Universitat de les Illes Balears

Facultat de Ciències

Memòria del Treball de Fi de Grau

Reaprovechamiento de secuencias de la base de datos GenBank y su potencial para la realización de estudios de genética poblacional y filogeográficos

Guillermo Andrés Fernández Olivares

Grau de Bioquímica

Any acadèmic 2017-18

Treball tutelat per: Oscar Moya Mesa
*Departament de: Biologia - Genètica

S'autoritza la Universitat a incloure aquest treball en el Repositori Institucional per a la seva consulta en accés obert i difusió en línia, amb finalitats exclusivament acadèmiques i d'investigació	Autor		Tutor	
	Sí	No	Sí	No
	X		X	

Paraules clau del treball:
Bioinformática, *GenBank*, entradas, secuencias nucleotídicas, mtDNA, *Homo Sapiens*, filogeografía, genética de poblaciones, *geocoding*, Python, BioPython, GeoPy, localizaciones de texto, coordenadas geográficas, *country, isolation_source, lat_lon*.

Índice

1. Abstract

Durante las últimas décadas, el número de secuencias que se encuentran albergadas en la base de datos *GenBank* ha incrementado de forma exponencial, en gran medida por los diferentes avances que han tenido lugar en cuanto a los métodos de secuenciación, ya sea en la metodología en sí como en los costes económicos de su utilización. *GenBank* da la opción a los investigadores de incorporar metadatos de todo tipo asociados a la nueva secuencia nucleotídica, entre ellas su ubicación geográfica de origen. Sin embargo, gran parte de las entradas alojadas no tienen ningún tipo de información geográfica o, en el caso de aquellas que sí contienen, no se proporcionan con la precisión o formato adecuado para poder reutilizarlas en estudios filogeográficos y de genética de poblaciones.

En el siguiente trabajo, de carácter bioinformático, se tratará el reaprovechamiento de las secuencias de esta base de datos como forma de dar solución a la problemática descrita. Para ello, se llevará a cabo un caso práctico utilizando entradas de *GenBank* pertenecientes al mitogenoma de *Homo Sapiens* mediante la utilización de *scripts* en lenguaje Python. A partir de los resultados obtenidos, se propone una herramienta web de dominio público como forma de dar apoyo a los investigadores pertenecientes a las áreas científicas descritas anteriormente para que puedan llevar a cabo la reutilización de secuencias durante la realización de sus estudios.

During the last decades, the number of sequences that are contained into the *GenBank* database have increased exponentially, because of advances related to new sequencing methods, either methodologic improvements and reduced costs. Moreover, *GenBank* allows researchers to incorporate different types of metadata in association with the new sequence, including its original geographical location. Furthermore, most of the sequence records in this database do not have any type of geographical information or, in the case of those that do contain, do not have enough geographical precision to carry out phylogeographic and population genetics studies.

In this work, we are going to address the reuse of this database sequences to solve this problem. We are carrying out an exercise using *Homo Sapiens* mitogenome sequence records with Python *scripts*. From the obtained results, we propose a web open source tool to help other phylogeography and population genetics scientists to do their studies reusing sequences from *GenBank*.

2. Introducción

GenBank, la base de datos de dominio público perteneciente al Centro Nacional de Información Biotecnológica (NCBI), se caracteriza por ser ampliamente utilizada en muchos campos dentro del ámbito de la investigación científica a nivel mundial – por ejemplo, filogeografía, genética de poblaciones, biomedicina, etc. -, debido a que en ella se albergan millones de secuencias nucleotídicas de diferente tipo – concretamente, existen más de 200 millones albergadas en la actualidad. Este hecho ha sido posible gracias al acuerdo de participación internacional establecido entre el propio NCBI estadounidense y el Instituto Europeo de Bioinformática (EMBL-EBI) y la base de datos de DNA de Japón (DDBJ) (*International Nucleotide Sequence Collaboration – INSDC*) (1–3).

Es importante destacar que, dentro de las más de 260 millones de entradas alojadas en esta base de datos, cada una de las secuencias se encuentra asociada a toda una serie de metadatos que representan información como, por ejemplo, el tipo de secuencia nucleotídica (DNA, mtDNA, RNA, etc.), la especie a la cual pertenece y su etnia o su ubicación geográfica de

origen. (3). Gracias a estos metadatos, los cuales se encuentran albergados en diferentes campos, las entradas de *GenBank* pasan a tener un mayor valor a la hora de ser utilizadas por investigadores, mejorando así tanto los procedimientos como la calidad de los resultados experimentales. Sin embargo, uno de los principales problemas que existe actualmente en *GenBank* es la falta de protocolos que permitan albergar, de forma estandarizada, metadatos referentes a la información geográfica para cada una de las entradas. En este sentido, cada investigador que sube una nueva secuencia a *GenBank* puede incluir la información de interés en el campo que prefiera y de la forma que más adecuada le parezca (3).

A la hora de analizar las diferentes entradas que contienen información geográfica, se aprecia notablemente esta falta de estandarización ya descrita por diferentes motivos. Uno de los más importantes se debe a la variedad de formas en las que los investigadores asocian este tipo de información a la nueva secuencia dentro del atributo *country*, el campo más utilizado en este aspecto. A pesar de que su nombre indica el país de origen de la secuencia en cuestión, a la práctica se observa como este campo también permite incorporar nombres de regiones y ciudades. Esta falta de homogeneidad a la hora de incluir información geográfica sobre el lugar de recogida de la muestra es la causante de la aparición de diferentes grados de especificidad geográfica, situación que dificulta la reutilización de secuencias para estudios de genética de poblaciones y filogeografía (3–5).

La problemática descrita provoca la aparición de toda una serie de trabas para aquellos investigadores que trabajen con secuencias nucleotídicas asociadas a localizaciones de texto. Por una parte, esta situación provoca que el investigador realice un previo proceso de normalización de los datos de forma manual, lo cual representa una tarea ardua por el tiempo y esfuerzo que ha de realizar para unificar en un mismo formato la información geográfica de cada una de las entradas con las cuales está trabajando – especialmente cuando se encuentra manipulando un gran volumen de información. Por otra parte, el hecho que existan varios niveles de precisión geográfico para las localizaciones de texto normalizadas impide obtener coordenadas geográficas adecuadas para todos los casos. Concretamente, para aquellas localizaciones con nivel de precisión bajo se obtienen coordenadas geográficas poco específicas, dificultando considerablemente la creación de árboles filogeográficos y de análisis poblacionales (3,4). Para dar solución a esta última problemática, GenBank actualizó en el año 2005 los campos disponibles para albergar metadatos, incorporando un nuevo campo específico para incorporar las coordenadas geográficas de origen, con nombre lat_lon. Sin embargo, y a pesar de que el número de entradas con coordenadas geográficas aumentó respecto a años anteriores, el número de entradas que contienen este tipo de información geográfica es mucho menor en comparación a aquellas entradas que contienen localizaciones de texto dentro del campo country – de hecho, se estima que solamente un 1% del total de entradas de GenBank contienen información geográfica en formato de coordenadas (3,5).

A pesar de la existencia de esta diferencia notable entre entradas con localizaciones de texto y entradas con coordenadas geográficas, existen estrategias para poder reaprovechar las secuencias albergadas en las diversas entradas de GenBank, mediante la utilización de técnicas de programación basadas en el *geocoding*. En este sentido, la librería GeoPy, la cual se encuentra basado en lenguaje Python, pueden ser de gran utilidad en este aspecto, puesto que contiene un gran abanico de funciones para llevar a cabo la obtención de coordenadas geográficas a partir de países, regiones, ciudades y municipios de alrededor del mundo. Este hecho se consigue gracias a que esta librería Python trabaja con buscadores de geolocalización y con la base de datos de dominio público GeoNames, la cual alberga más de once millones de localizaciones de texto (6,7).

Ahora bien, la falta de estandarización de los datos geográficos dentro del campo *country* no es la única problemática existente en cuanto a la información de este tipo presente en las entradas de *GenBank*. En este sentido, existe un gran número de entradas en esta base de datos que no contienen ningún tipo de información geográfica, tanto en el atributo *country* como en el atributo *lat_lon*. De hecho, existe un mayor número de entradas que no contienen este tipo de información geográfica en comparación a las que sí tienen este tipo de información, la cual es menor. Este hecho impide su reutilización en estudios filogeográficos. Uno de los principales motivos que se barajan a la hora de explicar por qué existe un gran número de entradas sin información geográfica en *GenBank* consiste en que muchos de los investigadores que suben nuevas secuencias a esta base de datos no consideran importante añadir localizaciones de texto o coordenadas geográficas en los campos *country* y *lat_lon*, respectivamente, debido a que no es una información relevante desde su punto de vista (8).

A pesar de que, anteriormente, se ha planteado el *geocoding* como estrategia de reaprovechamiento de secuencias de *GenBank*, se ha visto que no es de utilidad en el caso de las entradas que no contienen ningún dato geográfico, motivo por el cual se requiere de estrategias alternativas para poder llevar a cabo el reaprovechamiento en este caso descrito. Una de las soluciones que se plantea ante esta problemática consiste en leer y extraer la ubicación geográfica de origen de la publicación donde aparece la secuencia de interés. Sin embargo, y a pesar de que hay investigadores que recurren a esta técnica para la recuperación de localizaciones de texto, existen varios inconvenientes que son importantes de destacar. Por una parte, un número considerable de entradas que se encuentran alojadas en esta base de datos no tienen un enlace que permita redirigir a la publicación de origen en PubMed – se estima que aproximadamente la mitad de las entradas pertenecientes a GenBank se encuentran en esta situación descrita. Por otra parte, extraer localizaciones de texto mediante este sistema se convierte en una tarea que consume mucho tiempo y esfuerzo, especialmente cuando se trabaja con muchas secuencias que no contienen ningún tipo de información geográfica asociada. Por tanto, es una estrategia que puede ser de interés en el caso de que el investigador trabaje con un número realmente reducido de secuencias nucleotídicas, pero no cuando el número de secuencias es elevado (8,9).

A causa del elevado tiempo y recursos necesarios para llevar a cabo el reaprovechamiento de secuencias de forma manual, existen líneas de investigación bioinformáticas que buscan desarrollar *software* (programas NER - *Named Entity Recognition*) que ejecute de forma automática dicha búsqueda para, posteriormente, asociar la localización de texto encontrada con su correspondiente entrada de *GenBank*. Sin embargo, y a pesar de que ha mejorado su eficacia de forma progresiva a lo largo de los últimos años, actualmente no tiene un nivel de efectividad que permita su aplicación adecuadamente. En este sentido, las mejoras que se dan son de carácter leve, por lo que es necesario la mejora constante de estos programas bioinformáticos hasta llegar al nivel de efectividad deseado (8–10).

Ahora bien, ¿por qué es importante que las entradas de GenBank contengan información geográfica adecuada en los diferentes campos disponibles para ello? ¿Por qué es necesario el desarrollo de métodos bioinformáticos enfocados en el reaprovechamiento de secuencias nucleotídicas? Para poder llevar a cabo de forma adecuada estudios basados en filogeografía y genética de poblaciones, reutilizando secuencias ya disponibles en las bases de datos, evitando cuando sea posible la secuenciación *de novo*. La filogeografía se define como el campo científico que estudia los linajes geográficos entre especies de vertebrados e invertebrados, además de microorganismos de diverso tipo (8). Por su parte, la genética de poblaciones se define como el campo que tiene como finalidad comprender las diferencias que existen entre diferentes poblaciones a nivel genético, contribuyendo notablemente a la biología evolutiva (11). Ambos campos, al estar fuertemente influenciados por las

ubicaciones geográficas, requieren de la presencia de entradas con secuencias nucleotídicas asociadas con localizaciones de texto o, preferiblemente, con coordenadas geográficas (8,12).

