les types de scénarios de travail de cas

simulés testés, à l'exception des échantillons d'objets touchés avec des quantités d'ADN d'entrée infimes qui étaient bien inférieures au seuil de sensibilité estimé des tests.

3.3.4 Outil de test d'échantillons de mélange et de déconvolution

La déconvolution des mélanges est un domaine de recherche actif et plusieurs des outils commerciaux ont été développés pour aider à l'interprétation mixte des profils STR médico-légaux.

Cependant, les outils de séparation des mélanges pour les SNP font actuellement défaut, il est très difficile de séparer les mélanges lors de l'utilisation de méthodes de génotypage SNP basées sur CE (Capillary Electrophoresis). Outre l'augmentation de la capacité de Multiplexage, cela a fourni l'autre motivation pour développer des tests SNP basés sur MPS, qui permettent le séquençage des nucléotides entourant la variante d'ADN et fournissent des informations sur le nombre de lectures par allèle. (Voir annexe 7).

Un organigramme a été conçu qui indique les outils et les tableaux (c'est-à-dire outil de déconvolution de mélange pour deux personnes, table de seuil de lecture, etc.) pour mieux comprendre comment traiter un échantillon inconnu (considérer comme un contributeur unique ou une source d'ADN mixte) en utilisant à la fois des tests et des systèmes de séquençage pour générer correctement un Profil génotypique HIrisPlex-S MPS à partir d'un échantillon d'ADN pour une utilisation ultérieure avec l'outil Web de prédiction HIrisPlex-S pour obtenir des probabilités de couleur des yeux, des cheveux et de la peau.(Voir annexe 7).

3.3.5 Tests de spécificité et d'endommagement/dégradation de l'ADN :

Cinq espèces animales ont été testées avec le HPS-MPS-MiSeq et le dosages HPS-MPS-ION. Les échantillons comprenaient un chat, un chien, un cochon, une souris et chimpanzé (à des entrées d'ADN de 1 ng). Le nombre de lectures de séquençage générées pour les cinq espèces avec les deux tests HPS-MPS-MiSeq et HPS-MPS-ION.

En utilisant le test HPS-MPS-MiSeq,

31 (76%) des 41 variantes d'ADN ont révélé des lectures de séquençage chez le chat,

34 (83%) en cochon,

40 (98 %) chez la souris et

2 (1 %) chez le chien,

Tandis que le chimpanzé a produit un profil de génotype de 39 (95 %) des 41 variantes d'ADN HIrisPlex

Avec le test HPS-MPS-ION, 21 (51 %) variantes d'ADN ont donné des résultats de séquençage chez le chat,

20 (49 %) chez le porc,

31 (76 %) chez la souris,

28 (68 %) chez le chien,

Le chimpanzé a produit un profil complet des 41 variantes d'ADN de HIrisPlex.

Dans l'ensemble, le nombre moyen de lectures des espèces non humaines était bien inférieur à celui attendu pour une entrée d'ADN humain de 1n.

Pour HPS-MPS-MiSeq, le chat a donné 2 fois moins en moyenne d'échantillon de lecture, tandis que le porc et la souris ont donné 4 fois moins.

Pour HPS-MPS-ION, le chat a donné 100 fois moins en moyenne d'échantillon de lecture, tandis que le porc (10 fois), la souris (30 fois) et le chien ont donné (8 fois) moins.

Ces observations couplées aux profils partiels générés peuvent servir d'outil pour aider distinguer les échantillons humains et non humains lors de l'évaluation d'un échantillon d'ADN inconnu sur une scène de crime.
Cependant, étant donné que le FDP serait généralement effectué sur des échantillons d'ADN de scène de crime après le profilage STR, l'ADN humain est déjà détecté et confirmé dans chaque cas.
Pour préparer des échantillons qui testeraient l'effet des dommages à l'ADN sur les performances du test HPS-MPS-MiSeq conçu en interne, des aliquotes d'un échantillon d'entrée d'ADN de 500 pg ont été soumises à un rayonnement ultraviolet (UV) pendant 0 s, 5 min, 10 min et 20 min. Même après 10 minutes d'exposition à la lumière UV, un profil HPS 41-SNP complet a été obtenu avec une couverture moyenne de 2040 lectures. Après 20 minutes d'exposition aux rayons UV, 5 SNP : rs28777 SLC45A2, rs4959270 EXOC2, rs12896399 SLC24A4, rs1426654 SLC24A5 et rs3212355 MC1R ont affiché une chute en raison d'une dégradation présumée.
Ces résultats indiquent la robustesse du test HPSMPS-MiSeq pour faire face aux dommages simulés de l'ADN.
Pour HPS-MPS-ION, les tests de dégradation n'ont pas été effectués sur ces échantillons d'ADN artificiellement endommagés.

Cependant, les preuves préliminaires de la capacité de ce test à traiter l'ADN naturellement dégradé proviennent de l'analyse d'une série d'échantillons d'ADN extraits d'os qui ont passé environ 1 à 78 ans dans le sol, où l'efficacité HPS-MPS-ION s'est avérée comparable à celle du kit d'amplification PCR GlobalFiler sur les mêmes échantillons. Cependant, il convient de noter que la quantité maximale d'ADN utilisée pour analyser les STR était de 15µl alors que seulement 6µl ont été utilisés pour HPS-MPS-ION, ce qui a fait une différence significative dans les échantillons faibles. Des profils HPS complets ont été obtenus à partir d'aussi peu que 50 pg d'ADN avec un seuil de couverture de 200 lectures. Cependant, les performances de trois SNP : rs1545397 et rs1470608 dans OCA2 et rs10756819 dans BNC2 étaient légèrement plus faibles par rapport à d'autres marqueurs.

3.3.6 Test de concordance :
Ainsi que la coordination IUPUI US (Indiana University – Purdue University Indianapolis) et Erasmus MC Rotterdam laboratoires, cinq sites partenaires, avec une expérience MPS variable, ont été impliqués dans les tests de concordance des deux tests HPS-MPS, 3 pour HPS-MPS-ION et 2 pour HPS-MPS-MiSeq.
Au cours de la phase initiale des tests de concordance, il est devenu évident qu'il y avait un besoin de lignes directrices pour évaluer les seuils de lecture et l'évaluation des données pour l'interprétation d'ADN à source unique ou mixte.Si une seule source était indiquée, les utilisateurs étaient également invités à fournir un profil prévu final.
Par conséquent, ces directives d'interprétation ont été conçues pour aider les utilisateurs du test HPS-MPS à appeler le génotype et à séparer les mélanges à l'aide de la sortie du pipeline d'analyse HPSMPS. (Pour plus de détails voir annexe 8).

4. Deuxième étude : Architecture génétique du visage

4.1 Matériels et méthodes :

Le visage humain joue un rôle central dans la vie quotidienne, la communication, l'identification mutuelle, l'attirance. Il représente un ensemble multidimensionnel de phénotypes corrélés, la plupart du temps symétriques, complexes avec l'héritabilité élevée

Pouvoir prédire un trait phénotypique depuis des données génomiques semble difficile pour certains, pour cela un séquençage du génome complet du phénotype est étudié en détails puis modélisé en traits statistique. Plusieurs études ont contribué à la recherche de l'architecture génétique du visage. Dans ce travail, white, et al ;Bonfante, et *al;* Claes. Et *al ;*Qiao, Et *al,*quimontrent qu'il existe plusieurs signaux associés à une variation faciale et qui sont significatifs à l'échelle du génome.

Ce qui a aussi favorisé ce projet fut Identification d'individus par prédiction de traits à l'aide de données de séquençage du génome entier par Christoph Lippert, et al. Leurs analyses en permis de mieux comprendre comment les traits morphologiques complexes sont façonnés par des actions génétiques individuelles et coordonnées.

4.1.2<u>Echantillon et Recrutement</u> :

Auteur	Population	Effectif N
White et al 2020	ÉU /GB	Américain =4680 Anglais =3566
Bonfante et al 2021	Sud-amérique	Brésil =693, Mexique =1265 Pérou=1285 Chili=1891 Colombie=1853
Quoi et a 2018	Eurasiatique	Européen=86 Han Taizhou=154 Ouighur=694

Tableau 04 : échantillon selon chaque auteur et selon différentes populations

4.1.3 Génotypage, imputation et contrôle de qualité :

L'ADN génomique utilisé est extrait d'échantillon sanguin, salivaire ou reste humain qui ont été génotypé à l'aide de plate-forme de génotypages SNP à l'échelle du génome d'Illumina HumanHAP550 utilisé par white et al,2020, le kit coreExome SNP et illumina OmniExpress8 Beadchip sont utilisés par LuQiao ,2018, Bonfante,2021 et white et al,2020). (Voir annexe9pour Protocol illumina).

Les génotypes ont été "harmonisés" à l'aide genotype harmonizer (V.1.4.20) avec une taille de fenêtre 200 SNP, un minimum de 10 variantes avec une fréquence des allèles mineurs. Par la suite la phase d'imputation est nécessaire durant ce processus (voir annexe 10). Cette dernière a été effectué par le serveur d'imputation SHAPEIT2 et par Sanger Imputation server (v.0.0.6). Les résultats imputés sont ensuite combinés. Enfin un contrôle de qualité est effectué à l'aide de PLink v.107 basé sur l'assemblage du génome GRCh37 ou la fréquence allélique de chaque SNP a été comparée à la fréquence allélique du SNP dans l'ensemble de données 1000G Phase 3. Après application de ces filtres (génotypages imputation et contrôle de qualité) les SNP retenus pour une analyse plus approfondie sont comme suit : 636195 SNP pour population Latine, 747619 SNP pour la population américaine et 8629873 SNP pour la population anglaise avec 203 Lead SNP et 810648 SNP pour la population Ouighour.

4.1.4 Extraction des caractéristiques du visage (phénotypage) :

Un système 3dMD face et Vectra H1 ont été utilisé pour recueillir des images 3D de haute résolution. Ces images sont ensuite converties par logiciel 3dMD en format fichier OBJ. Ces dernières sont ensuite exportées vers un ordinateur équipé d'un programme de nettoyage, de balayage pour le recadrage et la découpe, l'enlèvement des cheveux, des oreilles et des polygones dissociés. (Figure.7)

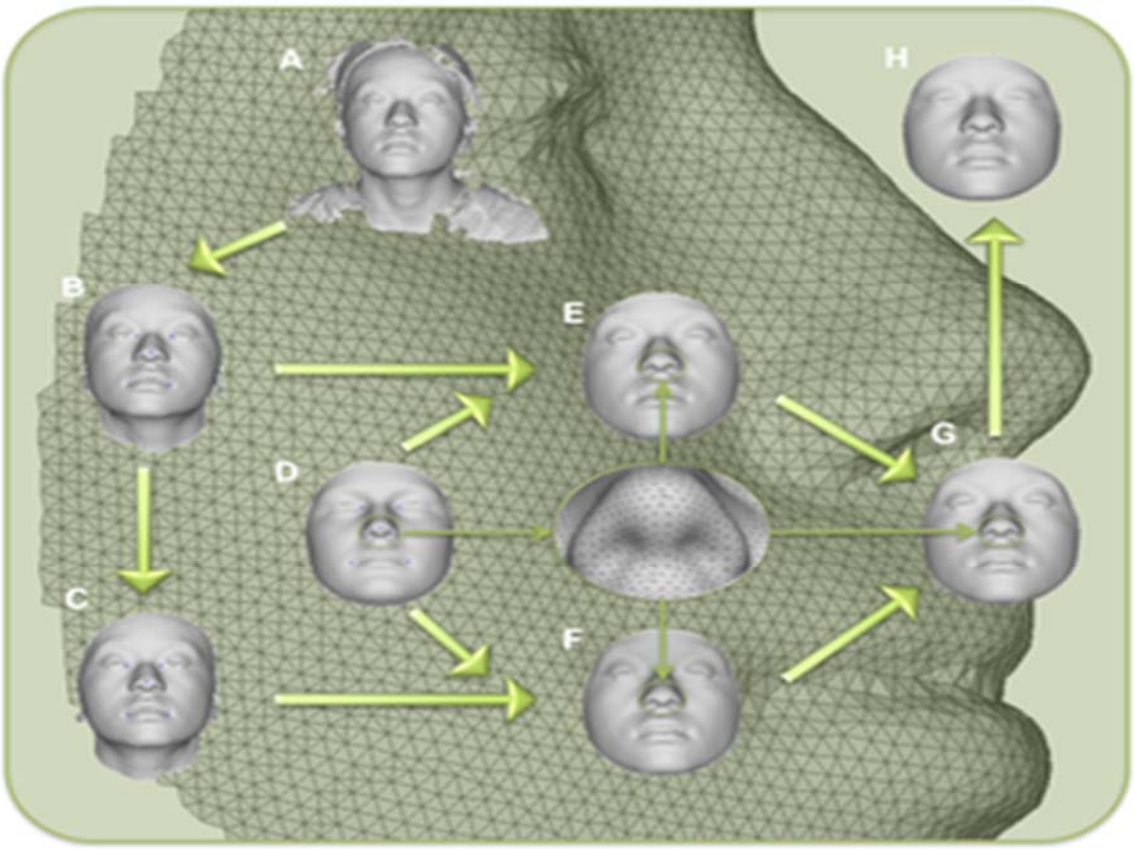
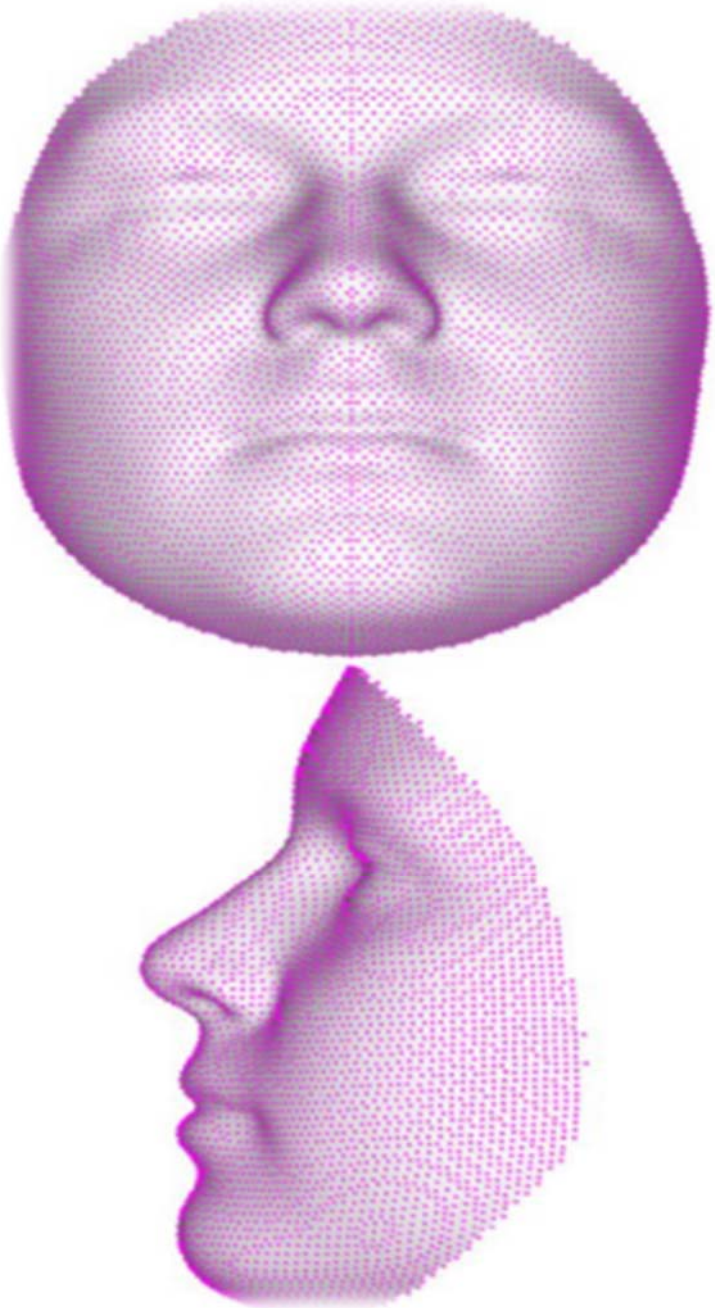


Fig.7 : Flux de travail pour le traitement 3D de l'analyse du visage**.** A) surface d'origine, B) taillée pour exclure les parties non faciales, C) réfléchie pour faire l'image miroir, D) masque anthropométrique de quasi-repères, E) remappé, F) réfléchi remappé, G) symétrique, H) reconstruit modélisation de la forme faciale 3D à partir de l'ADN (Claes et al 2014).

Une cartographie de maillage dense de 7016 sommets homologue a été placé sur les images (échantillons américain et anglais), et 32251 sommets au total pour le maillage de référence eurasiatique en utilisant la boîte à outils MeshMonk (Figure.8).

A



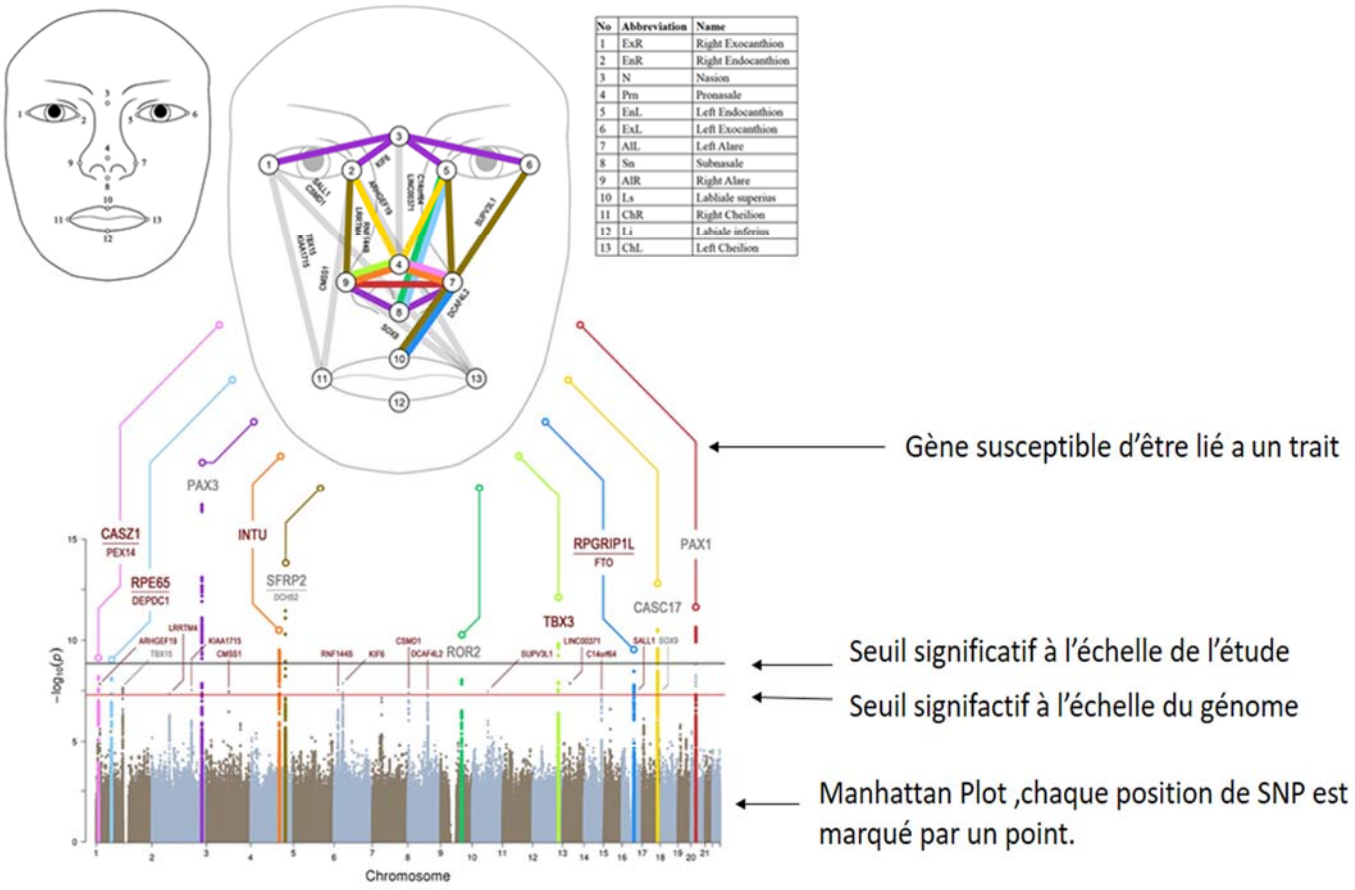
B

Fig.8 : Enregistrement du modèle facial, Chaque point magenta représente un sommet unique N = 7 160 pour le visage
A : Visage de face
B : Visage de profil
(Claes et al ,2019).

Ces images ont permis de placer des repères qui seront utilisés dans l'étude de morphologie du visage (Fig.9 et 9.1).

Auteur	Nombre de repères
LuQiao et al	15
xiong et al	13
Bonfante et al	19+22semi-repères
white et al	19

Tableau 05 : nombre de repère selon différents auteurs



No	Abbreviation	Name
1	ExR	Right Exocanthion
2	EnR	Right Endocanthion
3	N	Nasion
4	Prn	Pronasale
5	EnL	Left Endocanthion
6	ExL	Left Exocanthion
7	AlL	Left Alare
8	Sn	Subnasale
9	AlR	Right Alare
10	Ls	Labiale superius
11	ChR	Right Cheilion
12	Li	Labiale inferius
13	ChL	Left Cheilion

Fig 9 :Caractéristiques du profil du visage montrant une association significative à l'échelle du génome (xiong et al 2018).

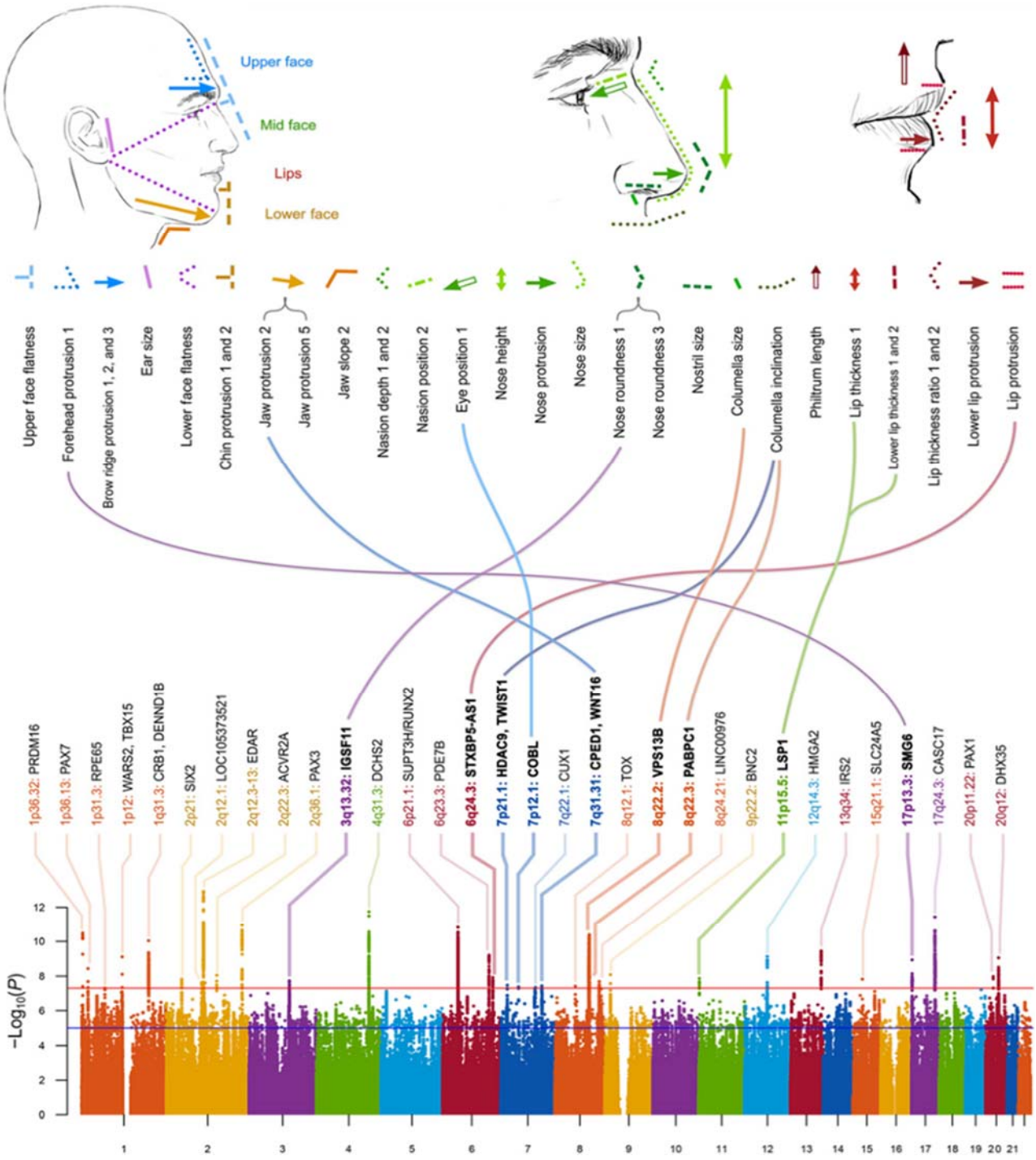


Fig. 9.1 : 32 caractères mesuré dans les individus Candela et parcelle Manhattan
Les dessins indiquent les caractéristiques pour lesquelles les 32 caractères énumérés ci-dessous ont été mesurés chez les individus CANDELA
-Parcelle de Manhattan à partir de la découverte de méta-analyse GWAS (N = 10.115 Européens) pour 78 distances euclidien entre 13 repères faciaux. (Bonfante et al,2021)

Une approche guidée par les données de segmentation facial regroupe les vertices qui sont fortement corréler. Le coefficient RV définit les covariances entre les quasi-repère pour ensuite construire une matrice de similaritéAinsi, un espace de forme a été construit pour chaque segment facial, indépendant des autres segments.

Les scores obtenus sont analysés par une analyse Procruste puis aligné dans un système de coordonnée pour générer des variables qui capture des formes biologiques.(Voir schéma Fig.10).

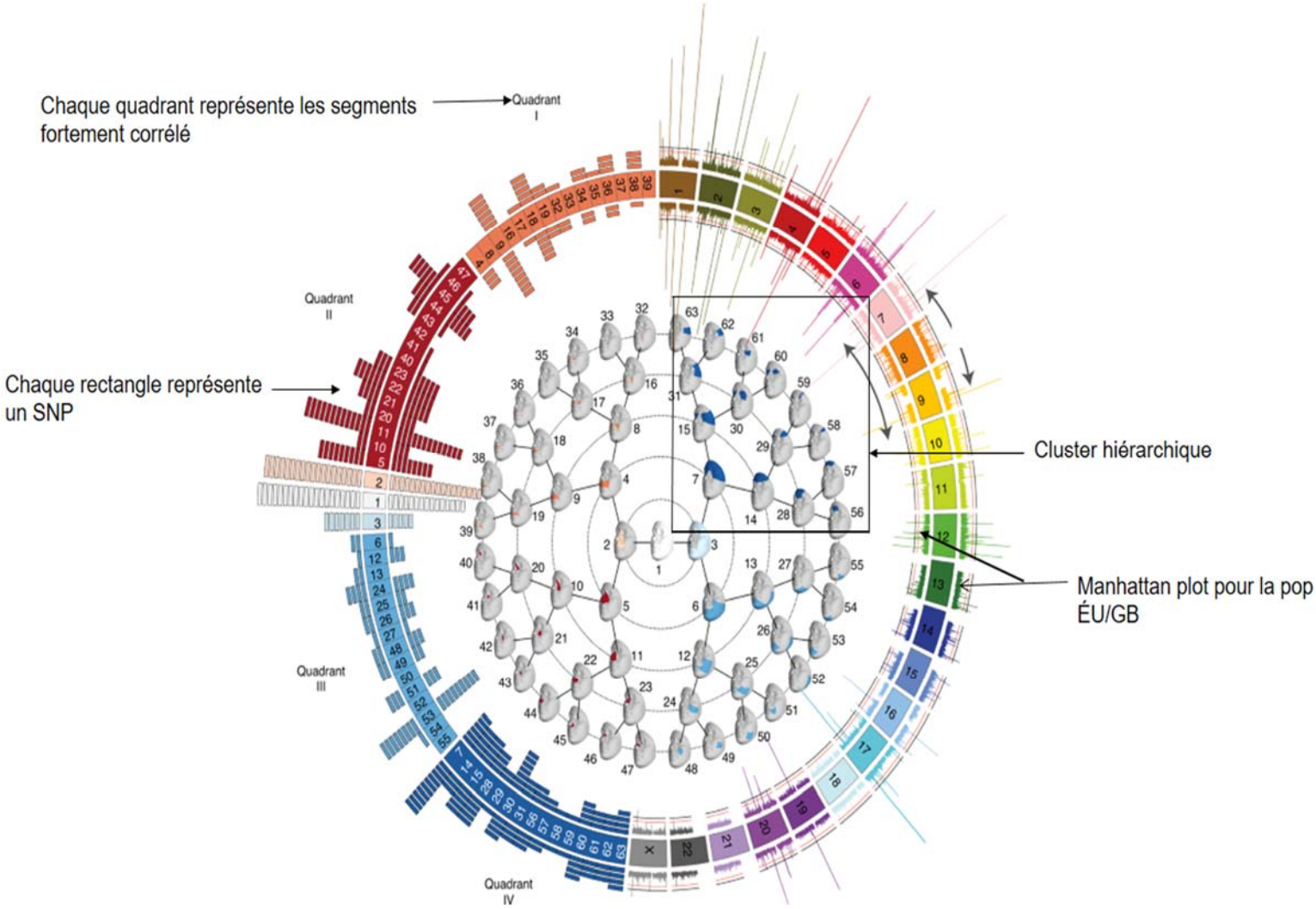


Fig.10 : Résultats globaux des méta-analyses menées par les États-Unis et le Royaume-Uni L'histogramme disposé sur le côté gauche représente le nombre de SNP principaux Significatifs à l'échelle du génome atteignant leur valeur P la plus faible dans chaque segment, chaque rectangle représentant un SNP.(white et al , 2020).

Une analyse en composantes principales est par la suite performée en superposition pour capturer les variations phénotypiques du visage pour chaque segment. Une extraction d'une signature 3D ou vecteur de correspondance (CV) permet de mesurer les distances entre un visage de saisie et un visage de référence. Il permet aussi de transformer un vecteur 3D en un vecteur 1D facilitant alors le calcul de l'analyse des composantes principales. Par la suite chaque SNP est associé à une variation phénotypique en utilisant l'analyse des corrélation canonique CCA. Cette analyse permet d'extraire la combinaison linaire de l'ensemble des PCquiest maximalement corrélé avec un SNP. Ce qui donne une valeur corrélative entre le PC et SNP. L'analyse des corrélations canonique évite la présélection de PC individuels (Weinberg et al 2018).
Le test RAO-F fournit la valeur P prédictive.

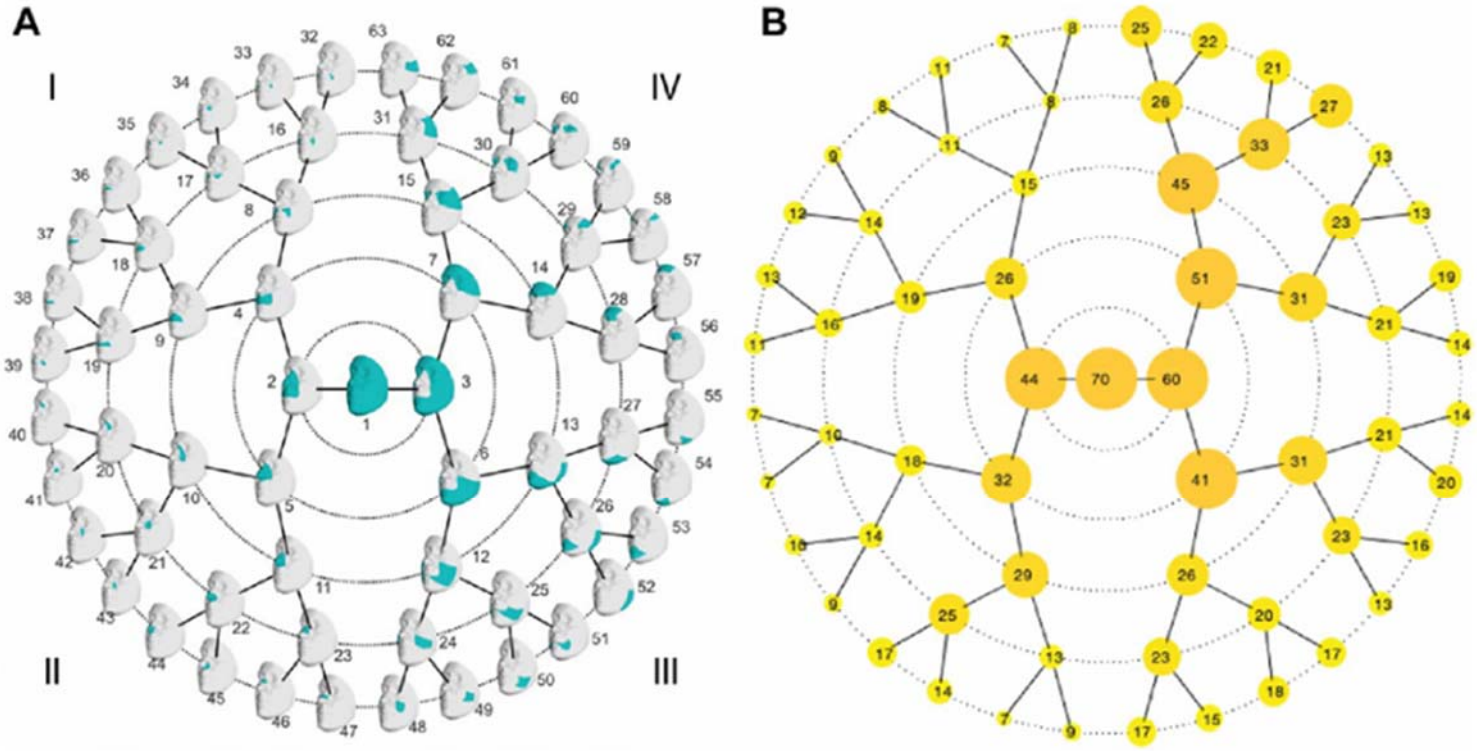


Fig.11 : Le nombre de composantes principales retenues après l'analyse parallèle pour chaque segment facial. (White et al ,2020)

A) Segmentation facial obtenue par cluster hiérarchique
B) Nombre de PC pour chaque segment

Les chiffres romans représentent les quadrants de segments facials

4.1.5GWAS et Méta-analyse :

Une série de GWAS et méta-analyse ont été mené pour tester l'association génétique entre les différents SNP autosomal avec les phénotypes du visage.

