FACULTAD DE CIENCIAS AGRARIAS

UNIVERSIDAD NACIONAL DE ROSARIO

VARIANTES MICROSATÉLITES HUMANAS: CREACIÓN DE UNA BASE DE DATOS (LOCAL Y REMOTA) CON ACCESO WEB, Y EL APORTE DE DATOS DE SECUENCIACIÓN DE SEGUNDA GENERACIÓN

PABLO SEBASTIÁN VÉLEZ

TRABAJO FINAL PARA OPTAR AL TÍTULO DE ESPECIALISTA EN BIOINFORMÁTICA

TUTORA: DÉBORA ARCE

AÑO 2021

TÍTULO DEL TRABAJO FINAL

Pablo Sebastián Vélez

Licenciatura en Bioquímica Clínica – Universidad Nacional de Córdoba

Este Trabajo Final es presentado como parte de los requisitos para optar al grado académico de Especialista en Bioinformática, de la Universidad Nacional de Rosario y no ha sido previamente presentada para la obtención de otro título en ésta u otra Universidad. El mismo contiene los resultados obtenidos en investigaciones llevadas a cabo en CEPROCOR, durante el período comprendido entre Febrero de 2018 y Marzo de 2019, bajo la dirección de Farm. (Dra.) Débora ARCE.

Lic. Pablo Sebastián Vélez

Farm. (Dra.) Débora ARCE.

Defendida: 02 de Noviembre de 2021.

Agradecimientos.

A mis hermanos Silvana y Néstor, sin ellos no hubiera podido emprender este desafío.

A la Dra. Débora Arce quien desde un principio me apoyó con esta idea, cuyo origen también vino en parte desde ella.

Al Dr. Raúl Badini que aprobó y favoreció todo lo relacionado con esta formación de posgrado, en una clara apuesta para mejorar recursos humanos desde lo institucional.

A todo el grupo de Biología Molecular del Ceprocor: la Dra. Andrea Belaus, el Dr. Juan Rondan Dueñas y el Dr. Guillermo Giaj-Merlera, que me animaron a realizar esta especialización para sumar principalmente otro enfoque o manejo de datos a los temas del grupo, además de la mejora en el ámbito profesional en lo personal.

A todo el personal Administrativo, de Recursos Humanos y autoridades del Ceprocor y demás integrantes de comisiones de evaluación, que me facilitaron no solo los recursos para viajar, establecerme y demás gastos, sino que también tuvieron muy buenas consideraciones con este posgrado.

A todos los compañeros que pasaron por el grupo de Biología Molecular: Ing. Agr. Fernando Del Inocenti y Lic. Aylen Makhoul que me ayudaron con algunas tareas importantes.

A la Ing. Laura Morcillo que me facilitó los recursos para alojar el sitio web de STRs humanos.

A todos los docentes y personal administrativo de la Facultad de Ciencias Agrarias y CIFASIS, autoridades e integrantes de la Comisión de la Especialización en Bioinformática quienes promovieron en todo momento un excelente ámbito académico, atendiendo todas las demandas que se presentaron.

A todos mis compañeros de especialización con quienes he compartido muy buenos momentos, de estudio y de muchos otros.

A las evaluadoras de este trabajo final: Lic. Dra. Estefanía Mancini, Dra. Romina Martinelli y Bioq. Dra. Angela Rosaria Solano, que me aportaron lo necesario para que este trabajo sea mucho mejor.

A Celeste Correa, quien me ayudo en muchos análisis.

Dedicado a mi hijo Gonzalo, los días todos.

Abreviaturas, acrónimos y símbolos.

En el presente trabajo final se utilizan las siguientes abreviaturas, acrónimos y símbolos.

ADN: Ácido Desoxirribo Nucleico.

API: (del inglés) Application Programming Interface.

BAC: (del inglés) Bacterial artificial chromosome

BD: Base de datos.

BED: (del inglés) Browser Extensible Data.

CODIS: (del inglés) Combined DNA Index System.