Existen varios ejemplos prácticos que ilustran claramente la importancia que tiene que el conjunto de entradas alojadas en *GenBank* contenga información geográfica de elevada especificidad. Uno de ellos hace referencia al campo de la epidemiología, relacionado estrechamente con la filogeografía. Un epidemiólogo que trata de rastrear la leptospirosis en ratas hace uso de secuencias nucleotídicas para elaborar un estudio filogeográfico que dé explicación al aumento de infecciones de esta bacteria zoonótica en humanos en un país determinado. En el caso en el que este epidemiólogo utilice secuencias con información geográfica imprecisa (por ejemplo, secuencias con información geográfica estatal), únicamente podría llegar a la conclusión de que la propagación de la enfermedad tiene un carácter estatal. Sin embargo, si este epidemiólogo conociera las ubicaciones de cada una de secuencias con las que trabaja con un nivel de especificidad elevado, podría concluir que la propagación de la enfermedad se da en una región determinada y no en todo el país. Esta diferencia no solamente mejoraría la eficacia de tratamiento hacia la población, también supondría un ahorro de costes a nivel sanitario considerable, puesto que solamente se trata a la población de una región en concreto en lugar de toda la población a nivel estatal (8)

Otro ejemplo que ilustra la importancia de incorporar información geográfica precisa junto a las secuencias albergadas en las entradas de GenBank hace referencia al estudio tanto de la vía de propagación de un virus patógeno en humanos entre diferentes países como el origen de su propagación, además de estimar las posibles vías de su difusión a nivel geográfico – tal y como ilustran estudios realizados con el virus patógeno H5N1 y H7 influenza A (12).

Llegados a este punto, existe la posibilidad de aplicar diferentes tipos de metodología experimental en estudios filogeográficos. De esta forma, existen estudios que recurren a la reutilización de secuencias de GenBank para la creación de árboles filogeográficos que permitan conocer la propagación de virus patógenos a nivel geográfico o que permitan elucidar el origen de su aparición, entre otros. Cabe destacar que dichos estudios, a la hora de llevar a cabo esta metolodogía, aplican un previo filtrado para escoger aquellas entradas que contienen ubicaciones geográficas de la especificidad adecuada. Incluso se han llevado a cabo estudios donde han ido un paso más allá en este aspecto, aplicando un proceso de geoconversión para obtener las coordenadas geográficas de cada una de las localizaciones de texto seleccionadas (13–16).

Por otro lado, existe otra metodología dentro de la filogeografía que complementan la reutilización de secuencias de GenBank con la secuenciación de novo de cara a realizar el tipo de estudios descritos anteriormente (17,18). Incluso existe la posibilidad de llevar a cabo este tipo de estudios únicamente aplicando la secuenciación *de novo* de las muestras tomadas inicialmente. Uno de los estudios más destacados que reflejan la combinación de ambas metodologías se relaciona con uno de los brotes de ébola más agresivos que ha acontecido durante esta última década en Guinea, en el cual se combinaron estas dos formas de manipulación de secuencias nucleotídicas para, finalmente, concluir que dicho brote era causado por una variante diferente del virus del ébola en comparación a los que aparecieron en la República Democrática del Congo y Gabón con anterioridad (18).

Relacionado con el parágrafo anterior, la metodología de secuenciación que se lleve a cabo durante la fase experimental cobra mucha importancia. En este sentido, dichos métodos pueden basarse en *Sanger* o en *Next Generation Sequencing* (NGS). Sin embargo, y a pesar de que las técnicas NGS han experimentado una evolución a lo largo de los años – coincidiendo, además, con la evolución que ha sufrido la filogeografía en las últimas décadas – los nuevos métodos de secuenciación masiva no son utilizados de forma amplia por los

grupos de investigación que se dedican a este último campo descrito, debido a aspectos como, por ejemplo, las dificultades que surgen a la hora de manipular los datos generados a causa del gran volumen de secuencias resultantes del proceso – lo cual obliga a tener sistemas de almacenamiento de gran capacidad y herramientas bioinformáticas potentes – o la falta de consenso sobre los protocolos de preparación de bibliotecas, entre otros. En este sentido, existe la necesidad de seguir mejorando estos puntos débiles descritos, ya que la evolución de estos métodos puede ayudar a la evolución de la filogeografía (13,14). No obstante, los métodos NGS sí que se utilizan con frecuencia en los estudios relacionados con la genética de poblaciones, puesto que estos permiten identificar multitud de polimorfismos de un solo nucleótido (SNPs) como marcadores complementarios a los microsatélites, con la finalidad de mejorar la comprensión de la distribución de las variantes génicas en la población humana. De hecho, a lo largo de los últimos años, se ha observado como este aumento de marcadores ha permitido mejorar la efectividad en los estudios de este campo mencionado (15,16).

3. Materiales y métodos

3.1. Búsqueda bibliográfica

Para llevar a cabo el estudio bibliográfico, se han llevado a cabo un análisis del estudio de arte mediante el uso de diferentes motores de búsqueda especializados en publicaciones científicas a partir de las palabras clave que se detallan en el anexo 1. La selección de los artículos de interés ha sido en base a la lectura del *abstract* y adecuación del tema tratado con los objetivos que se llevan a cabo en este trabajo.

3.2. Autoaprendizaje en programación (Python)

Python es un lenguaje de programación *open source* multiplataforma que se caracteriza por su sencillez tanto a nivel de escritura como de comprensión, debido a que su sintaxis permite ejecutar todo tipo de funciones en un menor número de líneas de código en comparación a otros lenguajes de programación como, por ejemplo, Java o C++. Algunas de las características sintácticas que ayudan enormemente a que Python sea fácil de escribir y comprender por otros programadores son la capacidad de identar el código y de aplicar la programación orientada a objetos, además de la posibilidad de añadir líneas de código a lo largo del *script* por parte del programador. Por todos estos motivos, Python es uno de los lenguajes de programación que más soporte tiene a nivel mundial, gracias a la gran comunidad de programadores provenientes de un gran abanico de ámbitos que utilizan este lenguaje para llevar a cabo sus proyectos (17,18). Gracias al gran número de programadores mencionado, existen miles de librerías y paquetes Python que contienen multitud de funciones de gran utilidad para resolver todo tipo de situaciones y problemas. En este sentido, algunos de los más destacados a nivel científico con *MatPlotLib* (contiene un gran número de funciones de cara a la creación de gráficas y figuras para la representación de resultados experimentales), *Numpy* (contiene funciones relacionadas con cálculos con matrices) o *SciPy* (contiene funciones de utilidad para llevar a cabo cálculos algebraicos lineales) (19–21).

Sin embargo, dentro de las ciencias biológicas, el más destacable es *BioPython*, una librería que contiene multitud de módulos con funciones de gran utilidad para llevar a cabo la lectura y escritura de secuencias nucleotídicas, además de su alineación basado en BLAST, la creación de árboles filogeográficos o el análisis estructural de proteínas en tres dimensiones. No obstante, también incorpora toda una serie de funcionalidades que permiten acceder a las bases de datos online con información biológica relevante, de tal forma que se pueda importar dicha información para poder llevar a cabo las funciones de programación necesarias. *BioPython* continua con su desarrollo en la actualidad, gracias a que representa una

colaboración internacional de código abierto llevado a cabo por numerosos desarrolladores (22).

Para poder realizar el procedimiento experimental, previamente se llevó a cabo una fase de autoaprendizaje sobre conceptos básicos de programación en lenguaje Python, mediante la visualización de videotutoriales en la plataforma *YouTube* y posterior puesta en práctica de ejemplos de forma autónoma (existen varios canales que tienen listas de reproducción a modo de curso como, por ejemplo, *deividcoptero*, entre otros). Una vez asimilados los conceptos básicos de programación, se realizaron diversas consultas en varias páginas web y foros sobre nociones más avanzadas para poder llevar a cabo el desarrollo y mejora de los *scripts* que se describen más adelante, entre los cuales destaca el foro *stackoverflow*, uno de los más conocidos y utilizados por desarrolladores que trabajan con este lenguaje de programación. Además de dicho foro, se han realizado consultas en *Sixth Researcher*, un blog especializado en bioinformática que contiene material didáctico para aplicar funciones Python a problemas biológicos. Cabe mencionar el blog *Jarroba.com*, el cual fue de gran utilidad a la hora de aprender funciones más avanzadas en cuanto a la manipulación de listas y diccionarios, elementos que han sido muy utilizados a la hora de manipular localizaciones de texto y coordenadas geográficas durante la ejecución de los *scripts*. Por último, otra página web que ha sido de utilidad en cuanto a complementar nociones de programación ha sido *LibrosWeb*, el cual contiene entradas sobre programación en el lenguaje mencionado.

3.3 Características técnicas del equipo

Para la realización de la fase experimental, se ha utilizado un ordenador con las siguientes características a nivel de hardware: procesador Intel(R) Core(TM) i7-7500 CPU @ 2.7 GHz a 2.9 GHz; 16GB RAM, x64-bit Windows 10 Home.

3.4. Descarga de entradas pertenecientes a *GenBank*

En primer lugar, se ha realizado la descarga de los registros de *GenBank* que contienen entradas referentes a DNA de primates, mediante el programa *FileZilla Client 3.30.0*. Cada una de las entradas a descargar contienen, además de la secuencia nucleotídica, toda una serie de información distribuida en varios atributos dentro del campo */source*, los cuales pueden estar relacionados con la taxonomía, la localización geográfica, el tipo de secuencia, patologías, sexo etc. Esta información recibe el nombre de metadatos, y es una parte imprescindible para que la entrada tenga un valor biológico determinado, clave a la hora de utilizarlo en investigación. Al ejecutar el software descrito, se ha introducido el enlace web de *GenBank* donde se encuentran almacenados estas entradas para, posteriormente, seleccionar aquellas que contienen el prefijo "*PRI"* y que se encuentran en formato .seq.gz (es decir, gbpri1.seq.gz y consecutivos) – el conjunto de entradas descargadas hace referencia a la *release* 224 – febrero 2018. Una vez descargados, se han almacenado en la partición del disco duro correspondiente al programa *Python*, de tal forma que sean fácilmente accesibles durante el proceso experimental.

3.5. Ejecución *script* de extracción de metadatos de las entradas de *GenBank*

Una vez finalizada la descarga del conjunto de entradas de *GenBank*, se procede a realizar la extracción de metadatos de los archivos gbpri.seq.gz ya descargados. Para ello, se ha utilizado el *script* Python creado por Óscar Moya Mesa (su funcionamiento se describe en la figura 3). Sin embargo, es importante tener instalados los paquetes Python adecuados antes de su ejecución, pues estos proporcionan todas las funcionalidades necesarias para trabajar con los archivos mencionados. Concretamente, es necesario tener instalados los paquetes *BioPython* y *Numpy* (para el procedimiento experimental realizado, se ha utilizado *BioPython 1.71* y *Numpy 1.14.3*).