Validité de la méta-analyse

La méta-analyse utilisée consiste en 3 étapes réalisées séparément

1. Identification : la composante CCA de l'étape d'identification identifie le trait phénotypique le plus corrélé avec chaque SNP. (Production de valeur P par test ROA-F)
2. Vérification : projection des PC vers les trait pour avoir une valeur univarié, Permet de vérifier ou reproduire l'association SNP-Trait. (Production de valeur P par regression univariante)
3. Méta-analyse : les deux valeursP de l'étape d'identification et de vérification sont

Les statistiques sommaires de tous les SNP et de tous les phénotypes faciaux de la méta-analyse à effet fixe de variance inverse ont été effectué à l'aide du logiciel PLink (le chromosome X n'est pas inclus dans le méta-analyse) (White et al, 2020). Au total 5 loci dit hautement significatif pour la population Sud-Américaine ,15 loci pour la population américaine et anglaise et 6 loci pour la population ouighour.

4..1.6Enrichissement des pics GWAS par enhanceurs spécifique à travers la chronologie du développement du visage :

Les index SNP identifié par le GWAS se produisent généralement dans la région non-codante. L'évaluation d'enrichissement des SNP dans ces régions au sein des enhanceurs identifié par ChromHMM annotations à 25 états de chromatine (voir annexe 11) dans 150 tissus humains, type de cellule au stade embryonnaire, fœtale et adulte (plusieurs stades de développements crâniofaciales). Un programme Gregor est utilisé pour évaluer l'enrichissement de ces SNP. Un enrichissement 2fois /4fois des SNP a été trouvé au GWAS du visage dans tous les enhenceurs crâniofaciales et les enhenceurs spécifique au crâniofaciales $P<10^{-50}$ avec correction Bonferroni. Ces enhenceurs spécifique n'ont été trouvé sur aucun autre tissu. La présence de ses enhenceurs suggère que ces SNP jouent un rôle réglementaire spécifique au développement facial. Les échantillons du post et pré embryonnaire ont un enrichissement plus élevé $P<10^{-20}$ C.B. Les résultats de l'analyse de GREAT indiquent que certains gènes proches des pics de GWAS sont impliqués dans le développement du visage et des membres. Un lead SNP principal significatifs à l'échelle du génome a montré une activité préférentielle dans d'autre type de cellules dérivées in vitro (Bonfante et al 2021).

4.2Résultats :

4.2.1 Association des SNP a la forme du visage : Avec l'ensemble de phénotypage et multiples analyses faites, il est en mesure de fournir une compréhension plus claire de l'architecture génétique de la variation faciale. Cette étude représente le plus grand balayage du génome, examinant les phénotypes de forme faciale chez l'homme à partir d'image 3D. Une multitude de loci génétique montrent l'association suggestive à l'échelle de l'étude. Étant donnée ces SNP ont été extrait de différents individu et de différentes ethnicités, ces derniers varient d'une population a une autre.

4.2.2 Population EU/GB :

15 loci sont impliqué dans une variété de segments du visage, et plusieurs de ces loci affectent des segments dans plus d'une une région faciale Figure.12.

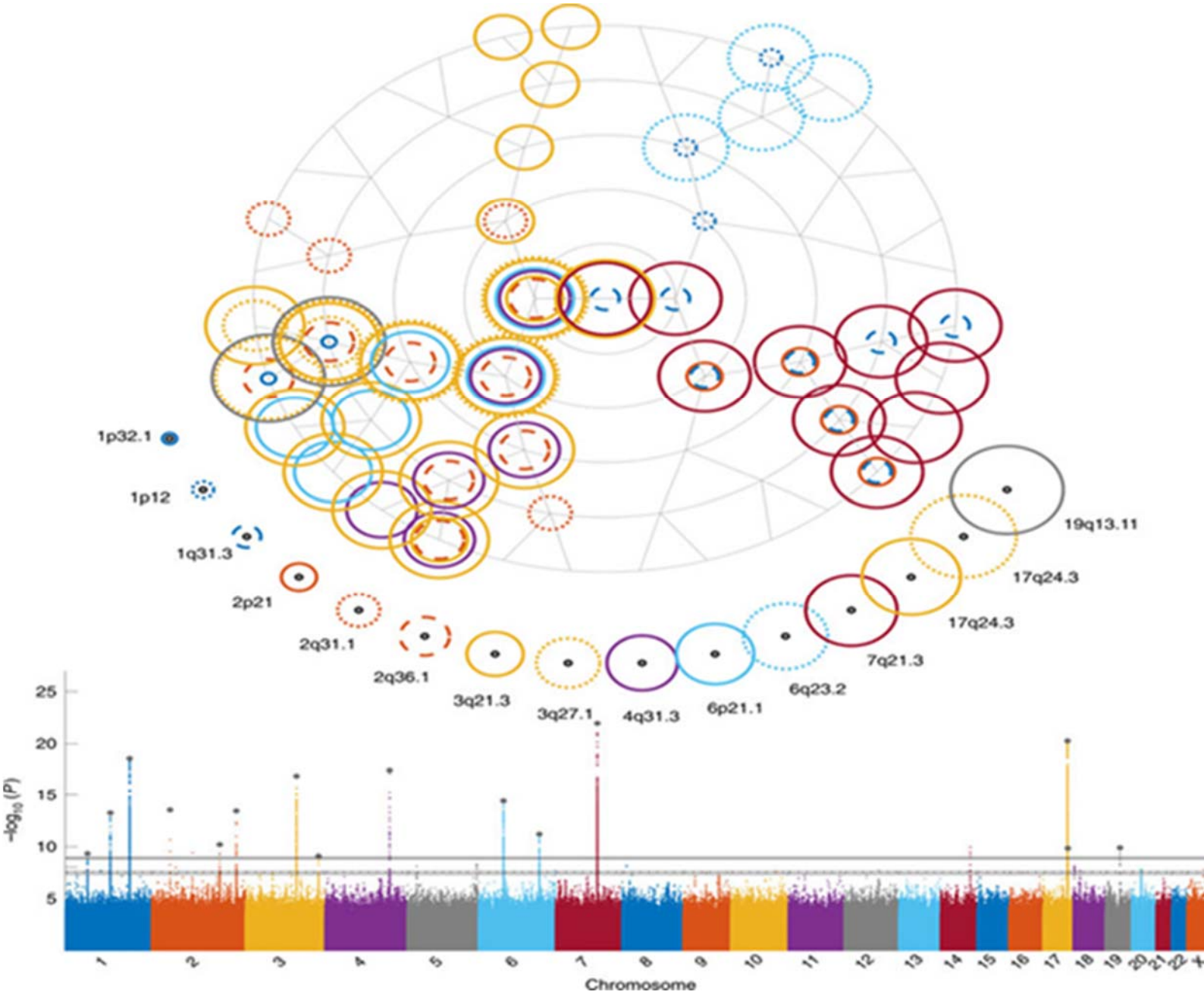


Fig. 12 : les cercles concentriques représentent les loci atteignant une signification à l'échelle du génome pour chaque segment et les chromosomes qui se chevauchent
Par-dessous un Manhattan plot avec les différents SNP ayant atteint le seuil significatif a l'échelle du génome et de l'étude (claes ,2018).

	SNP/GÈNE	TRAIT	VALEUR P
ÉU	-rs3936018/ TBX15 -rs7513680/ TBX15 - rs13117653/ MSX1	-Front - zone des joues autour des coins de la bouche - la lèvre inférieure latérale et mandibule	-8.01×10^{-58} $-7{,}03\times10^{-13}$ $-4{,}2\times10^{-18}$
GB	-rs7513680/ TBX15 -rs rs3910659/ MSX1 -rs76244841PRDM16et rs62443772 GLI3	-zone des joues autour des coins de la bouche - la lèvre inférieure latérale et mandibule - la columelleà la commissure orale	$-3{,}26\times10^{-15}$ $-4{,}45\times10^{-9}$. $-5{,}35\times10^{-16}$

Tableau 06 : SNP significatif a l'échelle du génome pour la population anglaise et américaine.

Le reste des Loci significatifs à l'échelle du génome et résultats de réplication sont indiqués dans le tables ci-dessous et

Segment	SNP 1			SNP 2			Test statistic	P value
	rsID	Location	Gene annotation	rsID	Location	Gene annotation		
6	rs10838269	11:44378010	*ALX4*	rs11175967	12:66321344	*HMGA2*	23.9422	9.94×10^{-7}
9	rs76244841	1:2775953	*PRDM16*	rs62443772	7:42131949	*GLI3*	16.5745	4.68×10^{-6}
11	rs6740960	2:42181679	*PKDCC*	rs6795164	3:133885925	*SLCO2A1*	16.3707	5.21×10^{-5}
22	rs7373685	3:128107020	*GATA2*	rs7843236	8:121980512	*SNTB1*	15.7837	7.10×10^{-5}

Tableau 07 : loci significatif à l'échelle du génome (Claes ,2018)

4.2.3 POPULATION EURASIATIQUE :

6 régions ont révélé des signaux significatifs à l'échelle du génome

Population euroasiatique

Chromosome	SNP	Valeur P	Trait
11q24	Rs11868752	$P=7.70 \times 10^{-9}$	Distance entre le canthus ext et int.
5q35.3	RS118078182	$P= 2.64\times10^{-10}$	Distance de Nasion point-Pronasale-Subnasale.
2p16.3	Rs17868256	$P= 1.07\times10^{-9}$	Joue et mâchoire chez les femmes.
20p12.3	Rs3920540	$P= 4.09\times10^{-8}$	Forme nasal chez les femmes.
3p12.2	Rs61672954	$P= 2.61\times10^{-7}$	Des faces latérales chez les deux sexes.
4p15.1	Rs60159418	9.78×10^{-10}	Bouche-menton chez les hommes.

Tableau 08 : résultat des SNP significatif à l'échelle du génome pour la population eurasiatique

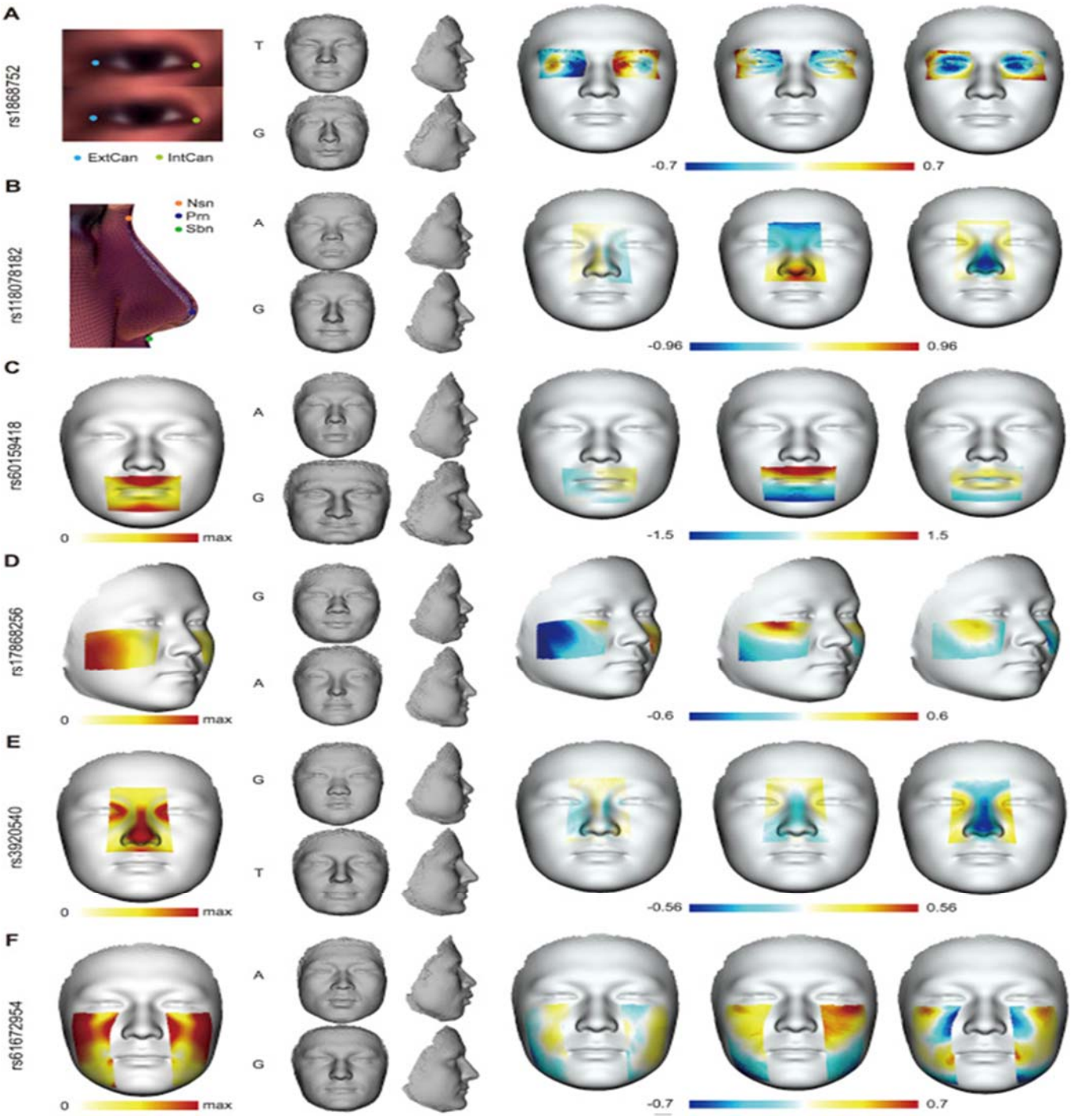


Fig.13 : Le premier visage montre l'effet général sur la caractéristique correspondante comme le déplacement de repères ou de mailles. Le panneau intermédiaire de quatre faces miniatures donne les extrapolations vers la tendance Han sur le dessus, ou la tendance européenne au-dessous, avec l'allèle associé étiqueté sur le côté gauche (LuQiao,2018).

1.A/

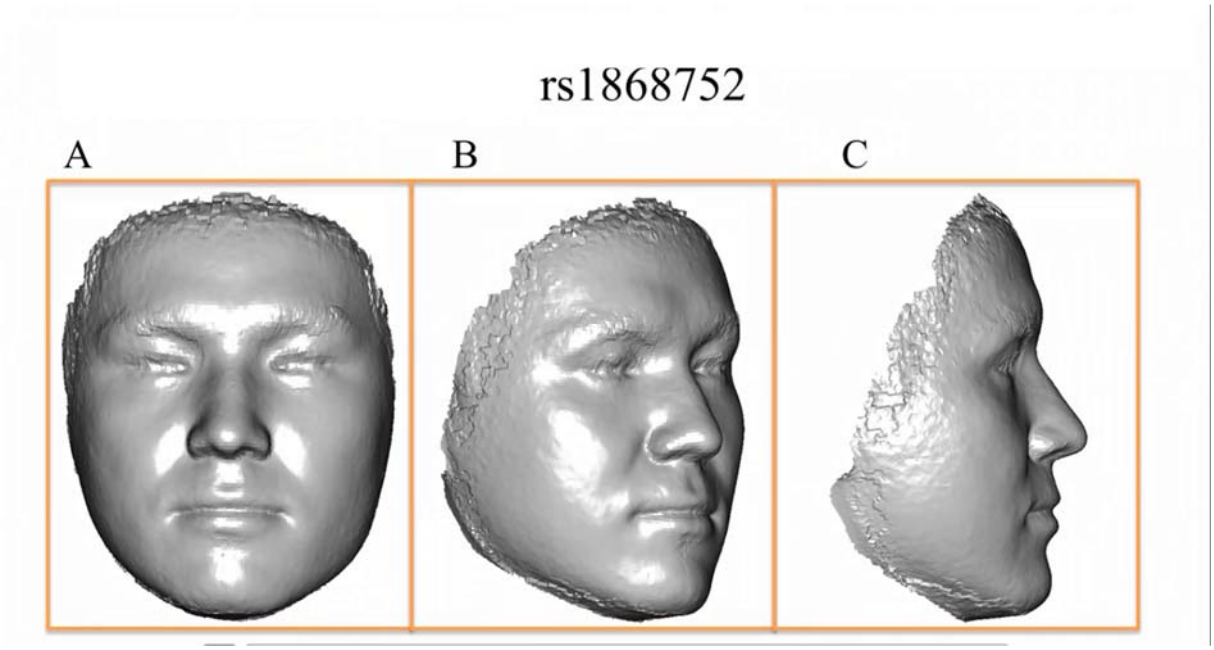


Fig.13.1 : Le visage extrapole l'effet de rs1868752T donnant des yeux plus étroits (plus petite distance ExtCan-IntCan) et un plus petit nez tendant vers la population Han

2.A/

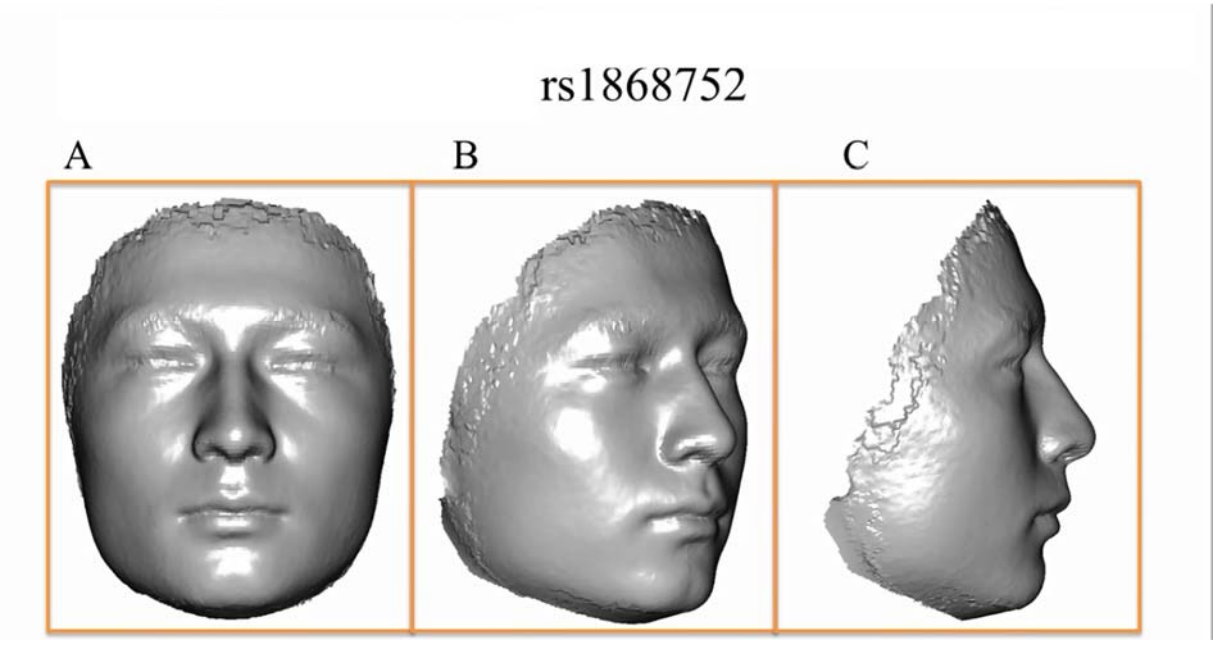


Fig 13.2 :/ Le visage extrapole l'effet de rs1868752, l'allèle G semble être associé à la crête élevée de nez tendant vers la population Européenne

1.B/

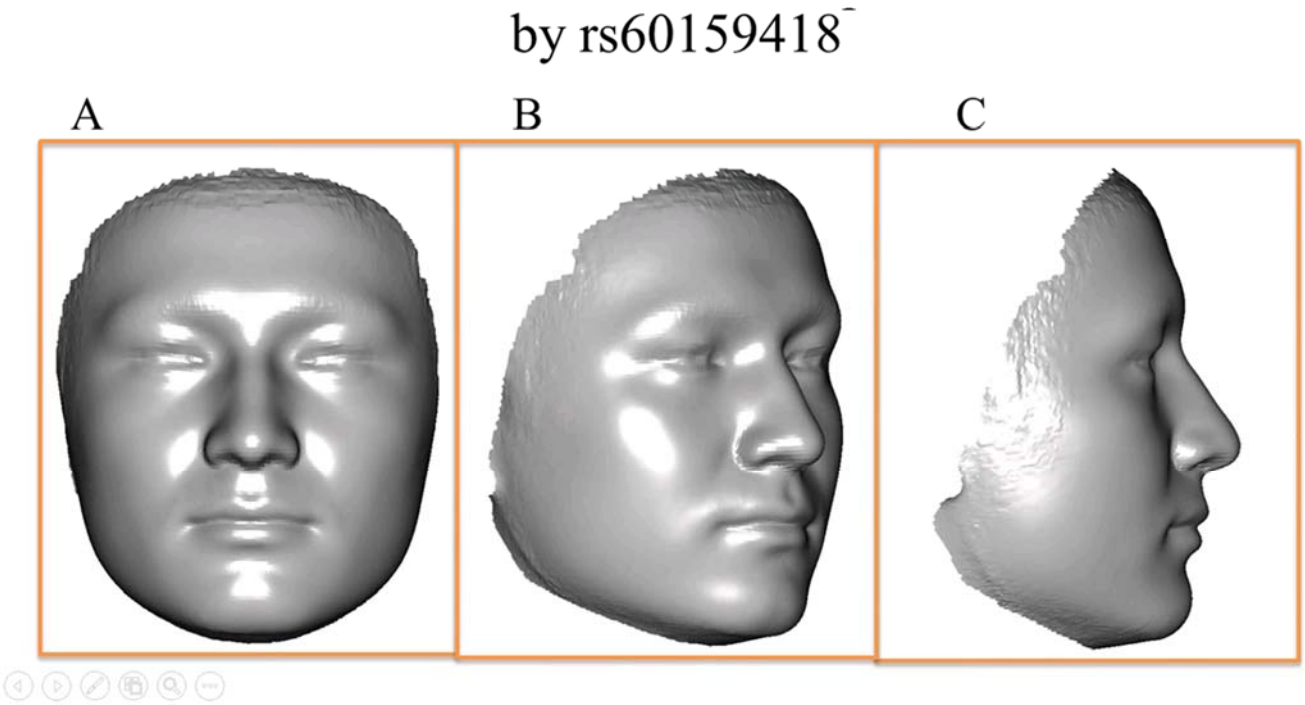


Fig13.3 rs60159418G chez les mâles la forme principale a changé la bouche. Allèle G semblait encastrée la zone de la bouche du plan du visage tendant vers la population Han

2.B/

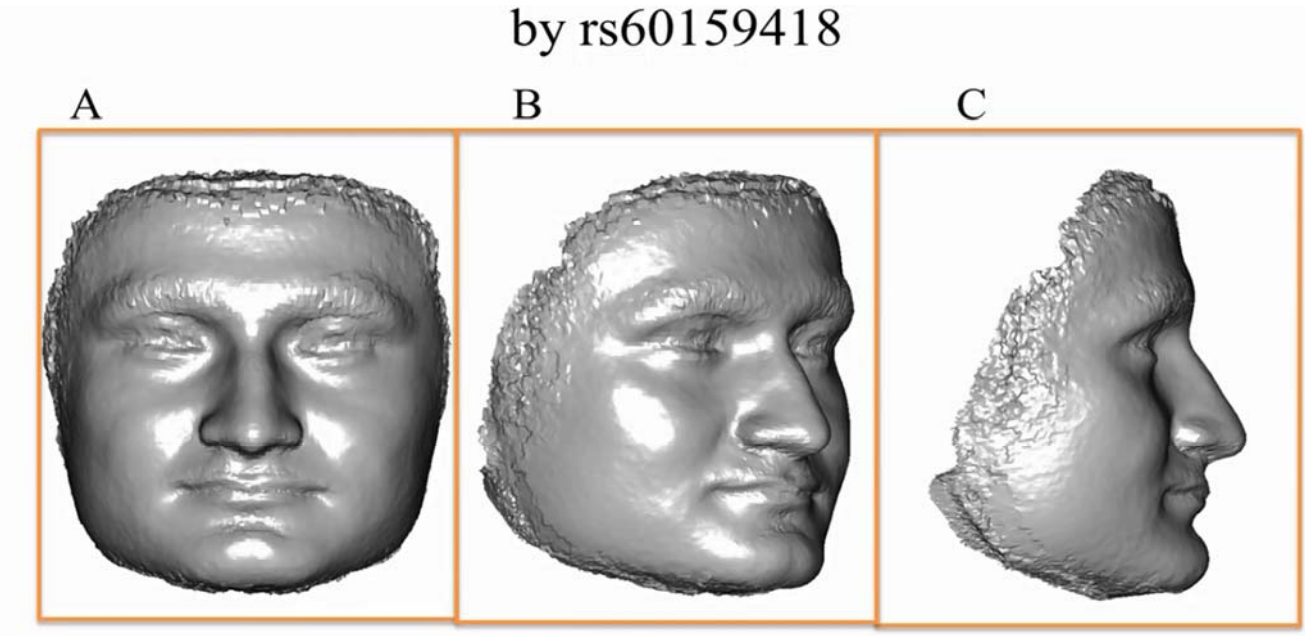


Fig.13.4rs60159418A la courbe bouche-menton s'est pliée convexement du plan facial et inclinaient du nez tendant vers la population européenne

1.C/

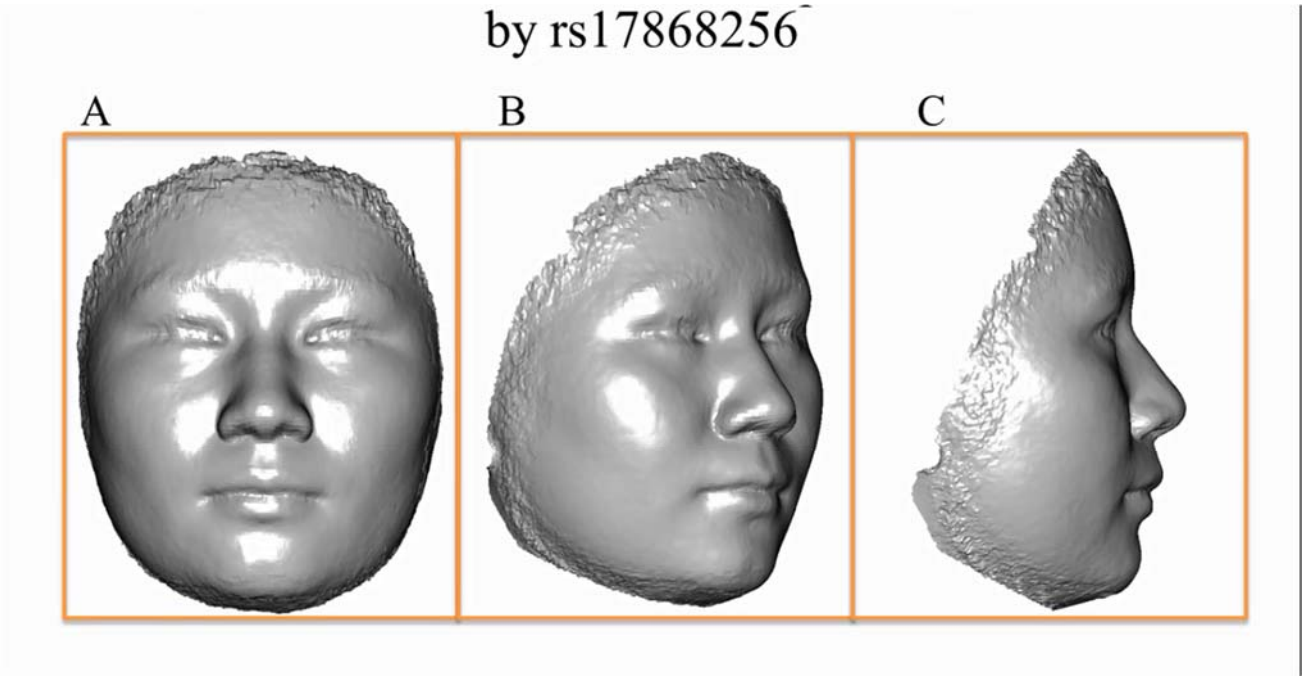


Fig 13.5 / rs17868256avec allèle G associé à des joues latéralement élargies, ce qui rend le visage plus large tendant vers la population Han

2.C/

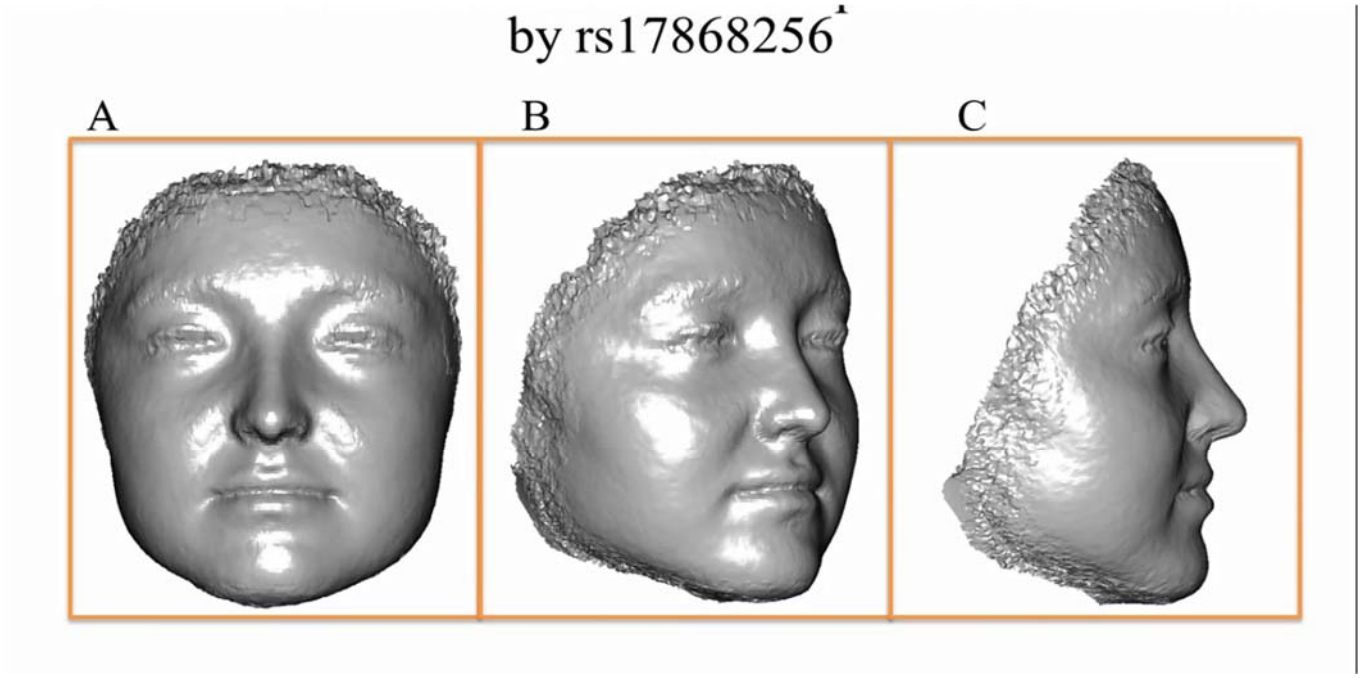


Fig 13.6 /:rs17868256 A semble donné au visage un aspect plus étroit tendant vers la population Européenne.

4.2.4 POPULATION CANDELA : Cinq loci sans aucune preuve de réplication dans la méta-analyse européenne incluent des gènes et des dispositifs réglementaires d'importance potentielle pour le développement crâniofaciales.
La région associée en 1p12 inclut une région d'introgression adaptative archaïque, avec un haplotype de Dénisovien commun dans les Amérindiens affectant particulièrement l'épaisseur de lèvre. Cette région génomique a été précédemment rapportée pour être associée aux dispositifs de l'oreille externe et du visage (méta-analyse européenne), mais c'est la première fois qu'elle est associée à l'épaisseur de lèvre. Le signal de sélection dans cette région s'est chevauché avec une région d'introgression de l'homme archaïque, très probablement Dénisovien, soulignant la possibilité d'introgression adaptative.