FBI: (del inglés) Federal Bureau of Investigation.

HGNC: (del inglés) HUGO Gene Nomenclature Committee.

HUGO: (del inglés) Human Genome Organization.

ID: Identificador.

ISFG: (del inglés) International Society for Forensic Genetics.

LINEs: (del inglés) Long Interspersed Elements

LTR: (del inglés) Long Terminal Repeats

MAF: (del inglés) Minor Allele Frequency.

NCBI: (del inglés) National Center for Biotechnology Information.

NGS: (del inglés) Next Generation Sequencing.

NT: Nucleótido.

PB: Par de bases.

PCR: (del inglés) Polymerase Chain Reaction.

RFLP: (del inglés) Restriction Fragment Length Polymorphism.

SINEs: (del inglés) Short Interspersed Elements

SQL: (del inglés) Structured Query Language.

SSR: (del inglés) Single Sequence Repeat.

STR: (del inglés) Short Tandem Repeat.

STS: (del inglés) Sequence-Tagged Sites.

SNP: (del inglés) Single Nucleotide Polymorphism.

SNV: (del inglés) Single Nucleotide Variant.

SWGDAM: (del inglés) Scientific Working Group on DNA Analysis Methods.

UCSC: (del inglés) University of California Santa Cruz

VCF: (del inglés) Variant Call Format.

VNTR: (del inglés) Variable Number Tandem Repeat.

Glosario

BED (del inglés *Browser Extensible Data*): El formato BED proporciona una forma flexible de definir las líneas de datos que se muestran en una pista de anotación. Las líneas BED tienen tres campos obligatorios y nueve campos opcionales adicionales. El número de campos por línea debe ser coherente en cualquier conjunto de datos en una pista de anotación. El orden de los campos opcionales es vinculante: los campos con números inferiores siempre deben rellenarse si se utilizan campos con números superiores.

E-VALUE: El valor esperado (E) es un parámetro que describe el número de *hits* que uno puede "esperar" al azar cuando busca en una base de datos de un tamaño particular.

FASTA: Es un formato basado en texto para representar secuencias de nucleótidos o secuencias de aminoácidos (proteínas), en las que los nucleótidos o aminoácidos se representan usando códigos de una letra.

FASTQ: es un formato basado en texto para almacenar tanto una secuencia biológica (de nucleótidos por lo general) y sus puntuaciones de calidad correspondientes en formato PHRED.

GAP: En un alineamiento bioinformático se permite “abrir” brechas en las secuencias para mejorar el emparejamiento.

HIT: Hallazgo exitoso.

MATCH: En un alineamiento bioinformático cuando dos ó más secuencias tienen máximo emparejamiento, caso contrario se considera MISMATCH.

MISMATCH: En un alineamiento bioinformático el emparejamiento entre dos ó más secuencias no es máximo.

INDEL: Contracción de los términos INserción y DELeción y se refiere a pérdida o ganancia de nucleótidos en una secuencia.

MAF: frecuencia del alelo minoritario, si este es menor al 1 % ese sistema/marcador se considera monomórfico.

MICROVARIANTE: Es un término usado en ciencias forenses para el uso de STRs y se refiere a la aparición de repeticiones incompletas dentro del núcleo.

NGS: Secuenciación de "nueva generación" o "segunda generación", para distinguirlos de los métodos anteriores como la secuenciación de Sanger. Esta tecnología generalmente se caracteriza por ser altamente escalable, lo que permite secuenciar todo el genoma a la vez. Por lo general, esto se logra fragmentando el genoma en pedazos pequeños, muestreando aleatoriamente un fragmento y secuenciando usando una variedad de tecnologías.

PHRED: Es una medida de la calidad de la identificación de las nucleobases generadas por secuenciación automática de ADN.

SCORE: Puntaje o magnitud numérica que informa el grado de éxito de una tarea.