Cabe destacar la utilización del programa *JetBrains PyCharm Community Edition 2017 3.3*. a la hora de trabajar con el script mencionado - corriendo en él Python 2.7 - debido a que es un IDE (Entorno de Desarrollo Integrado) multiplataforma que contiene una gran variedad de herramientas útiles a la hora de programar. En este sentido, dicho IDE facilita el *debug* - la trazabilidad de errores conociendo el valor de las variables en tiempo de ejecución - además del *highlighting*, función que permite visualizar de forma más clara el código que se escribe en el *script* – se resalta en un color las funciones, en otro color los objetos, en otro color diferente los elementos pertenecientes a una lista, en otro color diferente los comentarios que incluye el programador a lo largo del *script*, etc. Además, *PyCharm* facilita tanto la instalación como la actualización de las librerías y paquetes Python necesarios para desarrollar dichos *scripts*, lo cual facilita y permite optimizar las horas de trabajo del programador. De hecho, es uno de los IDE más recomendados dentro del ámbito de la bioinformática, pues paquetes y librerías específicos como *Anaconda*, *MatPlotLib* o *Numpy* se integran de forma realmente efectiva en el entorno de trabajo de *PyCharm* (23).

Así pues, se ha podido poner en funcionamiento la extracción de metadatos, la cual sigue el procedimiento descrito en la figura 1. Durante el proceso descrito, se han analizado los campos */country*, */lat_lon*, */isolation_source*, */geographic_location*, */race*, */village* y */zipcode.*

3.6. Desarrollo y ejecución script de conversión *geocoding*

Una vez finalizada la extracción de metadatos descrita, se procede al desarrollo de un *script* de conversión de localizaciones de texto obtenidas para el atributo *country* a coordenadas geográficas basado en la librería *GeoPy*, la cual aporta funciones relacionadas con el *geocoding* (funcionamiento del *script* descrito en el anexo 2). Para llevar a cabo este proceso, se ha utilizado el programa *JetBrains PyCharm Community Edition 2017 3.3*, corriendo la misma versión de Python descrito en el apartado anterior, y se ha procedido a la descarga del paquete *GeoPy* (para el procedimiento experimental realizado, se ha utilizado *GeoPy 1.14*).

3.7. Mapa de entradas con datos geográficos del *geocoding*

A partir de los resultados obtenidos de la extracción de metadatos para el atributo *lat_lon* y del proceso de conversión basado en *geocoding*, se procede a representar un mapa de entradas con todo el conjunto de coordenadas geográficas obtenidas en estos dos procesos experimentales descritos. Para ello, se ha utilizado *Excel 2016* para almacenar adecuadamente los resultados descritos para, posteriormente, ser representados en un mapa geográfico mediante el programa web *Google Maps Engine*.

3.8. Tratamiento y organización de resultados

Una vez realizada la extracción de metadatos, se ha procedido a la organización de los resultados y a la obtención de las tablas y gráficas correspondientes mediante *Excel 2016*. Cabe mencionar que los cálculos estadísticos, además de los sumatorios de los datos organizados, se han llevado a cabo mediante las diferentes funciones que posee el programa descrito. Cabe destacar que este mismo proceso se ha llevado a cabo una vez finalizada la ejecución del *script* de conversión basado en *geocoding*, posterior a la realización de la extracción de metadatos ya descrita.

3.9. Estudio sobre reutilización de secuencias de GenBank

Con la finalidad de poder comprobar el grado de reutilización de secuencias de *GenBank* frente a la secuenciación *de novo* o la combinación de ambas de forma simultánea en filogeografía, se lleva a cabo un estudio bibliográfico en *PubMed Central*, una base de datos perteneciente a la NCBI que contiene miles de publicaciones científicas de acceso gratuito (*full text articles*). Dicho estudio comienza con la búsqueda en dicha base de datos mediante

la utilización de las palabras clave "*phylogeography homo sapiens*", debido a que interesa conocer la cuestión planteada inicialmente en estudios filogeográficos llevados a cabo para la especie nombrada. De esta forma, se consigue obtener un total de 5199 resultados.

A partir de estos 5199 resultados mencionados, se leen los títulos de las publicaciones resultantes, descartando aquellos que hacen referencia a *reviews*, debates y perspectivas de futuro sobre filogeografía, además de aquellas que no se encuentran dentro del criterio de búsqueda planteado anteriormente (de esta forma, se descartan estudios filogeográficos relacionados con otras especies, además de microorganismos patógenos y virus). Por su parte, se seleccionan aquellas que sí hacen referencia a estudios de la rama descrita aplicados en *Homo Sapiens* – concretamente, 20 publicaciones. De estas 20, se lee el apartado referente a los materiales y métodos de experimentación llevados a cabo en cada uno de ellos, de tal forma que se han ido clasificando en tres grupos: (1) secuenciación de novo, (2) reutilización de secuencias públicas y (3) ambas (ver figura 1).

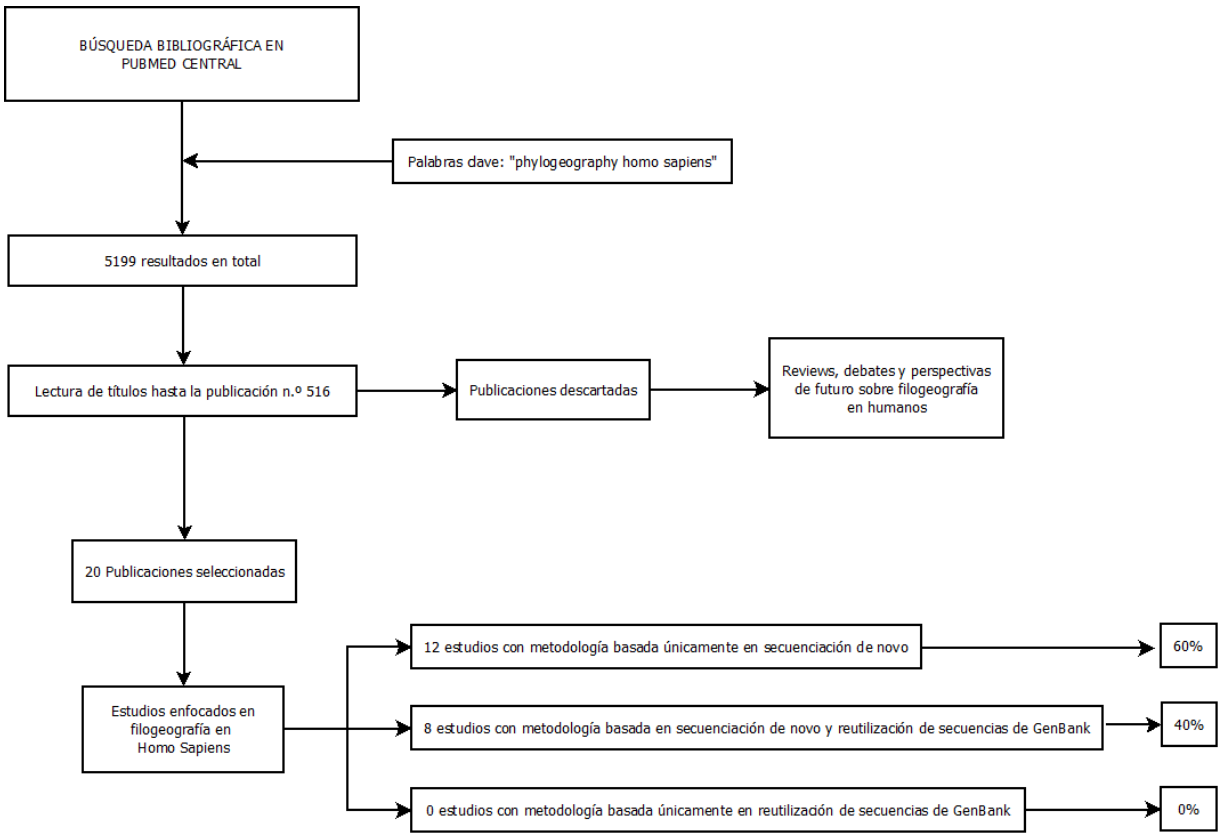

Figura 1: diagrama de flujo donde se representa el procedimiento llevado a cabo en el estudio bibliográfico, además de los criterios de búsqueda y filtro de resultados aplicados

4. Resultados experimentales

4.1. Distribución de la información geográfica por atributos en las entradas de GenBank relacionadas con el mitogenoma de Homo Sapiens

Tabla 1. Distribución de la información geográfica por atributos en las entradas de mitogenoma de *Homo Sapiens* de *GenBank* (febrero 2018)

Información	N.º de entradas	% de entradas
Entradas de mtDNA de *Homo Sapiens* analizadas en total	42597	100
Entradas de mtDNA de *Homo Sapiens* con información geográfica en el atributo ***/country***	16229	38.1
Entradas de mtDNA de *Homo Sapiens* con información geográfica en el atributo ***/isolation_source***	5571	13.08
Entradas de mtDNA de *Homo Sapiens* con información geográfica en el atributo ***/lat_lon***	309	0.73
Entradas de mtDNA de *Homo Sapiens* con información geográfica en el atributo ***/geographic_location***	0	0
Entradas de mtDNA de *Homo Sapiens* con información geográfica en el atributo ***/race***	0	0
Entradas de mtDNA de *Homo Sapiens* con información geográfica en el atributo ***/village***	0	0
Entradas de mtDNA de *Homo Sapiens* mates con información geográfica en el atributo ***/zipcode***	0	0

Tabla 2. Tipo de información almacenada dentro del atributo */isolation_source* para las entradas de mtDNA de *Homo Sapiens* de *GenBank*

Información	N.º de entradas	% de entradas
Entradas de mtDNA de *Homo Sapiens* con información geográfica en el atributo ***/isolation_source***	5571	100
Información sobre etnias	3297	59.18
Información sobre pacientes (etnia y enfermedad o sólo enfermedad)	1076	19.31
Información sobre ubicaciones geográficas	650	11.67
Otro tipo de información	468	8.4
Información sobre halogrupos	80	1.44

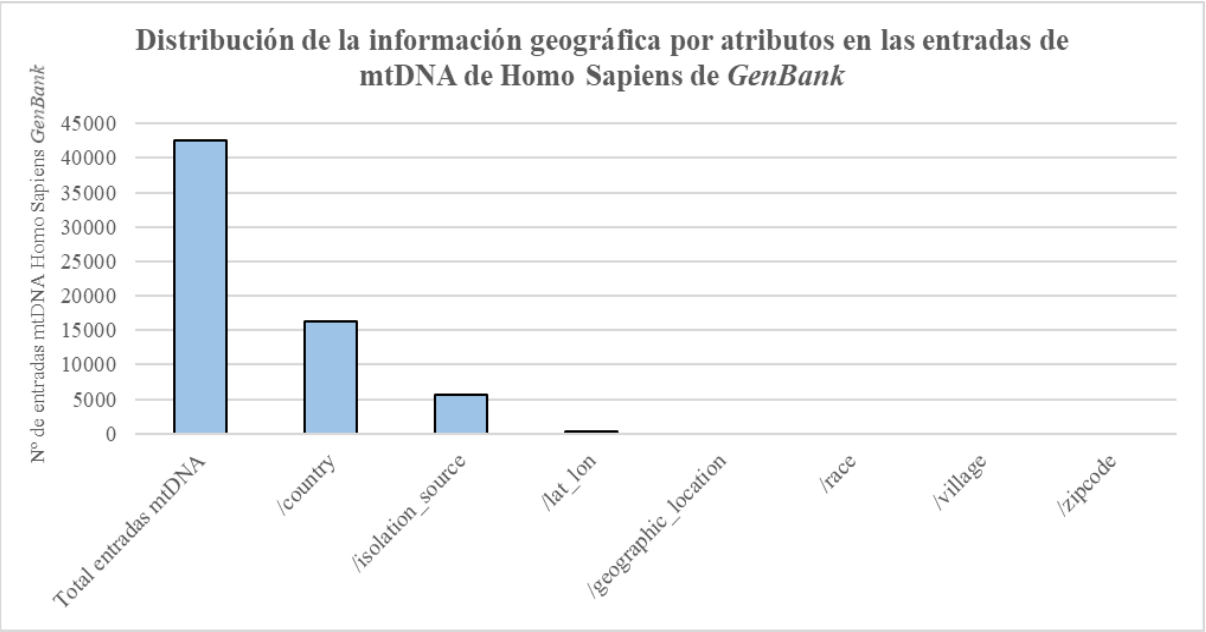

Figura 2. Distribución de la información geográfica por atributos en las entradas de mtDNA de *Homo Sapiens* de GenBank (febrero 2018).

