English version following the French version.

TITRE DU PROJET : Diversité génomique de la population française (POPGEN)

RESPONSABLE : Emmanuelle Génin, Inserm – Unité Génétique, Génomique fonctionnelle et Biotechnologies (UMR 1078), Brest

RESUME

Contexte

Le développement des technologies de séquençage à haut débit a ouvert de nouvelles perspectives pour le séquençage du génome humain avec l'identification de tous les variants génétiques d'un individu. Dans chaque génome, on retrouve plus de 4 millions de différences par rapport à la séquence de référence, i.e. les variants. La majorité de ces variants sont neutres, c'est-à-dire sans impact sur le phénotype d'un individu, mais certains d'entre eux peuvent induire une maladie. Des recommandations ont été proposées pour déterminer l'implication de ces variants dans des pathologies et limiter le nombre de faux positifs. Dans ces recommandations, l'accent a été mis sur la nécessité de comparer les variants trouvés chez des patients avec des données témoins appariées autant que possible sur les origines géographiques des ascendants.

Il existe plusieurs grands panels de données de référence d'ores et déjà disponibles, comme « 1000 Genomes Project Consortium », « Exome Sequencing Project » ou « Exome Aggregation Consortium » et, qui fournissent plusieurs informations sur les variants présents dans des exomes ou des génomes entiers d'individus originaires de différentes populations, ainsi que leur fréquence respective. La population européenne y est relativement bien représentée, mais il y peu d'informations sur la région géographique d'origine de ces individus. De précédentes études utilisant les puces de SNP ont montré qu'il existe des différences de fréquences alléliques entre des populations issues d'Europe continentale ; ces différences pouvant induire des résultats faux-positifs dans les études d'association. Ces différences de fréquence allélique pour les variants communs sont également détectables entre régions d'un même pays, comme plusieurs études menées, dans différentes populations européennes, dont la France, l'ont montré. La description de ces schémas de variations génétiques stratifiées à l'échelle de la région géographique d'origine des ascendants est donc essentielle pour permettre un meilleur appariement entre les cas et les témoins dans les tests d'association. C'est d'autant plus important quand l'intérêt se porte sur les variants rares car ces derniers sont, pour la majorité, des variants apparus récemment dans des populations locales et qui ne se sont pas propagés dans de larges régions géographiques.

Avoir accès aux séquences d'ADN d'individus partageant une ascendance génétique commune avec un patient a un intérêt lors de l'analyse de génomes à visée de diagnostic. Connaître la distribution des fréquences alléliques des zones géographiques d'où sont issus les ascendants des patients sera donc indispensable pour aider à sélectionner les variants potentiellement impliqués dans une maladie ; qui, dans un second temps, devront être testés lors d'essais fonctionnels.

Un premier projet de ce genre ciblant le grand-ouest français est actuellement mené à l'Institut du Thorax de Nantes. Les premiers résultats ont montré que l'utilisation de ce panel d'individus améliore le filtrage des variants pour les analyses de génomes de patients partageant une ascendance génétique commune. Compte-tenu de ces résultats préliminaires, des efforts pour couvrir les autres régions géographiques françaises s'imposent.

Pour fournir un tel panel de la population générale française, le projet POPGEN va sélectionner 10 000 individus suivis dans la cohorte CONSTANCES ayant des ascendants originaires de différentes régions françaises, autres que celle du grand-ouest. Cette étude est un des quatre projets pilotes du

plan France Médecine Génomique 2025 (FMG) qui a pour objectifs d'introduire le séquençage du génome dans la pratique clinique quotidienne pour accélérer et affiner la pose de diagnostics.

Objectifs

L'objectif principal de cette recherche est de constituer un panel de référence de génomes séquencés provenant d'individus représentatifs de la population métropolitaine française, pour aider à identifier et interpréter les variants génétiques potentiellement impliqués dans des maladies.

Les objectifs secondaires de cette recherche sont :

- Imputer les variants génétiques des individus uniquement génotypés, à partir des 4000 génomes entiers ;
- Décrire et étudier la diversité génétique de la population française et les mécanismes à l'origine de celle-ci ;
- Développer des méthodes pour la réalisation de tests d'association avec des témoins appariés sur l'origine géographique de leurs ascendants.

Méthodes

Les participants seront recrutés parmi les volontaires de la cohorte CONSTANCES qui ont consenti à la transmission de leurs données et ont renseigné l'année et la commune de naissance de leurs parents et grands-parents. Parmi ces individus, 15000 seront sélectionnés aléatoirement sur les critères suivants :

- les lieux de naissance des 4 grands-parents sont connus et regroupés dans une zone géographique,
- la répartition homogène des regroupements en France métropolitaine.

Les volontaires sélectionnés recevront un courrier contenant : un courrier d'invitation, une notice d'information, un formulaire de consentement éclairé dupliqué et un kit d'auto-prélèvement salivaire. S'ils acceptent de participer, ils devront dater et signer le consentement éclairé et réaliser l'auto-prélèvement salivaire. Ils renverront leur échantillon salivaire ainsi qu'un exemplaire daté et signé du consentement via une enveloppe prépayée. Ces participants auront leur ADN salivaire extrait et génotypé au moyen d'une puce de SNP. Parmi ces 10000 individus, 4000 auront un séquençage du génome entier.