SQL: Lenguaje de consulta estructurado, es un lenguaje específico de dominio utilizado en programación y diseñado para administrar datos almacenados en un sistema de administración de bases de datos relacionales, o para el procesamiento de flujo en un sistema de administración de flujo de datos relacionales. Es particularmente útil en el manejo de datos estructurados, es decir, datos que incorporan relaciones entre entidades y variables.

SCRIPT: Archivo de texto plano ejecutable mediante un programa que lo interpreta.

STR: Secuencias que poseen repeticiones de un motivo o unidad y pueden alcanzar de 7 nt hasta 2.500 nt.

STS: son secuencias de ADN cortas (40 a 5.000 pares de bases) que tienen una sola aparición en el genoma y cuya ubicación y secuencia de bases son conocidas.

VCF: Es un formato de archivo de texto (muy probablemente almacenado de forma comprimida). Contiene líneas de metainformación (con el prefijo "##"), una línea de encabezado (con el prefijo "#") y líneas de datos que contienen información sobre una posición en el genoma y la información del genotipo en muestras para cada posición (campos de texto separados por pestañas) Los campos de longitud cero no están permitidos, se debe usar un punto (".") en su lugar. Para garantizar la interoperabilidad entre plataformas, las implementaciones compatibles con VCF deben ser compatibles con las convenciones de nueva línea LF (\n) y CR + LF (\r\n).

FORMATOS

Títulos primarios en negrita y mayúscula: **3. MATERIALES Y MÉTODOS, 3.1 DATOS**

Títulos secundarios en negrita: **3.1.2 Datos de marcadores STSs**

Anglicismos y otros extranjerismos, en cursiva: *software, core, polymerase, in silico.*

Archivos en general, entre comillas: “human-UniSts-primers-2p-replaced.txt”

Archivos ejecutables *scripts* en cursiva y entre comillas:

“fase1_vs_(exon ó intrón)_v2_all.pl”

Resumen.

Los microsatélites son un tipo de variante ampliamente distribuidos en el genoma humano y otros organismos procariotas y eucariotas. Su elevado polimorfismo y tasa de mutación los ha convertido en candidatos predilectos como marcadores moleculares de uso en disciplinas como genética de poblaciones, en ciencias forenses y de la salud.

Muchos microsatélites contenidos en STSs usados para crear los primeros mapas físicos del genoma humano quedaron registrados en las primeras bases de datos públicas, y que, al provenir de distintos grupos de trabajo, se generaron distintas nomenclaturas que apuntan a las mismas regiones, estableciéndose un elevado solapamiento.

Además, la información para microsatélites de estas bases de datos originales solo contempló una pequeña proporción del total de variantes que asciende a casi 700.000 (320.000 MAF > 1%) según un amplio catálogo de microsatélites generado en el año 2014 y actualizado en el año 2020.

A partir de estos datos y de los disponibles más antiguos en los sitios de UCSC y de Ensembl se construyó una base de datos con funcionamiento local y remoto, cuya finalidad es reunir datos de secuenciación con tecnologías de segunda generación junto con la información antigua y enriquecer todo lo que está disponible para microsatélites, y que se encuentra fragmentado en diferentes sitios.

También se generó nueva información usando dos herramientas: StraitRazor y lobSTR que detectan microsatélites en archivos de secuenciación de segunda generación de 27 genomas disponibles del Proyecto 1000 Genomas, con un total de 323 marcadores más usados en ciencias forenses y de la salud, entre otros elegidos según posean elevada heterocigosidad o pertenezcan a cromosomas involucrados en aneuploidías.

A lo sumo un 12,8 % de la información de STSs contienen STRs, lo que implica que aquellos datos de STSs en sitios como Ensembl y UCSC, que provienen originalmente de UniSTS, son insuficientes aunque complementarios para describir STRs. Además, se evidencia que hay poco más de 5% de las secuencias de cebadores que tienen errores respecto del fragmento blanco. Esto significa que es necesario hacer correcciones para muchos STRs incluidos en STSs, y generar nueva información de casi la totalidad de los STRs humanos. Hasta ahora se ha conseguido un catálogo de variantes ambicioso de casi 700.000, pero se necesita aun un mejor detalle de cada variante: nomenclatura y descripción general de la estructura del STR y de cada variante alélica.