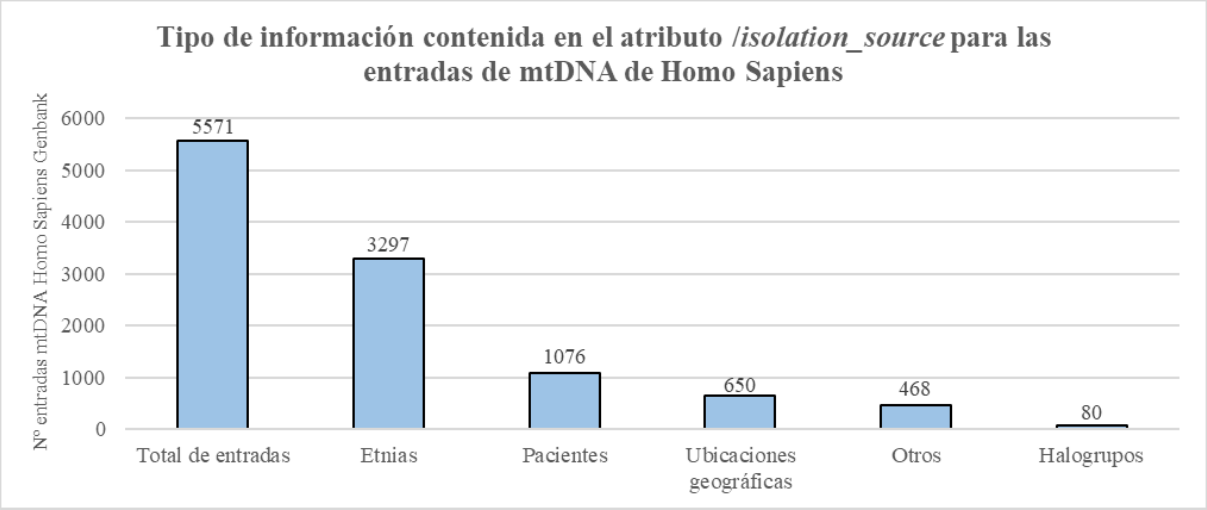

Figura 3. Tipo de información que se encuentra contenida dentro del atributo /isolation_source para las entradas de mtDNA de *Homo Sapiens* de GenBank.

Tal y como se muestra en la tabla 1 y la figura 1, se ha conseguido obtener un total de 42597 entradas de GenBank relacionadas con el mitogenoma de *Homo Sapiens*. De todo este conjunto de entradas, un 38,1% contiene información geográfica en el atributo *country*, mientras que un 13.08% contiene información en el atributo *isolation_source*. Además, un 0.73% de dichas entradas contienen información sobre ubicaciones geográficas en el atributo *lat_lon*. Por último, cabe destacar que no se encuentran entradas dentro de todo este conjunto que contengan información geográfica en los atributos *geographic_location*, *race*, *village* y *zipcode.*

Por su parte, la tabla 2 y la figura 2 muestra cómo, del total de entradas que contienen información en el atributo *isolation_source*, un 59.18% contiene información relacionada con las etnias, mientras que un 19.31% contiene información relacionada con la etnia de los pacientes o con pacientes que poseen determinadas enfermedades (Diabetes, Obesidad, Alzheimer, Parkinson, etc.). Además, un 11.67% de las entradas con información en este

atributo contienen localizaciones de texto referentes a continentes, países, regiones y ciudades. Por último, un 1.44% del total de entradas descritas contienen información referente a halogrupos. Cabe destacar que, dentro de todo este conjunto de entradas, un 8.4% contienen información diferente a la descrita anteriormente.

4.2. Nivel de especificidad geográfica de las entradas de *GenBank* relacionadas con el mitogenoma de *Homo Sapiens*

Tabla 3. Nivel de especificidad geográfica para las entradas de genoma mitocondrial de *Homo Sapiens* alojadas en la base de datos *Genbank* (febrero 2018).

Información	N.º de entradas	% de entradas
Entradas de mtDNA de *Homo Sapiens* analizadas en total	42597	100
Entradas de mtDNA de *Homo Sapiens* sin información geográfica (nivel 0)	18534	43.5
Entradas de mtDNA de *Homo Sapiens* con información geográfica de nivel 1	16229	38.1
Entradas de mtDNA de *Homo Sapiens* con información geográfica de nivel 2	6862	16.11
Entradas de mtDNA de *Homo Sapiens* con información geográfica de nivel 3	954	2.24

Tabla 4. Nivel de especificidad geográfica para las entradas de genoma mitocondrial de *Homo Sapiens* alojadas en la base de datos *Genbank* (febrero 2018).

Información	N.º de entradas	% de entradas
Entradas de mtDNA de *Homo Sapiens* analizadas en total	42597	100
Entradas de mtDNA de *Homo Sapiens* con información geográfica imprecisa (nivel 0 + 1)	34763	81.61
Entradas de mtDNA de *Homo Sapiens* con información geográfica específica (nivel 2 + 3)	7816	18.35

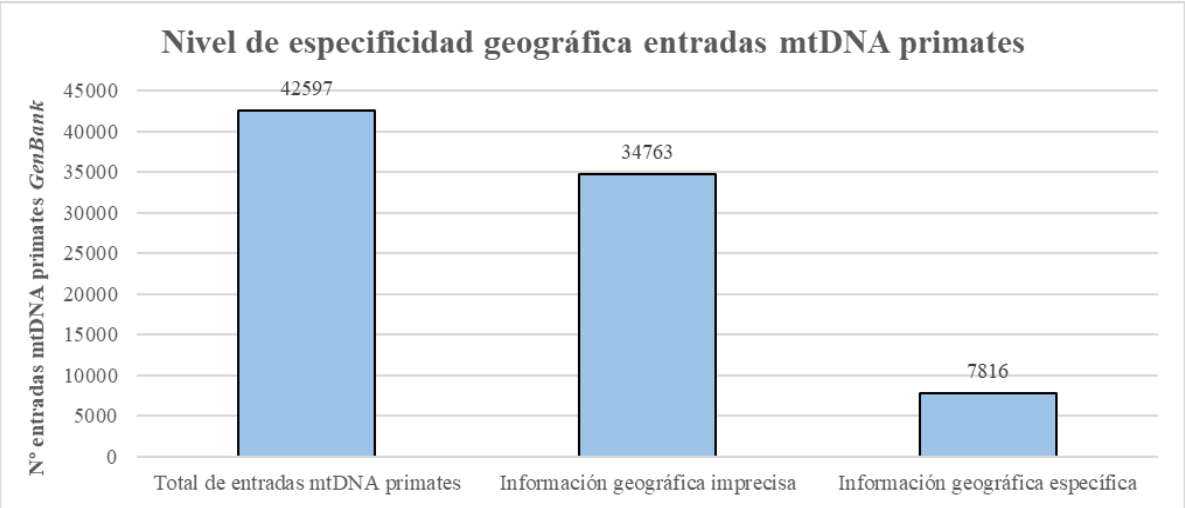

Figura 4. Nivel de especificidad geográfica para las entradas de genoma mitocondrial de *Homo Sapiens* alojadas en la base de datos GenBank (febrero 2018).

Tal y como refleja la tabla 3, se ha conseguido obtener un total de 42597 entradas de *GenBank* relacionadas con el mitogenoma de *Homo Sapiens*. De todo este conjunto de entradas, un 43.5% no contienen ningún tipo de información geográfica (nivel 0), mientras que el resto de las entradas contienen localizaciones de texto de diferente grado de especificidad. Dentro de este último grupo, un 38.1% solamente contienen localizaciones de texto de nivel 1 de especificidad geográfica (referentes solamente al país de origen); mientras que un 16.11% de este conjunto de entradas descritas contienen localizaciones de texto de nivel 2 de especificidad geográfica (referentes al país de origen, además de la región o la ciudad en cuestión). Por su parte, solamente un 2.24% de dichas entradas contienen localizaciones de texto pertenecientes a un nivel 3 de especificidad geográfica (referentes al país de origen, además de la región y la ciudad en cuestión).

Por otra parte, la tabla 3 y la figura 2 representa el porcentaje de entradas con información geográfica imprecisa o específica (81.61% y 18.35%, respectivamente). Cabe destacar que la información referente a las tablas y gráfica mencionadas provienen del atributo /*country*.

4.3. Comparación de uso entre localizaciones de texto y coordenadas geográficas a la hora de proporcionar información geográfica a las entradas de *GenBank*

Tabla 5. Comparación cuantitativa entre aquellas entradas que contienen localizaciones de texto y aquellas que contienen coordenadas geográficas

Información	N.º de entradas	% de entradas
Entradas de mtDNA de *Homo Sapiens* analizadas en total	42597	100
Entradas de mtDNA de *Homo Sapiens* con localizaciones de texto	16229	38.1
Entradas de mtDNA de *Homo Sapiens* con coordenadas geográficas	309	0.73

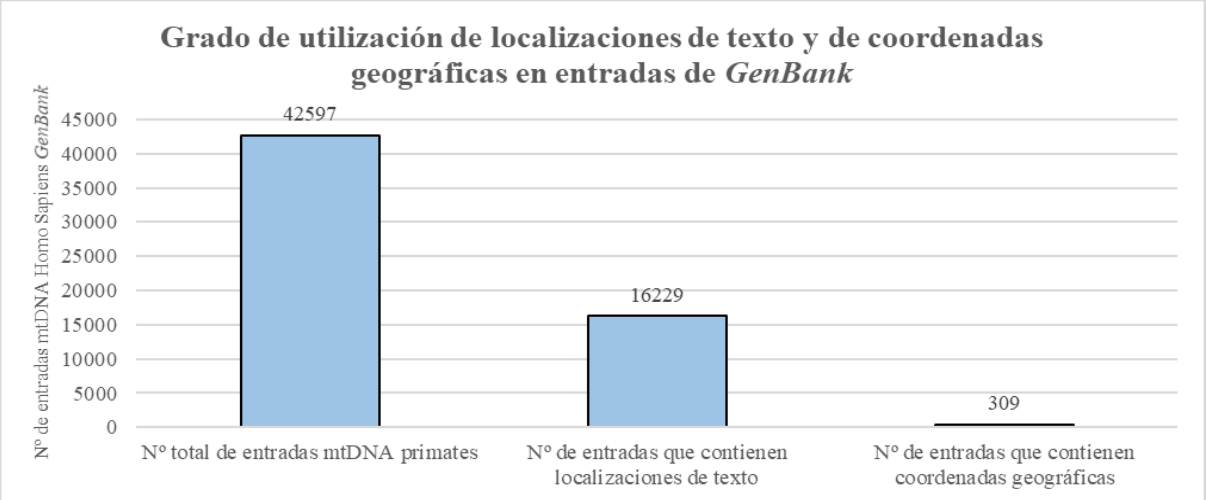

Figura 5. Grado de utilización de diferentes formatos de información geográfica para las entradas de genoma mitocondrial de *Homo Sapiens* alojadas en la base de datos GenBank (febrero 2018).