SNP	**GÉNE**	**TRAIT**
RS3790553	WARS2/TBX151p12	Epaisseur de lèvres supérieur
RS10225796	COBL7p12.1	position des yeux
RS143566339	BABPC18q22.3	l'inclinaison du columelle
RS907613	LSP111p15.5	l'épaisseur des lèvres et lèvres inferieur

Tableau 09 : SNP significatif a l'échelle du génome pour la population CANDELA

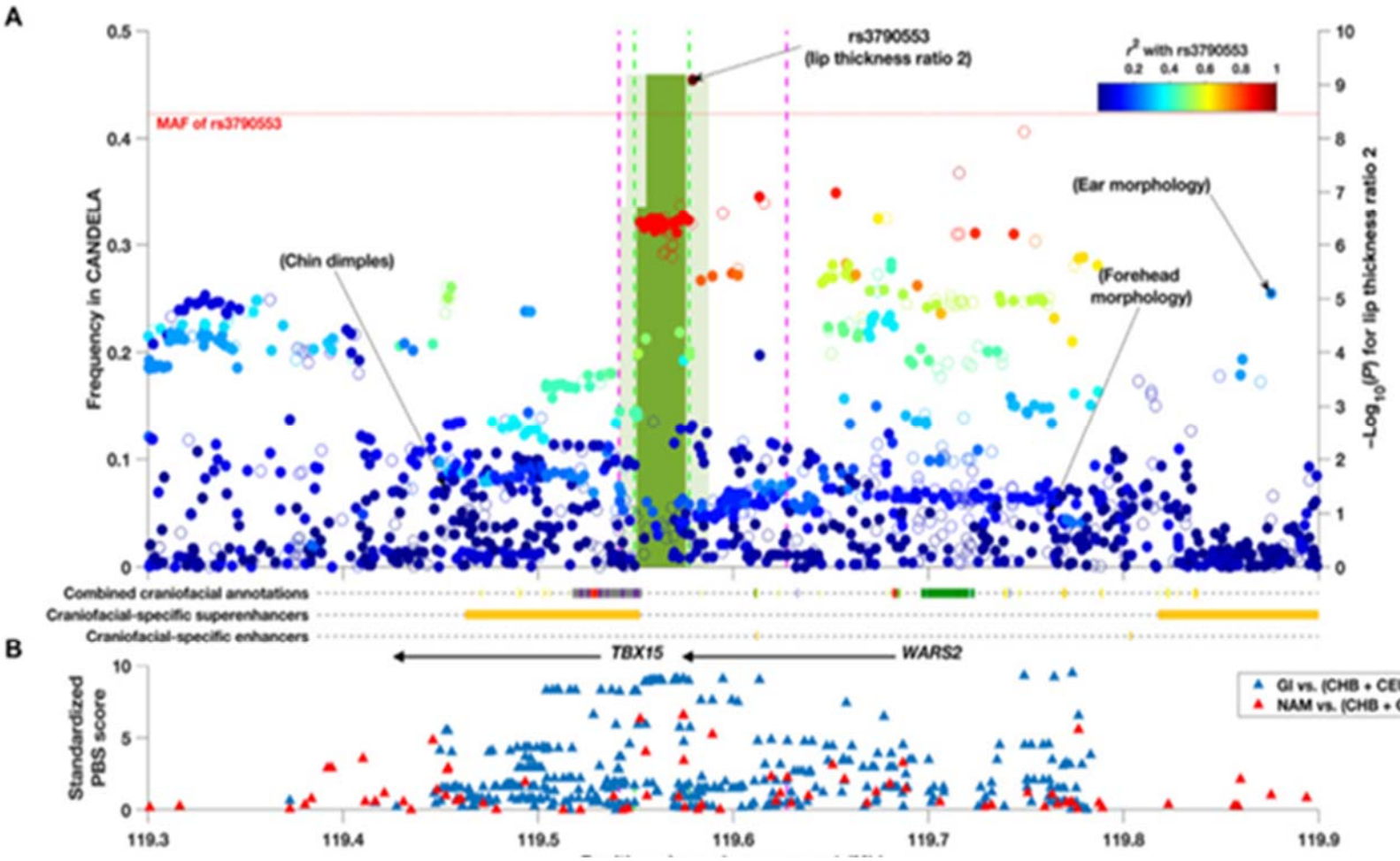
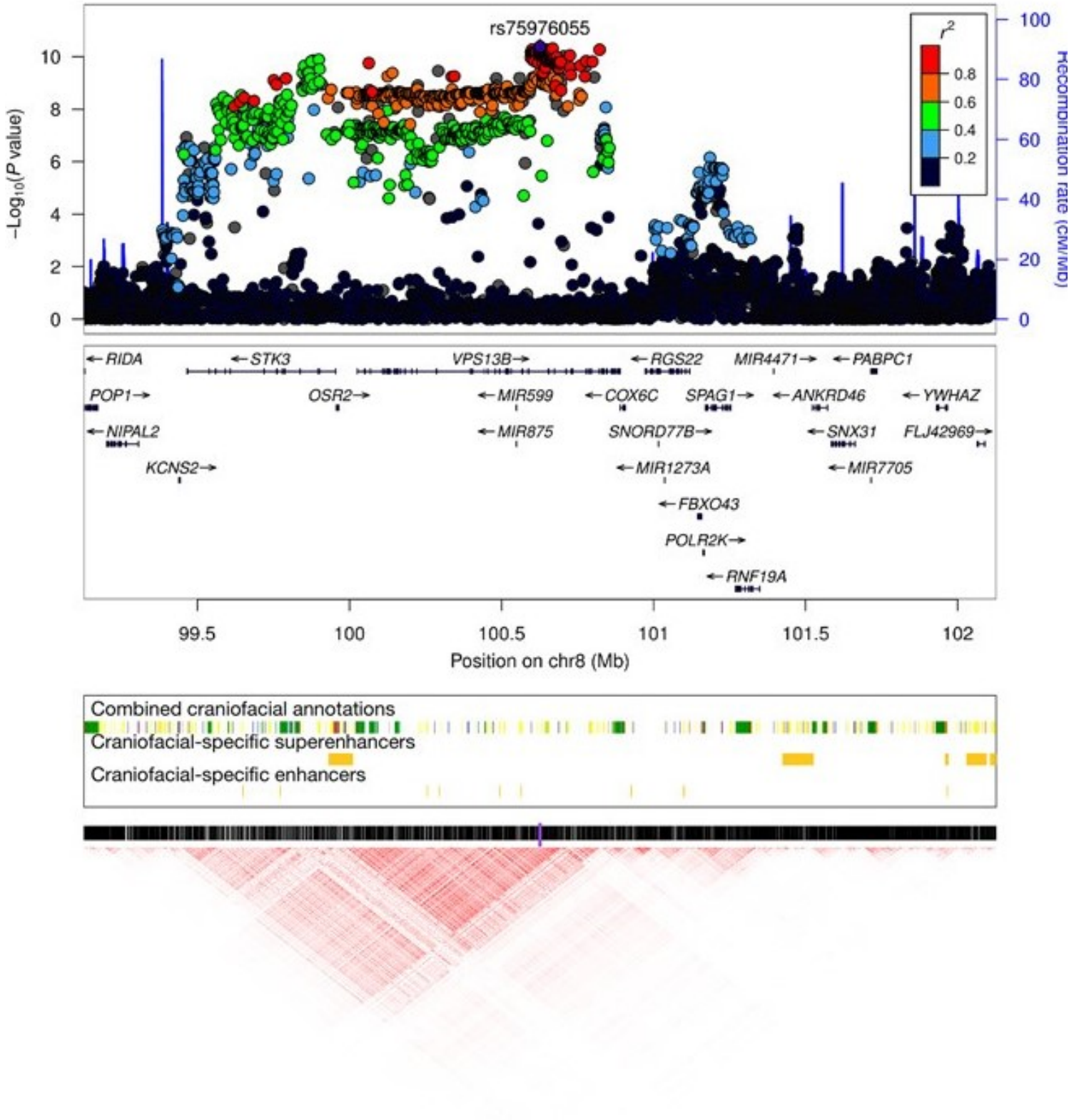


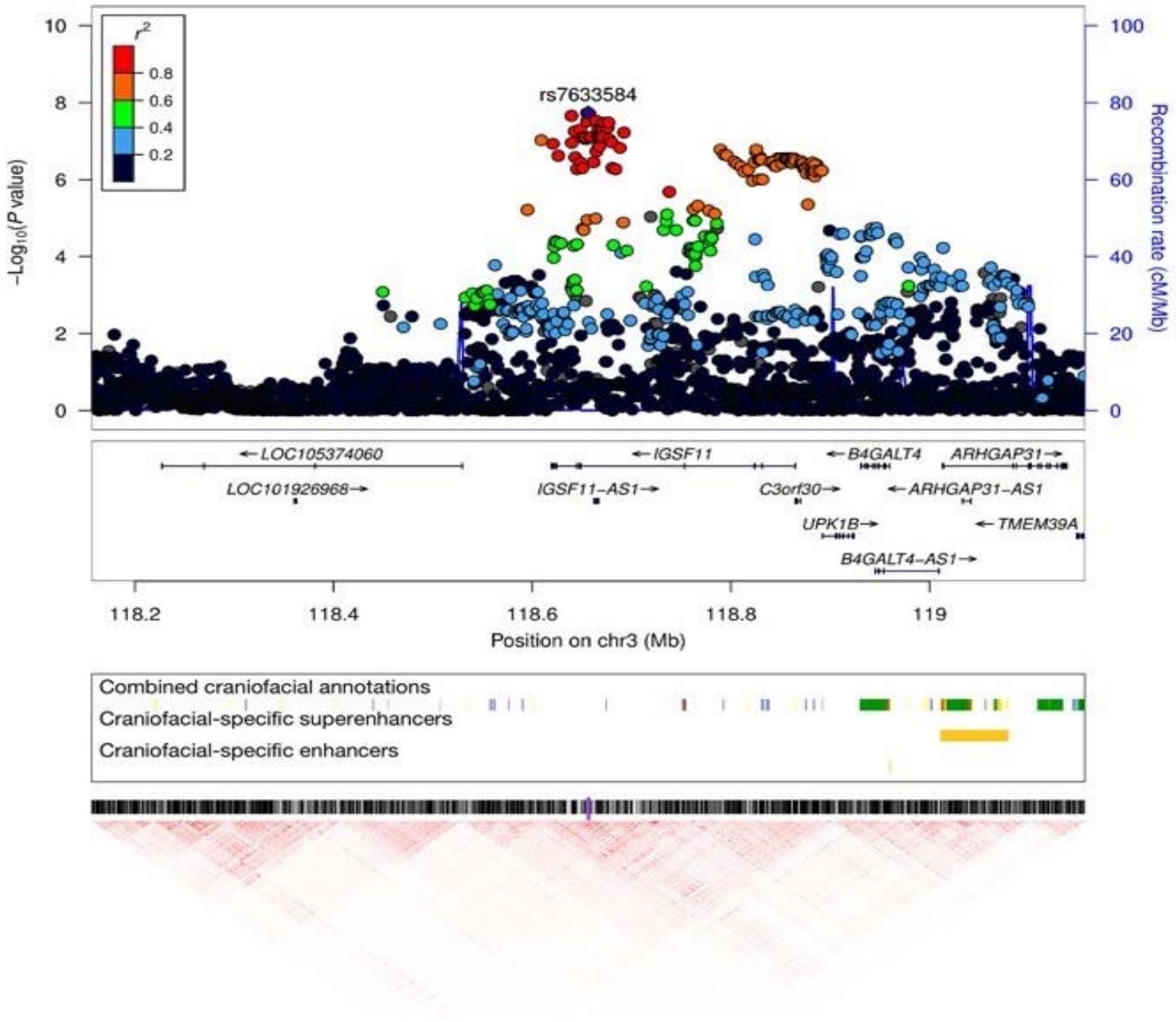
Fig.14 :Preuve d'association, d'introgression de Dénisovien et de sélection dans la région WARS2-TBX15 (Bonfanteet al,2021).

Fig.14.1 Caractéristiques de quatre régions génomiques nouvellement associées à des caractéristiques faciales dans l'échantillon CANDELA et se reproduisant chez les Européens. (Bonfante et al, 2021). Chaque image est subdivisée en trois parties : en haut Manhattan plot et les différents gènes sur le chromosome. Au milieu l'enrichissement des enhanceurs et en bas le déséquilibre de liaison

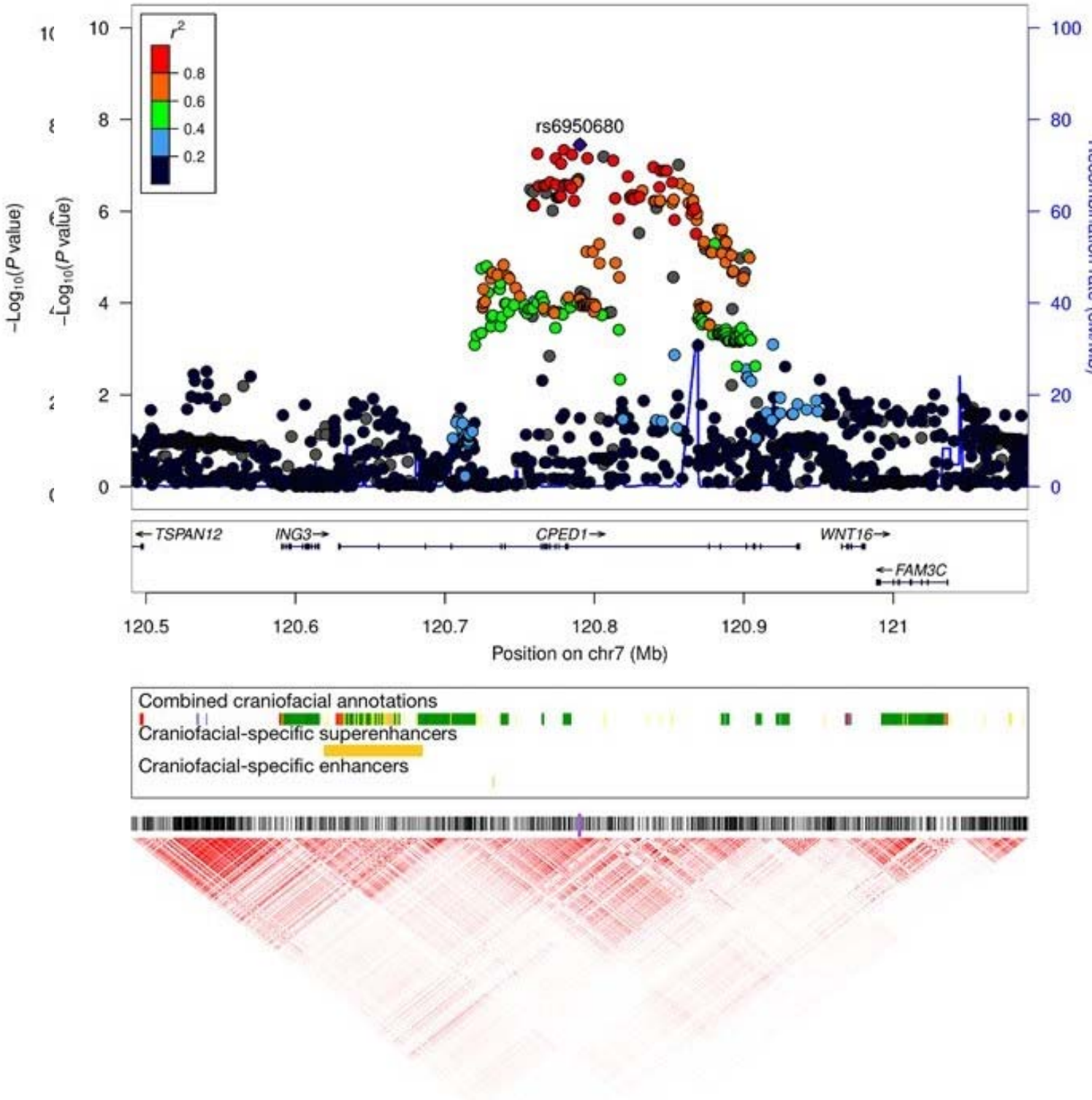
8q22.2 taille de la columelle



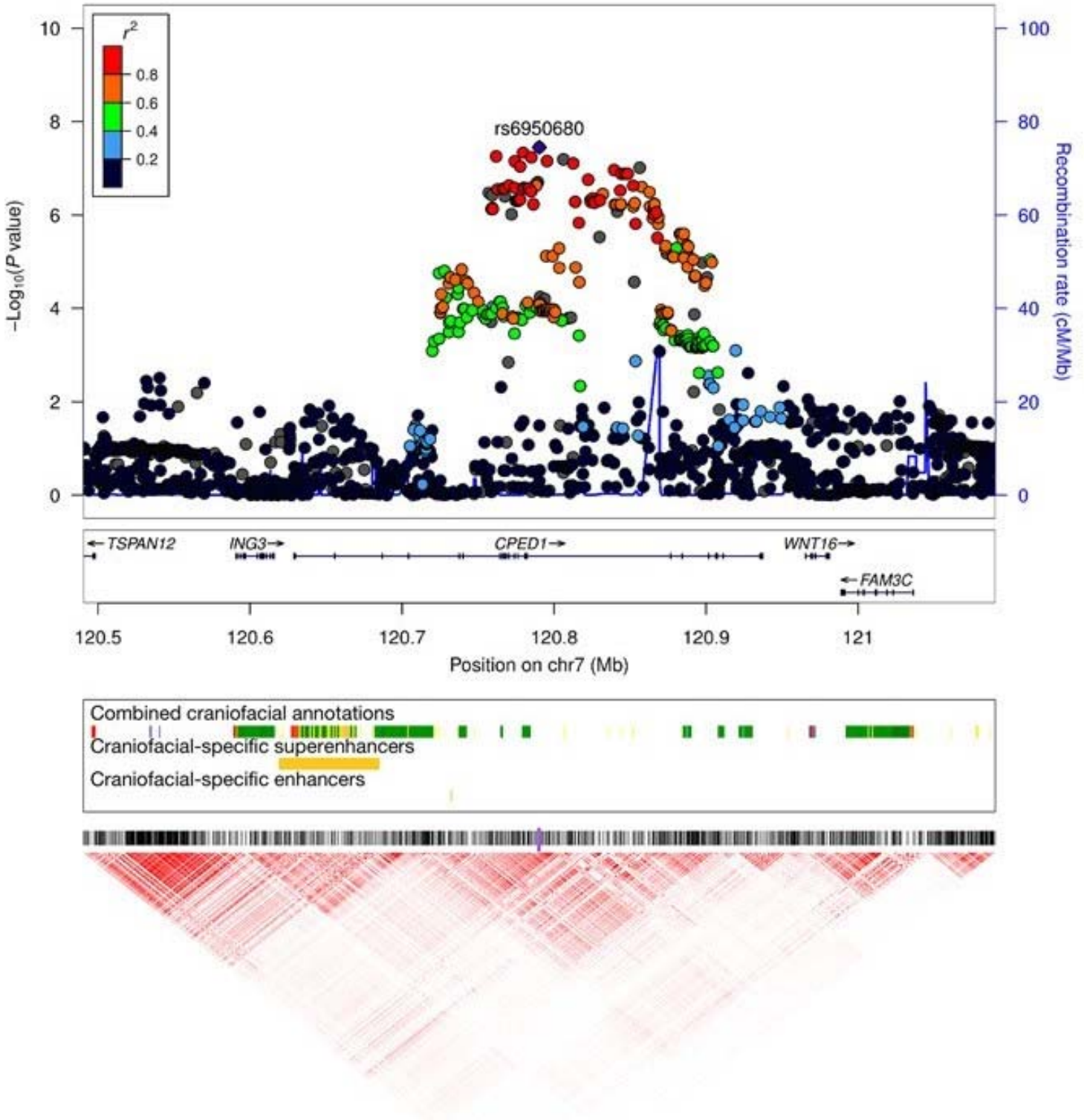
3q13.12rondeur du nez



7p21.1 inclinaison de la columelle



7q31.31 protrusion de la mâchoire



4.2.5 Effet de l'ascendance, l'âge et le sexe sur la morphologie faciale.

L'ascendance, l'âge et le sexe joue un rôle majeur sur certain trait morphologique facial lors de la prédiction d'individu. Ces données peuvent intensifier la configuration des composants principaux lors de de la modélisation. Selon (Claes, et al,2014) le sexe et l'ascendance emphatisaient la plupart des composites faciaux basés sur la ADN.

Afin de fournir une ressource pour évaluer l'ascendance continentale dans une grande variété d'études génétiques, un ensemble de 128 marqueurs informatifs d'ascendance (AIMs) ont été identifiés et validés. Les marqueurs ont été choisis pour l'information, la distribution à l'échelle du génome et la reproductibilité du génotype sur deux plates-formes (TaqMan® des analyses et Illumina). Un panel de référence a été construit à partir du Projet de diversité du génome humain (dit HGDP) et du Projet 1000 Genome puis des mélanges d'ascendances ont été analysés. Les deux références ont été utilisées pour estimer les proportions d'ascendance pour chaque individu.

Les données génotypiques du HGDP (52 populations) et du projet 1000 Genome (26 populations) ont été fusionnées sur la base des positions rsid

de dbSNP. La collection résultante de 57 214 SNP informatifs uniques sur l'ascendance a été utilisée pour l'analyse de l'ascendance en utilisant le logiciel ADMIXTURE 1.23. Au total cinq composantes d'ascendance ont pu être prédit européenne (EUR), africaine (AFR), est-asiatique (EAS), d'Asie centrale et du Sud (CSA) et d'Amérique du Nord (AMR). Ensembles de marqueurs informatifs ancestraux pour déterminer l'origine continentale et les proportions d'admixture dans les populations communes en Amérique (Kosoy et al 2009).

Dans le but de la prédiction d'âge, deux systèmes candidats ont été nominé plus plausible : l'horloge épigénétique et la longueur de télomère.

L'horloge épigénétique ou également l'âge de méthylation (Horvath, 2013 ; Hannum et al., 2013) sont peut-être les prédicteurs les plus fiable, les 2 montrent une corrélation d'âge élevé avec un écart de 3.6 et 4.9 ans respectivement. L'horloge Horvath est un prédicteur multi-tissues basés sur le niveau de méthylation de 353 sites CPG d'illumina 27K tandis l'horloge Hannum utilise 7 sites CPG d'illumina 450K. Cette technique consiste à former une moyenne pondérée de 353 CPG horloge qui est ensuite transformée en âge ADNm en utilisant une fonction étalonnage. (Jylhäv,2017)

Les télomères sont des séquences d'ADN répétitive qui plafonnent les chromosomes qui raccourcissent chaque fois que les cellules se divisent ce qui fait que la longueur des télomères est un marqueur populaire de vieillissement. (Jylhäv, 2017)

Pour prédire le sexe à partir du génome, il est nécessaire d'abord d'estimer le nombre de copies du chromosome X (CCN_chrX) et du chromosome Y (CCN_chrY). Les mâles sont censés avoir une copie du chromosome X, tandis que les femelles sont censées avoir deux copies du chromosome X. Le nombres de copies des chromosomes sexuels permettent de prédire le sexe. La prédiction du sexe basée sur les règles suivantes : les individus avec CCN_chrY $\leq$ 0,25 étaient préditscomme féminins, quelle que soit la valeur de CCN_chrX. Les individus avec CCN_chrY > 0,25 ont été prédits comme étant de sexe masculin. Lors de la prédiction du genre auto déclaré à partir du sexe, les règles basées sur le nombre de copies de chromosomes (CCN) ont obtenu une précision de 99,6 % (Lippertet al, 2017).

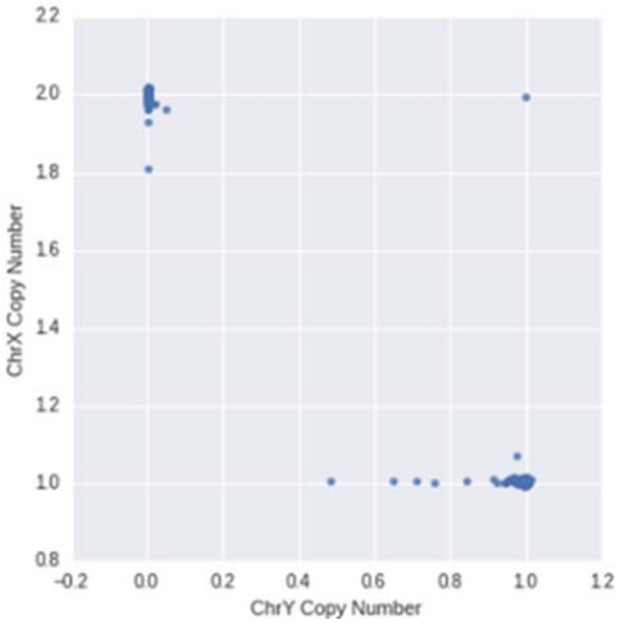


Fig.15 : Représente la distribution des estimations du nombre de copies du chromosome X par rapport au chromosome Y. et obtention des règles finales pour prédire le sexe, (Lippert et al, 2017

5. Troisième étude : Résultat final et discussion

5. Résultats finals :

5.1 Prédiction de profil d'individu connu et inconnu :

Dans les sections précédentes nous avons abouti à la connaissance de multiple SNP qui nous permet la prédiction du visage combiné avec la couleur des yeux, des cheveux et de la peau. Ce dernier a été utilisé dans un panneau regroupant un ensemble de données qui sont ensuite mis dans un modèle prédictif.

Les visages gris sont des visages extrêmes des axes des composants principaux du visage qui sont aux nombre de cinq :

-la forme du visage (ovale-long)

-saillie de la mâchoire (arcade orbitaire, nez, milieu du visage)

-surface globale (convexité, concavité)

-visage inférieur (retranchement, projection)

-taille du nez et projection

Des méthodes pour visualiser et quantifier la différence faciale ont été mise en place afin de pouvoir systématiquement exprimer les effets de variables de prédicteur imputées (RIP) basées sur la réponse particulière sur le visage en résultats anatomiquement interprétables. Ceux-ci sont basés sur la comparaison des visages dans le sens pair, comme la comparaison du RIP-S le plus féminin avec les faces de consensus transformées RIP-S les plus masculines en utilisant trois mesures fondamentales : le rapport de zone, le déplacement normal et le rapport de courbure. Ces deux ratios et un déplacement ainsi que des distances et des angles inter-repères particuliers peuvent ensemble être appelés « paramètres de changement de forme du visage » (FSCPs) (Claes et al, 2014).

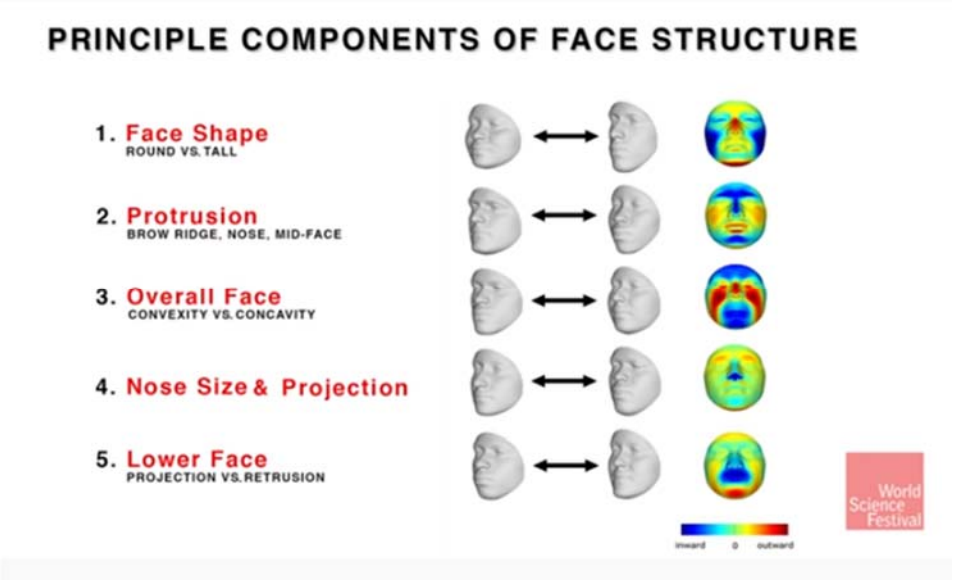


Fig.16 : les structures principales du visage. (shriver,2015)

5.2 Validation croisé et machine-learning : (Apprentissage automatique)

Afin de parvenir à une prédiction il est nécessaire d'apprendre à l'ordinateur à comprendre quelle partie du génome contrôle les différentes parties du facies et de sa coloration, ce qui permet à l'ordinateur de créer un algorithme de prédiction ; l'apprentissage automatique (dit machineLearning). Ce dernier utilise des méthodes informatiques pour "apprendre et mémoriser" des informations directement à partir de données sans se baser sur une équation prédéterminée comme modèle puis ensuite le tester. Ce processus, qui consiste à décider si les

résultats numériques quantifiant les relations hypothétiques entre les variables sont acceptables en tant que descriptions des données, est appelé validation. En général, une estimation de l'erreur du modèle est effectuée après l'entraînement, mieux connue sous le nom d'évaluation des résidus. Dans ce processus, une estimation numérique de la différence entre les réponses prédites et les réponses originales est effectuée, également appelée erreur d'apprentissage. Une fois que l'ensembles des données SNP, age sexe ascendance et méta-analysé global (préexistante dans le système) sont fourni à l'ordinateur,ces derniers sont divisés en training et testing sets pour apprentissage surveillé.

Évaluation de l'importance des caractéristiques en utilisant un algorithme et une foret aléatoire pour classer les SNP selon leurs importances pour la prédiction.Les résultats de SNP connus et nouveaux sont ensuite examinés. La foret aléatoire est entrainer à classer les individus par phénotype en se basant sur les SNP intégré au system.

5.3 : Machine-learning pour la prédiction des trois traits de pigmentation à partir de l'ADN

Actuellement, l'apprentissage automatique ou machine-learning en anglais (ML) est devenu une méthode puissante et largement utilisée pour résoudre les problèmes de classification et de clustering.

Une comparaison systématique de classificateurs d'apprentissage automatique (ML) populaires, a été effectué entre la régression logistique multinomiale (MLR) largement utilisé dans le FDP et trois d'autres, à savoir les machines à vecteurs de support (SVM), la forêt aléatoire (RF) et réseaux de neurones artificiels (ANN), qui ont montré de bonnes performances en dehors de la prédiction EVC. Comme exemples, utilisation del'œil, catégories de couleur de cheveux et de peau en tant que phénotypes et génotypes basés sur les marqueurs ADN IrisPlex, HIrisPlex et HIrisPlex-S Pour la couleur des yeux, la couleur des cheveux et la couleur de la peau, 6 SNP ont été appliqués du modèle IrisPlex pour la prédiction de la couleur des yeux ; les 22 SNP utilisés pour la prédiction de la couleur des cheveux à partir du modèle HIrisPlex, et les 36 SNP appliqués pour la prédiction de la couleur de la peau à partir du modèle HIrisPlex-S, respectivement.

Chaque classificateur nécessite différentes étapes de réglage et des hyperparamètres, les valeurs réglées dépendent à chaque fois de l'ensemble de données d'apprentissage.

Afin de comparer les performances des différents classificateurs de la (ML), pour chaque modèle, ils ont calculé la sensibilité, la spécificité, la valeur prédictive positive (VPP), la valeur prédictive négative (VPN), l'aire sous la courbe (ASC), la matrice de confusion et la précision globale.

Presque tous les modèles de prédiction de pigmentation précédemment établis étaient basés sur la MLR parce que les performances les plus élevées ont été obtenues avec MLR

Aucune des autres méthodes ML n'a surpassé la méthode conventionnelle MLR pour prédire la couleur des yeux, des cheveux et de la peau sur la base des marqueurs ADN IrisPlex, HIrisPlex et HIrisPlex-S, respectivement.

Cependant, pour le moment, et avec les prédicteurs d'ADN de pigmentation établis actuellement disponibles, la MLR reste la méthode de classification préférée pour prédire les traits de pigmentation catégoriques à partir de l'ADN.

Les futures études de prédiction de pigmentation basées sur ML utilisant des listes allongées de prédicteurs d'ADN qui sont déjà disponibles à partir de GWAS à grande échelle pour les

cheveux, la couleur de la peau et la couleur des yeux pourraient améliorer les performances globales de prédiction de l'apparence. (Maria-Alexandra et al,2021).

5.4 Formation de modèles prédictifs par validation croisée :

Évaluation de chaque modèle prédictif à l'aide de k-fold Cross Validation. Les données sont divisées en training data 80% et testing 20%. Les assignations sont aléatoiresà partir des données d'origine.

Étant donné que le CV utilisé est de type k-fold, le training data sont attribué aléatoirement a des sous-ensembles.

Après comparaison de plusieurs modèle linaire, le ridge regression et le K-nearest neighbour ont été choisi étant simple et efficace sur le plan informatique. K a été fixé à 10 pour chaque plis k-1 a été utilisé pour trainings et et le reste au testing set de sorte que chaque individu a été utilisé au training 9 fois et une fois au testing. Pour chaque répétition., des ajustements de réglages pour modèle prédictif ont été choisis et cela à partir de 5-plis nested CV durant la phase de training. Après que les paramètres de réglages sont déterminés, le modèle subit un réentrainement en utilisant l'ensemble des modèles des trainings sets. Le modèle entrainé final est utilisé pour prédiction sur le testing set.Enfin pour chaque plis la prédiction est évalué en utilisant une métrique de qualité qui est évalué au testing set.

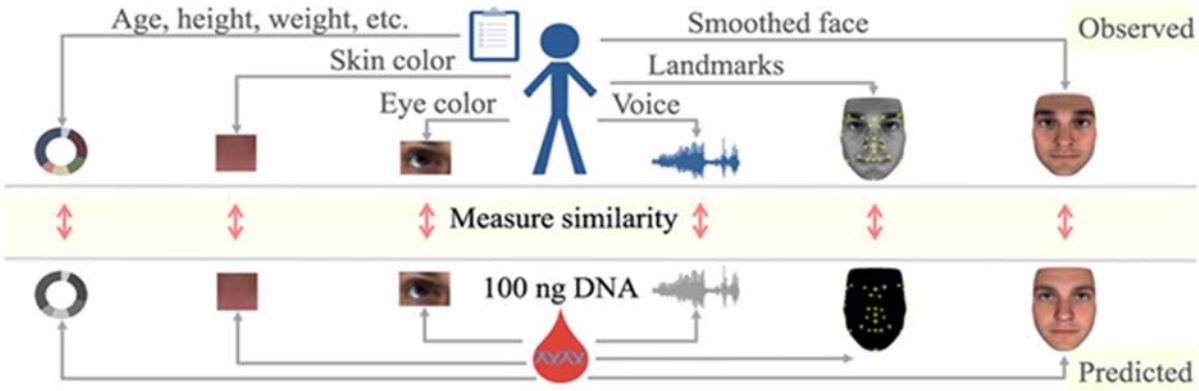


Fig.17. Aperçu de l'approche expérimentale (Lippert et al,2017)

Pour évaluer le rendement du processus et la re-idendification un scénario a été mis au point qui teste la probabilité d'inclure la vraie personne dans un sous-ensemble de 10 personnes d'un pool aléatoire de 100 personnes choisi parmi un ensemble d'échantillon. La fig.18 présente la capacité à assurer qu'une personne est au sommet M à partir d'un pool de taille N>M. L'individu correct a été classé dans le haut M= 10 de N= 100, 88% du temps, montrant la capacité d'enrichir pour les personnes d'intérêt.