Los STRs se encuentran homogéneamente distribuidos, por esa razón encontramos un 56,12% del total en regiones génicas (99,4 % en intrones), lo que fue útil para nombrar muchos marcadores y no usar la notación de SNPs. Sin embargo, los SNPs modifican las regiones STRs y deben ser tenidas en cuenta en la nomenclatura, configuración de la estructura del nucleo STR y parámetros de herramientas bioinformáticas.

Otro evento genómico incluido en este estudio e incorporado a la base de datos son los segmentos múltiples. Hemos visto más del 40 % STRs del cromosoma Y que son múltiples intra y inter-cromosómicamente. Evento que explica porque el cromosoma Y posee tantos STRs con más de una alelo por individuo. También es útil para explicar inespecificidad del uso de herramientas bioinformáticas. El diseño para detectar una determinada región y usarlo como marcador también requiere el conocimiento y buen uso de estos segmentos. La especificidad depende de ese diseño.

En el sitio http://arrobasistemas.com/humstrs2/index.html se muestran las características de los 319 STRs usados en este trabajo, y se contemplan aquellos usados en Ciencias Forenses y de la Salud, entre otros. En el detalle de cada marcador se reúne información proveniente de los sitios UCSC y Ensembl, además de datos generados con las herramientas lobSTR y StraitRazor sobre 27 genomas.

El detalle de cada STR está comprendido entre los siguientes grupos: Nombres, Ubicación Genómica, Eventos Genómicos Asociados, Salida de Tándem Repeat Finder, Parámetros Poblacionales, Asignaciones Alélicas, Alineamientos, SNPs (+/- 100 pb alrededor del núcleo) y por último la Estructura del núcleo (190 estructuras fueron definidas). Se menciona el recurso usado para cada grupo.

Se estableció el uso de dos herramientas para la detección de STRs debido principalmente al bajo número de lecturas (en promedio de 11 a 15) hallados en los genomas secuenciados e implicó establecer una doble asignación alélica. Las coincidencias de las herramientas dependieron del grado de complejidad de la estructura del STR, del número de secuencias detectadas por cada herramienta, del ajuste de parámetros teniendo en cuenta SNPs y segmentos múltiples, estas mismas variaciones también produjeron diferentes asignaciones alélicas en muchos casos, debido a situaciones individuales, otros cambios no tenidos en cuenta en los parámetros e incluso errores debido a una anotación incorrecta de algún SNP.

Considerando el tartamudeo y habiendo establecido el grado de coincidencia de las herramientas se consiguió que un 76,9 % de asignaciones alélicas totales (8.710) sean idénticas entre ambas herramientas, o un 65,4 % en condiciones un poco más estrictas. El restante aún puede ser corregidas con la intervención del analista. Aquellas secuencias obtenidas con StraitRazor (cuyas asignaciones alélicas coincidieron con lobSTR) fueron usadas para establecer las estructuras.

Aún faltan modificaciones o nuevas herramientas que detecten aquellos STRs de configuración del núcleo complejas, o mayor información de esas estructuras para el universo completo de STRs humanos. También se requiere que las herramientas (solas o en conjunto con otras) puedan calificar las detecciones y asignaciones alélicas que consiguen. La calidad de esas asignaciones dependerá de todo lo expuesto anteriormente, además de las condiciones de calidad y tecnologías usadas en la secuenciación.

Se pudo chequear en un simple análisis (y algunos hallazgos con SNPs) que el ensamblaje de trabajo: HG19 y el vigente HG38 poseen errores de anotación respecto del alelo de referencia (el anotado no es el más frecuente) tanto para SNPs y como para STRs de estructuras simples. Esta situación no es fácilmente demostrable con STRs complejos.