Tal y como refleja la tabla 4 y la figura 3, existe un total de 16229 entradas de *GenBank* que contienen información geográfica en formato de localización de texto, mientras que un total de 309 entradas de mitogenoma de *Homo Sapiens* que contienen coordenadas como información geográfica. Cabe destacar que la información referente a estas tablas y gráfica provienen del atributo /*country* y del atributo */lat_lon*.

4.4. Conversión de localizaciones de texto a coordenadas geográficas de entradas de *GenBank* mediante el desarrollo y uso de *scripts* Python basados en *geocoding*

Tabla 6. Comparación cuantitativa entre aquellas entradas que contienen localizaciones de texto y aquellas que contienen coordenadas geográficas una vez ejecutado el script de *Geocoding* en lenguaje Python.

Información	N.º de entradas	% de entradas
Entradas de mtDNA de *Homo Sapiens* analizadas en total	42597	100
Entradas de mtDNA de *Homo Sapiens* con localizaciones de texto	2124	13.09
Entradas de mtDNA de *Homo Sapiens* con coordenadas geográficas	14414	86.91

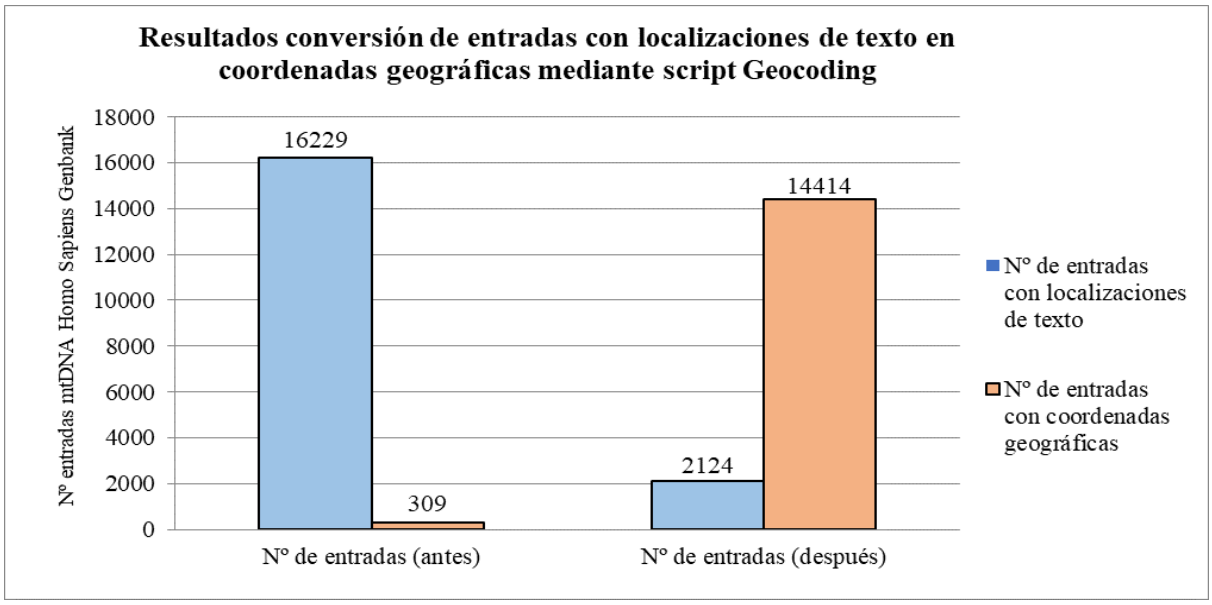

Figura 6. Comparación cuantitativa entre aquellas entradas que contienen localizaciones de texto y aquellas que contienen coordenadas geográficas una vez ejecutado el script de *geocoding* en lenguaje Python.

Tal y como reflejan la tabla y figura 6, el porcentaje de entradas que han dado un resultado satisfactorio en la ejecución del *script* Python de *geocoding* es de un 86.91%, mientras que un porcentaje más reducido de entradas (13.09%) han reportado algún tipo de error durante el proceso de conversión de localizaciones de texto a coordenadas geográficas.

4.5. Mapa de coordenadas a partir de los resultados obtenidos en el proceso de *geocoding*

Ilustración 1: Mapa geográfico con las coordenadas pertenecientes a las entradas con información lat_lon y a las localizaciones de texto convertidas mediante el script de geocoding.

La ilustración 1 representa el mapa geográfico generado a partir de las coordenadas geográficas contenidas en el atributo *lat_lon* (extracción de metadatos) y a partir de aquellas que proceden de la conversión *geocoding* de localizaciones de texto del atributo *country*. Es posible acceder a una versión interactiva del mapa siguiente el enlace que se adjunta a continuación. En ella es posible hacer *zoom*, desplazarse y ver cuántas secuencias hay asociadas a las coordenadas geográficas representadas (www.google.com/maps/d/edit?mid=1Fd-rjEWo3W6pez_ItO3ZruWK37gAkjHn&ll=9.320721545582412%2C0&z=2).

4.6. Reutilización de secuencias de GenBank frente a la secuenciación de novo en estudios de filogeografía en *Homo Sapiens*

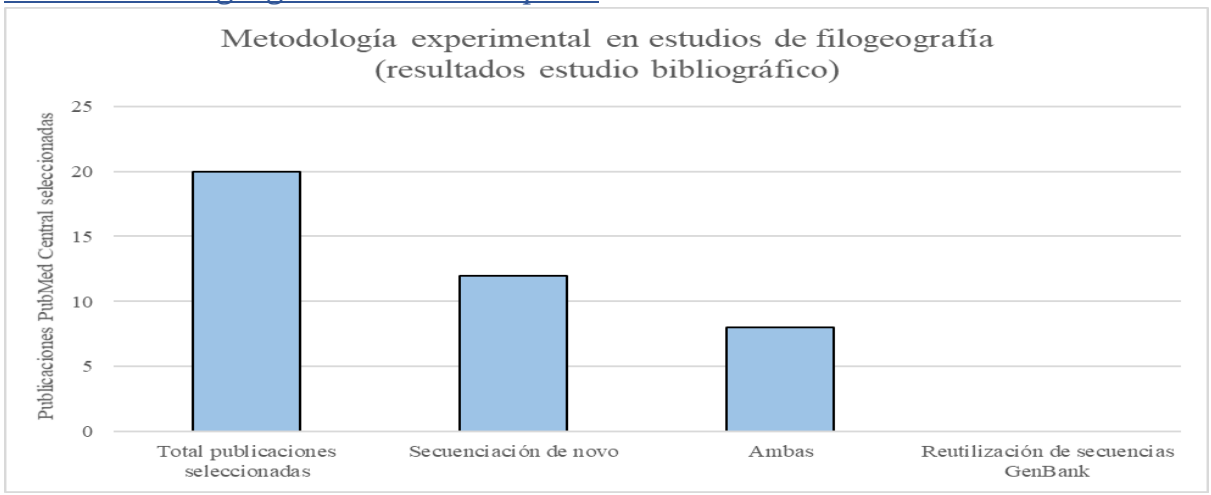

Figura 7: comparación entre metodologías experimentales aplicadas en estudios de filogeografía (a partir del estudio bibliográfico llevado a cabo - ver figura 1).

Tal y como refleja la figura 7, la metodología experimental que más se aplica en las publicaciones analizadas es la secuenciación *de novo* - 60% del total seleccionadas (24,25,26–33,34), seguido de la utilización tanto de la metodología descrita anteriormente como de la reutilización de secuencias de *GenBank* - 40% respecto el total (35–42). Por su parte, no se ha encontrado ninguna publicación dentro de las seleccionadas que recurra únicamente a la reutilización de secuencias de la base de datos descrita.

5. Discusión

Los resultados obtenidos una vez finalizado el análisis bioinformático muestran cómo, para las entradas de mtDNA de *Homo Sapiens* alojadas en *GenBank*, el atributo más utilizado para almacenar metadatos, entre los estudiados, es *country*, seguido de *isolation_source* y de *lat_lon* (figura 2). El tipo de información que se encuentra en *country* son localizaciones en formato texto, los cuales hacen referencia a la ubicación geográfica donde se obtuvo la secuencia nucleotídica. Dichas localizaciones pueden abarcar tanto áreas geográficas de gran extensión – países, estados, regiones, comunidades autónomas, etc. - como áreas de poca superficie – ciudades, municipios, pueblos, etc. Además, estas ubicaciones de texto descritas pueden encontrarse tanto de forma individual como de forma conjunta, provocando así la aparición de diferentes niveles de precisión geográfica. Cabe destacar que esta observación se cumple en otros muchos estudios similares al caso práctico llevado a cabo, ya que un número elevado de entradas alojadas actualmente en esta base de datos contienen información geográfica dentro del atributo mencionado (5).

Otro de los atributos que se encuentra presente en un elevado número de entradas en este caso práctico es *isolation_source*, el cual contiene información relacionada con la etnia de origen de la persona que proporcionó la muestra de la que proviene la secuencia nucleotídica. Sin embargo, a pesar de hacer referencia a este tipo de metadatos, el análisis de resultados ha reflejado que en este atributo también se proporciona información sobre pacientes que padecen una determinada patología (diabetes, obesidad, Parkinson, Alzheimer, etc.) y con halogrupos de diferente tipo, aunque muchas otras entradas contienen información que no se engloba en los tipos de información descritos en la figura 3. Por tanto, estos resultados observados para *isolation_source* dan a ver la elevada variedad de información que contiene este campo. No obstante, dicha variedad de información es de gran utilidad para llevar a cabo estudios de genética de poblaciones, no solamente porque contiene información de utilidad para llevar a cabo análisis de distribución alélica y genotípica entre poblaciones, también para llevar a cabo estudios dentro del ámbito de la biomedicina. De hecho, los metadatos alojados en este campo permiten estudiar cómo acontecen las enfermedades autosómicas recesivas en diferentes poblaciones étnicas y las diferencias que existen entre ellos en este aspecto. Además, dichos metadatos permiten comprender porqué existen diferencias en cuanto a la incidencia, la prevalencia y la mortalidad de los diferentes tipos de cáncer para cada uno de los grupos étnicos presentes. De esta forma, se pueden llevar a cabo mejoras en cuanto a la detección y prevención tanto de enfermedades autosómicas recesivas como del cáncer para cada población étnica (43,44).