Perspectives

Ce panel de référence de la population française aidera à identifier et interpréter les variants génétiques potentiellement impliqués dans des maladies, dans le cadre du plan FMG2025.

Le projet POPGEN sera le support pour le développement du programme transversal Variabilité génomique, mené dans le cadre du plan FMG par l'Inserm.

TITLE OF THE PROJECT: Population Genomic Diversity of France (POPGEN)

HEAD OF THE TEAM: Emmanuelle Génin, Inserm (French National Institute of Health and Medical Research) - Genetics, functional genomics and biotechnology Unit, Brest (France)

SUMMARY

Context

The development of high throughput sequencing technologies has opened-up new possibilities to sequence the human genome and identify all genetic variants in individual genome. In each genome, more than four million differences from the reference sequence, i.e. variants, are found and the real challenge is to find "the needles in stacks of needles". A majority of these variants are neutral with no impact on individual phenotype but some of them could impair individual health and lead to disease. Guidelines have been proposed for implicating sequence variants in diseases and limiting false positive reports. In these guidelines, emphasis is put on the requirement to compare the distributions of variants between patients and large control datasets matched as closely as possible to the patients in terms of ancestry.

Several large reference datasets such as those from the 1000 Genomes Project Consortium, the Exome Sequencing Project (ESP) or the Exome Aggregation Consortium (ExAC) are publicly available and give information on the variants found in the exomes or whole genomes of individuals with ancestries in different populations and their respective frequencies. Europeans are well represented in this database. However, there is no information on the geographic region in Europe where individuals are originating from, except for the Finns that are considered separately from the rest of Europe. Previous studies using SNP-chips have shown however that there exist some differences in allele frequencies between continental European populations and that these differences could lead to false positive results in association studies. These allele frequency differences at common variants are also detectable between regions within a country as found by several studies on different European populations, including France. It is very important to describe these geographic fine-scale stratification patterns to allow an efficient matching of cases and controls in association studies. This is especially true when the interest is on rare variants as these variants are, for the majority, young variants that have recently appeared in local populations and not spread over large geographic regions.

Having access to DNA sequences from individuals that share common ancestry with patients is also of interest when analysing individual genomes for diagnoses. Information regarding allele frequency distribution in the same geographic areas where patients have ancestry will therefore be necessary to help select the variants that are the most likely involved in disease and should hence be tested in functional assays.

A first project focusing on the western part of France is ongoing at Institut du Thorax in Nantes. Preliminary results have shown that using this panel of individuals improve variant filtering in the genomes of patients with similar ancestries. Efforts should therefore be made to cover other geographic regions of France.

To provide such a general population panel for France, the POPGEN project will select 10,000 individuals from CONSTANCES cohort with ancestry in different regions of France outside western part. These participants will have their DNA extracted from salivary kits and genotyped using SNP-chip. Among these 10,000 individuals, 4,000 individuals will have their whole genome sequenced. This study is one of the four pilot projects of the France Genomic Medicine plan (FMG 2025). The FMG2025 plan aims at introducing genome sequencing in the routine clinical practice to accelerate and improve diagnoses.

Objectives

The POPGEN primary objective is to develop a reference panel of whole genome sequences of individuals, representatives of the metropolitan French population, which could be used in genetic studies to highlight variants associated with diseases.

The secondary objectives are:

- Imputation of genetic variants in the 6,000 genotyped individuals who will not be sequenced using the 4,000 WGS to achieve a total sample size of 10,000 individuals;
- Study of the population genetics mechanisms that have shaped the genomic diversity of the French population;
- Development of methods to select ancestry matched controls for association testing.

Methods

Participants will be recruited among CONSTANCES cohort volunteers who agreed transmission of their data and provided information about their parents and grandparents year and place of birth. Among all those individuals, 15,000 will be randomly selected according the following criteria:

- The place of birth of the four grandparents is known and clustered in a restricted geographic area,
- Homogeneous distribution in metropolitan France.

Selected volunteers will receive a letter containing (1) an invitation letter (2) an information leaflet explaining the objectives and the ins-and-outs of the study, (3) two copies of the informed consent form and (4) one saliva self-collection device. Volunteers will be asked to read the information leaflet and, if they agree to participate, to date and sign the informed consent form and, to collect their saliva with the self-collection device. They then send back their saliva sample and a dated and signed copy of the informed consent form in the pre-paid return envelope. These participants will have their DNA extracted from salivary kits and genotyped using SNP-chip. Among these 10,000 individuals, 4,000 individuals will have their whole genome sequenced.

Perspectives

The primary outcome of the project will be a catalogue of genetic variants found in the whole genome data with their frequency and their distribution in the different regions of France. This reference panel of whole genome sequences of individuals which could be used in genetic studies to highlight variants associated with diseases, as specified the France Genomic Medicine plan 2025.

POPGEN project will be the support to develop the cross-cutting scientific program “Genomic variability”, connected to the France Genomic Medicine plan 2025, led by Inserm.