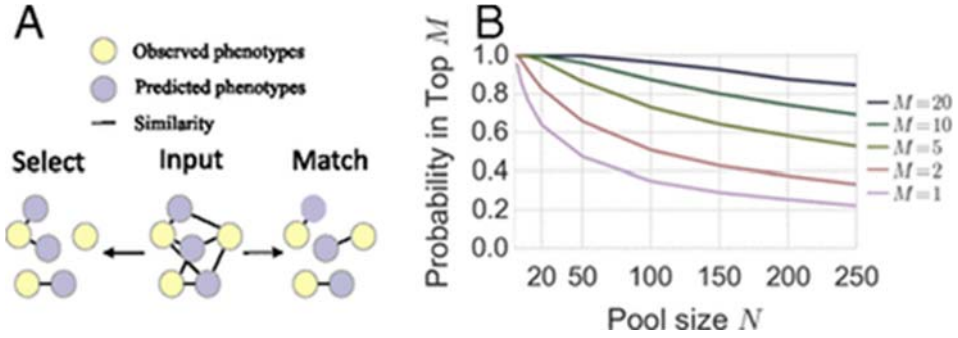


Fig.18 : représentation schématique entre l'individu sélectionné et le sous-ensemble.(Lippert et al,2017)

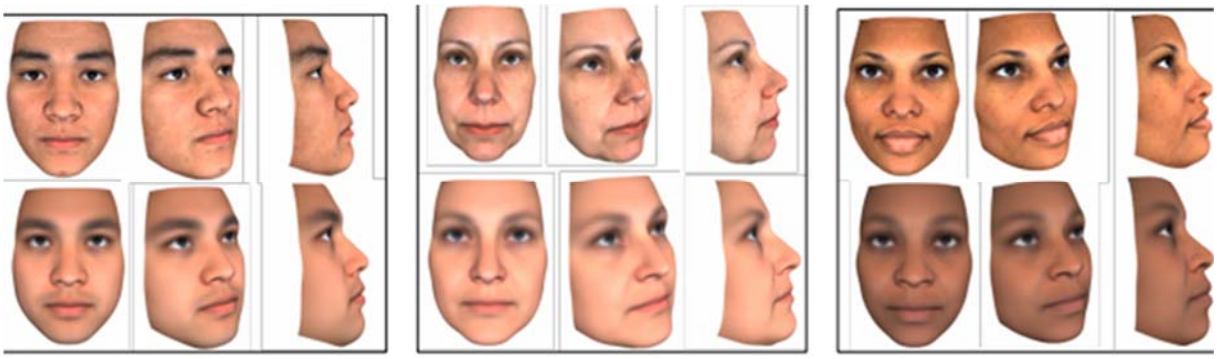


Fig.19. Comparaison entre un Scanner et le résultat d'unePrédiction 3D. La ligne supérieure de chaque panneau représente le visage observé (rotation de 0, 45 et 90 degrés) et la ligne inférieure représente le visage prédit (rotation de 0, 45 et 90 degrés). b.90 degrés), et la ligne du bas de chaque panneau représente le visage prédit (0 degré, 45 degrés et 90 degrés de rotation). (Lippert et al,2017)

5.5 Prédiction de l'inconnu par Parabon Nanolab :

Lorsqu'il s'agit de la prédiction d'un individu inconnu c'est a dire qu'il ne figure pas dans le fichier national d'empreinte génétique le travail est envoyé a des société privé telle que Parabon Nano Labs En utilisant des données génomiques provenant de grandes populations de sujets connus, Snapshot produit des modèles statistiques qui identifient les régions fortement associées à certains traits médico-légaux. Il utilise ensuite ces modèles pour prédire l'apparence physique d'individus inconnus En commençant par de grands ensembles de données composés d'un phénotype (trait) d'intérêt et de données génotypes pour des milliers de sujets, l'équipe de bio-informatique effectue une analyse statistique à grande échelle sur des millions de SNP individuels et des milliards de combinaisons de celles-ci pour identifier les ensembles de ces marqueurs génétiques qui s'associent au trait donné. Ce processus d'extraction peut prendre des semaines sur des centaines, parfois des milliers, d'ordinateurs. En fin de compte, les SNP qui ont le plus de chances de contribuer à la variation observée dans le caractère cible sont recueillis pour être utilisés dans des modèles prédictifs. La phase de modélisation affine encore cet ensemble de SNP à un ensemble final qui prédit le plus précisément le trait cible dans un cadre d'algorithmes d'apprentissage automatique. Les modèles sont validés par rapport aux données retenues pour ces tests et calibrés avec toutes les données disponibles avant d'être installés dans l'architecture Snapshot.

Bien que l'ADN puisse en révéler beaucoup sur l'apparence d'un sujet, les informations sur des caractéristiques telles que l'âge, l'indice de masse corporelle (IMC) ou la présence de poils faciaux ne sont pas disponibles dans le code génétique d'une personne Le test Snapshot de Parabon ne permet pas encore de prédire l'âge de l'individu inconnu. Tous les composites sont générés pour apparaître à l'âge de ~25 ans avec un IMC moyen. Des progressions d'âge ou d'autres modifications peuvent être effectuées sur l'image composite à la demande de l'agence cliente, sur la base des descriptions des témoins ou du temps qui s'est écoulé depuis le crime.

Les services d'art médico-légal instantané fournissent un moyen d'intégrer ces informations dans un composite instantané lorsqu'elles sont disponibles à partir de sources non génétiques

tel progression de l'âge, d'altération de l'IMC et d'accessoirisassions, qui peuvent inclure l'ajout de poils du visage, lunettes, piercings,

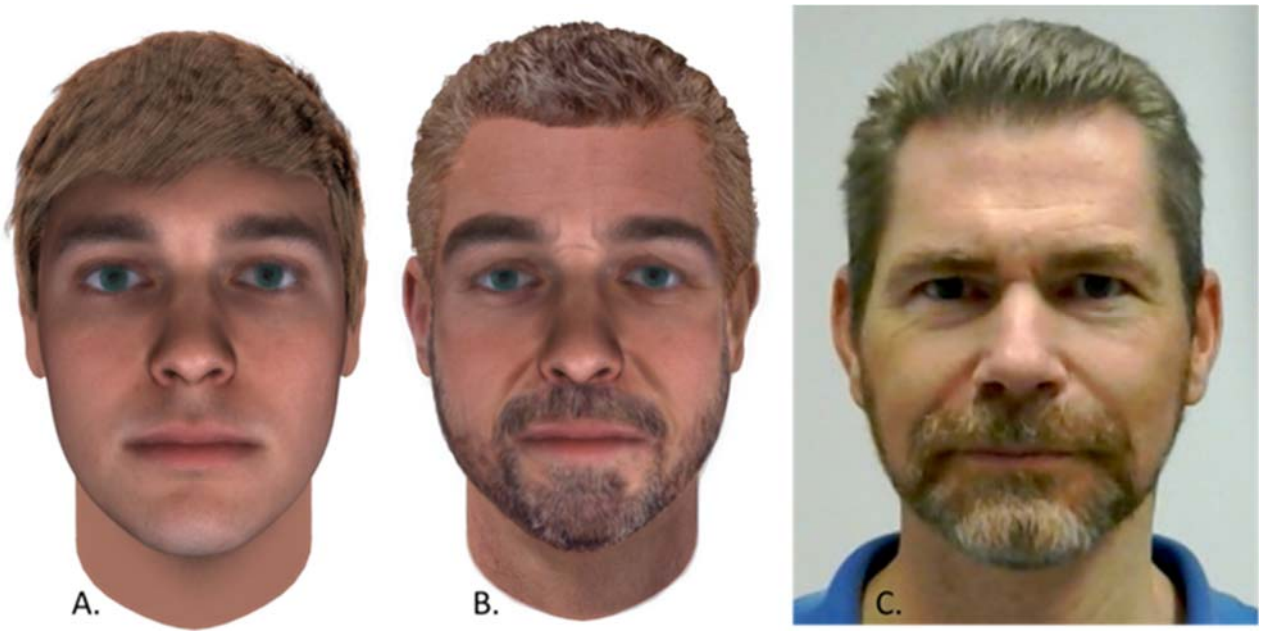


Fig.20.Progression de l'âge par rapport à une personne réelle
Composite (A) montré après la progression de l'âge à 50 ans, y compris une barbe (B) par rapport au sujet réel (C) (http://snapshot. parabon-nanolabs.com/artwork)

Les capacités de Parabon suscitent le scepticisme. Il est difficile d'évaluer le système de Parabon car le code informatique n'est pas ouvert, et la méthodologie n'a pas été publiée avec un examen complet et minutieux.

Discussion :
En 1991 Atchley et Hall ont résumé l'un des problèmes majeurs de la biologie contemporaine en disant qu'il fallait "comprendre comment des structures morphologiques complexes apparaissent au cours du développement et comment elles se développent" Dans leurs travaux, ces auteurs décrivent une "chorégraphie complexe du développement" dans laquelle les facteurs génétiques intrinsèques, les facteurs épigénétiques et les interactions entre les deux constituent le génotype de la progéniture. Ce dernier s'engage avec l'environnement pour finalement produire un trait morphologique complexe composé d'éléments distincts .Nous savons maintenant que les facteurs génétiques intrinsèques qui contribuent en fin de compte aux traits morphologiques complexes ne consistent pas seulement en des variantes uniques modifiant la structure et/ou la fonction des protéines, mais aussi en des variantes non codantes et des interactions entre les variantes, chacune affectant de multiples tissus et point de développement.
L'équipe de recherche dirigée par Dr claes, (Claes et al, 2018) a précédemment décrit dans l'article Genome-wide mapping of global-to-local genetic effects on human facial shape une approche du phénotypage du visage basée sur une approche guidée par les données qui a facilité l'identification et la réplication de 15 loci impliqués dans la variation morphologique du visage Dans cette étude, une application de techniques multivariées permet de découvrir de nouvelles perspectives biologiques dans l'architecture génétique du visage humain. Il est maintenant possible d'identifier 203 Signaux significatifs à l'échelle du génome (120 également

significatifs à l'échelle de l'étude), situés dans 138 bandes cytogénétiques, associés à une morphologie faciale normale multivariée

Des recherches approfondies ont élucidé l'association des variantes d'ADN avec les traits de pigmentation humaine. La capacité de prédire les aspects de l'apparence d'un individu inconnu à partir de son ADN à l'aide des SNP EVC pourrait fournir des informations comparables à la déclaration d'un témoin oculaire. Plusieurs approches, telles que les méthodes de rapport d'association de vraisemblance, un processus de guide de prédiction et un classificateur bayésien basé sur des rapports de vraisemblance sont disponibles pour prédire la couleur des yeux. Cependant, le système IrisPlex qui utilise une approche de prédiction de probabilité basée sur un modèle pour la prédiction de la couleur des yeux bleus et bruns s'est imposé comme l'outil le plus recherché dans la communauté médico-légale pour déterminer la couleur des yeux la plus probable.

Les deux premières études qui ont effectué l'iris (œil) à base d'ADN prédiction de couleur des yeux ont été publiées en 2007 par Frudakis et al. Utilisent 33 SNP du gène OCA2, ce qui a leur permis de classer 8%des couleurs des yeux observés parmi >1000 échantillons. Sulem et al, intègrent dans le premier GWAS sur les traits de pigmentation humaine, ont utilisé 9 SNP de 6 régions génomiques (SLC24A4, KITLG, 6p25.3, TYR, OCA2–HERC2 et MC1R) qu'ils ont identifiées avec des associations de la couleur des yeux parmi plusieurs milliers d'Européens, pour prédiction catégorique de la couleur des yeux.

En 2008, trois études parallèles ont signalé le gène HERC2 comme le gène de la couleur des yeux le plus important, (Sturm et al ; Eiberg et al) Ont mis en évidence HERC2 rs12913832 comme prédicteur majeur de la couleur des yeux. Les auteurs ont obtenu la prévalence ajustée la précisions moyennes des prédictions exprimées en aire sous la courbe caractéristique de fonctionnement du récepteur (AUC) d'environ 0,93 pour marron et 0,91 pour la couleur des yeux bleu, tandis que pour la couleur des yeux intermédiaires, l'AUC était considérablement plus petite à 0,73., (où 0,5 signifie prédiction aléatoire et 1.0 signifie une prédiction complètement précise); la plus grande partie de la valeur prédictive de la couleur des yeux a été fournie par leHERC2 rs916977 seul.

Le système IrisPlex est adapté pour prédire avec précision les couleurs des yeux bleus et bruns, mais est incapable de prédire les couleurs des yeux intermédiaires/verts au même degré en raison de limitations générales dans l'explication génétique de ces couleurs d'yeux non bleus et non bruns avec des marqueurs d'ADN existants. Ce test est très sensible, fournissant 6- Profils SNP jusqu'à environ 30 pg d'entrée d'ADN (voir généralité).

Depuis mai 2012, le gouvernement néerlandais a autorisé l'utilisation du système validé de prédiction de la couleur des yeux IrisPlex, qui est effectué à l'Institut médico-légal des Pays-Bas (NFI). La validation médico-légale du développement du test IrisPlex a été publiée la même année, démontrant que le test est entièrement compatible avec toutes les directives SWGDAM.

Le premier système de test ADN pour prédire toutes les couleurs de cheveux catégoriques en combinaison avec prédiction catégorique de la couleur des yeux, a été développé et publié en 2013. Ce système HIrisPlex comprend un seul test multiplex de génotypages de 24 SNP prédictifs de la couleur des yeux et des cheveux, comprenant les 6 d'IrisPlex, ainsi que deux modèles de prédiction, un pour la couleur des cheveux et le modèle IrisPlex précédent pour la couleur des yeux, ont été utilisés pour la prédiction de la couleur des cheveux. (Voir généralité).

Le test HIrisPlex a fourni des profils complets de 24 SNP jusqu'à environ 60 pg d'entrée d'ADN. Avec ce modèle, les valeurs d'ASC de 0,92 ont été obtenu pour le rouge, 0,85 pour le noir, 0,81 pour le blond et 0,75 pour le brun comme couleur des cheveux basé sur >1600 individus.

En 2014, Une étude de validation du test HIrisPlex a été publiée, démontrant qu'il est entièrement compatible avec tous les SWGDAM des lignes directrices.

Les auteurs ont concentré sur l'adéquation du système HIrisPlex en tant qu'outil robuste pour récupérer des informations sur la couleur des yeux et des cheveux à partir d'échantillons d'ADN provenant de restes humains âgés de près de 70 ans et enterrés pendant la majeure partie de cette période. Dans ce cas le système HIrisPlex a été précédemment appliqué à des échantillons d'ADN extraits de restes squelettiques anciens et contemporains. La faisabilité du test HIrisPlex a été illustrée par l'analyse de la dent prélevée sur le cadavre du général Władysław Sikorski., personnage historique de l'histoire polonaise décédé en 1943 dans un accident d'avion. La couleur des yeux bleus et des cheveux blonds obtenue à partir de HIrisPlex a été confirmée positivement par des rapports documentés fiables. Récemment, les marqueurs ADN HIrisPlex et les modèles de prédiction ont été également utilisé pour prédire la couleur des yeux bleus et les informations sur la couleur des cheveux blonds à partir des restes squelettiques identifiés comme étant ceux du roi Richard III d'Angleterre (1452-1485), les auteurs ont révélé pour le squelette une probabilité de 96 % d'avoir les yeux bleus ainsi qu'une probabilité de 77 % d'avoir les cheveux blonds. De plus, deux échantillons squelettiques provenaient de deux frères, comme le confirme le génotypages STR lorsqu'ils sont analysés avec l'échantillon de référence d'une sœur vivante. La sœur vivante des deux frères a confirmé la couleur des yeux et des cheveux prévus par HIrisPlex.

Il convient de noter qu'il n'est pas toujours possible d'obtenir des résultats satisfaisants lorsque le système HIrisPlex est appliqué à des échantillons très anciens et dégradés, ainsi que les conditions de conservation et de stockage de l'échantillon influencent également les résultats obtenus avec le système HIrisPlex. De plus comme la couleur des cheveux est un trait plus variable et complexe, par conséquent, sa prédiction présente plus de difficultés. L'un des problèmes apparents est le changement de couleur des cheveux en fonction de l'âge, passant du blond pendant la petite enfance au blond foncé ou au brun à la fin de l'enfance et à l'âge adulte. C'est la principale raison pour laquelle les blondes et les la couleur des cheveux bruns est prédite avec moins de précision avec le système HIrisPlex que le rouge et le noir. Un autre problème avec relativement peu de connaissances est la perte de couleur des cheveux en fonction de l'âge, c'est-à-dire le blanchiment ou le grisonnement. Récemment, il a été découvert que le SNP rs12203592 dans le gène IRF4 est associé au grisonnement des cheveux.

Avec le développement de systèmes de prédiction de la couleur des yeux et des cheveux tels que IrisPlex et HIrisPlex, la prédiction des EVC est devenue populaire et l'étape logique était d'ajouter la couleur de la peau. Plusieurs études ont publié des SNP associés à la couleur de la peau. En utilisant les connaissances disponibles sur les SNP associés à la couleur de la peau, une extension du système HIrisPlex a été développée.

Le système HIrisPlex–S (S pour skin) consiste en un deuxième ensemble de 17 SNP ciblés avec un multiplex, en plus au multiplex ciblant les 24 SNP de HIrisPlex, et trois modèles de prédiction pour la couleur des yeux, des cheveux et de la peau, respectivement. L'échelle Fitzpatrick qui est la classification numérique de la couleur de la peau humaine a été utilisée pour le phénotypage. Les six catégories de l'échelle Fitzpatrick (Type I-Type VI) ont été

classées en cinq catégories pour le test HIrisPlex-S : très pâle, pâle, intermédiaire, foncé et foncé/noir.
La prédiction de la couleur de la peau est plus complexe car sa variation se fait entre les populations de différents continents tels que l'Europe, l'Asie, l'Afrique, alors que la couleur des yeux et des cheveux varie principalement dans les populations européennes.
Des études sur la couleur de la peau à l'aide de GWAS peuvent être effectuées au sein de groupes continentaux, mais comme la variation entre les groupes continentaux était moindre, la liste des gènes est également assez limitée. Cependant, le modèle de prédiction de la couleur de la peau du HIrisPlex-S a atteint une précision de 0,75. Pour très pâle, 0,73 pour pâle, 0,75 pour moyen, 0,84 pour foncé et 0,98 pour foncé/noir exprimé en AUC.
Un total de 36 SNP de 15 gènes ont été utilisés pour la prédiction de la couleur de la peau. Ces 36 marqueurs comprenaient les 17 SNP de l'ensemble nouvellement développé et 19 des 24 SNP (à l'exclusion de N29insA, rs1805005, rs1805009, Y152OCH et rs4959270) du test HIrisPlex.
Nous présentons ici des solutions de séquençage massivement parallèle (MPS) pour le système HIrisPlex-S sur deux plates-formes MPS couramment utilisées en médecine légale, Ion Torrent et MiSeq, qui couvrent les 41 variantes d'ADN dans un seul test, respectivement. De plus, nous présentons la validation médico-légale du développement des deux tests HPS-MPS. Le test Ion Torrent MPS, basé sur la technologie Ion AmpliSeq, a illustré la génération réussie de profils génotypiques HIrisPlex-S complets à partir de 100 pg d'ADN de contrôle d'entrée, tandis que le test MiSeq MPS basé sur une conception interne a produit des profils complets à partir de 250 pg d'entrée d'ADN.
Comme pour l'IrisPlex et le HIrisPlex, la détection et l'interprétation du mélange étaient difficiles. Il a été possible de déduire la présence probable d'un mélange avec des rapports de mélange, mais le typage STR est toujours la méthode de choix pour l'interprétation du mélange avant d'autres tests.
De plus, des résultats précis ont été obtenus à partir d'échantillons médico-légaux simulés tels que la salive, les cheveux (bulbe), le sang, le sperme et l'ADN tactile(traces) en faible quantité, ainsi que des échantillons d'ADN artificiellement endommagés, des tests de concordance et des échantillons de nombreuses espèces, ont révélé la robustesse, l'efficacité et la capacité des deux versions du Test HIrisPlex-S MPS pour produire des résultats qui motivent les applications médico-légales. Un pipeline d'analyse bio-informatique intégré, les données MPS peuvent désormais être analysées et un fichier généré pour être téléchargé sur l'outil Web en ligne HIrisPlex-S accessible au public (https://hirisplex.erasmusmc.nl).
Enfin, avec cette étude, nous prévoyons que le système HIrisPlex-S, peut être appliqué à des fins d'identification des victimes de catastrophes et des personnes disparues.
Ceci dit la présente étude et les applications du système HPS, peuvent nous permettre de conclure que l'outil HIrisPlex-S pourrait être utilisé pour répondre à des questions non seulement en médecine légale, mais aussi dans une perspective historique, anthropologique et même évolutive, car il commencé à faire déjà. (Kayser et al ,2015 ; Chaitanya ,2016).
La construction de modèle prédit est basée par l'identification de la relation génotype-phénotype qui est souvent réalisée par GWAS, où les SNP identifiés avec des associations significatives sont considérés comme des marqueurs candidats pour la construction de modèle. L'étape de construction de modèle est basée généralement sur des statistiques (modèle de régression), ou des techniques d'apprentissage automatique (machine-learning) par ses différents classificateurs. La validation de modèle se fait par la comparaison des valeurs de

résultats prédites aux valeurs réellement observé pour estimer la précision de la prédiction.et ceci se fait grâce à une société privée Parabon Nanolabs, les scientifiques de Parabon utilisent des algorithmes d'apprentissage automatique pour combiner l'ensemble sélectionné de SNP en une équation mathématique complexe pour l'architecture génétique du trait. Les données SNP d'un nouvel individu inconnu peuvent ensuite être connectées à cette équation pour produire une prédiction du trait chez cet individu.(http://www.parabon-nanolabs.com)

Témoignage 01 :

En 1997 une étudiante en photographie de 26 ans, a été retrouvée sur son lit à l'intérieur de son appartement de Costa Mesa. Elle avait été violée et étranglée. Les détectives ont parcouru la scène, recueillant plus de 130 échantillons d'ADN et 265 éléments de preuve. Mais n'ont pas retourné de correspondance. 20 après le meurtre, les détectives ont décidé de rouvrir l'affaire. Utilisant des preuves ADN recueillies sur les lieux, Parabon NanoLabs a fourni à la police un profil instantané d'un suspect possible.

Les estampes correspondaient à celles recueillies auprès de Hernandez Tellez lors d'une arrestation dans une affaire de violence familiale en 2000. Il a ensuite été condamné, Beckman un des officiers de police a déclaré ¨La ressemblance entre l'instantané composite et la photo de réservation d'Hernandez Tellez était étrange¨ du sens qu'ils sont tout à fait identiques. (Fry,2017).

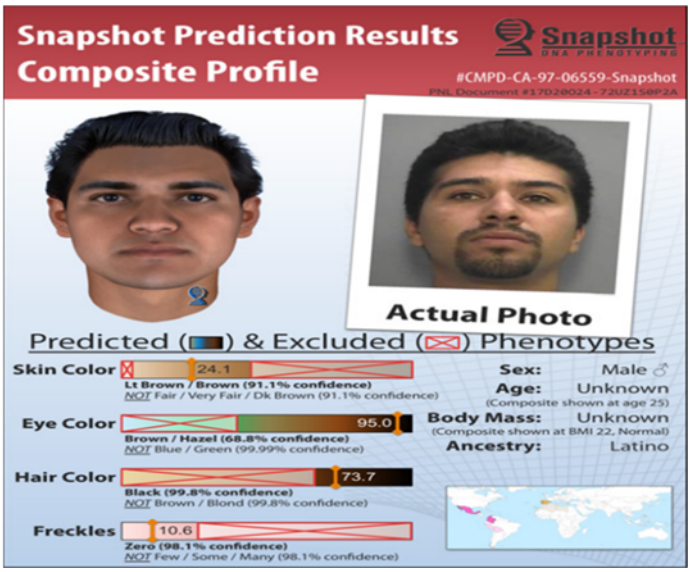


Fig.21. Résultat d'une prédiction Snapshot.
(http://www.parabon-nanolabs.com)

Une autre préoccupation générale est que les traits physiques visibles peuvent ne pas toujours être authentiques ou inhérents, car les individus peuvent modifier leurs caractéristiques physiques. Il existe plusieurs façons de modifier l'apparence physique d'un individu, telles que l'utilisation de lentilles de contact colorées, la teinture des cheveux, la coiffure, le bronzage solaire et l'auto bronzage de la peau, les chirurgies esthétiques, entre autres. Cependant, il convient de noter qu'une fois l'apparence simulée, les auteurs devraient conserver l'apparence fabriquée tout au long, au moins jusqu'à la fin de l'enquête pour rester à l'abri des soupçons. De plus, tous les documents contenant des informations sur les EVC, tels que les passeports, les cartes d'identité, les permis de conduire doivent être falsifiés pour correspondre à l'apparence simulée, ce qui n'est pas très facile et parfois peu pratique.

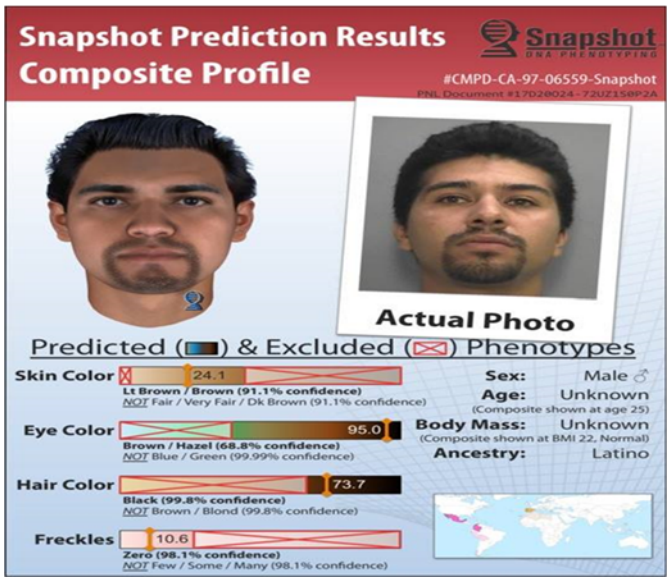


Fig.21.1. L'impact de la pilosité facial
(Modifiée par nous-mêmes).

Témoignage02 :

Le 7 aout 1986, des restes ont été découverts par des travailleur d'enfouissement de déchets de Chersterfield ÉU qui déchargeaient les déchets d'une ancienne station de transfert de la rue Scholl dans la ville. La police n'a récupéré qu'une jambe, un pied et un torse dans la décharge. La tête et les mains de la victime étaient portées disparues, ce qui compliquait la capacité des enquêteurs à identifier la personne. En raison du manque de technologie pour élaborer le profil génétique de ces restes l'affaire a pris du retard et est devenue une affaire classée.
Deux mois au paravent une adolescente Christy Lynn Floyd de 16 ans a été porté disparu dans la région de Chesterfield. À ce temps la police avait rejeté la disparition comme un fugueur. La police de l'époque croyait que la benne à ordures avait été ramassée à l'origine derrière ce qui était alors le bâtiment Emrick Chevrolet, qui se trouvait à moins de 3km de l'appartement de la mère de la disparu dans le bloc 2300 de Grace Street. Mais personne n'avait fait le lien.
Ce n'est qu'en 2019 les détectives ont fait appel aux services d'une société appelée Parabon NanoLabs. Celle-ci a développé un « Snapshot » de ce à quoi ils pensaient que la personne aurait ressemblé en fonction de l'âge, du sexe, des caractéristiques faciales, de la race, jusqu'à la tache de rousseur et cela depuis les restes retrouvés. Le détective Chris Humphries responsable de l'affaire a déclaré que dans les 30 minutes qui ont suivant la publication du communiqué de presse, le téléphone a littéralement explosé avec des appels téléphoniques de tout le pays. L'un des premiers appels qu'ils ont reçus provenait d'un détective à la retraite de Richmond, Mark Williams, que la famille Atkins (famille de la victime) avait contacté il y a plusieurs années Un échantillon d'ADN a été fourni par le neveu de la victime, espérant qu'il correspondrait à sa tante. Des tests médico-légaux supplémentaires effectués par le laboratoire de Floride ont confirmé que les restes étaient ceux de Floyd. (Richmond ,2020)

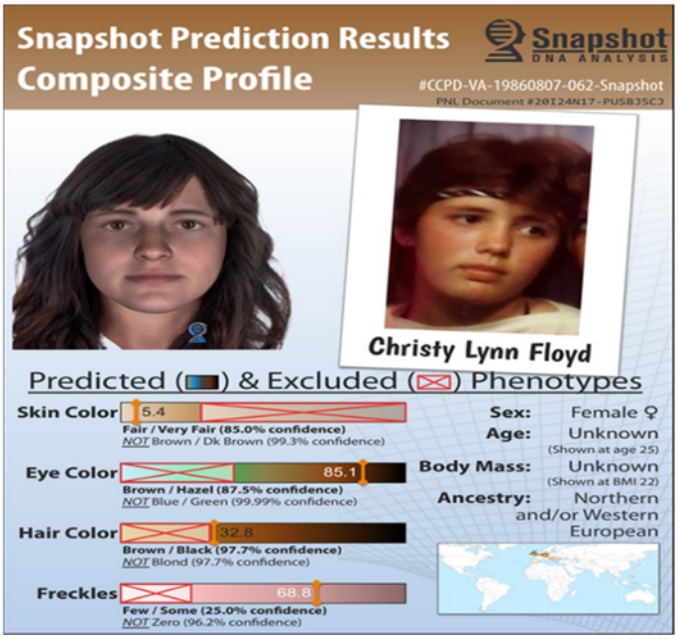


Fig.22. Profil Snapshot prédictif de ce que Christy Lynn Floyd aurait pu ressemblé (http://www.parabon-nanolabs.com)

Témoignage 03

En juin 1977 le corps en décomposition d'une femme nue a été retrouvé dans un fossé. La femme, qui n'avait aucune pièce d'identité, était probablement là depuis des semaines, a déclaré un médecin légiste à l'époque. Elle avait entre 40 et 55 ans et mesurait 5 pieds 3 pouces avec des cheveux blonds foncé. Après une enquête approfondie, l'affaire s'est enrayée. Pendant trois décennies, la police n'a fait aucun progrès dans l'affaire. Environ une décennie plus tard, en février 2017, la police du comté de New Castle a soumis l'ADN a Parabon NanoLabs Les détectives ont envoyé les informations sur la femme à un artiste médico-légal de Parabon Nanolabs, qui a créé un croquis. Le laboratoire a également produit une image numérique de Heiser et envoyé des informations génétiques aux bases de données d'ascendance pour créer un arbre généalogique potentiel. La police du comté de New Castle a suivi ces pistes et obtenu des échantillons d'ADN de proches parents." Le médecin légiste en chef, le Dr Gary Collins, a confirmé l'identité de Heiser grâce à l'analyse et a également confirmé la décision rendue par un médecin légiste du Delaware 40 ans plus tôt :
mort de Heiser était un homicide (hughes ,2021)

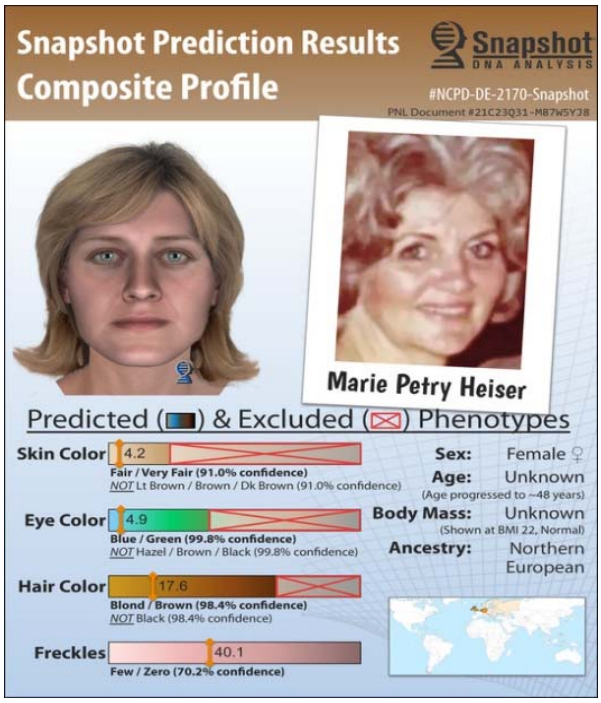


Fig.23 :Portrait prédictif de Heiser (victime)
https://snapshot.parabon-nanolabs.com/posters

Témoignage 4 :

BROWNWOOD, Texas - Le suspect de meurtre Ryan Derek Riggs n'a jamais été sur le radar jusqu'à ce qu'un croquis composite du tueur - tiré d'une nouvelle méthode qui utilise l'ADN pour prédire les apparences physiques, a déclaré jeudi le shérif du comté de Brown, Vance Hill. Riggs, 21 ans, est détenu à la prison du comté de Brown pour meurtre. Il aurait agressé sexuellement puis tué Chantay Blankinship, résidente de Lake Brownwood. Hill a déclaré que son bureau et le bureau du procureur de district ont dépensé 4 000 $ pour embaucher une entreprise avec un laboratoire en Virginie pour utiliser l'ADN pour générer l'image - et que 4 000 $ sont le meilleur argent que son bureau n'ait jamais dépensé. Les enquêteurs affirment que Riggs a jeté le corps de la femme de 25 ans dans une zone rurale. Une autopsie a déterminé qu'elle était morte des suites d'un « larynx écrasé avec des blessures contondantes au cou, au visage et au torse ». Une lame de tondeuse à gazon a été trouvée sur les lieux du crime. Les blessures à la tête et au torse étaient compatibles avec l'utilisation d'une telle lame, a déclaré Hill. Riggs a avoué avoir tué Blankinship lors d'un service religieux mercredi. Ses parents l'ont ensuite conduit au centre d'application de la loi et l'ont remis, a déclaré Hil Lors d'une conférence de presse, Hill a crédité Parabon Nanolabs en Virginie d'avoir joué un rôle clé dans la résolution d'un meurtre survenu il y a 18 mois, le 15 mai 2016. Le laboratoire utilise une méthode connue sous le nom de phénotypage pour déterminer la couleur de la peau, des yeux et des cheveux, ainsi que l'ascendance. Dans les heures qui ont suivi la publication du croquis composite, Hill a déclaré qu'ils avaient plusieurs suspects - y compris Riggs. Hill a déclaré que le sergent Scott Bird, du bureau du shérif, a vu la nouvelle technique et technologie de phénotypage dans une émission de télévision criminelle. Bird ne se souvenait pas du nom de l'émission. L'ADN provenait du sperme et de la lame de la tondeuse à gazon, a déclaré Hill. Un profil a été produit en quatre semaines environ. Pendant ce temps, Hill a déclaré que les autorités avaient déjà eu des contacts avec Riggs dans le cas d'un déversement illégal dans la région. Selon les autorités, les parents de Riggs lui avaient donné de l'argent pour aller jeter les ordures, et il a choisi de les jeter illégalement. Les ordures provenaient de la maison de Riggs. L'affaire a été présentée à un grand jury du comté de Brown jeudi.