Por otra parte, existe un número reducido de entradas dentro del caso práctico llevado a cabo que contiene metadatos dentro del campo *lat_lon* (figuras 2 y 5). Este atributo se caracteriza por almacenar las coordenadas geográficas referentes a la ubicación de origen de la secuencia nucleotídica. En este sentido, se observa que los metadatos almacenados en los atributos *country* y *lat_lon* son de tipo geográfico. Sin embargo, la precisión geográfica de las coordenadas geográficas es mucho mayor que la de las localizaciones de texto – aunque estas puedan tener información del país, región y ciudad de forma conjunta – lo cual provoca que la información almacenada en *lat_lon* se considere bien referenciada en un contexto geográfico,

lo cual es de gran interés en estudios de filogeografía y de genética de poblaciones a escala global (5).

Ahora bien, el hecho de que las coordenadas geográficas sean más precisas a la hora de definir el lugar de muestreo que las localizaciones de texto no significa que estas últimas no puedan ser utilizadas en estudios de filogeografía y de genética de poblaciones. No obstante, para que dichas localizaciones puedan ser utilizadas con tal propósito, es necesario que cumplan un mínimo de precisión geográfica, puesto que una mayor especificidad de estas permite que los resultados obtenidos sean más fiables y, por ende, las conclusiones del estudio sean sólidas. Sin embargo, tras llevar a cabo el análisis bioinformático de las entradas de *GenBank* referentes al mitogenoma humano, se observa como un elevado porcentaje de las secuencias analizadas están asociadas a información geográfica imprecisa (figura 4). Esto se debe a que dichas entradas contienen únicamente el nombre del país de origen en el campo *country* o a que, simplemente, no contienen ningún metadato alojado en este atributo.

Uno de los principales motivos que puede dar explicación a este hecho hace referencia a los hábitos del personal investigador a la hora de subir nuevas secuencias a *GenBank*. En este aspecto, muchos de los investigadores que llevan a cabo la incorporación de nuevas entradas a esta base de datos no tienen en cuenta incluir metadatos geográficos en el campo correspondiente ya que, desde su punto de vista o por el tipo de análisis que realizaron con estas secuencias, no consideran necesario asociar la secuencia con su ubicación de origen. Ahora bien, también hay investigadores que llevan a cabo este trabajo incluyendo la ubicación de origen correspondiente. Sin embargo, las localizaciones de texto que asocian a la nueva secuencia únicamente describen el país de origen de esta, provocando así que las entradas resultantes también tengan información geográfica imprecisa (8). Una solución que puede ayudar a resolver este problema consiste en instaurar todo un conjunto de buenos hábitos estandarizados dentro de la comunidad científica internacional relacionados con el proceso de actualización de las bases de datos de *GenBank*. Estos buenos hábitos no solo harían hincapié en la importancia de incluir ubicaciones precisas a la hora de desempeñar este trabajo, también lo harían en cuanto a promover la inclusión de coordenadas geográficas en *lat_lon* para las nuevas entradas presentes, debido a que este tipo de información geográfica se caracteriza por tener una precisión muy elevada.

Tal y como se ha comentado anteriormente, el número de entradas que contienen localizaciones de texto (entradas con metadatos en el campo *country*) es significativamente superior al número de entradas que contienen coordenadas geográficas (entradas con metadatos en el campo *lat_lon*). A pesar de esta situación, existe la posibilidad de aplicar técnicas de programación basadas en *geocoding* para revertir esta situación – proceso por el cual las localizaciones geográficas son convertidas a coordenadas (figura 5). De hecho, existen estudios dentro del ámbito de la filogeografía donde se ha aplicado esta estrategia para poder aumentar la precisión geográfica de las entradas con las cuales se trabaja, por lo que su aplicabilidad es viable si se desean obtener mejores resultados experimentales (referencia). Así pues, otro caso práctico llevado a cabo a partir de las mismas entradas de mitogenoma humano de *GenBank* ha sido el crear y ejecutar un *script* en lenguaje Python con funciones de *geocoding* a partir de la utilización de la librería *GeoPy*. En este sentido, tal y como muestran la tabla y figura 6, el *script* de *geocoding* ha dado un resultado satisfactorio con las localizaciones de texto recopiladas tras el análisis del atributo *country* – ver apartado 3.5 de este trabajo - concretamente, el número de entradas que contiene coordenadas geográficas ha aumentado alrededor de un 74%.

Ahora bien, del análisis de los resultados obtenidos tras el proceso de conversión se pueden extraer varios aspectos de interés. Uno de ellos consiste en que muchas de las coordenadas geográficas resultantes tienen como origen localizaciones de texto imprecisas, ya que estas hacen referencia únicamente al país de origen de la secuencia nucleotídica (por ejemplo, *'Denmark', 'Japan', 'Finland', 'USA', 'Zambia', 'Iran', 'Australia', 'Peru', 'Spain'*). Debido a su baja precisión geográfica, las coordenadas que se obtienen de estas localizaciones de texto no tienen una especificidad elevada en cuanto a la ubicación real, por lo que existe la posibilidad de que la fiabilidad de los resultados obtenidos en estudios de filogeografía y de genética de poblaciones a partir de dichas coordenadas no sea suficientemente sólida. Por tanto, se observa que no es lo mismo obtener las coordenadas para *'Spain'* que para *'Spain, Balearic Islands, Palma'*, puesto que el área demográfica que representa la primera localización es mucho más amplia que la segunda, obteniendo coordenadas mucho más acordes al origen real para el último caso.

Relacionado con el parágrafo anterior, es importante recalcar que se han observado coordenadas geográficas en los resultados experimentales que no concuerdan con la localización de texto de procedencia, a pesar de que dichas localizaciones tengan un nivel de precisión adecuado. Esto es debido a errores o limitaciones de la librería *GeoPy* a la hora de ubicarlas correctamente. Un ejemplo de este hecho se da en la localización *'Solomon Islands, Gela'*, donde las coordenadas resultantes se ubican en Italia en lugar de las Islas Solomon. Otro ejemplo que ilustra esta problemática se da en la ubicación de texto *'Italy, Piedmont'*, donde las coordenadas resultantes se sitúan en la región de Alabama, en Estados Unidos. Es recomendable revisar, una vez ejecutado el *script*, si dichas coordenadas guardan relación con la localización de texto en cuestión, evitando así asociarlas de forma errónea con la entrada de *GenBank* de interés. Aun así, estos casos descritos son muy puntuales, puesto que la gran mayoría de las localizaciones de texto de dicho nivel de especificidad no reportan este problema descrito.

Del análisis mencionado se pueden extraer algunas conclusiones en referencia a los resultados fallidos en el *geocoding*. Una de ellas consiste en la terminología que se encuentra presente en las localizaciones de texto. En este sentido, se aprecia como aquellas que utilizan términos como *'Northern', 'Southern', 'Eastern',* y *'Western'*, seguido del país en cuestión, reportan errores durante la ejecución del *script* (por ejemplo, *'Spain, Southern Spain'*). Un motivo que puede dar explicación a este hecho podría ser la falta de entradas con este tipo de terminología en la base de datos con la cual trabaja *GeoPy* (dicha base de datos corresponde a *GeoNames*). Se confirmó también experimentalmente en este trabajo como la supresión de dicha terminología mediante funciones específicas de listas Python permite recuperar estas localizaciones y obtener, por tanto, sus respectivas coordenadas geográficas. No obstante, esta modificación descrita conlleva connotaciones negativas en las coordenadas resultantes puesto que, finalmente, acaban teniendo como origen localizaciones de texto imprecisas– únicamente contienen información del país de origen, por lo que se vuelve a la misma situación descrita anteriormente. Cabe destacar que la única excepción que se ha encontrado a este hecho durante el caso práctico ha sido con la localización de texto *'France, Southern France'*. Sin embargo, se observa como las coordenadas geográficas derivadas de dicha ubicación no coinciden con el sur de Francia, sino que lo hacen con una isla presente en el Índico (Territorios Australes Franceses).

Similar al caso descrito anteriormente, también se reportan problemas cuando las localizaciones de texto de las entradas analizadas contienen términos que representan las características del territorio en cuestión – por ejemplo, islas, ríos, etc.). Así pues, las palabras clave *'isle'* o *'river'* y sus palabras complementarias como *'upper'* y *'lower'* contenidas en dichas localizaciones – por ejemplo, *'Russia, Upper Anadyr River'*, *'Italy, Isle of Elba'* – no permiten una buena ejecución del *script* de *geocoding*. Mediante la supresión de estos términos con funciones específicas de edición de listas, se ha conseguido revertir el problema y, por tanto, obtener las coordenadas geográficas deseadas. Al contrario de lo que se describe en el parágrafo anterior, se observa como en estos casos las coordenadas obtenidas son de una precisión adecuada, pues su localización de texto de origen se basa también en nombres referentes a áreas demográficas más pequeñas.

Paralelamente, se reportan problemas en la ejecución del *script* de *geocoding* relacionados con el contenido que poseen ciertas localizaciones de texto de *per se*. En este sentido, se aprecia como algunas localizaciones con información geográfica precisa causan anomalías en cuanto a la obtención de sus correspondientes coordenadas. A pesar de ello, no existe un patrón común en el contenido de éstas que se defina como la causante de los fallos reportados, aunque algunos casos se deben a que la presencia conjunta del país y región de origen genera conflicto durante el proceso de ejecución en PyCharm. Esta última problemática puede solventarse eliminando de la localización el nombre del país, mediante la utilización de la función *split* – separa los nombres del texto por las comas (,) o dos puntos (:) en listas de ubicaciones – y la función *remove* – elimina elementos de una lista por su valor. Un ejemplo que permite visualizar lo explicado en este parágrafo hace referencia a la localización de texto *'Iraq, Marsh Arab'*, para la cual el *script* de *geocoding* devuelve las coordenadas geográficas correspondientes a la localización una vez que se le ha eliminado el nombre *Iraq*. Otro ejemplo hace referencia a la localización *'Solomon Islands, Malaita, Kwaio'*, donde la supresión de la ciudad y/o la región y/o de la ciudad de origen mediante la estrategia ya descrita permite obtener las coordenadas geográficas referentes a la ubicación en cuestión con un nivel de especificidad adecuado. Ahora bien, es importante mencionar que existen localizaciones de texto que no se pueden recuperar a pesar de aplicar las funciones Python ya mencionadas, por lo que existe la posibilidad de que dichas localizaciones no se encuentren albergadas en la base de datos *GeoNames*.

Ante los problemas descritos que han aparecido durante la ejecución del *script* de *geocoding*, ¿qué se puede hacer para que no existan coordenadas geográficas con un nivel bajo de precisión? ¿Qué solución se puede aplicar para minimizar el número de errores durante el proceso? ¿Qué solución puede darse para aquellas coordenadas geográficas que no corresponden con su localización de texto de origen? Por una parte, es necesario que las localizaciones de texto que se incluyen dentro del atributo *country* contengan más información además del país de origen de la secuencia nucleotídica – esta solución se relaciona con los buenos hábitos necesarios de instaurar en la comunidad científica descritos anteriormente. Por otra parte, una posibilidad para reducir el número de errores descrito sería el de establecer un formato estandarizado de escritura para completar la información geográfica dentro del campo *country* en cada una de las nuevas entradas de *GenBank* – concretamente, se destacaría el no incluir tanto los términos como las abreviaturas que generan conflicto en el *script;* además de que la sintaxis de las localizaciones esté escrita de tal forma que no originen errores como los descritos previamente como, por ejemplo, escribir las ubicaciones en inglés. Por último, una posibilidad para solucionar la problemática mencionada en cuanto a la no concordancia geográfica entre algunas coordenadas y sus localizaciones de texto de origen puede ser una mejora en las bases de datos que forman parte de *GeoNames* que permita, de una u otra forma, que no se produzcan estos errores – aunque

una sintaxis más clarificadora para la localización de texto podría ser de ayuda para reducir el número de casos que sufren esta situación.

A partir de los resultados experimentales obtenidos tras el proceso experimental de *geocoding*, se ha creado un mapa interactivo mediante la aplicación web *Google Maps Engine* (ilustración 1). En este sentido, cada una de las pestañas azules que se visualizan en el mapa contiene el nombre de la ubicación (localización de texto de origen), sus coordenadas geográficas (coordenadas *geocoding*) y el número de secuencias/entradas que contienen dicha ubicación. Además, todo el conjunto de pestañas se encuentra organizadas en una lista, las cuales se ordenan de mayor a menor número de secuencias. La creación de este mapa tiene como objetivo crear una base para el futuro desarrollo de una herramienta web que ponga a disposición pública cada una de las secuencias nucleotídicas con sus respectivas coordenadas geográficas para todo aquel personal investigador que necesite datos con los que llevar a cabo sus estudios de filogeografía y genética de poblaciones. En este sentido – y como una de las mejorar futuras que se pueden llevar a cabo – sería muy interesante incorporar a cada una de las pestañas opciones y filtros para poder descargarse las secuencias nucleotídicas correspondientes a la ubicación geográfica en cuestión.

Ahora bien, al tener cada pestaña un número de secuencias determinado, ¿cuál es el número mínimo de secuencias para llevar a cabo estudios de filogeografía y de genética de poblaciones? En este aspecto, ¿qué regiones del mapa interactivo podrían ser de utilidad para llevar a cabo este propósito? En el caso de la filogeografía, por normal general, se recomienda utilizar cuantas más secuencias mejor, ya que se facilita que los resultados obtenidos sean adecuados para generar conclusiones sólidas. Sin embargo, el número mínimo a utilizar depende del tipo de estudio filogeográfico que se lleve a cabo y de las incógnitas que el investigador quera resolver, provocando así que no exista un número mínimo establecido para todos los casos, generándose así controversia. Ahora bien, existen variables que obligan a aumentar el número de secuencias por cada ubicación geográfica. Por ejemplo, si la variedad del conjunto de ubicaciones es elevada, se recomienda aumentar el número mínimo en el estudio. Por su parte, si el estudio implica a diferentes especies, el número mínimo se reduce – al contrario que si se trabaja con secuencias pertenecientes a una sola especie, donde dicho valor se ha de aumentar considerablemente (45,46).

En el caso de la genética de poblaciones, el número mínimo de secuencias a utilizar para cada loci durante el estudio depende del tipo de marcadores que se utilicen. Así pues, como norma general, aquellos estudios que recurran al uso de microsatélites requieren de un número mínimo igual a 30, mientras que aquellos que usen SNPs necesitan únicamente de 10 como mínimo. La reducción de dicho valor para este último marcador se debe a que cada SNP tiene su propio historial evolutivo. A pesar de ello, existe controversia en este aspecto, puesto que hay investigadores que defienden que no hay una barrera o límite establecido para todos los casos, argumentando que el número mínimo depende del tipo de población y de la finalidad del estudio poblacional. Aun así, existen programas bioinformáticos en la actualidad – tanto web como *software* bioestadístico – que tienen capacidad para calcular el número mínimo de secuencias adecuado a partir de las características del estudio, basándose en el uso de algoritmos matemáticos (47,48).

Anteriormente se ha mencionado la incorporación de secuencias nucleotídicas para la mejora del mapa interactivo creado previamente. Sin embargo, existen más estrategias futuras que pueden ser de utilidad para mejorar esta herramienta web. Una de ellas sería asociar cada una de las secuencias albergadas en el mapa a su correspondiente *pmid* de NCBI, de tal forma que permitan acceder a la publicación donde aparece la secuencia de interés. Para llevarlo a cabo, es necesario crear nuevos *scripts* que permitan tanto la extracción como la organización de

estos datos. Otra de las mejoras que se puede llevar a cabo consiste en incluir secuencias localizadas en el mapa a partir de la etnia o raza que tengan asociada, de tal forma que se enriquezca la información proporcionada públicamente a los investigadores para llevar a cabo sus estudios de genética de poblaciones. Por último, para cada una de las entradas analizadas de metadatos que no contienen ningún tipo de información geográfica, se puede obtener el nombre del responsable de la publicación y su email para enviarle una petición formal, solicitando que actualicen sus entradas de *GenBank* mediante la incorporación de información geográfica relacionada con las secuencias utilizadas en su estudio. De esta forma, si han accedido a esta petición, la siguiente *reléase* de *GenBank* tendrá más entradas con información geográfica por lo que, repitiendo el proceso experimental llevado a cabo en este trabajo, se podrían obtener más datos para incorporar al mapa interactivo, facilitando así que más investigadores especializados en filogeografía y en genética de poblaciones puedan encontrar de utilidad la herramienta creada.

No obstante, hay que tener en cuenta la perspectiva actual en cuanto al grado de reutilización de secuencias alojadas en *GenBank* para analizar el impacto que puede tener la herramienta web descrita. Para ello, se ha llevado a cabo un estudio bibliográfico sobre el estado del arte en esta cuestión. A partir de los resultados descritos en la figura 1, la reutilización de secuencias de *GenBank* como única metodología para llevar a cabo estudios de filogeografía no se lleva a cabo. En su lugar, existen publicaciones donde se recurre a esta metodología como forma de complementar los resultados filogeográficos obtenidos a partir de secuencias generadas por secuenciación *de novo*. Incluso un porcentaje mayor de publicaciones muestran como únicamente recurren a la secuenciación *de novo* como metodología para realizar dichos estudios.

Una posible explicación ante esta tendencia a secuenciar las muestras en lugar de reutilizar las ya existentes en *GenBank* se podría relacionar con la problemática descrita a lo largo del trabajo en cuanto al elevado porcentaje de entradas de esta base de datos con un nivel de precisión geográfica bajo, además del tiempo y esfuerzo que hay que dedicar tras su descarga para normalizar los datos. Ante esta situación, es posible que los grupos de investigación en filogeografía se decanten por secuenciar las muestras y georreferenciarlas adecuadamente para ganar tanto en optimización del tiempo de trabajo como en calidad de los resultados experimentales. En cualquier caso, la muestra tomada para llevar a cabo el estudio es pequeña y, aunque es de utilidad para hacerse una idea preliminar de la situación actual, lo ideal sería realizar un estudio más amplio en este aspecto.

En relación con las técnicas de secuenciación *de novo* descritas, existen líneas de trabajo por desarrollar dentro de la filogeografía y la genética de poblaciones. Gran parte de los análisis filogeográficos se han llevado a cabo hasta la fecha mediante secuenciación de *Sanger*, ya que permite generar múltiples fragmentos genómicos superpuestos de forma muy específica. Sin embargo, la irrupción de los métodos de secuenciación NGS ha provocado que los investigadores involucrados en estas áreas se planteen su utilización para realizar dichos estudios. Sin embargo, los métodos de NGS tienen como contrapartida que no permiten llevar a cabo un control adecuado sobre qué regiones del genoma se han secuenciado, hecho que dificulta la reducción del genoma a fragmentos ortólogos. Actualmente, se busca mejorar la eficacia y los costes de NGS, de tal forma que sea rentable la secuenciación completa de genomas, situación en la cual podría ser viable la aplicación de estos métodos para llevar a cabo estudios filogeográficos. Ahora bien, estas mejoras han de ir acompañadas no solamente de mejoras en cuanto al análisis de datos, también han de ir acompañadas de mejoras en cuanto a los sistemas de almacenamiento informáticos – mientras que el conjunto de datos de secuenciación *Sanger* puede tener 500 secuencias, en NGS este valor puede ser cuatro millones superior (13).

Conclusiones

A continuación, se enumeran las conclusiones a las cuales se llega tras haber llevado a cabo el análisis del estado del arte y el trabajo experimental bioinformático:

1. Gran parte de las entradas alojadas en *GenBank* tienen un nivel de precisión geográfica bajo, ya sea por contener únicamente el país de origen o por no incluir ningún tipo de información geográfica.
2. A pesar de que las coordenadas geográficas son el formato ideal para georreferenciar las secuencias nucleotídicas en estudios de filogeografía y de genética de poblaciones, la mayoría de las entradas de *GenBank* con información geográfica lo tienen en forma de localización de texto.
3. Las técnicas de *geocoding* son de gran utilidad de cara a obtener coordenadas geográficas a partir de localizaciones de texto, aunque es necesario mejorar la eficacia de esta metodología para minimizar los errores durante el proceso – concretamente, la no obtención de coordenadas o la no concordancia entre las coordenadas resultantes y la localización de texto de origen.
4. El campo *isolation_source* representa una fuente de metadatos de gran interés para estudios de genética de poblaciones en humanos, puesto que la asociación de la etnia/raza de origen a cada secuencia nucleotídica no solamente permite llevar a cabo el estudio de la distribución de alelos entre poblaciones, también permiten llevar a cabo avances en biomedicina relacionados con la comprensión sobre cómo se dan las enfermedades autosómicas recesivas y los diferentes tipos de cáncer entre etnias.
5. Los problemas asociados con la falta de precisión geográfica de muchas de las entradas de *GenBank*, además de los inconvenientes asociados con la normalización de datos geográficos para todo el conjunto de secuencias a utilizar, han podido jugar un papel importante para que la secuenciación *de novo* se haya establecido como la principal metodología para llevar a cabo estudios sobre filogeografía en lugar de la reutilización de secuencias de dicha base de datos.
6. El desarrollo de una herramienta web de acceso libre que albergue las secuencias nucleotídicas por su ubicación geográfica puede ayudar a que los investigadores pertenecientes a las áreas de filogeografía y de genética de poblaciones recurran a la reutilización de secuencias como metodología a la hora de llevar a cabo sus respectivos estudios.
7. A pesar de que la secuenciación de *Sanger* es un estándar en estudios de filogeografía, los avances en NGS pueden ayudar a que esta nueva metodología de secuenciación se implante como opción real para llevar a cabo estudios de filogeografía y de genética de poblaciones.

Bibliografía

1. GenBank and WGS Statistics [Internet]. [citado 29 de mayo de 2018]. Disponible en: https://www.ncbi.nlm.nih.gov/genbank/statistics/
2. International Nucleotide Sequence Database Collaboration | INSDC [Internet]. [citado 29 de mayo de 2018]. Disponible en: http://www.insdc.org/
3. Tahsin T, Weissenbacher D, Jones-shargani D, Magee D, Vaiente M, Gonzalez G, et al. Original article Named entity linking of geospatial and host metadata in GenBank for advancing biomedical research. 2017;(May 2018):1-16.
4. Tahsin T, Weissenbacher D, O'Connor K, Magge A, Scotch M, Gonzalez-Hernandez G. GeoBoost: accelerating research involving the geospatial metadata of virus GenBank records. Bioinformatics. 2017;34(December 2017):1606-8.