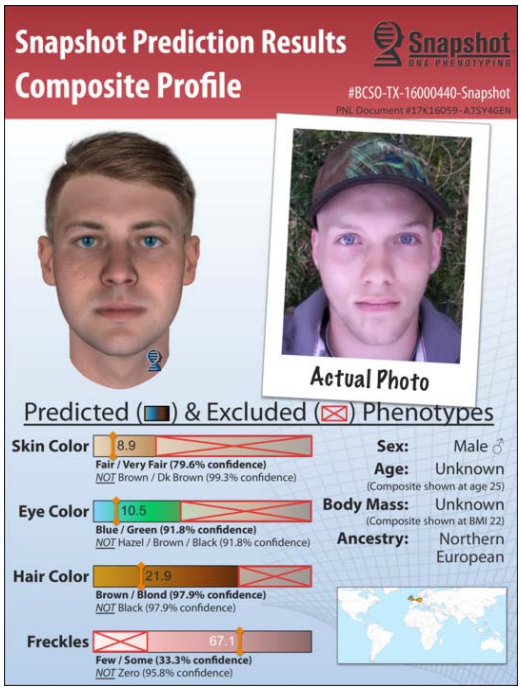


Fig.24 : portrait prédictif de Ryan Derek Riggs (suspect)
https://snapshot.parabon-nanolabs.com/poster

Conclusion :

La façon dont l'ADN code nos EVC peut être différente chez les personnes de différents groupes d'échantillon. Actuellement, la capacité de prédire les Européens modernes sera meilleure que celle d'autres groupes parce que la majorité des bases de données génétiques sont dominées par des sujets d'ascendance européenne. Alors que l'utilisation des approches d'apprentissage automatique de plus en plus sophistiquées sur des bases de données plus grandes (et plus représentatives sur le plan ethnique), la capacité à prédire l'apparence à partir de l'ADN est susceptible de s'améliorer considérablement. Cela implique que beaucoup plus d'informations génétiques prédictives du faciès restent à découvrir dans de futures études, ce qui - au cas où cela serait suffisant - pourrait fournir la condition préalable à des applications pratiques de la prédiction des informations faciales humaines à partir de données génomiques telles que la criminalistique ou l'anthropologie. D'autre part, la capacité croissante de révéler des renseignements personnels à partir de données génomiques peut également avoir des implications éthiques, {article 5 de la loi d'analyse génétique humaine LAGH (RS 810.12 Loi fédérale du 8 octobre. Suisse) cite que les analyses génétiques et prénatales ne sont admissibles qu'à condition que la personne concernée y ait consenti librement après avoir reçu des informations suffisantes, qui seront largement discutées par les divers intervenants en plus des progrès génomiques et technologiques réalisés ici et qui seront réalisés dans les études futures.

ANNEXE 1

Validation du développement du systèmeIrisPlex : Détermination de la couleur de l'iris bleu et brun pour le renseignement médico-légal

Matériels et méthodes :

Une gamme d'échantillons de fluides corporels et des tissus humains ont été obtenus auprès de donneurs ayant des phénotypes de couleur des yeux connus sous un consentement éclairé. Les échantillons de test comprenaient des échantillons de source unique et multiple, des échantillons de cas non humains et simulés (sang, salive, sperme, cheveux, éléments touchés). Des réactions en double ont été effectuées sur tous les échantillons, sauf indication contraire. L'ADN a été extrait de tous les échantillons à l'aide du kit QIAamp DNA Mini (Qiagen, Hagen, Allemagne) conformément aux directives du fabricant et quantifié à l'aide du kit QuantifilerHuman DNA Quantification Kit (AppliedBiosystems Inc., Foster City, États-Unis).

Conception et protocole multiplex :

L'IrisPlex se compose de 6 SNP, rs12913832 (HERC2), rs1800407 (OCA2), rs12896399 (SLC24A4), rs16891982 (SLC45A2 (MATP)), rs1393350 (TYR) et rs12203592 (IRF4). Tous les détails des marqueurs et séquences d'amorces, y compris de légères modifications par rapport au protocole précédemment publié(Walsh et al,2012), sont répertoriés dans le tableau 1. Le protocole consiste en une seule PCR multiplexe en deux étapes utilisant 1 ml d'extrait d'ADN génomique (concentrations variables) et des amorces dans une réaction de 12 ml qui comprend 1 tampon PCR, 2,7 mM de MgCl2, 200 mM de chaque dNTP et utilise des conditions de thermocyclage ajustées pour une spécificité accrue :

(1) 95 8C pendant 10 min,(2) 33 cycles de 95 8C pendant 30 s et 61 8C pendant 30 s , (3) 5 min à 61 8C. Ceci est suivi d'une purification du produit et d'une autre réaction d'extension de base unique (SBE) multiplex à l'aide du kit ABI Prism1 SNaPshot (AppliedBiosystems) tel que publié précédemment . Tous les produits nettoyés ont été analysés sur l'analyseur génétique ABI 3130xl (AppliedBiosystems) avec POP-7 sur un réseau capillaire de 36 cm. Les paramètres d'exécution ont été optimisés pour augmenter la sensibilité, avec une tension d'injection de 2,5 kV pendant 10 s et un temps d'exécution de 500 s à 60°C.

Table 1
SNP markers used in the IrisPlex assay for eye colour prediction, including multiplex genotyping details.

SNP-ID	Prediction rank	Chr position	Gene	PCR product (bp)	Reward PCR primer (5'–3') and reverse PCR primer (5'–3')	Each primer conc. (μM)	Extension primer (5'–3') with t-tail for length differentiation	Primer direction	Primer length no tail (bp)	Total primer length (bp)	Conc. (μM)	Tm (°C)	Alleles detected
rs12913832 (1F and 1R)	1	15 – 28003813	HERC2	87	TGGCTCTCTGTCTGATGC GCCCCTGATGATGATAGC	0.416	TTGACA	Reverse	17	41	0.2	55	T/C
rs1800407* (2F and 2R)	2	15 – 28003813	OCA2	127	TGAAAGCTGCCTCTGTCT CGATGACAGACAGACATGATGA	0.416	CATACCCGCTCTCCC	Forward	16	23	0.1	57.3	G/A
rs12896399 (3F and 3R)	3	14 – 91845486	SLC24A4	104	CTCCGATCCAATTCTTTCT CTAGCCCTCCGCTCTGATG	0.416	TCTTTGCG TCACTA TTTTCCG	Forward	23	53	0.15	54.5	G/T
rs16891982 (4F and 4R)	4	5 – 33987480	SLC45A2 (MATP)	128	TGGAAGTTGTGCTAGACCAGA CGAACAGCACTCGACGTTG	0.416	AAACACGGAGTTGATGCA	Forward	18	29	0.5	55.9	G/C
rs1393350 (5F and 5R)	5	11 – 88650804	TYR	80	TTCCTCACTCCCTTCTCTGCC GCGAAGCTGAATGATAACACG	0.416	TTTGTAAAAGA CCACACAGATTT	Reverse	23	47	0.1	55.6	T/C
rs12203592* (6F and 6R)	6	6 – 341321	IRF4	115	ACAGCGCAGCTGATCTCTTC GCTAAACCTGCCACCAAAGC	0.416	TTTGCTCGCTAAAAGAAGG	Forward	19	35	0.3	55.2	C/T

* Changes made compared to initially published protocol [5] alternative direction of single base extension SNP primer used.

Tableau 1 : Marqueurs SNP utilisés dans le test IrisPlex pour la prédiction de la couleur des yeux, y compris les détails du génotypage multiplex.(Walsh et al,2010)

Sensibilité, équilibre de hauteur de pic et cohérence :

La sensibilité du test a été précédemment publiée à 31 pg, pour les profils 6-SNP complets (Walsh et al,2010). Ici, ils ont répété cette étude pour évaluer l'impact des modifications mineures apportées à l'essai. Des échantillons de 3 individus ont été préparés par dilution en série à partir de 500 pg et remesurés avec le kit QuantifilerHuman DNA Quantification en double. Les phénotypes de couleur des yeux marron, bleu et intermédiaire ont été génotypés à 500, 250, 125, 62, 31 et 16 pg d'ADN avec le test IrisPlex pour évaluer la sensibilité. Pour assurer la cohérence du génotypage avec l'ABI3130xl, des échantillons d'ADN de 3 individus avec des génotypes et des phénotypes de couleur des yeux différents ont été amplifiés à la concentration optimale de 250 pg et exécutés 25 fois. Les plages de hauteur des pics homozygotes et hétérozygotes ont été notées pour chaque locus. La hauteur de pic moyenne à chaque allèle et les rapports de hauteur de pic des hétérozygotes pour chaque locus SNP ont également été déterminés à des entrées d'ADN de 1000, 250 et 100 pg. Le rapport de hauteur de pic d'hétérozygotes a été calculé en divisant la hauteur du pic de l'allèle de poids moléculaire inférieur (premier pic au locus SNP) par la hauteur de pic de l'allèle de poids moléculaire supérieur (deuxième pic au locus SNP) pour tous les SNP.

Études de mélange :

Des mélanges d'ADN ont été préparés dans des rapports de 1:1, 1:5 et 1:10, avec une concentration totale de 500 pg. Des combinaisons d'ADN d'individus aux yeux bleus:brun, bleu:bleu et brun:brun ont été créées pour évaluer l'impact de plusieurs donneurs d'échantillons sur la capacité des systèmes IrisPlex à prédire la couleur des yeux.

Spécificité de l'espèce :

La spécificité de l'amplification a étéexaminéengénotypant une variété d'échantillons d'ADN animal avec le test IrisPlex. Les échantillons comprennent le chat, le chien, la souris, le rat, le bovin, le porc et le poulet à des quantités d'ADN de 100 ng et un échantillon de chimpanzé de 1 ng. L'ADN a été obtenu commercialement auprès de Novagen, Inc. (Madison, WI) pour tout sauf l'échantillon de chimpanzé, qui a été utilisé et décrit ailleurs.

Reproductibilité :

Des tests de concordance ont été effectués sur 40 échantillons d'ADN de concentrations variables et typés par trois laboratoires indépendants, dont deux n'avaient aucune expérience préalable avec le test IrisPlex.

Échantillons de type cas et études de stabilité :

Des échantillons de cas simulés ont été génotypés à l'aide du test IrisPlex dans le cadre d'un test d'aptitude à l'aveugle. Les échantillons comprennent le sang, le sperme, la salive, les cheveux, les échantillons inhibés (inhibition intentionnelle de l'hème) et les objets touchés. Pour l'étude de stabilité, l'ADN d'un individu a été soumis à un traitement à la DNase (10 unités) à des intervalles de 5, 10, 20, 30, 40 et 50 min et génotypé pour tester la stabilité du dosage dans des conditions dégradées, souvent observées chez échantillons de dossiers.

Modèle de prédiction de la couleur des yeux :

Le modèle de prédiction de la couleur de l'œil humain utilisé a déjà été publié(Walsh et al,2010) est un composant essentiel du système IrisPlex. Des probabilités de prédiction

statistique sont attribuées à chacun des allèles SNP obtenus à partir du test de génotypage IrisPlex, basé sur 3804 individus néerlandais dans l'ensemble de construction de modèles d'une étude précédente (Fan Liu et al, 2009). Lorsqu'il est combiné dans le modèle, il donne trois estimations de probabilité pour la couleur des yeux marron, bleu et intermédiaire du donneur inconnu, la probabilité la plus élevée étant le phénotype prédit. Une macro Excel a été créée pour permettre un calcul simplifié des estimations de probabilité, et se trouve dans le tableau supplémentaire 3 (Walsh et al,2010).

Études de population :

Le test IrisPlex a été réalisé sur les échantillons HGDP-CEPH H952 qui comprend 952 individus de 51 populations mondiales et a été précédemment publié par ce groupe (Walsh et al,2010).

ANNEXE 2

Le système HIrisPlex pour la prédiction simultanée de la couleur des cheveux et des yeux à partir de l'ADN

Matériels et méthodes :

Sujets, images et classification de la couleur des cheveux et des yeux :

Des échantillons d'ADN et des informations sur la couleur des cheveux ont été collectés auprès de 1551 sujets européens vivant en Pologne (n = 1093), en République d'Irlande (n =339) et en Grèce (n = 119). Tous les participants ont donné leur consentement éclairé. L'étude a été approuvée en partie par le Comité d'éthique de l'Université Jagellonne, numéro KBET/17/B/2005 et la Commission de bioéthique du Conseil régional des médecins de Cracovie numéro 48 KBL/OIL/2008. Les phénotypes de couleur des cheveux et des yeux ont été collectés par une combinaison d'auto-évaluation et classement professionnel par observateur unique (données polonaises). L'évaluateur professionnel (AKK) pour l'ensemble de données polonais est un médecin (dermatologue) qui a évalué la couleur des cheveux lors de l'observation et de l'interrogatoire d'individus dans des circonstances où les cheveux étaient teints ou gris. Pour l'auto-évaluation du phénotype de la couleur des cheveux, il a été demandé aux individus de remplir le questionnaire, la couleur de leurs cheveux pendant leur vingtaine, et à quel âge les cheveux gris/blanc ont commencé à apparaître (collection irlandaise), cela a évité les effets du grisonnement des cheveux et blanchiment au phénotypage. La collecte d'échantillons en Irlande comprenait des images photographiques des yeux et des cheveux à haute résolution. Dans une brève description, les images des cheveux et des yeux ont été prises à l'aide d'un Nikon D3100 avec un objectif macro AF-S Micro Nikkor 60 mm, l'ouverture, la vitesse d'obturation et la sensibilité ISO ont été fixées à f = 22, 1/125 et 200 respectivement. Un flash annulaire (modèle Speedlight SB-R200) et une distance moyenne de 7 cm de l'œil et de l'arrière de la tête ont été utilisés pour l'imagerie des cheveux. Cela a assuré un échantillonnage cohérent et des conditions d'éclairage régulées, y compris des réglages d'objectif d'une focale fixe de 0,2 et 0,23. Tous les individus ont été invités à remplir un questionnaire contenant des informations de base, telles que le sexe et l'âge, ainsi que des données concernant le phénotype de pigmentation des yeux et des cheveux. Cependant, en raison du fait que de nombreux Irlandais avaient les cheveux teints ou gris, des classifications de couleur de cheveux autodéclarées ont été utilisées pour cet

ensemble dans la formation des modèles. Pour la collection grecque, un écouvillon buccal a été prélevé sur chaque individu et un questionnaire auto-déclaré concernant les informations sur la couleur des cheveux et des yeux a été collecté. Pour l'ensemble irlandais et grec, la couleur des cheveux a été classée en 7 catégories : blond (5,9 %), brun clair (34 %), brun foncé (45,2 %), auburn (5,7 %), blond-roux (1,3 %), rouge (2,2%) et noir (5,7%). Pour l'ensemble de données polonais, ces données ont été collectées comme indiqué précédemment (Susan Walsh et al,2012) et la couleur des cheveux a été classée en 7 catégories : blond (13,7 %), blond foncé (44,2 %), brun (22,6 %), auburn (1 %), blond-roux (3,9 %), rouge (3,8 %) et noir (10,8 %)). Pour les analyses de prédiction de la couleur des cheveux, nous avons regroupé le blond et le blond foncé dans une catégorie blonde (42,6%), marron clair et marron foncé dans une catégorie marron (39,3%) et auburn, blond-roux et rouge dans une catégorie rouge (8,8 %) avec le noir comme quatrième catégorie supplémentaire (9,3 %). La couleur des yeux a été classée en 3 catégories bleu, brun et intermédiaire (y compris le vert). Le terme catégorie dans ce contexte fait référence au regroupement de couleurs phénotypiques similaires dans un groupe pour les séparer d'un autre groupe de couleurs, c'est-à-dire la catégorie blonde, la catégorie noire. Tableau 1affiche le nombre de phénotypes de couleur des cheveux et des yeux, y compris le sexe, dans les 3 populations échantillonnées. Notamment, les cheveux roux dans la population polonaise et la couleur des yeux verts dans la population irlandaise ont été intentionnellement enrichis en raison de leur rareté, donc les deux phénotypes ne reflètent pas les fréquences naturelles de la population.

Table 1
Phenotype frequencies according to hair and eye colour categories (including sex) for the full combined set of individuals from Poland, Ireland and Greece.

Hair colour	Blond	Dark blond*/light brown	Dark brown	Brown red/auburn	Blond red	Red	Black	Total	Eye colour – blue	Intermediate (green, heterochromia)	Brown	Total	Male	Female	Total
Poland	150	483*	247	11	43	41	118	1093	590	164	339	1093	449	644	1093
Ireland	16	111	158	23	6	10	15	339	172	90	77	339	77	262	339
Greece	11	45	49	3	0	0	11	119	13	15	91	119	51	68	119
Total	177	639	454	37	49	51	144	1551	775	269	507	1551	577	974	1551

* represents individuals who were reported as dark blond in the dark blond/light brown category.

Tableau.1. Fréquences phénotypiques selon les catégories de couleur des cheveux et des yeux (y compris le sexe) pour l'ensemble combiné des individus de Pologne, d'Irlande et de Grèce (Susan Walsh et al,2012)

Échantillons d'ADN et génotypage HIrisPlex :

L'ADN des échantillons polonais a été extrait comme décrit précédemment (Susan Walsh et al, 201). Des échantillons de salive prélevés sur des individus en Irlande ont été extraits à l'aide du kit d'isolement d'ADN Puregene (Qiagen, Hilden, Allemagne). Des écouvillons buccaux collectés sur des individus en Grèce ont été extraits à l'aide d'un protocole d'extraction biologique interne. L'ADN du sous-ensemble H952 du panel HGDP-CEPH qui représente 952 individus de 51 populations mondiales, a été acheté auprès du CEPH. En raison du manque d'ADN dans certains échantillons appartenant à l'ensemble HGDP-CEPH 952, 7 individus n'ont pas pu être génotypés par le test HIrisPlex, et donc le nombre final d'échantillons dans le monde était de 945.

Tous les échantillons ont été génotypés à l'aide du test HIrisPlex. Le test comprend 23 SNP et 1 polymorphisme d'insertion/délétion (INDEL), au total 24 variantes d'ADN, provenant de 11 gènes : MC1R , HERC2 , OCA2 , SLC24A4 , SLC45A2 , IRF4 , EXOC2 , TRYP1 , TYR , KITLG et PIGU/ASIP . De plus amples informations sur ces 24 marqueurs peuvent être trouvées dans le tableau 2, y compris les séquences d'amorces. Les 24 paires d'amorces PCR ont été conçues en utilisant les paramètres par défaut du programme Primer3Plus, qui est un logiciel de conception Web gratuit. Les fragments de PCR ont été conçus pour être aussi courts que possible pour prendre en charge l'ADN dégradé et, par conséquent, tous ont une longueur inférieure à 160 pb. Pour réduire la possibilité que des paires d'amorces interagissent entre elles, le programme Autodimera été utilisé pour analyser les séquences d'amorces. Les régions de séquence environnantes ont également été recherchées avec BLAST contre dbSNPpour réduire le risque qu'un emplacement d'amorce recouvre un site SNP interférent connu pour une liaison d'amorce efficace.

Table 2
Information about the 24 DNA variants of the HIrisPlex assay, including PCR and single base extension (SBE) primer sequences and concentrations.

Assay position	SNP	CHR	Position		Gene	Major Allele	Minor Allele	PCR primers		Concentration		Product size	SBE primers	Concentratio
1	N29insA	16	89985753	Exonic	*MC1R*	C	insA	MC1Rset1F	Set1	0.55 µm	GCAGGCATCCCAGAGAAGAC	117bp	CCCCAGCTGGGGCTGGCTGCCAA	1.3 µm
2	rs11547464	16	89986091	Exonic	*MC1R*	G	A	MC1Rset1R		0.55 µm	TCAGAGATGGACACCTCCAG		tttttttGCCATCGCCGTGGACC	0.1 µm
3	rs885479	16	89986154	Exonic	*MC1R*	C	T	MC1Rset2F	Set2	0.5 µm	CTGGTGAGCTTGGTGGAGA	158bp	tttttttttGATGCCCGCAACGGCT	1.25 µm
4	rs1805008	16	89986144	Exonic	*MC1R*	C	T	MC1Rset2R		0.5 µm	TCCAGCAGGAGGATGACG		tttttttACAGCATCGTGACCCTGCCG	0.375 µm
5	rs1805005	16	89985844	Exonic	*MC1R*	G	T	MC1Rset3F	Set3	0.5 µm	GTCCAGCCTCTGCTTCCTG	147bp	tttttttttTGGTGGAGAACG-CGCTGGTG	0.75 µm
6	rs1805006	16	89985918	Exonic	*MC1R*	C	A	MC1Rset3R		0.5 µm	AGCGTGCTGAAGACGACAC		ttttttttttttCTGCCTGGC-CTTGTCGGA	0.75 µm
7	rs1805007	16	89986117	Exonic	*MC1R*	C	T	MC1Rset4F	Set4	0.4 µm	CAAGAACTTCAACCTCTTTCTCG	106bp	ttttttttttttttCTCCATCTTC-TACGCACTG	1 µm
8	rs1805009	16	89986546	Exonic	*MC1R*	G	C	MC1Rset4R		0.4 µm	CACCTCCTTGAGCGTCCTG		ttttttttttttttttATCTGC-AATGCCATCATC	0.4 µm
9	Y152OCH	16	89986122	Exonic	*MC1R*	C	A						ttttttttttttttttCATCTT-CTACGCACTGCGCTA	0.6 µm
10	rs2228479	16	89985940	Exonic	*MC1R*	G	A						ttttttttttttttttttttC-TGGTGAGCCGCGAGCAAC	0.375 µm
11	rs1110400	16	89986130	Exonic	*MC1R*	T	C						ttttttttttttttttttCTTCTA-CGCACTGCGCTACCACACCA	0.3 µm
12	rs28777	5	33994716	Intronic	*SLC45A2*	A	C	rs28777_F	Set5	0.4 µm	TACTCGTGTGGGAGTTCCAT	150bp	ttttttttttttttttttttttCA-TGTGATCCTCACAGCAG	1.2 µm
								rs28777_R		0.4 µm	TCTTTGATGTCCCCTTCGAT			
13	rs16891982	5	33987450	Exonic	*SLC45A2*	G	C	Rs16891982_F	Set6	0.4 µm	TCCAAGTTGTCGTCAGACCAGA	128bp	ttttttttttttttttttttttttt-AAACACGGAGTTGATGCA	1 µm
								Rs16891982_R		0.4 µm	CGAAAGAGGAGTCGAGGTTG			
14	rs12821256	12	87852466	Intergenic	*KITLG*	A	G	rs12821256_F	Set7	0.4 µm	ATGCCCAAAGGATAAGGAAT	118bp	ttttttttttttttttttttttt-GGAGCCAAGGGCATGTTACTACGGCAC	0.1 µm
								rs12821256_R		0.4 µm	GCAGCCAAGGGCATGTTACT			
15	rs4959270	6	402748	Intergenic	*EXOC2*	C	A	Rs4959270_F	Set8	0.4 µm	TGAGAAATCTACCCCCACGA	140bp	ttttttttttttttttttttttt-GGAACACATCCAAACTATGACACTATG	0.375 µm
								Rs4959270_R		0.4 µm	GTGTTCTTACCCCCTGTGGA			
16	rs12203592	6	341321	Intronic	*IRF4*	C	T	rs12203592_F	Set9	0.4 µm	AGGGCAGCTGATCTCTTCAG	126bp	ttttttttttttttttttttttttt-tTCCACTTTGGTGGGTAAAAGAAGG	0.3 µm
								rs12203592_R		0.4 µm	GCTTCGTCATATGGCTAAACCT			
17	rs1042602	11	88551344	Exonic	*TYR*	G	T	rs1042602 _F	Set10	0.4 µm	CAACACCCATGTTTAACGACA	124bp	ttttttttttttttttttttttttt-ttttttCAATGTCTCTCCCAGATTTCA	1.25 µm
								rs1042602 _R		0.4 µm	GCTTCATGGCCAAAATCAAT			
18	rs1800407	15	25903913	Exonic	*OCA2*	G	A	rs1800407_F	Set11	0.4 µm	AAGGCTGCCTCTGTTCTACG	124bp	ttttttttttttttttttttttttt-ttttttttGCATACCCGCTCTCCC	0.1 µm
								rs1800407_R		0.4 µm	CGATGAGACAGCAGCATGATGA			
19	rs2402130	14	91870956	Intronic	*SLC24A4*	A	G	rs2402130_F	Set12	0.4 µm	ACCTGTCTCACAGTGCTGCT	150bp	ttttttttttttttttttttttttttttt-ttttttttGAACCATACCGAGCCCCTG	0.75 µm
								rs2402130_R		0.4 µm	TTCACCTCGATGACGATGAT			
20	rs12913832	15	26039213	Intronic	*HERC2*	C	T	rs12913832_F	Set13	0.4 µm	TCAACATCAGGGTAAAA-ATCATGT	150bp	ttttttttttttttttttttttttttttt-ttttttttttTAGCGTGCAGAACTTGACA	1.2 µm
								rs12913832_R		0.4 µm	GGCCCCTGATGATGATAGC			
21	rs2378249	20	32681751	Intronic	*ASIP/PIGU*	T	C	rs2378249_F	Set14	0.4 µm	CGCATAACCCATCCCTCTAA	136bp	ttttttttttttttttttttttttttttt-ttttttttttttCCACACCTCTCCTCAGCCCA	0.18 µm
								rs2378249_R		0.4 µm	CATTGCTTTTCAGCCCACAC			
22	rs12896399	14	91863416	Intergenic	*SLC24A4*	T	G	Rs12896399_F	Set15	0.4 µm	CTGCGATCCAATTCTTTGT	125bp	ttttttttttttttttttttttttttttt-ttttttttttttttttttttttttttttttCTTTAGCT-CACTATATTTTTCGG	1.125 µm
								Rs12896399_R		0.4 µm	GACCCTCTGTCAGACCCAGT			
23	rs1393350	11	88860094	Intronic	*TYR*	C	T	Rs1393350_F	Set16	0.4 µm	TTTCTTTATCCCCCTGATCC	124bp	ttttttttttttttttttttttttttttt-ttttttttttttttttttttttttttttttCATTTCTA-AAAGACCACACAGATT	1.1 µm
								Rs1393350_R		0.4 µm	GCGAAGCTGAATGATGAACACG			
24	rs683	9	120899005	Exonic	*TYRP1*	T	G	rs683_F	Set17	0.4 µm	CAGAAACCACCTCGTTGAA	138bp	ttttttttttttttttttttttttttttt-ttttttttttttttttttttttttttttttGCTTTG-AAAGTATGCCTAGAACCTTAAT	0.175 µm
								rs683_R		0.4 µm	TCAAAGCGTTTCCCAGCTT			

Tableau 2. Informations sur les 24 variantes d'ADN du test HIrisPlex, y compris les séquences et les concentrations d'amorces de PCR et d'extension de base unique (SBE).(Susan Walsh et al,2012)

Pour le génotypage de la population, des quantités d'ADN génomique allant de 300 pg à 3 ng dans des formats de 1 l ont été amplifiées par individu dans un volume de réaction de 10 consistingl composé de 1× tampon PCR, 2,5 mM de MgCl 2 , 220 M de chaque dNTP et

1,75 U AmpliTaq Gold ADN polymérase (AppliedBiosystems Inc., Foster City, CA), y compris les concentrations d'amorces PCR trouvées dans le tableau 2 . Le thermocyclage a été réalisé sur le système 96 puits GeneAmp ® PCR 9700 (AppliedBiosystems) dans les conditions suivantes

(1) 95 °C pendant 10 min, (2) 33 cycles de 95 °C pendant 30s et 61°C pendant 30 s, (3) 5 min à 61 °C. Les produits PCR ont été nettoyés avec ExoSAP-IT (USB Corp., Cleveland, OH), tel que recommandé par le fabricant. Après élimination des dNTP et des amorces non incorporées. Le dosage multiplex SBE (single base extension) a été réalisé en utilisant 2l de produit avec 1l de mélange réactionnel du kit ABI SNaPshot (AppliedBiosystems, Foster City, CA) dans un volume réactionnel total de 5l. Les séquences d'amorces d'extension de base unique (SBE) et les concentrations utilisées dans le test peuvent être trouvées dans le tableau 2. Les conditions de thermocyclage étaient les suivantes : 96 °C pendant 2 min et 25 cycles de 96 °C pendant 10 s, 50 °C pendant 5 s et 60 °C pendant 30 s. Les produits ont été nettoyés à l'aide de SAP (USB Corp.), en suivant les directives du fabricant et 1l de produit nettoyé a été exécuté sur l'analyseur génétique ABI 3130xl (AppliedBiosystems) avec POP-7 sur un réseau capillaire de 36 cm en suivant les directives de préparation des échantillons du kit SNaPshot, cependant les paramètres d'exécution de 2,5 kVpour une tension d'injection de 10 s et un temps d'exécution de 500 s à 60 °C ont été utilisés pour une sensibilité accrue.

Pour les études de sensibilité du dosage, les résultats de génotypage de deux individus différents ont été évalués à partir de dilutions en série d'échantillons d'ADN d'entrée de 500 pg, 250 pg, 125 pg, 63 pg et 31 pg. Chaque résultat a été étudié pour l'abandon allélique, qui comprend des pics inférieurs au seuil de 50 rfu qui ne peuvent pas être appelés. La détermination de la sensibilité était basée sur la production d'un profil complet dans chaque réplicat à un niveau d'entrée d'ADN particulier.

Variantes d'ADN HIrisPlex et leur utilisation pour la prédiction de la couleur des yeux/des cheveux, y compris dans un échantillon mondial :

Le dosage HIrisPlex se compose des 24 variantes d'ADN (23 SNP et 1 INDEL), 6 de ces marqueurs, rs12913832 (HERC2), rs1800407 (OCA2), rs12896399 (SLC24A4), rs16891982 (SLC45A2 (MATP)), rs1393350 (TYR) et rs12203592 (IRF4) sont tirés du système IrisPlex qui est déjà bien établi(Walsh et al,2010) et sont utilisés pour la partie prédiction de la couleur des yeux du système HIrisPlex. Les résultats de ces 6 SNP lorsque leur allèle mineur est entré dans l'outil de prédiction HIrisPlex sont utilisés pour prédire la couleur des yeux de l'individu à l'aide du modèle IrisPlex tel que publié précédemment.Avec la probabilité la plus élevée des trois catégories, marron, bleu ou intermédiaire étant la couleur des yeux prédite.

Les 22 variables d'ADN utilisées pour la prédiction de la couleur des cheveux sont Y152OCH, N29insA, rs1805006, rs11547464, rs1805007, rs1805008, rs1805009, rs1805005, rs2228479, rs1110400 et rs885479 du MC1R gène, rs1042602 (TYR), rs4959270 (EXOC2), rs28777 (SLC45A2 (MATP)), rs683 (TYRP1), rs2402130 (SLC24A4), rs12821256 (KITLG), rs2378249 (PIGU/ASIP), rs12913832 (HERC2), rs1800407 (OCA2), rs16891982 (SLC45A2 (MATP12492) et rs basé sur IRF notre précédente publication pour la prédiction de la couleur des cheveux. Lorsque leurs allèles mineurs sont entrés dans l'outil

de prédiction HIrisPlex, ils sont utilisés pour prédire la couleur des cheveux de l'individu à l'aide du modèle de prédiction des cheveux HIrisPlex développé dans cet article. Parmi les quatre catégories de couleur de cheveux blond, brun, rouge et noir, la valeur de probabilité la plus élevée est indicative de la couleur de cheveux prévue suivant les directives publiées dans ce document et décrites dans la section suivante.

Pour la prédiction de la couleur des cheveux dans le monde entier, nous avons évalué les performances du test HIrisPlex sur 945 échantillons provenant de 51 populations de l'ensemble HGDP-CEPH. Le package MapViewer 7 (Golden Software, Inc., Golden, CO, États-Unis) a été utilisé pour tracer les catégories de couleurs de cheveux prédites et la distribution des génotypes SNP sur la carte du monde. Un graphique de mise à l'échelle multidimensionnelle non métrique (MDS) a été produit pour illustrer les distances F ST par paires des 24 SNP de couleur des yeux et des cheveux entre les populations, en utilisant SPSS 17.0.2 pour Windows (SPSS Inc., Chicago, États-Unis). L'analyse de la variance moléculaire (AMOVA) (Excoffier 1992) a été réalisée en utilisant Arlequin v3.11 . Une évaluation de seuil des probabilités de prédiction pour chaque catégorie de couleur de cheveux a également été réalisée, y compris un seuil de probabilité de prédiction de couleur des yeux et des cheveux combinés dans l'inférence d'un individu non européen aux cheveux noirs et aux yeux bruns. Pour l'évaluation d'un changement de couleur de cheveux en fonction de l'âge, une corrélation de Pearson a été calculée et le graphique tracé à l'aide de SPSS 17.0.2 pour Windows (SPSS Inc., Chicago, États-Unis).