5. Gratton P, Marta S, Bocksberger G, Winter M, Trucchi E, Kühl H. A world of sequences: can we use georeferenced nucleotide databases for a robust automated phylogeography? J Biogeogr. 2017;44(2):475-86.
6. geopy · PyPI [Internet]. [citado 29 de mayo de 2018]. Disponible en: https://pypi.org/project/geopy/
7. GeoNames [Internet]. [citado 29 de mayo de 2018]. Disponible en: http://www.geonames.org/
8. Scotch M, Sarkar IN, Mei C, Leaman R, Cheung K-H, Ortiz P, et al. Enhancing phylogeography by improving geographical information from GenBank. J Biomed Inform. diciembre de 2011;44:S44-7.
9. Weissenbacher D, Sarker A, Tahsin T, Scotch M, Gonzalez G. Extracting geographic locations from the literature for virus phylogeography using supervised and distant supervision methods. AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci. 2017;2017:114-22.
10. Tahsin T, Beard R, Rivera R, Lauder R, Wallstrom G, Scotch M, et al. Natural language processing methods for enhancing geographic metadata for phylogeography of zoonotic viruses. AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci. 2014;2014:102-11.
11. Population Genetics [Internet]. Wikipedia, the free encyclopedia. 2018. Disponible en: https://en.wikipedia.org/wiki/Population_genetics
12. Tahsin T, Weissenbacher D, Rivera R, Beard R, Firago M, Wallstrom G, et al. A high-precision rule-based extraction system for expanding geospatial metadata in GenBank records. J Am Med Informatics Assoc. 2016;23(5):934-41.
13. McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol. 2013;66(2):526-38.
14. Carstens B, Lemmon AR, Lemmon EM. The Promises and Pitfalls of Next-Generation Sequencing Data in Phylogeography. Syst Biol. 2012;61(5):713-5.
15. Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, Tishkoff SA, et al. Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics 2005 22. 2005;2(2):81.
16. Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, et al. Application of SNPs for population genetics of nonmodel organisms: New opportunities and challenges. Mol Ecol Resour. 2011;11(SUPPL. 1):123-36.
17. Bassi S. A Primer on Python for Life Science Researchers. PLoS Comput Biol. 2007;3(11):e199.
18. Maria Nattestad. For bioinformatics, which language should I learn first? – OMGenomics [Internet]. 2017 [citado 3 de junio de 2018]. Disponible en: http://omgenomics.com/programming-languages/
19. Matplotlib: Python plotting — Matplotlib 2.2.2 documentation [Internet]. [citado 3 de junio de 2018]. Disponible en: https://matplotlib.org/
20. NumPy — NumPy [Internet]. [citado 3 de junio de 2018]. Disponible en: http://www.numpy.org/
21. SciPy - Wikipedia, la enciclopedia libre [Internet]. [citado 3 de junio de 2018]. Disponible en: https://es.wikipedia.org/wiki/SciPy
22. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 1 de junio de 2009;25(11):1422-3.
23. Scientific & Data Science Tools - Features | PyCharm [Internet]. [citado 1 de junio de 2018]. Disponible en: https://www.jetbrains.com/pycharm/features/scientific_tools.html
24. Hernández CL, Reales G, Dugoujon JM, Novelletto A, Rodríguez JN, Cuesta P, et al. Human maternal heritage in Andalusia (Spain): Its composition reveals high internal complexity and distinctive influences of mtDNA haplogroups U6 and L in the western and eastern side of region. BMC Genet. 2014;15:1-16.
25. Batini C, Hallast P, Zadik D, Delser PM, Benazzo A, Ghirotto S, et al. Large-scale recent expansion of European patrilineages shown by population resequencing. Nat Commun. 2015;6(May).
26. Maca-Meyer N, González AM, Pestano J, Flores C, Larruga JM, Cabrera VM. Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. BMC Genet. 2003;4:1-11.
27. Der Sarkissian C, Brotherton P, Balanovsky O, Templeton JEL, Llamas B, Soubrier J, et al. Mitochondrial genome sequencing in mesolithic North East Europe unearths a new sub-clade within the broadly distributed human haplogroup C1. PLoS One. 2014;9(2).
28. González AM, Larruga JM, Abu-Amero KK, Shi Y, Pestano J, Cabrera VM. Mitochondrial lineage M1 traces an early human backflow to Africa. BMC Genomics. 2007;8:1-12.
29. Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Rogalla U, Perkova M, et al. Origin and post-glacial dispersal of mitochondrial DNA haplogroups C and D in Northern Asia. PLoS One. 2010;5(12):1-9.
30. Rajkumar R, Banerjee J, Gunturi HB, Trivedi R, Kashyap VK. Phylogeny and antiquity of M macrohaplogroup inferred from complete mt DNA sequence of Indian specific lineages. BMC Evol Biol. 2005;5:1-8.
31. Brisighelli F, Álvarez-Iglesias V, Fondevila M, Blanco-Verea A, Carracedo Á, Pascali VL, et al.

Uniparental Markers of Contemporary Italian Population Reveals Details on Its Pre-Roman Heritage. PLoS One. 2012;7(12).
32. Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, Scozzari R, et al. Where West Meets East: The Complex mtDNA Landscape of the Southwest and Central Asian Corridor. Am J Hum Genet. 2004;74(5):827-45.
33. Shi H, Zhong H, Peng Y, Dong YL, Qi X Bin, Zhang F, et al. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol. 2008;6(Figure 1):1-10.
34. Rai N, Chaubey G, Tamang R, Pathak AK, Singh VK, Karmin M, et al. The Phylogeography of Y-Chromosome Haplogroup H1a1a-M82 Reveals the Likely Indian Origin of the European Romani Populations. PLoS One. 2012;7(11):1-7.
35. Marrero P, Abu-Amero KK, Larruga JM, Cabrera VM. Carriers of human mitochondrial DNA macrohaplogroup M colonized India from southeastern Asia. BMC Evol Biol. 2016;16(1):1-13.
36. Larruga JM, Marrero P, Abu-Amero KK, Golubenko M V., Cabrera VM. Carriers of mitochondrial DNA macrohaplogroup R colonized Eurasia and Australasia from a southeast Asia core area. BMC Evol Biol. 2017;17(1):1-15.
37. Pennarun E, Kivisild T, Metspalu E, Metspalu M, Reisberg T, Moisan JP, et al. Divorcing the Late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa. BMC Evol Biol. 2012;12(1):1.
38. Abu-Amero KK, Gonźlez AM, Larruga JM, Bosley TM, Cabrera VM. Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. BMC Evol Biol. 2007;7:1-15.
39. Chaubey G, Karmin M, Metspalu E, Metspalu M, Selvi-Rani D, Singh VK, et al. Phylogeography of mtDNA haplogroup R7 in the Indian peninsula. BMC Evol Biol. 2008;8(1):1-12.
40. Pereira L, Silva NM, Franco-Duarte R, Fernandes V, Pereira JB, Costa MD, et al. Population expansion in the North African Late Pleistocene signalled by mitochondrial DNA haplogroup U6. BMC Evol Biol. 2010;10(1):390.
41. Secher B, Fregel R, Larruga JM, Cabrera VM, Endicott P, Pestano JJ, et al. The history of the North African mitochondrial DNA haplogroup U6 gene flow into the African, Eurasian and American continents. BMC Evol Biol. 2014;14(1):1-17.
42. García O, Fregel R, Larruga JM, Álvarez V, Yurrebaso I, Cabrera VM, et al. Using mitochondrial DNA to test the hypothesis of a European post-glacial human recolonization from the Franco-Cantabrian refuge. Heredity (Edinb). 2011;106(1):37-45.
43. Genetic Risk, Race and Ethnicity - Cancer Fighters Thrive [Internet]. 2014 [citado 7 de junio de 2018]. Disponible en: http://www.cancerfightersthrive.com/genetic-risk-race-ethnicity/
44. Octavio-Aguilar P, Ramos-Frías J. Aplicación de la genética de poblaciones en el ámbito de la medicina. Biomedica. 2014;34:171-9.
45. What is an acceptable sample size for a phylogeographic study? [Internet]. 2013 [citado 7 de junio de 2018]. Disponible en: https://www.researchgate.net/post/What_is_an_acceptable_sample_size_for_a_phylogeographic_study
46. Can someone advise in the field of phylogeography and sample... [Internet]. 2014 [citado 7 de junio de 2018]. Disponible en: https://www.researchgate.net/post/Can_someone_advise_in_the_field_of_phylogeography_and_sample_size
47. What is the minimum sample size for population genetic study using SNPS? [Internet]. Research Gate. 2014. Disponible en: https://www.researchgate.net/post/What_is_the_minimum_sample_size_for_population_genetic_study_using_SNPS
48. What is an ideal sample size for studies of phylogeography... [Internet]. 2017 [citado 7 de junio de 2018]. Disponible en: https://www.researchgate.net/post/What_is_an_ideal_sample_size_for_studies_of_phylogeography_and_population_structure_using_2b-RAD-Seq_data

Anexos

Glosario

Vocabulario	Definición
Atributo	Variable que contiene una propiedad de interés (país, región, etnia, coordenadas geográficas, etc.).
Coordenadas geográficas	Conjunto de números que representan ubicaciones precisas de todo tipo. Dichos números pueden ir acompañados de letras y símbolos.
GenBank	Base de datos perteneciente a NCBI que contiene numerosas entradas con secuencias nucleotídicas de diverso tipo.
Geocoding	Proceso por el cual se convierte una localización de texto a coordenadas geográficas (latitud, longitud).
Librería	Conjunto de implementaciones funcionales organizadas en módulos.
Localización de texto	Ubicación geográfica en formato de texto que contiene el nombre del país, la región, la ciudad, el municipio, etc.
Microsatélites	Secuencias de DNA que contienen repeticiones consecutivas de un determinado fragmento nucleotídico. Son de gran utilidad para llevar a cabo, por ejemplo, estudios de genética de poblaciones.
Módulo	Archivo que contiene todo un conjunto de variables y funciones
mtDNA	DNA mitocondrial, ya sea de forma completa o parcial, el cual puede ser de gran utilidad como marcador para estudios de filogeografía y de genética de poblaciones.
Script	Programa formado por líneas de comando basados en un determinado lenguaje de programación, el cual tiene como finalidad llevar a cabo una operación.
SNPs	Polimorfismos de un solo nucleótido, los cuales son secuencias de DNA que varían su pauta de lectura en un único nucleótido. Son de gran utilidad para llevar a cabo, por ejemplo, estudios de genética de poblaciones.

Script extracción de metadatos

El siguiente enlace *cloud* permite acceder al *script* de extracción de metadatos creado por Óscar Moya Mesa (https://www.dropbox.com/sh/nnskb0te5wmwffq/AACAmjLAy-A9wWLbMNwUwt_Pa?dl=0).

Script *geocoding*

El siguiente enlace *cloud* permite acceder al *script* de *geocoding* creado por Guillermo Andrés Fernández Olivares (https://www.dropbox.com/sh/nnskb0te5wmwffq/AACAmjLAy-A9wWLbMNwUwt_Pa?dl=0).

Tabla resultados *geocoding* y mapa interactivo

El siguiente enlace *cloud* permite acceder a los resultados obtenidos tras ejecutar el *script* de *geocoding*, los cuales han sido utilizados para crear el mapa interactivo descrito en este trabajo de final de grado (https://www.dropbox.com/sh/nnskb0te5wmwffq/AACAmjLAy-A9wWLbMNwUwt_Pa?dl=0).