Modélisation de prédiction pour la couleur des cheveux :

Pour développer un modèle de prédiction de la couleur des cheveux en utilisant des échantillons de plusieurs sites avec différents niveaux de couleur des cheveux en raison de leur position en Europe, Europe centrale, occidentale et méridionale, nous avons pris un sous-ensemble aléatoire de 80 % des échantillons de chaque site, Pologne (n = 875), l'Irlande (n = 272) et la Grèce (n = 96). Ce sous-ensemble de 80 % a été utilisé pour entraîner le modèle et était basé sur la régression logistique multinomiale (MLR), telle que publiée précédemment par Liu et al, 2009. En bref, les individus ont été classés en fonction de leurs phénotypes capillaires et ont été divisés en 4 catégories, Blond (n = 529), Brun (n = 490), Rouge (n = 109) et Noir (n = 115). Pour leurs génotypes, 22 des 24 variations d'ADN HIrisPlex (comme décrit ci-dessus) ont été utilisées pour tester la différenciation de la couleur des cheveux et l'utiliser dans le modèle de prédiction. En entrant l'allèle mineur de chaque variante d'ADN, y compris son phénotype et en appliquant la MLR, des valeurs alpha et bêta sont générées qui forment le noyau du modèle de prédiction. Ce modèle permet ensuite la prédiction probabiliste d'une catégorie de couleur de cheveux d'un individu uniquement sur la base d'entrée des 22 allèles mineurs variantes dans l'outil de prédiction de couleur de cheveux HIrisPlex. Pour évaluer l'effet des nuances claires et foncées de la couleur des cheveux qui peuvent être attribuées respectivement au blond et au noir, une approche similaire a été utilisée qui combinait les individus regroupés dans la catégorie claire (blond, n = 529) versus une catégorie sombre (noir, n = 115). Les individus aux cheveux roux ont été omis (n = 109) de cette analyse car leur couleur résultante est basée sur une mutation cumulative MC1R et non sur le spectre continu du clair au foncé (c'est-à-dire du blond au noir). Individus aux cheveux bruns (n = 490) ont été omis, car seuls les extrêmes de lumière et d'obscurité étaient requis. Par conséquent, en utilisant cette approche de modèle à deux volets, une couleur de cheveux prédite est générée avec une indication approximative de la

couleur claire ou foncée (c'est-à-dire brun clair, brun foncé) en raison de l'influence des génotypes couramment associés aux catégories clair/foncé, respectivement de blond et de noir. Les 20% supplémentaires de l'ensemble de données combiné (total n = 308), c'est-à-dire de Pologne (n = 218), d'Irlande (n = 67) et de Grèce (n = 23), a été utilisé pour évaluer la précision du modèle de prédiction en termes de prédiction de couleur de cheveux finale correcte ou incorrecte en fonction de la catégorie de couleur, de la nuance et de l'utilisation du guide de prédiction de la couleur des cheveux qui est décrit en détail dans la section 3 , et un une évaluation des seuils optimaux de catégorie a été entreprise. Les étapes à suivre lors de l'acquisition d'une prédiction basée sur la couleur et la nuance sont décrites dans un guide fourni ci-dessous.

Number	DNA variant	Gene	Blond(beta)	Blond(p)	Brown(beta)	Brown(p)	Black(beta)	Black(p)	Red(beta)	Red(p)
1	N29insA	MC1R	–	–	–	–	–	–	-21.9731	0.994026
2	rs11547464	MC1R	–0.947299	0.081175	–0.4007191	0.441688	–16.782634	0.995907	–2.8866	4.42E–08
3	rs885479	MC1R	0.272536	3.36E–01	0.1938828	0.460717	2.29E–01	0.575679	0.315529	0.707292
4	rs1805008	MC1R	–0.57034	0.003874	–0.3058868	0.097798	–5.66E–01	0.084668	–3.02472	2.20E–16
5	rs1805005	MC1R	0.20689	2.28E–01	0.2382036	0.128146	–1.57E–01	0.539306	–0.86742	0.025064
6	rs1805006	MC1R	1.718508	0.045418	2.1268136	0.009857	–1.70E+01	0.996356	–2.43626	0.001714
7	rs1805007	MC1R	–0.53542	0.030279	–0.1503278	0.508278	–1.32E+00	0.009567	–3.59956	2.20E–16
8	rs1805009	MC1R	0.550547	5.60E–01	0.5309897	0.49513	–4.70E–01	0.693758	–4.25774	4.14E–08
9	Y152OCH	MC1R	–	–	–	–	–	–	–19.3501	0.992969
10	rs2228479	MC1R	–0.025643	8.83E–01	–0.1128742	0.483857	1.98E–01	0.413966	–0.61967	0.110936
11	rs1110400	MC1R	–0.366071	0.338334	–0.5920858	0.123046	6.63E–01	0.21252	–1.67775	0.009302
12	rs28777	SLC45A2	0.566568	0.414238	0.3138274	0.561428	4.85E–01	0.468883	–0.41607	0.743869
13	rs16891982	SLC45A2	0.863795	0.194837	0.2562763	0.618846	6.29E–01	0.326034	0.891013	0.522114
14	rs12821256	KITLG	–0.434962	0.020898	–0.1743193	0.32142	–6.87E–01	0.056556	0.406751	0.312582
15	rs4959270	EXOC2	–0.251437	0.019073	–0.1555227	0.120958	–2.71E–01	0.104087	–0.34639	0.107774
16	rs12203592	IRF4	1.741377	2.20E–16	1.0810914	2.22E–16	8.80E–01	2.35E–06	0.071132	0.773323
17	rs1042602	TYR	0.125113	0.24551	0.141479	0.155781	–4.52E–02	0.779493	–0.3842	0.071464
18	rs1800407	OCA2	–0.204189	0.331948	–0.0048133	0.97935	–3.53E–01	0.202517	0.223931	0.580501
19	rs2402130	SLC24A4	0.354085	0.00797	0.2752735	0.023746	4.36E–02	0.820086	–0.08861	0.724429
20	rs12913832	HERC2	1.372353	2.20E–16	0.6797949	6.83E–10	1.19E+00	6.65E–13	0.754729	0.004319
21	rs2378249	PIGU/ASIP	0.088319	0.526489	0.1828612	0.154928	–1.64E–01	0.449722	–0.72184	0.002302
22	rs683	TYRP1	0.197865	0.066913	0.168184	0.08995	1.58E–03	0.992081	0.129235	0.540918

Tableau 1 : Évaluation de la contribution de chaque variante d'ADN HIrisPlex pour la prédiction de la couleur des cheveux dans le modèle en termes de valeurs bêta et de probabilité (p). Les valeurs générées reflètent une évaluation de la catégorie binaire de la prédiction de la couleur, c'est-à-dire blond par rapport au non-blond, brun par rapport au non-brun, etc. Les valeurs p les plus basses (et donc les plus statistiquement significatives) pour chaque catégorie sont mises en évidence pour les variantes d'ADN respectivement associées.(Susan Walsh et al,2012)

Highest probability category approach	Observed categories				
No threshold p value	Red	Blond	Brown	Black	Total prediction(n)
Red predicted	17(89.5%)	1(5.3%)	1(5.3%)	0(0%)	19(100%)
Blond predicted	8(3.7%)	123(57.2%)	68(31.6%)	16(7.4%)	215(100%)
Brown predicted	2(6.3%)	5(15.6%)	24(75%)	1(3.1%)	32(100%)
Black predicted	1(2.4%)	2(4.8%)	27(64.3%)	12(28.6%)	42(100%)
Total phenotype(n)	28	131	120	29308 Total	

Highest probability category approach	Observed categories				
>0.7p threshold	Red	Blond	Brown	Black	Total prediction(n)
Red predicted	8(100%)	0(0%)	0(0%)	0(0%)	8(100%)
Blond predicted	2(1.8%)	70(63.6%)	33(30%)	5(4.6%)	110(100%)
Brown predicted	0(0%)	0(0%)	3(75%)	1(25%)	4(100%)
Black predicted	0(0%)	0(0%)	4(80%)	1(20%)	5(100%)
Total phenotype(n)	10	70	40	7127 Total	
Undefined	**18(64.3%)**	**61(46.6%)**	**80(66.7%)**	**22(75.9%)**	**181 (58.8%)**

	Observed categories					
Prediction guide approach	Red	Blond	d-blond/l-brown	D-Brown	Black	Total prediction (n)
Red predicted	24(80%)		0 6(20%)	0(0%)	0(0%)	30(100%)
Blond/d blond predicted	2(1.7%)	32(26.4%)	52(43%)	30(24.8%)	5(4.1%)	121(100%)
L-brown/d-brown predicted	2(1.3%)	9(6%)	59(39.6%)	58(38.9%)	21(14.1%)	149(100%)
Black/d-brown predicted	0		0 1(12.5%)	4(50%)	3(37.5%)	8(100%)
Total phenotype (n)	28	41	118	92	29308 Total	

Tableau 2 :Précisions de prédiction de la couleur des cheveux HIrisPlex obtenues à partir d'un ensemble de tests de modèles distincts de 308 individus de Pologne, d'Irlande et de Grèce (les individus n'ont pas été pris en compte pour la construction du modèle de prédiction pour lequel un ensemble différent de 1243 individus a été utilisé) en utilisant deux approches : l'approche de la catégorie de probabilité la plus élevée (avec et sans seuils) et l'approche du guide de prédiction (voir Fig.1 pour le guide de prédiction)(Susan Walsh et al,2012)

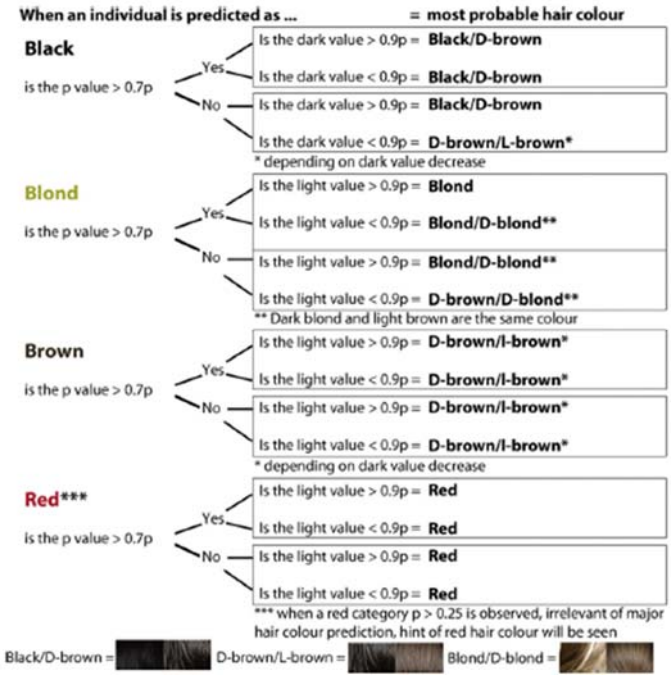


Fig. 1. Guide de prédiction HIrisPlex sur la façon d'interpréter les probabilités individuelles de couleur de cheveux et de nuance de cheveux. d-Brown signifie brun foncé et l-brown signifie brun clair.(Susan Walsh et al,2012)

Fig. 2. Quatre exemples divers d'individus européens (A–D) illustrant l'application du système HIrisPlex, y compris les résumés des résultats finaux pouvant être utilisés à des fins de rapport. Chaque individu représente des images de cheveux et d'yeux à haute résolution, qui affichent leurs phénotypes réels de couleur des yeux et des cheveux. La couleur des yeux et des cheveux, ainsi que les catégories de nuances de cheveux sont affichées avec leursProbabilités respectives dérivées du génotypage HIrisPlex et entrées dans le modèle de prédiction de la couleur des yeux et des cheveux, la catégorie affichant la probabilité la plus élevée pour chaque trait mis en évidence. y compris les précisions de prédiction actuellement connues pour une certaine couleur de cheveux, sur la base de l'ensemble de tests de 308 individus de 3 populations européennes utilisés dans cette étude. En ce qui concerne la couleur des yeux, la déclaration est complétée par les directives suivantes pour la prédiction de la couleur des yeux décrites précédemment (Walsh et al,2010) en utilisant les précisions d'évaluation de seuil précédemment produites à partir d'un ensemble de tests de > 3800 individus de 7 populations européennes.(Susan Walsh et al,2012).

ANNEXE 3

Le système HIrisPlex-S pour la prédiction de la couleur des yeux, des cheveux et de la peau à partir de l'ADN : introduction et validation du développementmédico-légale (ancienne étude)

Matériels et méthodes :

Une sélection d'échantillons de fluides corporels et des tissus ont été collectés en interne auprès de personnes ayant un consentement éclairé, y compris des échantillons de source unique et multiple et des échantillons de cas simulés (salive, sang, sperme, écouvillonnages vaginaux et objets touchés). L'ADN a été extrait de tous les échantillons à l'aide du kit QIAamp DNA Mini (Qiagen, Hagen, Allemagne) selon les directives du fabricant, ou un protocole de relargage standard interne (non publié). Tous les échantillons d'ADN extraits ont été quantifiés à l'aide du kit QuantifilerHuman DNA Quantification (AppliedBiosystems Inc., Foster City, USA) conformément aux directives du fabricant.

Conception multiplex et protocole de génotypage

Un ensemble de 36 SNP de 16 gènes a été choisi comme ensemble final de prédicteurs de SNP pour la couleur de la peau, qui incluent 19 des 24 SNP inclus dans le test HIrisPlex précédemment développé (Walsh et al,2010) ;(Walsh et al,2012), que nous avons laissé inchangés dans la conception du système HIrisPlex-S. Le deuxième test de génotypage nouvellement développé du système HIrisPlex-S comprend 17 SNP provenant de 7 gènes : ANKRD11 rs3114908, OCA2 rs1800414, BNC2 rs10756819, HERC2 rs2238289, SLC24A4 rs17128291, HERC2 rs6497292, HERC2 rs21229038, rs1426654, ASIP rs6119471, OCA2 rs1545397, RALY rs6059655, OCA2 rs12441727, MC1R rs3212355 et DEF8 rs8051733. Primer3Plus, un logiciel de conception Web accessible au public, a été utilisé pour concevoir les 17 paires d'amorces et leur extension de base unique (SBE) respective.en utilisant les paramètres par défaut du logiciel. Pour chaque SNP, à la fois vers l'avant et des amorces inverses SBE ont été conçues et testées, et la plus appropriée des deux a finalement été incluse dans le système multiplex. Pour assurer une séparation complète par électrophorèse capillaire entre les 17 SBE produits, des queues poly-T de longueurs différentes ont été ajoutées à l'extrémité 5' des amorces SBE (tableau 1). En ordre pour prendre en charge les échantillons dégradés, les fragments PCR ont été conçus pour être aussi courts que possible et la PCR les tailles des fragments ont été limitées à moins de 150 pb si possible (le plus grand fragment est de 170 pb). Avant le labo tests, toutes les séquences d'amorces ont été analysées à l'aide du logiciel AutoDimerpouréviterinteractions d'amorces. Tous les détails des marqueurs, les séquences d'amorces et les concentrations d'amorces sont fournis dans le tableau 1. L'amplification par PCR des 17 SNP a été réalisée dans un seul test PCR multiplex dans un volume total de 10l, contenant des amorces de PCR dans des concentrations spécifiées, 1l d'extrait d'ADN génomique (concentrations variables), tampon PCR 1X (AppliedBiosystems), 2,5 mM MgCl2 (AppliedBiosystems), 220 M de chaque dNTP (Roche, Mannheim, Allemagne) et 1,75 U AmpliTaq Gold ADN polymérase (AppliedBiosystems). Un thermocycleurGeneAmp PCR System (AppliedBiosystems) a été utilisé pour toutes les amplifications, avec les paramètres suivants : 95oC pendant 10 minutes, 33 cycles de 95oC pendant 30 secondes et 62oC pendant 30 secondes, 62oC pendant 5 minutes. Les produits de PCR amplifiés ont été purifiés avec ExoProStar – S (GE Healthcare Europe GmbH,

Eindhoven, Pays-Bas) et incubés à 37oC pendant 45 minutes et à 80oC pendant 15 minutes. La SBE a été réalisée pour les 17 produits PCR simultanément dans une seule réaction multiplex en utilisant 2l du produit PCR purifié et 1l du mélange réactionnel SNaPshotReady avec les paramètres de thermocycleur suivants : 96 oC pendant 2 minutes ; 25 cycles de 96oC pendant 10 secondes, 50oC pendant 5 secondes, 60oC pendant 30 secondes. Les produits étendus ont ensuite été purifiés en utilisant 1 l de phosphatase alcaline de crevette (SAP) (USB Corporation) et incubés à 37oC pendant 45 minutes et à 75oC pendant 15 minutes.

Enfin, les produits SBE purifiés ont été analysés sur l'analyseur génétique ABI 3130xl (AppliedBiosystems) avec le polymère POP-7 sur un capillaire de 36 cm sous une tension d'injection de 2,5 kV pendant 10 secondes et avec un temps de fonctionnement de 500 secondes à 60 °C. Le logiciel Gene Mapper v3.7 (AppliedBiosystems) a été utilisé pour l'appel des allèles et l'analyse des résultats.

Assay Position	PCR Primers	Bases	Sequence	Concentration	Product Size	SBE Primers	Gene	NCBI direction	HPS model input	Bases	Sequence	Concentration
1	rs3114908_F	20	CAGAACACAGCCACACCCTA	0.4 µm	118bp	rs3114908_R	ANKRD11	Rev	C/T	29	TTTTTTTTTTAGAGAAGGGTCAAGCACTT	0.12 µm
	rs3114908_R	20	CATAAAGGGGTCACCAGCAA	0.4 µm								
2	rs1800414_F	20	GCTGCAGGAGTCAGAAGGTT	0.4 µm	145bp	rs1800414_R	OCA2	Rev	T/C	43	TTT TTT TTT TTC AGA ATC CCG TCA GAT ATC CTA	0.2 µm
	rs1800414_R	20	GGGACAAACGAATTGAGGAA	0.4 µm								
3	rs1075681 9_F	20	AAAGCAAGCTCATGTTTCCA	0.4 µm	145bp	rs1075681 9_F	BNC2	Fwd	A/G	35	TTTTTTTTTTTTGGACCAGTTATTTTGGGTTTGGA	1.7 µm
	5681 9_R	23	GAAAAACACCAA	0.4 µm								
4	rs2238289_F	21	GGAACATGAAGATTTCCCAGT	0.4 µm	112bp	rs2238289_F	HERC2	Rev	C/T	53	TTT TTT TTT TTT TTT TTT TTG AGA TTG GAA GAT TGG AGC C	0.5 µm
	rs2238289_R	22	CTGATTCAGGTCTGCTGTCACT	0.4 µm								
5	rs1712829 1_F	20	CCAGCACTGCCAAAATAACA	0.4 µm	129bp	rs1712829 1_R	SLC24A4	Fwd	T/C	58	TTT TTT TTT TTT TTT TTT TTT CAA TGT GCA CTG GAT TAA AAG TC	1 µm
	rs1712829 1_R	20	CTCTTTGGACCCATCACCTC	0.4 µm								
6	rs6497292 _F	21	TCTGCTGTAGAACCAATGTCC	0.4 µm	150bp	rs6497292 _R	HERC2	Fwd	T/C	61	TTT TTT TTT TTT TTT TTT TTT TTT TTG TCT CCT GTG TCT TCA TCC T	0.2 µm
	rs6497292 _R	21	GAATTGCACCTGTAGCTCCAT	0.4 µm								
7	rs1129038 _F	20	ATGTCGACTCCTTTGCTTCG	0.4 µm	137bp	rs1129038 _F	HERC2	Rev	A/G	70	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT GAG CCA GGC AGC AGA GC	0.4 µm
	rs1129038 _R	20	ACACCAGGCAGCCTACAGTC	0.4 µm								

8	rs1667394_F	24	CAGCTGTAGAGAGAGACTTTGAGG	0.4 µm	130bp	rs1667394_R	HERC2	Rev	C/T	73	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT GCA GCA ATT CAA AAC GTG CAT A	0.2 µm
	rs1667394_R	20	CACCATTAAGACGCAGCAAT	0.4 µm								
9	rs1126809_F	27	TGTTTCTTAGTCTGAATAACCTTTTCC	0.4 µm	100bp	rs1126809_F	TYR	Fwd	A/G	77	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TGT ATT TTT GAG CAG TGG CTC C	0.05 µm
	rs1126809_R	20	GGTGCATTGGCTTCTGGATA	0.4 µm								
10	rs1470608_F	26	TTTCTTGTGTTAACTGTCCTTACAAA	0.4 µm	145bp	rs1470608_F	OCA2	Rev	A/C	62	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTC ATT CTC TCT TAA AAA TAT TAA TTT GCA CC	4 µm
	rs1470608_R	23	GGAAAATATGTTAGGGTTGATGG	0.4 µm								
11	rs1426654_F	20	TTCAGCCCTTGGATTGTCTC	0.4 µm	123bp	rs1426654_F	SLC24A5	Fwd	A/G	86	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TGT CTC AGG ATG TTG CAG GC	0.16 µm
	rs1426654_R	26	TGAGTAAGCAAGAAGTATAAGGAGCA	0.4 µm								
12	rs6119471_F	20	GCAGGAGAATTGCTGGAACT	0.4 µm	170bp	rs6119471_R	ASIP	Fwd	G/C*	91	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TGA AGG AAG AGT GAA AAT GCG TAA	1 µm
	rs6119471_R	21	AACCCGAAGGAAGAGTGAAAA	0.4 µm								
13	rs1545397_F	24	GGTATAGGATTATTTGGGGAATGA	0.4 µm	144bp	rs1545397_F	OCA2	Fwd	A/T	97	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT GTA CAA CTT TGT GAA TAT ACT AAA ATA C	1 µm
	rs1545397_R	28	TGGAGATATAGAATTCACACAACATAAA	0.4 µm								
14	rs6059655_F	20	GTGAGGAAATCGAGGCTCAG	0.4 µm	112bp	rs6059655_R	RALY	Fwd	A/G	76	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGC TGATGCCCTGAGCA	2 µm
	rs6059655_R	20	AGGAGAAAGCTGCAGATCCA	0.4 µm								
15	rs1244172_F	20	GGGAAGAGACAGCTCCATGT	0.4 µm	137bp	rs1244172_F	OCA2	Fwd	A/G	106	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TGG CTC AGT GTG GCC TT	0.5 µm
	rs1244172_R	20	ACAATCCTGGGAGGTACACG	0.4 µm								
16	rs3212355_F	20	GAGTGAACCCAGGAAGATGC	0.4 µm	144bp	rs3212355_R	MC1R	Fwd	T/C	113	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TCC GAA GCC CAG CAG G	1.5 µm
	rs3212355_R	20	CATCAAAGGCAGACCTCTCG	0.4 µm								
17	rs8051733_F	19	AGGCGGTGGTCTCTCTCTC	0.4 µm	124bp	rs8051733_R	DEF8	Fwd	T/C	115	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTC ACC CTG CCT GTC TCG	1.6 µm
	rs8051733_R	20	TTGCAACAGGAGGGTCTAGG	0.4 µm								

Tableau 1. Le nouveau test 17-plex du système de test ADN HIrisPlex-S avec les séquences d'amorces PCR et SBE et la concentration (Lakshmi Chaitanya et al, 2018).

Tests de sensibilité et optimisation pour l'équilibre de la hauteur des pics :

La sensibilité du nouveau test 17-plex a été évaluée pour déterminer la concentration minimale d'ADN d'entrée nécessaire pour obtenir des profils SNP complets. Pour cela, l'ADN de trois individus a été dilué et quantifié en double avec le kit QuantifilerHuman DNA Quantification (AppliedBiosystems) en suivant les recommandations du fabricant à 500 pg, 250 pg, 125 pg, 63 pg et 32 pg, et enfin génotypé. Un seuil de 50 unités de fluorescence relative (RFU) a été utilisé pour appeler les pics alléliques. De plus, une évaluation des hauteurs moyennes des pics homozygotes et hétérozygotes du 17-plex de 20des réplicats à 250 pg d'ADN en utilisant 10 individus différents ont été effectués et les erreurs standard des moyennes ont été calculées.

Etudes de mélanges et tests de reproductibilité :

Des échantillons d'ADN de deux individus de génotype connu à une concentration d'ADN génomique de 500 pg ont été mélangés dans les rapports de 1:1, 1:5 et 1:10. Des mélanges d'ADN ont été utilisés pour évaluer la capacité du nouveau 17-plex à détecter des mélanges d'ADN. Des tests de reproductibilité ou de concordance ont été effectués sur 30 échantillons de différentes concentrations d'ADN, génotypes SNP et phénotypes de couleur des yeux, des cheveux et de la peau par cinq laboratoires indépendants, dont quatre n'avaient aucune expérience préalable avec le test HIrisPlex-S. Les cinq laboratoires impliqués dans les tests de concordance étaient

i) le Laboratoire médico-légal pour la recherche sur l'ADN (FLDO) du Centre médical universitaire de Leiden (LUMC), ii) le groupe de recherche de la Division des traces biologiques de l'Institut médicolégal des Pays-Bas (NFI), iii) le Centre de biotechnologie Malopolska de l'Université Jagellonne en Pologne, iv) le Département de biologie de l'Université d'Indiana PurdueUniversity Indianapolis (IUPUI) aux États-Unis, et v) le Département d'identification génétique du Centre médical universitaire Erasmus MC de Rotterdam dans le Pays-Bas. Ce dernier laboratoire qui a développé le test avec une contribution égale de l'IUPUI, a fourni aux quatre autres laboratoires des échantillons d'ADN, des réactifs, un protocole écrit, les génotypes du test HIrisPlex et une macro Excel pour la prédiction statistique de la couleur des yeux, des cheveux et de la peau. Selon le laboratoire, les fragments SBE ont été séparés et génotypés sur une machine ABI 3100 utilisant POP-4, et une machine ABI 3130xl (ou ABI 3500) utilisant le polymère POP-7. Les résultats ont été analysés à l'aide du logiciel GeneMarker (Soft Genetics) ou GeneMapper v4.0 (AppliedBiosystems).

Spécificité d'espèce, tests de stabilité et analyse d'échantillons de cas simulés :

L'origine des espèces des échantillons prélevés sur les lieux d'un crime est souvent inconnue et ils peuvent représenter une contamination provenant de différentes sources biologiques. Par conséquent, une variété d'espèces non humaines a été génotypée avec le nouveau test 17-plex pour évaluer sa spécificité d'amplification humaine.

CommercialementADNdisponibledes échantillons d'ADN de poulet, de chat, de souris, de porc, de bovin, de rat et de chien à 3 ng d'entrée et un échantillon de chimpanzé à 1 ng d'entrée ont été utilisés. L'ADN a été obtenu commercialement auprès de Novagen, Inc. (Madison, WI) pour tout sauf l'échantillon de chimpanzé, qui a été décrit ailleurs(Chaitanya et al, 2018) . Le matériel biologique humain collecté sur les lieux du crime peut contenir de l'ADN dégradé. Pour tester l'influence de l'ADN dégradé sur les performances du nouveau 17-plex, l'ADN d'un individu avec une concentration de 250 pg a été exposé à la lumière ultraviolette à des intervalles de 0 seconde, 30 secondes, 60 secondes, 5 minutes, 10 minutes, 15 minutes, 20 minutes et 30 minutes en utilisant un Bio-Link (VilberLourmat) à une force de 50 J/cm2 et génotypé. des échantillons de cas simulés ont été génotypécomprenant du sang frais, du sang séché, du sperme, un écouvillon vaginal, de la salive, du mucus nasal, des échantillons contenant de l'hématine inhibitrice de PCR, des mélanges d'ADN et des traces d'ADN de faible qualité et quantité avec le nouveau test 17-plex ainsi que le Test HIrisPlex 24 plex et dans des conditions de test d'aptitude à l'aveugle . Tous ces échantillons ont été quantifiés par PCR à l'aide du kit QuantifilerHuman DNA Quantification (AppliedBiosystems) conformément aux directives du fabricant.

Études de population :

Le nouveau test 17-plex a été utilisé pour génotyper l'ensemble d'échantillons HGDP-CEPH H952 à partir de 51 populations mondiales après élimination des duplications d'échantillons et des échantillons associés. En raison de la pénurie d'ADN, 762 individus HGDPCEPH ont finalement été utilisés pour la prédiction de la couleur de la peau en utilisant le modèle de prédiction récemment établi en combinaison avec les génotypes du test HIrisPlex 24-plex établi précédemment pour ces échantillons. Le progiciel MapViewer 7 (Golden Software, Inc., Golden, CO, États-Unis) a été utilisé pour tracer respectivement les distributions génotypiques des 17 SNP sur la carte du monde. De plus, nous avons utilisé ces données de génotype HGDP-CEPH pour générer une carte globale de prédiction de la couleur de la peau disponible avec le système de test ADN HIrisPlex-S. Afin de combler les lacunes géographiques pour la prédiction globale de la couleur de la peau à l'aide du modèle HirisPlex-S et de l'ensemble HGDP-CEPH décrit ci-dessus, 36 SNP spécifiques à la couleur de la peau ont ététéléchargé à partir de 777 individus de 8 populations (IBS, FIN, ACB, CLM, GBR, KHV, PEL, PUR) du projet 1000 génomes et inclus dans la carte globale de prédiction de la couleur de la peau.

Modèles catégoriques de prédiction de la pigmentation :

Le modèle de prédiction de la couleur de la peau humaine utilisé ici a été récemment décrit ailleurs. En bref, les catégories de couleur de peau dans le cadre de ce modèle de prédiction sont basées sur une échelle de Fitzpatrick établie dermatologiquement pour la couleur de la peau et la sensibilité au soleil. Un dermatologue a examiné l'imagerie de la couleur de la peau et un questionnaire sur la capacité des participants à bronzer, ce qui a conduit à la définition d'une affectation à l'échelle de Fitzpatrick, qui a ensuite été convertie en une catégorie de couleur de peau spécifique pour ce modèle de prédiction de la couleur de la peau comme suit : Échelle de Fitzpatrick I = Couleur de peau très pâle, II=Pâle, combinés III & IV = Intermédiaire, V=Foncé, VI=Foncé à Noir. Ce modèle de couleur de peau complète le système HIrisPlex-S, qui est désormais capable de prédire la couleur des yeux à l'aide des coefficients du modèle IrisPlex, la couleur des cheveux à l'aide des coefficients du modèle

HIrisPlex et la couleur de la peau à l'aide des coefficients du modèle HIrisPlex-S. Cet outil de prédiction de la couleur de la peau, ainsi que les outils de prédiction de la couleur des yeux et des cheveux précédemment publiés, sont mis à la disposition du public via une interface Web facile à utiliser à l'adresse https://hirisplex.erasmusmc.nl/. De plus, il y a eu des ajouts dans le nombre d'individus utilisés dans les modèles de prédiction de la couleur des yeux et des cheveux avec cette publication, ce qui explique les légères différences potentielles dans les sorties de probabilités de couleur des yeux et des cheveux par rapport aux modèles de prédiction de la couleur des yeux et des cheveux précédemment améliorés disponibles. Via le site Web HIrisPlex comme décrit précédemment par Walsh et al. En particulier, 278 individus pour la couleur des yeux et 277 pour la couleur des cheveux ont été également inclus dans une population basée aux États-Unis contenant des individus dont le lieu de naissance parental était en dehors des États-Unis dans les pays suivants (Nigeria, Mexique, Colombie, Inde, Bangladesh, Palestine, Canada, Chine, Honduras, Allemagne, Philippines, Russie, Soudan, Japon, Arabie saoudite, Pakistan, El Salvador, Espagne, Haïti, Corée du Sud, Vietnam). Ces individus sont également inclus dans le modèle de prédiction de la couleur de la peau. Ainsi, en août 2017, les ensembles de données de validation du modèle sous-jacents aux probabilités individuelles disponibles sur ce site Web se composent de 9466 individus pour la couleur des yeux, 1878 individus pour la couleur des cheveux et 1423 individus pour la peau Couleur. Lorsque les allèles d'entrée du modèle sont générés à partir des deux tests de génotypage multiplex et entrés dans l'outil de prédiction HIrisPlex-S, des probabilités individuelles pour 3 couleurs d'yeux, 4 couleurs de cheveux et 5 couleurs de peau sont générées à partir de l'œil, des cheveux et de la peau sous-jacents. Modèles de prédiction des couleurs, respectivement. Des règles d'interprétation de la couleur des yeux basées sur la catégorie avec la valeur de probabilité la plus élevée ainsi que la couleur des cheveux basée sur un guide de prédiction ont été précédemment publiéesWalsh et al, tandis que l'interprétation de la couleur de la peau est décrite ici (voir ci-dessous).

		Not significantly impacted if second highest prediction is Very Pale
	Int.	
	If highest probability > 0.9p	Intermediate predicted (*unless Dark-Black is the second highest category, then prediction is Dark)
	If highest probability > 0.7p	Intermediate is predicted however it will be affected by the second highest category if it is > 0.15 p (will appear darker if Dark/Dark-Black and lighter if Very Pale/Pale)
		Intermediate is predicted, unlikely to be affected by the second highest category if it is < 0.15 p (*unless it is Dark-Black, then prediction is Dark)
	If highest probability > 0.5p	Prediction significantly affected by second category, and will be a mix of the two highest categories (darker if Dark/Dark-Black represents the second highest category)
		Not significantly impacted if second highest prediction is Very Pale/Pale
	Dark	
	If highest probability > 0.9p	Dark predicted
	If highest probability > 0.7p	Dark is predicted, unlikely to be affected by the second highest category if it is > 0.15 p (*unless it is Dark-Black, then prediction can be Dark-Black)
		Dark is predicted, unlikely to be affected by the second highest category if it is < 0.15 p (*unless it is Dark-Black, then prediction can be Dark-Black)
	If highest probability > 0.5p	Prediction significantly affected by second category, and will be a mix of the two highest categories (dark to black if Dark-Black represents the second highest category)
		Not significantly impacted if second highest prediction is Pale/Intermediate
	Dark-Black	
	If highest probability > 0.9p	Dark to Black predicted
	If highest probability > 0.7p	Dark to Black is predicted, unlikely to be affected by the second highest category if it is > 0.15 p
		Dark to Black is predicted, unlikely to be affected by the second highest category if it is < 0.15 p
	If highest probability > 0.5p	Prediction affected by second category, and will be a mix of the two highest categories (Will be lighter than Dark to Black if Dark represents the second highest category)

Tableau 2. Guide de prédiction préliminaire pour l'interprétation des valeurs et seuils de probabilité de couleur de peau obtenus avec l'ADN HIrisPlex-S(Lakshmi Chaitanya et al,2018)

ANNEXE 4

Les définitions :

SNaPshot : test multiplex utilisé pour analyser l'identité et les SNP associé à la FDP.**MPS :**MassivelyParallelSequencing (Ion Torrent et MiSeq)

TFS: ThermoFisherScientificest une multinationale américaine fournissant du matériel de recherche et d'analyse aux laboratoires. Elle fut créée en 2006 par la fusion de Thermo Electron Corporation avec Fisher Scientific.
Primer3Plus : Logiciel utilisé pour sélectionner des paires d'amorces pour détecter la séquence modèle donnée.
Bisearch : un algorithme de conception d'amorces pour les séquences d'ADN. Il peut être utilisé à la fois pour le bisulfite converti ainsi que pour les séquences originales non modifiées. Vous pouvez rechercher différents génomes avec les amorces conçues pour éviter les produits PCR non spécifiques grâce à notre méthode ePCR rapide. Ceci est particulièrement recommandé lorsque les amorces sont conçues pour amplifier les séquences traitées au bisulfite hautement redondantes.
Autodimer : La capacité de cribler rapidement des oligonucléotides courts d'ADN pour les interactions non spécifiques est essentielle pour le développement de tests de diagnostic à base d'acides nucléiques.
AmpliTaq Gold DNA Polymérase : sont des enzymes à démarrage à chaud qui utilisent une modification chimique pour inhiber l'activité enzymatique à température ambiante
FASTQ : est un format de fichier texte permettant de stocker à la fois des séquences des séquences nucléiques
Présentation d'un fichier FASTQ minimal :
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCAC
AGTTT
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65
Kapa master mix : quantification absolue des bibliothèques NGS
Nextera index : Préparation rapide de la bibliothèque, optimisée pour les petits génomes, les amplicons PCR et les plasmides
Le fluoromètreQubit : utilise des colorants fluorescents pour déterminer la concentration d'acides nucléiques ou de protéines dans un échantillon.
PhiX Control : PhageIllumina X est une bibliothèque fiable, liée à un adaptateur, utilisée comme contrôle pour les analyses de séquençage Illumina. La bibliothèque est dérivée du petit génome PhiX bien caractérisé, offrant plusieurs avantages pour le séquençage et l'alignement.
Ion Xpress™ Plus : Le kit de bibliothèque de fragments Ion Xpress™ Plus contient des réactifs de préparation d'échantillons pour le cisaillement enzymatique et la construction de bibliothèques pour jusqu'à 20 bibliothèques d'ADN (selon le type d'ADN d'entrée) pour le séquençage des semi-conducteurs. Conserver les composants à -20°C.
Le kit de quantification Ion Library TaqMan™ : est une solution flexible et facile à

utiliser pour la quantification précise de l'entrée de la bibliothèque pour le flux de travail de séquençage des semi-conducteurs Ion Torrent. Ce kit complet fournit un mélange TaqMan® qPCR et un standard de bibliothèque pour détecter et quantifier les quantités femtomolairesde bibliothèques de fragments d'ions., le kit de quantification Ion Library TaqMan prend en charge la plupart des plates-formes en temps réel. Ce kit universel offre une plus grande précision et spécificité par rapport aux méthodes sans sonde, offrant simplicité et rapidité au flux de travail de séquençage du système PGM™.

Le kit Ion PGM™ Hi-Q™ View OT2 : permet une préparation de modèle précise et reproductible pour les bibliothèques jusqu'à 400 paires de bases à l'aide du système Ion OneTouch™ 2 et est compatible avec le système Ion PGM™. En tant que composant intégral du flux de travail de séquençage des semi-conducteurs, le kit Ion PGM Hi-Q View OT2 est conçu pour produire des particules Ion Sphere™ de haute qualité et est conçu pour être utilisé en combinaison avec le kit de séquençage Ion PGM™ Hi-Q™ View.

Le kit de puces Ion 318™ v2 BC : contient 8 puces à code-barres pour le suivi des échantillons et les analyses de séquençage à l'aide du système Ion PersonalGenome Machine® (PGM™), le premier séquenceur à base de semi-conducteurs PostLight™. La puce Ion 318™ v2 BC détecte l'incorporation de bases induites par la polymérase et traduit ces informations sous forme numérique. En éliminant l'utilisation de systèmes chimiques et optiques produisant de la lumière, cette avancée révolutionnaire dans la technologie de séquençage de nouvelle génération rend le séquençage massivement parallèle abordable pour presque tous les laboratoires.

Le kit de séquençage Ion PGM™ Hi-Q™ View : contient des réactifs et des consommables pour un séquençage robuste et très précis de bibliothèques de 200 et 400 paires de bases à l'aide du système Ion OneTouch™ 2 combiné au système Ion PGM™. Ce kit offre la chimie de séquençage la plus avancée disponible pour les utilisateurs du système Ion PGM, et son prix économique apporte une capacité de séquençage de nouvelle génération à chaque laboratoire. Utilisez le kit de séquençage Ion PGM Hi-Q View pour maximiser vos performances d'appel de variante.

Le système Ion PersonalGenome Machine (PGM) : de Thermo Fisher Scientific utilise des puces à semi-conducteurs ioniques (Il existe trois options de puce à débit variable, permettant de traiter jusqu'à 5,5 millions de lectures (à une longueur moyenne de 200 pb), jusqu'à 2 Go de sortie et 4 à 7 heures de temps de traitement) qui contiennent des millions de micropuits, chacun superposé en dessous avec un capteur d'ions pour mesurer les changements de pH. L'ADN modèle lié aux billes remplit chaque puits, procédé par inondation de dNTP d'espèces individuelles. Si l'ADN polymérase incorpore un dNTP complémentaire, un ion hydrogène est libéré et ensuite détecté, ce qui entraîne un appel de base correspondant. Le cycle des lectures se produit en parallèle, permettant un séquençage précis et à haut débit.

BWA :BURROWS-WHEELER ALIGNER est un progiciel permettant de cartographier des séquences peu divergentes par rapport à un grand génome de référence, tel que le génome humain. Il se compose de trois algorithmes : BWA-backtrack, BWA-SW et BWA-MEM. Le premier algorithme est conçu pour les séquences Illumina pouvant lire jusqu'à 100 pb, tandis que les deux autres pour les séquences plus longues allaient de 70 pb à 1 Mpb. BWA-MEM et BWA-SW partagent des fonctionnalités similaires telles que la prise en charge de la lecture longue et l'alignement fractionné, mais BWA-MEM, qui est le dernier, est généralement recommandé pour les requêtes de haute qualité car il est plus rapide et plus précis. BWA-MEM a également de meilleures performances que BWA-backtrack pour les lectures Illumina 70-100bp.

SAM : SequenceAlignmentMap (est un format texte), Il est largement utilisé pour stocker des données telles que séquences de nucléotides générées par les technologies de séquençage de nouvelle génération. Il s'agit d'un format de texte délimité par des tabulations composé d'une section d'en-tête, qui est facultative, et d'une section d'alignement. S'il est présent, l'en-tête doit être antérieur aux alignements. Les lignes d'en-tête commencent par '@', contrairement aux lignes d'alignement. **BAM :**BinaryAlignmentMap L'équivalent binaire d'un fichier SAM qui stocke les mêmes données mais selon une représentation binaire compressée. **Samtools :** est un ensemble d'utilitaires permettant d'interagir avec et de post-traitement courte séquence d'ADN lire les alignements dans le SAM, BAM et Cram formats. Ces fichiers sont générés en sortie par des aligneurs de lecture courts tels que BWA. Des outils simples et avancés sont fournis, prenant en charge des tâches complexes telles que l'appel de variantes et la visualisation de l'alignement, ainsi que le tri, l'indexation, l'extraction de données et la conversion de format. **Picard-tools :** Un ensemble d'outils de ligne de commande (en Java) pour manipuler les données et les formats de séquençage à haut débit (HTS) tels que SAM/BAM/CRAM et VCF.
BCFtools : est un ensemble d'utilitaires qui manipulent les appels de variantes dans le format d'appel de variante (VCF) et son homologue binaire BCF. Toutes les commandes fonctionnent de manière transparente avec les VCF et les BCF, à la fois non compressés et compressés par BGZF.
VCF : VariantCall Format Il s'agit d'un fichier texte qui stocke des informations génomiques, en particulier des variations génétiques, par exemple des polymorphismes mononucléotidiques (SNP). Chaque fichier VCF est divisé en une section d'en-tête, qui fournit des métadonnées décrivant le contenu restant du fichier, et le corps, qui contient toutes les différentes variantes
Mpileup:empliment, c'est une représentation par colonne des lectures alignés sur la référence au niveau de base, où chaque ligne représente une position dans le génome
Java VarScan: détection de variants dans des données de séquençage massivement parallèles La version la plus récente , est écrite en Java, elle fonctionne donc sur la plupart des systèmes d'exploitation. Il peut être utilisé pour détecter différents types de variation (Variantes germinales) (SNP et dindels) dans des échantillons individuels ou des pools d'échantillons.
Langage R : Plus qu'un logiciel, R est à la fois un puissant langage de programmation et un outil permettant de réaliser des analyses statistiques et des représentations graphiques.
L'environnement R : comprend les fonctionnalités les plus courantes. Il s'enrichit d'une multitude de paquets (packages) spécialisés écrits par des utilisateurs en fonction de besoins spécifiques. Ces paquets sont un ensemble de fonctions qui ont la particularité d'être écrits le plus souvent directement dans le langage R et n'importe quel utilisateur un peu averti peut assez rapidement écrire ses propres fonctions.
Docker : est un logiciel libre permettant de lancer des applications dans des conteneurs logiciels.

ANNEXE 5

Conception de test HIrisPlex-S pour un séquençage parallèle massif à l'aide de MiSeq (HPS-MPS-MiSeq) :

Un protocole MPS personnalisé a été utilisé pour développer le test l'Illumina MiSeq ;

Chacun des paires d'amorces ont été conçues pour isoler entre 100 et 300 pb autour du variant d'intérêt en utilisant une paire d'amorces optimale proposée à partir de l'outil de conception Web gratuit Primer3Plus.

Ces amorces comprenaient également des séquences d'adaptateur spécifiques ; permettant ainsi aux fragments ou aux amplicons d'adhérer à la « pelouse de capture » de la cellule MiSeqflow.
Ces amorces comprenaient également des séquences d'adaptateur spécifiques ; Les conceptions d'amorces ont été vérifiées avec le programme Bisearch pour s'assurer que des amplicons uniques spécifiques ont été générés.

Enfin, le programme AutoDimer a été utilisé pour vérifier les interactions amorce-dimères et / ou amorce à amorce (y compris les interactions potentielles avec les séquences adaptatrices) au sein du multiplex.

Le tableau 01 répertorie les positions de l'assemblage du génome humain du Genome Reference Consortium GRCh37 (hg19) des 41 variantes utilisées dans le système HIrisPlex-S, y compris les conceptions de paires d'amorces avec des séquences d'adaptateur incorporées pour l'Illumina MiSeq protocole, nommé HPS-MPS-MiSeq
Pour s'adapter à la plage de température de plusieurs apprêts dans le multiplex, un programme PCR touchdown (utilisant un EppendorfMastercycler Nexus SX1) a appliqué les cycles suivants :

1) 94 ° C pendant 10 min,
2) 14 cycles de 94 ° C pendant 20 s et 64 ° C (avec une température de -0,6 ° C diminue à chaque cycle supplémentaire) pendant 1 min chacun (plage de toucher des roues de 64 ° C à 55,6 ° C),
3) 20 cycles de 94 ° C pendant 20 s et 57 ° C pendant 1 min et 68 ° C pendant 30 s,
4) 72 ° C pendant 3 min,
5) maintenir à 10 ° C.

La PCR multiplex unique avait un volume total de 10 µL contenant 1 µL d'ADN génomique (concentrations variables), des amorces (tableau 1), 1X PCR gold buffer (AppliedBiosystems), 2,5 mM MgCl2 (AppliedBiosystems), 220 µM de chaque dNTP (TFS) et 2 U AmpliTaq Gold DNA polymerase (AppliedBiosystems)

Le nettoyage à base de billes suivant a utilisé un rapport de 9 µL de billes AmPure XP (Beckman Coulter, Indianapolis, IN, USA) pour 5 µL de produit PCR.

Après avoir bien mélangé, les échantillons ont été incubés pendant 5 min pour permettre liaison des billes à l'ADN ; les échantillons ont ensuite été placés sur un support magnétique pendant 5 min. Tout sauf 5 µL du surnageant a été enlevé et jeté alors qu'il était sur lesupport,

puis lavé avec 200 µL d'éthanol à 70%. L'éthanol a été éliminé de la même manière et le lavage a été répété après 30 s.

Les échantillons ont été séchés à l'air pendant 2 à 5 minutes, remis en suspension dans 20 µL d'eau purifiée et soigneusement mélangés. Après une incubation de 2 minutes, les échantillons ont été placés sur un support magnétique pendant 1 minute, puis transférés sur une nouvelle plaque.

Le deuxième cycle d'amplification par PCR a ajouté des séquences d'index uniques à chaque échantillon afin de démultiplexer (séparer) les fichiers FASTQ de chaque individu après le séquençage. Dans chaque puits 5 µL de KAPA master mix (KAPA Biosystems, Wilmington, MA), 1 µL de chaque indice Nextera (à la fois avant et arrière pour un total de 2 µL), 2 µL de H20 et 1 µL d'ADN ont été ajoutés.

Les échantillons ont été placés sur le thermocycleur avec des cycles de :
1) 98 ° C pendant 2 min,
2) 12 cycles de 98 ° C pendant 30 s et 72 ° C pendant 30 s,
3) 72 ° C pendant 5 min,
4) à 15 ° C.

Un autre nettoyage des billes, comme décrit ci-dessus, a suivi cette réaction d'indexation. Pour séquencer avec succès les 96 échantillons en un seul cycle de séquençage, les produits ont été regroupés, dilués et quantifiés comme suit pour terminer la préparation de la bibliothèque. 5 µL de chaque échantillon ont été regroupés puis quantifiés à l'aide du fluoromètreQubit (TFS) en suivant les directives standard du fabricant.

Un calculateur de dilution interne a assuré une dilution précise à une concentration globale de la bibliothèque de 2 nM.

La dénaturation de la bibliothèque a utilisé 5 µL de 0,2 N NaOH à 5 µL de la bibliothèque 2 nM. Les tubes ont été centrifugés, puis incubés pendant 5 min à température ambiante.

La bibliothèque a été diluée à 10 pM avec 990 µL de tampon d'hybridation (Illumina, San Diego, CA) comme fourni avec le kit Illumina Nextera XT version 2 (Illumina, San Diego, CA) et diluée à 8 pM en utilisant 480 µL de bibliothèque et 120 µL de tampon d'hybridation avec vortex pulsé.

Pour des résultats de séquençage optimaux, un contrôle PhiX a été ajouté à 20% pour standardiser l'analyse.

Le contrôle PhiX a été préparé comme suit : 5 pi de la bibliothèque PhiX 4 nM ont été ajoutés à 5 µL de 0,2 N NaOH, vortexés, centrifugés et incubés pendant 5 min. Une dilution supplémentaire a combiné 10 µL de bibliothèque PhiX et 990 µL de tampon d'hybridation à une concentration finale de 20 pM. La dilution finale à 12,5 pM a ensuite été réalisée en utilisant 375 µL de la bibliothèque PhiX préalablement diluée et 225 µL de tampon d'hybridation.

Avant d'ajouter les échantillons à la cartouche MiSeq, le contrôle PhiX à 20% a été ajouté à la bibliothèque personnalisée (120 µLPhiX et 480 µL de la bibliothèque), avec 600 µL de la

bibliothèque et du contrôle combinés chargés dans la cartouche. La cartouche du kit de réactifs Illumina MiSeq v2 a ensuite été exécutée en mode de séquençage Nextera XT.

Le logiciel MiSeq Reporter (Illumina, San Diego, CA) démultiplexe les échantillons et exporte deux fichiers FASTQ appariés étiquetés avec leur nom d'échantillon respectif pour une utilisation dans les analyses en aval.

Table 1
Information on the 41 DNA variants used in the HIrisPlex-S system, including the primer pair designs with incorporated adapter sequences used for the HPS-MPS-MiSeq protocol, and their concentration (primer information for HPS-MPS-ION is not available due to the commercial design by ThermoFisher Scientific).

SNP	Gene	Chromosome	Position	Ref Allele	Alt Allele	Amplicon
rs796296176	MC1R	16	89985753	A insertion	–	MC1R Amplicon 1
rs11547464	MC1R	16	89986091	G	A	MC1R Amplicon 2
rs885479	MC1R	16	89986154	G	A	
rs1805007	MC1R	16	89986117	C	T	
rs1805008	MC1R	16	89986144	C	T	
rs201326893	MC1R	16	89986122	C	A	
rs1110400	MC1R	16	89986130	T	C	
rs2228479	MC1R	16	89985940	G	A	MC1R Amplicon 3
rs1805005	MC1R	16	89985844	G	T	
rs1805006	MC1R	16	89985918	C	A	
rs1805009	MC1R	16	89986546	G	C	MC1R Amplicon 4
rs28777	SLC45A2	5	33958959	C	A	SLC45A2 Amplicon 1
rs16891982	SLC45A2	5	33951693	C	G	SLC45A2 Amplicon 2
rs12821256	KITLG	12	89328335	T	C	KITLG Amplicon
rs4959270	EXOC2	6	457748	C	A	EXOC2 Amplicon
rs12203592	IRF4	6	396321	C	T	IRF4 Amplicon
rs1042602	TYR	11	88911696	C	A	TYR Amplicon
rs1800407	OCA2	15	28230318	C	T	OCA2 Amplicon 1
rs2402130	SLC24A4	14	92801203	G	A	SLC24A4 Amplicon
rs12913832	HERC2	15	28365618	A	G	HERC2 Amplicon 1
rs2378249	PIGU	20	33218090	G	A	PIGU Amplicon
rs12896399	SLC24A4	14	92773663	G	T	SLC24A4 Amplicon
rs1393350	TYR	11	89011046	G	A	TYR Amplicon
rs683	TYRP1	9	12709305	C	A	TYRP1 Amplicon
rs3114908	ANKRD11	16	89383725	T	C	ANKRD11 Amplicon
rs1800414	OCA2	15	28197037	T	C	OCA2 Amplicon 2
rs10756819	BNC2	9	16858084	G	A	BNC2 Amplicon
rs2238289	HERC2	15	28453215	A	G	HERC2 Amplicon 2
rs17128291	SLC24A4	14	92882826	A	G	SLC24A4 Amplicon
rs6497292	HERC2	15	28496195	A	G	HERC2 Amplicon 3
rs1129038	HERC2	15	28356859	G	A	HERC2 Amplicon 4
rs1667394	HERC2	15	28530182	C	T	HERC2 Amplicon 5
rs1126809	MC1R	16	89017961	G	A	MC1R Amplicon 5
rs1470608	OCA2	15	28288121	G	T	OCA2 Amplicon 3
rs1426654	SLC24A5	15	48426484	A	G	SLC24A5 Amplicon
rs6119471	ASIP	20	32785212	C	G	ASIP Amplicon
rs1545397	OCA2	15	28187772	A	T	OCA2 Amplicon 4
rs6059655	RALY	20	32665748	A	G	RALY Amplicon
rs12441727	OCA2	15	28271775	G	A	OCA2 Amplicon 5
rs3212355	MC1R	16	89984378	C	T	MC1R Amplicon 6
rs8051733	DEF8	16	90024206	A	G	DEF8 Amplicon

Table 1 (*continued*)

SNP	Forward Primer	Reverse Primer	Product Size with adapters (bp)	Input Concentration (μM)
rs796296176	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCAGGGATCCCAGAGAAGAC	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCAGAGATGGACACCTCCAG	184	0.7
rs11547464	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGGTGAGCTTGGTGGAGA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCCAGCAGGAGGATGACG	225	0.7
rs885479				
rs1805007				
rs1805008				
rs201326893				
rs1110400				
rs2228479	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTCCAGCCTCTGCTTCCTG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGCGTGCTGAAGACGACAC	214	0.7
rs1805005				
rs1805006				
rs1805009	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAAGAACTTCAACCTCTTTCTCG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCACCTCCTTGAGCGTCCTG	173	0.5
rs28777	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTTCAAAAGGCTTCCACTCA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTTTGATGTCCCCTTCGAT	195	0.6
rs16891982	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCAAGTTGTGCTAGACCAGA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGAAAGAGGAGTCGAGGTTG	195	0.4
rs12821256	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGGCCAAAGGATAAGGAAT	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGAGCCAAGGGCATGTTACT	185	0.6
rs4959270	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGAGAAATCTACCCCCACGA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGTTCTTACCCCCTGTGGA	207	0.4
rs12203592	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGGCAGCTGATCTCTTCAG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTTCGTCATATGGCTAAACCT	193	0.5
rs1042602	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAACACCCATGTTTAACGACA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTTCATGGGCAAAATCAAT	191	0.55
rs1800407	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAGGCTGCCTCTGTTCTACG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGATGAGACAGAGCATGATGA	191	0.35
rs2402130	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACCTGTCTCACAGTGCTGCT	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTCACCTCGATGACGATGAT	217	0.35
rs12913832	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGTTCTTCATGGCTCTCTGTG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGCCCCTGATGATGATAGC	163	0.45
rs2378249	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGCATAACCCATCCCTCTAA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCATTGCTTTTCAGCCCACAC	203	0.35
rs12896399	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGGCGATCCAATTCTTTGT	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACCCTGTGTGAGACCCAGT	192	0.4
rs1393350	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTCTTTATCCCCCTGATGC	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGGAAGGTGAATGATAACACG	191	0.6
rs683	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCACAAAACCACCTGGTTGAA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCCCAGCTTTGAAAAGTATGC	194	0.8
rs3114908	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGAACACAGCCACACCCTA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACAGGAATGGCAGCTTTGAG	166	0.2
rs1800414	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCTGCAGGAGTCAGAAGGTT	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGGACAAACGAATTGAGGAA	212	0.65
rs10756819	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGACCAGTTATTTTGGGTTTGGA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGTCATGACTAGAAAAACACCAA	143	0.4
rs2238289	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGAACATGAAGATTTCCCAGT	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTGATTCAGGTCTGCTGTCACT	179	0.25
rs17128291	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAGCACTGCCAAAATAACA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTCTTTGGACCCATCACCTC	196	0.4
rs6497292	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCTGCTGTAGAACCAATGTCC	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAATTGCACCTGTAGCTCCAT	217	0.4
rs1129038	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGTCGACTCCTTTGCTTCG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACACCAGGCAGCCTACAGTC	204	0.4
rs1667394	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGCTGTAGAGAGAGACTTTGAGG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAGCAATTCAAAACGTGCAT	184	0.4
rs1126809	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGTTTCTTAGTCTGAATAACCTTTTCC	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGTGCATTGGCTTCTGGATA	167	0.4
rs1470608	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTCTTGTGTTAACTGTCCTTACAAA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGAAAATATGTTAGGGTTGATGG	212	0.8
rs1426654	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCAGCCCTTGGATTGTCTC	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGAGTAAGCAAGAAGTATAAGGAGCA	190	0.8
rs6119471	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAAAGAAGTAGCTGTACTAGACGGGAT	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAACCCGAAGGAAGAGTGAAAA	130	0.25
rs1545397	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAAGTGTTCTGGAATTGGATACTGACAA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAATGGAGATATAGAATTCACACAACA	188	0.8
rs6059655	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGAGGAAATCGAGGCTCAG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGGAGAAAGCTGCAGATCCA	179	0.45
rs12441727	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGGAAGAGACAGCTCCATGT	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACAATCCTGGGAGGTACACG	204	0.35
rs3212355	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCCACCCTTCAGCACAGA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCATCAAAGGCAGACCTCTCG	211	0.8
rs8051733	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGCGGTGGTCTCTCTCTC	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTGCAACAGGAGGGTCTAGG	191	0.3

Tableau 01 : Informations sur les 41 variantes d'ADN utilisées dans le système HIrisPlex-S, y compris les conceptions de paires d'amorces avec séquences adaptatrices incorporées utilisées pour leprotocole HPS-MPS-MiSeq, et leur concentration (les Informations sur les amorces pour HPSMPSION ne sont pas disponibles en raison de la conception commerciale par ThermoFisher Scientific). (Breslin et al, 2019).

Conception de test HIrisPlex-S pour un séquençage parallèle massif utilisant Ion Torrent (HPS-MPS-ION)

Les amorces Ampliseq ont été conçues et testées pour de bonnes performances par TFS, Erasmus MC Rotterdam et l'Université Jagellonne.
La chimie (TFS) du kit de bibliothèque **Ion AmpliSeq 2.0** a été utilisée en suivant les directives du fabricant et en utilisant les adaptateurs de code à barres **Ion Xpress** (TFS).
20 cycles d'amplification, ainsi que les étapes d'incubation, ont été réalisés sur un Veriti 96-Well Thermal Cycler (TFS), les bibliothèques ont été quantifiées à l'aide du **TaqMan™ LibraryQuantitation Kit**(TFS) sur un **CFX96 Touch Real-Time PCR machine** (Bio-Rad, Hercules, CA, USA), puis normalisée et regroupée en conséquence.
La préparation du modèle a été effectuée à l'aide du **kit Ion PGM Hi-QView OT2** (TFS) en suivant les directives du fabricant.
Le séquençage de 48 échantillons par puce a été réalisé sur le **kit de pucesIon 318 ™ v2 BCE** (TFS) à l'aide du **kit de séquençage Hi-Q View** (TFS) **IonPersonalGenome Machine** (PGM) en suivant les directives du fabricant.
La version 5.2.2 de Torrent Suite a été utilisés pour le traitement initial des données et les appels de base, les fichiers FASTQ résultants ont été exportés et utilisés pour l'analyse du pipeline en aval.

ANNEXE 6

Appel de génotype et téléchargement de l'outil Web :

Par souci de cohérence, un pipeline a été conçu pour que les deux plates-formes soient évalués à l'aide des mêmes algorithmes pour générer les 41 appels de génotypes nécessaires à l'entrée du modèle de prédiction dans l'outil de prédiction HIrisPlex-S basé sur le Web.

Les scripts et fichiers nécessaires peuvent être téléchargés à partir du site Web HIrisPlex-S disponible à l'adresse https://hirisplex.erasmusmc.nl/hps/hps.

Les données brutes ont été alignées sur les séquences de référence humaines hg19 pour tous les amplicons en utilisant l'algorithme **BWA-MEM** (BURROWS-WHEELER ALIGNER)

Le fichier d'alignement/carte de séquence **SAM** (SequenceAlignmentMap) a été converti et trié à l'aide de **SAMtools** dans un fichier **BAM** (BinaryAlignmentMap) et lire les groupes ajoutés via **Picard Tools.**

L'appel de variante a été effectué par **BCFtools** à l'aide de **mpileup** d'appels (défini sur une profondeur de lecture de 8000), (à l'aide de l'appelant multi-allélique pour tous les sites -m -M) et de commandes de requête pour l'extraction SNP.

Enfin, l'applet **Java VarScan** a été utilisée pour détecter la présence ou l'absence de l'INDEL rs796296176 (variante 1 de HIrisPlex).

Un script utilisant les package par défaut du **langage et de l'environnement R** a été utilisé pour générer le fichier de téléchargement requis pour une utilisation sur le site de l'outil Web HIrisPlex.

Le pipeline peut générer des résultats HPS-MPS pour un maximum de 96 échantillons à la fois ; cependant, ce script est personnalisable pour inclure plus d'échantillons si vous le souhaitez. De plus, l'environnement nécessaire pour exécuter ce pipeline a également été empaqueté dans une image de conteneur Docker, accessible via **Docker Hub** sous suswalsh/hpsmps.

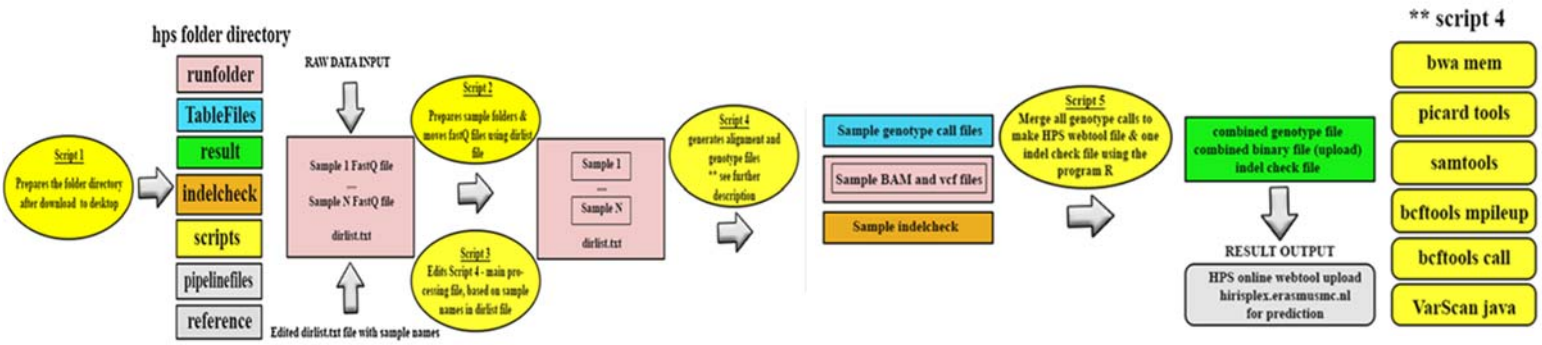


Fig. Exemple illustratif du pipeline MPS HIrisPlex-S utilisé pour évaluer et appeler des informations génotypiques à partir de données de séquençage HPS-MPS brutes et de génération de fichiers pour la saisie d'outils Web en ligne à l'aide d'un ensemble automatisé de scripts et de programmes.

ANNEXE 7

Outil de test d'échantillons de mélange et de déconvolution :

Le comptage des lectures de séquençage permet une évaluation quantitative avec des avantages pour la déconvolution du mélange, tandis que l'estimation de la hauteur des pics à l'aide des unités de fluorescence relative (RFU) générées à partir d'une analyse basée sur CE est semi-quantitative. Bien qu'il existe plusieurs critères pour détecter un mélange possible, en particulier des bilans de lecture inhabituels, il n'existe actuellement aucune directive pouvant indiquer un mélange à l'aide de données SNP autosomiques générées à partir de méthodes MPS.

Par conséquent, pour tester les performances du mélange des deux tests HPS-MPS en conjonction avec le pipeline d'analyse HPS-MPS, un outil de calcul du mélange a été conçu pour aider à la déconvolution du mélange à 2 personnes conçue spécifiquement pour la plate-forme MPS et le test MPS utilisés.

L'outil de mélange fonctionne sur la base des seuils minimaux de décompte de lecture, et un calcul d'entrée de ratio qui sépare les décomptes de lecture selon un ratio majeur : mineur, dans un mélange de 2 personnes, le tout basé sur la prémisse d'un profil STR disponible avant l'utilisation de ces outils FDP (c'est-à-dire en suivant la pratique courante de travail de cas).

En concevant l'outil de déconvolution du mélange autour de l'entrée d'un profil majeur : mineur à partir des données STR, les connaissances glanées à partir des comptes de lecture

hétérozygotes par variante et des comptes de lecture pour plusieurs scénarios de mélange à 2 personnes sont générées dans l'outil pour que l'utilisateur puisse décider à quel scénario leur échantillon ressemble le plus.

L'outil intègre également une plage à laquelle les hétérozygotes sont appelés à l'aide des informations de nombre de lectures de l'entrée 500 pg, comme décrit ci-dessus. Par exemple, pas tous les allèles hétérozygotes sont séquencés dans un rapport 50:50 pour une source unique d'échantillons, avec certains loci affichant un nombre de lectures plus élevé pour un allèle à un locus particulier.

Bien qu'il s'agisse d'un outil assez simple, il constitue la base de futures outils à automatiser davantage en utilisant ce processus/guide comme point de départ.

Une mise en garde à cet outil dans sa version actuelle est que tous les hétérozygotes étaient présents dans l'ensemble de données disponible. Donc certains HPS Variantes d'ADN telles que rs1805006 MC1R, rs1805007 MC1R, RS201326893 MC1R, RS1110400 MC1R, RS12821256 KITLG, RS12203592 IRF4, RS2378249 PIGU, RS2238289 HERC2, RS6119471 ASIP, rs6059655 RALY et rs3212355 MC1R n'ont pas leurs informations précises sur le nombre de lectures d'hétérozygotes incorporées dans cet outil à l'heure actuelle.

Pour surmonter ce déficit de données, un pourcentage conservateur de 45 :55 % la plage de déviation est actuellement appliquée pour ces variantes d'ADN HPS.

Référence et les comptes de lecture alternatifs sur chaque site sont comparés aux divers scénarios présentés dans l'outil pour déterminer le génotype profils pour les contributeurs majeurs et mineurs à l'échantillon et un le classement du meilleur scénario est généré avec une valeur et un code couleur du vert au rouge. Plus le nombre est vert et plus le nombre est bas, plus le scénario est probable.

Pour évaluer les performances de cet outil, des échantillons de mélange ont été séquencés avec les deux tests HPS-MPS à des ratios de mélange de 1:1, 1:2, 1:5 et 1:10 (x2) pour deux ensembles distincts d'individus (2 ensembles de 2 mélanges individuels), pour donner 10 types de mélanges qui ont été exécutés en double, N = 20 total pour chaque test MPS) avec différents phénotypes et génotypes.

Un évaluateur humain a été chargé d'utiliser l'outil pour déduire les profils des contributeurs sur une base variante par variante. Hormis la connaissance des ratios pour chacun des mélanges testés, l'évaluateur humain n'avait pas les génotypes des deux individus utilisés dans les mélanges à comparer, jusqu'à la fin de leur évaluation des génotypes séparés. Tous les contributeurs au mélange ont été quantifiés comme étant au-dessus des seuils de sensibilité pour l'entrée d'ADN, par conséquent, le risque d'abandon n'a pas été pris en compte dans cette évaluation. La plupart des scénarios (c'est-à-dire les profils majeurs et mineurs homozygotes pour l'allèle de référence, ou homozygote majeur et hétérozygote mineur) et donc les génotypes des deux individus, pourraient être séparés en utilisant cet outil.

Comme l'évaluation a été effectuée variante par variante, ces résultats sont présentés dans le tableau supplémentaire 8. Dans l'ensemble, 28 des 41 variantes HPS pourraient être

entièrement séparées en deux profils individuels dans 100% des échantillons sur les 40 échantillons analysés avec les deux tests MPS. Dans le cas des 13 autres variantes, trois variantes d'ADN ont entraîné des séparations de mélange incohérentes (plus de 20 erreurs ou plus de la moitié des échantillons testés) conduisant à des génotypes incorrects par personne à rs1805005 MC1R, rs4959270 EXOC2 et rs2402130 SLC24A4. Des appels de génotype incorrects ici signifient que le scénario le plus probable ne reflétait pas toujours le scénario d'ADN réellement préparé pour ces variantes. Cela pourrait être dû au fait que

i) l'échantillon préfabriqué ne reflétait pas le vrai rapport pour cette variante (c'est-à-dire que l'ADN de l'échantillon n'était pas exactement de 1:10 en ce qui concerne l'entrée d'ADN) ou ii) que plusieurs scénarios peuvent se chevaucher lors de la prise en compter.

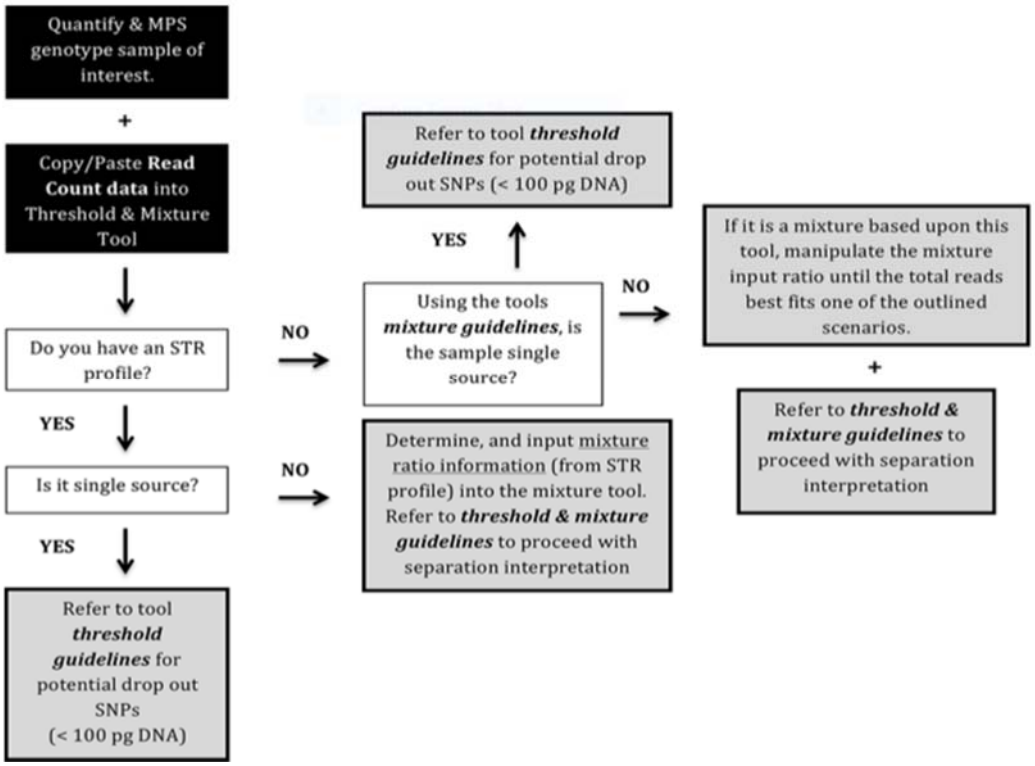
Le pourcentage d'écart pour cet outil a été calculé sur la base du déséquilibre allélique observé par variant dans son état hétérozygote où il peut y avoir une variation de 5 à 15 % du nombre de lectures dans la couverture du séquençage des allèles (c'est-à-dire le génotype GA appelé avec une profondeur de séquence de 100, l'allèle G appelé dans la séquence 40 fois, un allèle appelé 60 fois en fait un rapport 40:60, donc pour un rapport de 50:50 pour cet hétérozygote, un écart de lecture de 10 % s'applique).

Des précautions doivent être prises avec ces SNP lors de l'utilisation de cet outil pour accéder aux profils de mélange de 2 personnes à saisir dans l'outil Web de prédiction HIrisPlex-S et, par conséquent, ils ont été surlignés en rouge dans l'outil. Les 10 variantes finales ont montré un niveau d'erreur inférieur dans environ 10 à 19 cas ou 25 à 50 % du total des échantillons testés.

Enfin, il est fortement recommandé d'utiliser l'outil de mélange comme guide, tout en vérifiant manuellement à quel point les deuxièmes et troisièmes scénarii sont proches, car certains décomptes de lecture peuvent se situer entre deux scénarios, en raison de l'écart de pourcentage de lecture. Un génotypage supplémentaire d'un plus grand nombre d'individus sur ces sites erronés peut fournir une plage de lectures d'hétérozygotes plus claire pour aider à affiner le pourcentage d'écart pour les développements futurs de cet outil de déconvolution.

Il convient également de noter que les échantillons de mélange préparés dans un rapport 1:1 n'ont pas pu être appelés à l'aide de l'outil seul. Nous recommandons que dans les situations dans lesquelles les décomptes de génotypes diffèrent considérablement de tout scénario de prédiction donné, ou lorsque les informations de ratio ne fournissent toujours pas d'aide, qu'un visualiseur de génome tel que IntegrativeGenomics Viewer (IGV) soit utilisé pour visualiser le Brins d'ADN pour aider à la résolution des données de mélange. En tant qu'examen d'un type supplémentaire de scénario qui pourrait être rencontrés lors de l'interprétation du mélange d'un échantillon, nous avons testé les performances de l'outil de mélange sans connaître le rapport de mélange (c'est-à-dire aucune information de profil STR pour montrer le rapport mineur:majeur), en utilisant un échantillon de cas simulé à partir de sperme et de matériel vaginal mélangés. L'écouvillon vaginal de concentration d'ADN inconnue a été plongé dans une aliquote de sperme de concentration inconnue, et cet échantillon a été extrait pour l'ADN et passé à travers le pipeline HPS-MPS et l'outil de mélange conçu pour les deux tests HPS-MPS. Pour traiter avec succès cet échantillon, le ratio de contributeurs mineurs a été ajusté pour voir si un recompte de scénario approprié pouvait être apparié.

La déconvolution de ce mélange était possible pour les deux tests HPS-MPS sans information de ratio préalable une fois par Un ratio de contributeur mineur de 0,4 a été entré (ratio de 1 :2,5). Il est à noter que l'interprétation de l'examinateur humain est toujours nécessaire lors de la décision finale du génotype, en particulier avec les variantes gênantes mentionnées ci-dessus. Cependant, l'utilisation de cet outil a grandement facilité la déconvolution du mélange variante par variante. Dans certains scénarios, il peut ne pas être possible de diviser le profil et donc les options de génotype (c'est-à-dire signaler le profil mineur comme étant GA ou GG avec le majeur étant GG ou GA) si la séparation n'est pas facilement possible pour cette variante. Pour fournir une visualisation simple de la façon d'évaluer un échantillon en termes de source (mélange simple ou à 2 personnes) et de seuil de lecture (appels propres ou potentiel d'abandon d'allèle), un organigramme (Fig. 4) a été conçu qui indique les outils et les tableaux à utiliser pour mieux comprendre comment traiter un échantillon inconnu en utilisant à la fois des tests et des systèmes de séquençage. Un guide plus détaillé peut également être trouvé dans le matériel supplémentaire 1.



Organigramme d'interprétation de la sortie du pipeline HIrisPlex SMPS à considérer comme un contributeur unique ou une source d'ADN mixte établie à partir d'un profilage STR antérieur et des outils (c'est-à-dire un outil de déconvolution de mélange à deux personnes, un tableau de seuil de décompte, etc.)

ANNEXE 8

Test de concordance :

Les testeurs de concordance ont été invités à générer des données sur 16 échantillons inconnus dont la concentration varie de 6 pg à 25,4 ng d'entrée d'ADN, incluant ainsi des échantillons inférieurs aux seuils d'entrée d'ADN établis dans les tests de sensibilité de ces deux tests (voir le tableau supplémentaire 3 pour plus de détails). Des testeurs de concordance ont également été chargé d'exécuter les fichiers de séquence FASTQ bruts, produits par les séquenceurs, via le pipeline d'analyse HPS-MPS pour générer les appels de génotype et les fichiers d'informations de décompte nécessaires (pour plus d'informations sur le pipeline et ce qui est généré, veuillez consulter Matériel supplémentaire 2).

Enfin, il a été demandé aux testeurs de concordance d'utiliser l'outil de seuil et de mélange mis à leur disposition. (Tableau supplémentaire 2) pour générer l'interprétation du génotype de chaque échantillon. La figure 4 donne un aperçu de l'approche optimale pour échantillons d'ADN de source unique/mixte, et était la ligne directrice donnée au testeurs de concordance. Les testeurs ont utilisé cette approche pour chacun de leurs échantillons fichiers de résultats et résumé leurs résultats d'interprétation à des fins de comparaison avec l'IUPUI US et le laboratoire de développement Erasmus MC Rotterdam résultats.

Les résultats de l'étude de concordance sont disponibles dans le supplément Tableau 4, où le type de source et la concentration de l'échantillon sont donnés, ainsi que les appels d'interprétation corrects de chaque site (nombre et %) pour cet échantillon. Ce tableau montre le % de concordance entre le site de développement du test et les sites de concordance respectifs (IUPUI US pour le test HPS-MPS-MiSeq versus les deux sites de concordance MiSeq, et Erasmus MC Rotterdam pour le test HPS-MPS-ION versus les trois sites de concordance Ion Torrent). Les critères utilisés pour déclarer le génotype final par variant selon les seuils minimaux (par test/machine) et le scénario d'appels de génotypes finaux (y compris ceux utilisés dans les séparations de scénarios de mélange) sont décrits dans le tableau supplémentaire 2. La partie supérieure de ce tableau ne contient que les appels de génotypes corrects générés par les scénarios (simple vs mixte) mais ne prend pas en compte compte le seuil minimum nécessaire pour appeler un génotype variant en toute sécurité. HPS-MPS-MiSeq variait en termes de succès de génotypage de 49 % à 100 % (> 100 pg de succès moyen de génotypage d'entrée d'ADN est de 84 %) par rapport à 56 % à 100 % (> 100 pg de succès de génotypage moyen d'entrée d'ADN est de 92 %) pour HPS -MPS-ION. La partie inférieure du tableau montre que HPSMPS-MiSeq n'a pas fonctionné aussi bien que HPS-MPS-ION lors de la prise en compte du seuil de lecture minimum (qui peut également être trouvé dans le tableau supplémentaire 3 sous la colonne d'entrée d'ADN de niveau <50 pg). Ce seuil est nécessaire pour passer les critères de confiance d'appel de génotypage et les règles d'interprétation proposés par cette étude et tels que décrits dans le guide SupplementaryMaterial 1. Le résultat de la concordance HPS-MPS-MiSeq et son évaluation de l'interprétation variaient de 12 % à 85 % de concordance des résultats (> 100 pg de succès d'interprétation moyen de l'entrée d'ADN est de 33 %) avec les données de référence obtenues sur le site de développement du test IUPUI US, tandis que pour HPS-MPS-ION, il variait de 61 % à 100 % de concordance des résultats (> 100 pg de succès d'interprétation moyen de l'entrée d'ADN est de 88 %) avec les données de référence générées au Site de développement du test Erasmus MC Rotterdam. Notez que « 0 % » pour certains échantillons

indique qu'aucune donnée n'a dépassé le seuil minimal de lecture pour cet échantillon dans ce laboratoire lors de l'évaluation de l'interprétation. Dans l'ensemble, le test HPS-MPSION s'est bien comporté dans ce test de concordance dans les deux évaluations avec une concordance moyenne de 89 % avec les données de référence. Le test HPS-MPS-MiSeq a sous-performé, la concordance des données de référence des deux évaluations n'étant en moyenne que de 58 %. En raison de la conception interne et des multiples étapes nécessaires à la préparation de la bibliothèque du test MiSeq, il est possible qu'une dégradation des amorces se soit produite (en particulier compte tenu des petites amorces d'indexation nécessaires pour une étape intégrale dans le processus de préparation de la bibliothèque MiSeq) qui a affecté la diminution significative du nombre de lectures générées sur les sites de concordance MiSeq. Dans l'ensemble, les lectures moyennes pour le même échantillon d'ADN d'entrée de 100 pg (contrôle standard 9947A) exécutés par le site américain de référence IUPUI (qui avait une vaste expérience dans l'exécution de ce type particulier de conception interne) étaient environ le double (1185) de la moyenne lectures des deux sites testeurs (859 et 577) respectivement). Cela soutient la possibilité d'une dégradation de l'amorce HPS-MPS-MiSeq et/ou de l'échantillon pendant l'expédition du matériau vers les testeurs de concordance

ANNEXE 9

Matériel :

-Plink : LOGICIELPLINK est un ensemble d'outils d'analyse WGA avec des fonctionnalités de contrôle de la qualité qui sont utiles pour vérifier l'intégrité des données de la puce exome. PLINK a été conçu pour plusieurs systèmes d'exploitation ; (Illumina humanexome génotypage du regroupement de réseaux et contrôle de la qualité Yan Guo et al novembre 2014)

-R v3.02 : APPLICATION R est un langage de programmation statistique avec une excellente capacité à créer des figures. Des scripts R ont été fournis pour dessiner des chiffres de controle de qualité. R est conçu pour plusieurs systèmes d'exploitation (Yan Guo et al 2014)

-GenotypeHarmonizer (GH) : est un outil de ligne de commande pour harmoniser les ensembles de données génétiques en résolvant automatiquement les problèmes concernant le brin génomique et le format de fichier. GH résout le problème des brins inconnus en alignant les SSP A/T et G/C ambigus sur une référence spécifiée, en utilisant des modèles de déséquilibre de liaison sans connaissance préalable des brins utilisés.

-Shapeit2 v2. r790 /(v2.r644 : est une méthode rapide et précise pour l'estimation des haplotypes (akaphasing) à partir de données de génotype ou de séquençage

-Sanger imputation server V0.0.6 : service d'imputation du génotype et de mise en phase

-3dMD vectra H1 3dMDface : est une plate-forme stationnaire composée de 2 ou 3 caméras positionnées à des angles qui offrent des vues superposées du visage sous différents angles.

-Data-Driven : le data driven se base sur une approche qui consiste à prendre des décisions stratégiques sur la base d'une analyse et d'une interprétation des données.

-Analyse génerale des procustes/ analyseprocustéenne : est une technique pour comparer des formes. Elle est utilisée pour déformer un objet afin de le rendre autant que faire se peut semblable à une référence, ne laissant apparaître entre l'objet et la référence que les différences que les transformations autorisées.

-**Analyse canoniquedes corrélations** identifier et mesurer l'association entre deux groupes de variables

Analyses-en componantes principales PCA : réduit les dimensions d'une donnée multivariée à deux ou trois composantes principales, qui peuvent être visualisées graphiquement, en perdant le moins possible d'information.

GREGOR (GenomicRegulatoryElements and GwasOverlapalgoRithm) est un outil conçu pour évaluer l'enrichissement global des variantes associées à un caractère dans des éléments de régulation épigénomique annotés expérimentalement.

GREAT (GenomicRegionsEnrichment of Annotations Tool) : attribue une signification biologique à un ensemble de régions génomiques non codantes en analysant les annotations des gènes voisins.

Manhattan plot : Un diagramme de Manhattan illustre deux propriétés des résultats des études d'association pangénomiques, la localisation physique des SNP présentant des valeurs de p extrêmes, et le degré de corroboration de l'association d'un SNP par d'autres SNP voisins en déséquilibre de liaison.

QQ plot : un outil graphique permettant d'évaluer la pertinence de l'ajustement d'une distribution donnée à un modèle théorique.

Coefficient RV : il mesure le lien entre deux groupes de variables aléatoires en se basant sur la matrice de variance-covariance

MeshMonk : la boîte à outils MeshMonk produit un maillage dense de sommets sur toute la surface, facilitant ainsi des recherches plus complètes sur la variation de forme 3D

1000G phase 3 : Le moyen le plus simple de trouver des fichiers de séquences permet de rechercher des individus, des populations et des collections de données, et filtrer les fichiers par type de données et par technologies.

ChromHMM : aide à annoter le génome non codant, utilisé pour comprendre les interactions à longue distance de la chromatine, transcriptions naissantes, domaines topologiquement associés, préférences de liaison du facteur de transcription, et les perturbations des motifs réglementaires dans les essais de rapporteur massivement parallèles.

ANNEXE 10

2/Protocol Illumina

Étape 1, chargement des données dans GenomeStudio: □8 h pour 39 000 échantillons

Étapes 2 à 6, exécution du clusteringautomatique: □16 h

Étapes 7 à 22, QC sur les SNPs situés dans un génome haploïde : □4 h

Étapes 23 à 25, QC en fonction du score GenTrain: □4 h

Étapes 26 et 27, QC basé sur la séparation des grappes: □4 h

Étapes 28 à 30, QC en fonction de l'erreur mendélienne et de l'erreur de réplication (échelle avec le nombre de trios et d'échantillons dupliqués utilisés dans l'étude) : □4 h

Étape 31, QC basé sur d'autres critères : □4 h

Étapes 32 et 33, filtrage final des données : □20 min

Étape 34, appel de SNPsrares: □24 h

Étape 35, exportation des données, échelle avec la taille de l'échantillon : □4 h

Étapes 36 et 37, conversion de tous les SNPs en brin HG19 plus (échelle avec la taille de l'échantillon): □1 h

Étapes 38 et 39, vérification de l'inadéquation entre les sexes (échelle avec la taille de l'échantillon): □2 h

Étapes 40 à 43, vérification de l'inadéquation de la race (échelle avec la taille de l'échantillon): □2 h

Étapes 44 et 45, vérification de la conséité (échelle avec la taille de l'échantillon): □4 h

Étapes 46 à 48, vérification des valeurs aberrantes HWE : □2 h

Étapes 49 et 50, vérification de l'hétérozygosité et des valeurs aberrantes de consanguinité : □2 h

Étapes 51 et 52, vérification de la cohérence du génotype : □2 h

Étapes 53 à 56, vérification de la cohérence de la fréquence des allèles avec les données du projet 1000 Genomes par race : □2 h

Étape 57, vérification de la cohérence de la fréquence des allèles entre les lots (échelle avec le nombre de lots): □2 h

ANNEXE 11

3/Génotypage

Pour tous les ensembles de données, les échantillonsontétévalués pour vérifier la concordance entre le sexegénétique et le sexedéclaré, la présenced'aberrationschromosomiques, le tauxd'appel des génotypes et les effets de lot enutilisant PLINK 1.952. Les SNP ont été valués pour le tauxd'appel, les erreursmendéliennes, la déviation des proportions génotypiques de Hardy-Weinberg, et les différences de sexedans la fréquence des allèles et l'hétérozygotie, également enutilisant PLINK 1.9. Les génotypesontété "harmonisés" avec la phase 353 du projet 1000 génomes (1000G) enutilisant GenotypeHarmonizer (v1.4.20) avec unetaille de fenêtre de 200 SNP, un minimum de 10 variants, et un basé sur la fréquence des allèlesmineurs

L'imputation

Cetteméthodepermet de déduire des génotypes non-observésdans un ensemble de donnéesgénotypique à l'aided'haplotypeprovenant d'un ensemble de données de référence plus densément génotypé et pour combler le génotypemanquantcausés par l'utilisation de plate-forme de génotypages disparate.

Control de qualité

Les SNP ont été supprimés si la fréquence allélique dans l'ensemble de données de l'étude ne se situait pas dans une fourchette de 0,2| de l'une des super populations 1000G. Suppression des SNP ayant des positions en double, les insertions/délétions restantes, les variantes du nombre de copies et les génotypes haploïdes. Les individus ont été supprimés s'ils présentaient des valeurs d'hétérozygotie de ±3 écarts types par rapport à la moyenne. Les haplotypes ont été estimés en utilisant SHAPEIT2 (v2.r900).

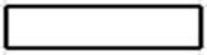
Analyse canonique

Des covariants ont été retiré (sexe, âge, poids, grandeur) parce qu'ils ne s'accommodent pas au CCA. Un autre test d'association est effectué ou multiple régression linéaire implémenter dans le PLink avec un modèle génétique additif s'ajustant à l'âge sexe IMC et l'opérateur gérant les points de repères aux composants principaux calculés à partir de SNP.

ANNEXE 12

4/25 états de chromatines

1	Active TSS	Red
2	Promoter Upstream TSS	Orange Red
3	Promoter Downstream TSS 1	
4	Promoter Downstream TSS 2	
5	Transcribed 5' preferential	Green
6	Strong transcription	
7	Transcribed 3' preferential	
8	Weak transcription	Lighter Green
9	Transcribed & regulatory (Prom/Enh)	Electric Lime
10	Transcribed 5' preferential and Enh	
11	Transcribed 3' preferential and Enh	
12	Transcribed and Weak Enhancer	
13	Active Enhancer 1	Orange
14	Active Enhancer 2	
15	Active Enhancer Flank	

16	Weak Enhancer 1	Yellow	
17	Weak Enhancer 2		
18	Primary H3K27ac possible Enhancer		
19	Primary DNase	Lemon	
20	ZNF genes & repeats	Aquamarine	
21	Heterochromatin Light	Purple	
22	Poised Promoter	Pink	
23	Bivalent Promoter	Dark Purple	
24	Repressed Polycomb	Gray	
25	Quiescent/Low	White	

Références :

- Bastien L , (2021) Machine Learning et Big Data : définition et explications https://www.lebigdata.fr/machine-learning-et-big-data

-Ballard.D. Et al, (2020). Séquençage parallèle massif en criminalistique avantage, enjeux,technicité et perspective.Revue international de médecine légale 134 page 1291-1303.

- Bonfante, B Betty Bonfante, Pierre Faux1, Nicolas Navarro, Javier Mendoza-Revilla, Morgane Dubied, Charlotte Montillot, Emma Wentworth, Lauriane PoloniCeferino Varón-González, Philippe Jones, Ziyi Xiong, Macarena Fuentes-Guajardo, Sagnik Palma, Juan Camilo Chacón-Duque, Malena Hurtado4, Valeria Villegas4, Vanessa Granja4, Claudia Jaramillo, William Arias, Rodrigo Barquera, Paola Everardo-Martínez, Mirsha Sánchez-Quinto, Jorge Gómez-Valdés, Hugo Villamil-Ramírez, Caio C. Silva de Cerqueira, Tábita Hünemeier, Virginie Ramallo, Fan Liu, Seth M. Weinberg, John R. Shaffer, Evie Stergiakouli, Laurence J. Howe, Pirro G. Hysi, Timothy D. Spector, Rolando Gonzalez-José, Lavinia Schüler-Faccini, Maria-Cátira Bortolini, Victor Acuña-Alonzo, Samuel Canizales-Quinteros, Carla Gallo, Giovanni Poletti, Gabriel Bedoya, Francisco Rothhammer, Christel Thauvin-Robinet, Laurence Faivre, Caroline Costedoat, David Balding, Timothy Cox, Manfred Kayser, Laurence Duplomb, Binnaz Yalcine, Justin Cotney, Kaustubh Adhikariet Andrés Ruiz-Linares (2021*)* Un GWAS en Amérique latine identifie de nouveaux loci de forme de visage, impliquant VPS13B et une région introgressée de Denisovan dans la variation faciale. Science Advances, Vol. 7, no.6.

-Chaitanya,L,R. (2016).Genetic Approches to Appearance and Ancestry , Improving FOrensic DNA Analysis. [These de Doctorat Erasmus University of Rotterdam].Sous la direction du (Professeur Dr H,A,P Pols) 315 pages .

-Chaitanya,L, Krystal Breslin , Sofia Zuñigac , Laura Wirken, Ewelina Pośpiech, Magdalena Kukla ,Bartoszek ,Titia Sijen ,Peter de Knijff ,Fan Liu, Wojciech Branicki ,Manfred Kaysera and SusanWalsh 2018).The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developemental valisation.Forensic Science International: Genetics (FSIGEN 1884)

- Peter Claes*, Jasmien Roosenboom, Julie D. White , Tomek Swigut, Dzemila Sero, Jiarui Li1, Myoung Keun Lee, Arslan Zaidi , Brooke C. Mattern , Corey Liebowitz , Laurel Pearson , Tomás González , Elizabeth J. Leslie, Jenna C. Carlson , Ekaterina Orlova, Paul Suetens, Dirk Vandermeulen1, Eleanor Feingold7,8, Mary L. Marazita, John R. Shaffer, Joanna Wysocka, Mark D. Shriver and Seth M. Weinberg, (2018). Genome-wide mapping of global-to-local genetic effects on human facial shape. Nature Genetics ,volume50, pages 414–423 .

- Claes,P. Peter Claes,Denise K. Liberton,Katleen Daniels,Kerri Matthes Rosana,Ellen E. Quillen,Laurel N. Pearson,Brian McEvoy,Marc Bauchet,Arslan A. Zaidi,Wei Yao,Hua Tang,Gregory S. Barsh,Devin M. Absher,David A. Puts,Jorge Rocha,Sandra Beleza,Rinaldo W. Pereira,Gareth Baynam,Paul Suetens,Dirk Vandermeulen,Jennifer K. Wagner,James S. Boster,Mark D. Shriver (2014). Modélisation de la forme faciale 3D à partir de l'ADN. Plot Genetic.

-Claes,P. (2015) .Predicting face from DNA TEDx Talks , [Vidéo]. YouTube. https://www.youtube.com/watch?v=Fii45aFKDl4&list=LL&index=25&t=335s

-Fry.H (2017) (Police identify suspect in 1997 cold-case rape and killing in Costa Mesa) https://www.latimes.com/socal/daily-pilot/tn-dpt-me-sudweeks-20170223-story.htm

- Yan Guo, Jing He, Shilin Zhao, Hui Wu, Xue Zhong, Quanhu Sheng, David C Samuels, Yu Shyr & Jirong Long. (2014) Illumina human exome genotyping array clustering and quality control Nature Protocols volume 9, pages2643–2662 .

-Gumpinger,A,C. , Damian Roqueiro 3, Dominik G Grimm , Karsten Borgwardt.,(2018).Methods and Tools in Genome-wide Association Studies Chapitre 5 (MIMB, volume 1819)

-https://snapshot.parabon-nanolabs.com/artwork

-https://snapshot.parabon-nanolabs.com/posters

-https://www.delawareonline.com/story/news/crime/2021/03/23/after-more-than-40-years-new-castle-county-homicide-victim-identified/6968087002/

-https://www.ktxs.com/archive/brown-county-sheriff-murder-suspect-wasnt-on-radar-until-composite-sketch-released

-https://hirisplex.erasmusmc.nl/

- Jylhäv.J , Nancy L Pedersen and Sara Hägg (2017). Biological Age Predictors. EBioMedicine ,VOLUME 21, Page 29-36

- Katsara ,M,A, Wojciech Branicki , susan walsh , Manfred Kayser, Michael Nothnagel.(2021) Evaluation of supervised machine-learning methods for predicting appearance traits from DNA VOLUME 53, 102507.

- Kayser,M.(2015) Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes VOLUME 18, P33-48.

- Christoph Lippert, Riccardo Sabatini, M. Cyrus Maher, Eun Yong Kang, Seunghak Lee, Okan Arikan, Alena Harley, Axel Bernal, Peter Garst, Victor Lavrenko, Ken Yocum, Theodore Wong, Mingfu Zhu, Wen-Yun Yang, Chris Chang, Tim Lu, Charlie W. H. Lee, Barry Hicks, Smriti Ramakrishnan, Haibao Tang, Chao Xie, Jason Piper, Suzanne Brewerton, Yaron Turpaz, Amalio Telenti, Rhonda K. Roby, Franz J. Och, et J. Craig Venter (2017). Identification d'individus par prédiction de caractères à l'aide de données de séquençage du génome PNAS 114 (38) 10166-10171.

-Liu .F ,atevan Duijn Johannes R.Vingerling AlbertHofman ,André G.Uitterlinden ,A. Cecile J.W.Janssens and ManfredKayser (2009). Eye color and the prediction of complex phenotypes from genotype . Current science volume 19,issue 5, pages R192-R193

-Nassir .R, Roman Kosoy, Chao Tian, Phoebe A White, Lesley M Butler, Gabriel Silva, Rick Kittles, Marta E Alarcon-Riquelme, Peter K Gregersen, John W Belmont, Francisco M De La Vega & Michael F Seldin (2009) An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels .BMC Genetics volume 10, Article number: 39

- LuQiao ,YajunYang, Pengcheng Fu SileHu ,HangZhou Shouneng Peng ,Jingze Tan, Yan Lu, Haiyi Lou, Dong sheng , Lu SijieWu ,Jing Guo, Li Jin ,Yaq Guan, Sijia Wang , ShuhuXu ,Kun Tang *al.*(2018). Variantes à l'échelle du génome de la différenciation eurasienne de la forme du visage et un modèle prospectif de prédiction faciale basée sur l'ADN. Journal de génetique Volume 45, Numéro 8,, Pages 419-432.

-Richmond .A.R, (2020) UPDATE: Human remains found in 1986 at Virginia landfill are identified as teen who disappeared at 16 .Times-Dispatch , https://newsadvance.com/news/state-and-regional/crime-and-courts/human-remains-found-in-1986-at-virginia-landfill-identified-as-16-year-old-girl/article_c92e5756-8f3f-5f34-9cda-fff0929ec20b.html

-Shriver.M.D,(2015) World Science Festival ,Genetics, it's written on your face , [Vidéo]. YouTube, https://www.youtube.com/watch?v=P_jKypC8X7o.

-. White.D, J , Karlijne Indencleef ,Sahin Naqvi 4,5, Ryan J. Eller6, Hanne Hoskens3,7, Jasmien Roosenboom8, Myoung Keun Lee8, Jiarui Li 2,3, Jaaved Mohammed4, Stephen Richmond 9 , Ellen E. Quillen 10,11, Heather L. Norton12, Eleanor Feingold13, Tomek Swigut4, Mary L. Marazita 8,13, Hilde Peeters7 , Greet Hens14, John R. Shaffer 8,13, Joanna Wysocka 4,15,16, Susan Walsh6, Seth M. Weinberg 8,13,17, Mark D. Shriver1 and Peter Claes.(2020) Insights into the genetic architecture of the human face.Génétique de la nature volume 53, pages45–53 .

-. White.D, J , Alejandra Ortega-Castrillón, Harold Matthews, Arslan A. Zaidi, Omid Ekrami, Jonatan Snyders, Yi Fan, Tony Penington, Stefan Van Dongen, Mark D. Shriver & Peter Claes (2019) MeshMonk: Open-source large-scale intensive 3D phenotyping .nature research.

-Walsh.S., Fan Liu , Andreas Wollstein, Leda Kovats , Arwin Ralf , Agnieszka Kosiniak-Kamysz , Wojciech Branicki, Manfred Kayser (2013) The HIirisPlex system for simultaneous prediction of hair and eye colour. Forensic science International:Genetics Volume 7, issue 1, page 98-115

-Walsh.S., Alexander Lindenbergh, Sofia B Zuniga, Titia Sijen, Peter de Knijff, Manfred Kayser, Kaye N Ballantyne (2011). Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence. Forensic science International:Genetics VOLUME 5, ISSUE 5, P464-471

-Walsh.S, Andreas Wollstein, Fan Liu, Usha Chakravarthy, Mati Rahu, Johan H Seland, Gisele Soubrane, Laura Tomazzoli, Fotis Topouzis, Johannes R Vingerling, Jesus Vioque, Astrid E Fletcher, Kaye N Ballantyne, Manfred Kayser (2012).DNA-based eye colour prediction across Europe with the IrisPlex system Forensic Science International: Genetics volume 6 issue 3, Page 330-340.

.

-Walsh.S . Lakshmi Chaitanya, Krystal Breslin, Charanya Muralidharan, Agnieszka Bronikowska, Ewelina Pospiech, Julia Koller, Leda Kovatsi, Andreas Wollstein, Wojciech Branicki, Fan Liu & Manfred Kayser,(2017)Global skin colour prediction from DNA .Human Genetics volume 136, pages847–863

- Xiong,Z. Ziyi Xiong, Gabriela Dankova, Laurence J Howe, Myoung Keun Lee, Pirro G Hysi, Markus A de Jong, Gu Zhu, Kaustubh Adhikari Dan Li, Yi Li, Bo Pan, Eleanor Feingold Mary L Marazita John R Shaffer Kerrie McAloney Shu-Hua Xu, Li Jin Sijia Wang Femke MS de Vrij, Bas Lendemeijer Stephen Richmond Alexei Zhurov, Sarah Lewis Gemma C Sharp, Lavinia Paternoster Holly Thompson Rolando Gonzalez-Jose, Maria Catira Bortolini Samuel Canizales-Quinteros, Carla Gallo Giovanni Poletti Gabriel Bedoya Francisco Rothhammer André G Uitterlinden, M Arfan Ikram, Eppo Wolvius, Steven A Kushner Tamar CE Nijsten, Robert-Jan TS Palstra, Stefan Boehringer Sarah E Medland Kun Tang, Andres Ruiz-Linares, Nicholas G Martin, Timothée D Spector, Evie Stergiakouli, Seth M Weinberg, Fan Liu Is a corresponding author, Manfred Kayser. (2019). Nouveaux loci génétiques affectant la variation de la forme du visage chez l’homme . elife26;8:e49898