Superado esto se podrá tener la mejor descripción de la totalidad de los STRs humanos. Este planteo también aplica al resto de STRs de los seres vivos.

Abstract.

Microsatellites are a type of variant widely distributed in the human genome and other prokaryotic and eukaryotic organisms. Their high polymorphism and mutation rate have made them favored candidates as molecular markers for use in disciplines such as population genetics, forensic and health sciences.

Many microsatellites contained in STSs used to create the first physical maps of the human genome were registered in the first public databases, and that, coming from different work groups, different nomenclatures were generated that point to the same regions, establishing a high overlap.

In addition, the information for microsatellites in these original databases only included a small proportion of the total variants, amounting to almost 700,000 (320,000 MAF > 1%) according to an extensive catalog of microsatellites generated in 2014 and updated in 2020. .

From these data and from the oldest data available at the UCSC and Ensembl sites, a database with local and remote operation was built, whose purpose is to gather sequencing data with second generation technologies together with the old information and enrich everything that is available to microsatellites, and that is fragmented in different places.

New information was also generated using two tools: StraitRazor and lobSTR that detect microsatellites in second-generation sequencing files of 27 genomes available from the 1000 Genomes Project, with a total of 323 markers most used in forensic and health sciences, among others chosen. depending on whether they have high heterozygosity or belong to chromosomes involved in aneuploidy.

At most 12.8% of the information on STSs contain STRs, which implies that those data on STSs in sites such as Ensembl and UCSC, which originally come from UniSTS, are insufficient but complementary to describe STRs. In addition, it is evident that there is little more than 5% of the primer sequences that have errors with respect to the target fragment. This means that it is necessary to make corrections for many STRs included in STSs, and to generate new information for almost all human STRs. So far, an ambitious variant catalog of almost 700,000 has been achieved, but even better detail of each variant is needed: nomenclature and general description of the structure of the STR and of each allelic variant.

The STRs are homogeneously distributed, for this reason we found 56.12% of the total in gene regions (99.4% in introns), which was useful for naming many markers and not using the SNP notation. However, the SNPs modify the STRs regions and must be taken into account in the nomenclature, configuration of the STR core structure and parameters of bioinformatics tools.

Another genomic event included in this study and incorporated into the database is multiple segments. We have seen more than 40% STRs of the Y chromosome that are multiple intra- and interchromosomally. Event that explains why the Y chromosome has so many STRs with more than one allele per individual. It is also useful to explain the non-specificity of the use of bioinformatic tools. The design to detect a certain region and use it as a marker also requires the knowledge and good use of these segments. The specificity depends on that design.

The site http://arrobasistemas.com/humstrs2/index.html shows the characteristics of the 319 STRs used in this work, and those used in Forensic and Health Sciences, among others, are considered. In the detail of each marker, information from the UCSC and Ensembl sites is gathered, in addition to data generated with the lobSTR and StraitRazor tools on 27 genomes.

The detail of each STR is comprised of the following groups: Names, Genomic Location, Associated Genomic Events, Tandem Repeat Finder Output, Population Parameters, Allelic Assignments, Alignments, SNPs (+/- 100 bp around the nucleus) and finally the Core structure (190 structures were defined). The resource used for each group is mentioned.

The use of two tools for the detection of STRs was established mainly due to the low number of reads (average 11 to 15) found in the sequenced genomes and involved establishing a double allelic assignment. The coincidences of the tools depended on the degree of complexity of the STR structure, the number of sequences detected by each tool, the adjustment of parameters taking into account SNPs and multiple segments, these same variations also produced different allelic assignments in many cases, due to to individual situations, other changes not taken into account in the parameters and even errors due to an incorrect annotation of some SNP.

Considering stuttering and having established the degree of coincidence of the tools, 76.9% of total allelic assignments (8,710) were identical between both tools, or 65.4% under slightly more stringent conditions. The rest can still be corrected with the intervention of the analyst. Those sequences obtained with StraitRazor (whose allelic assignments coincided with lobSTR) were used to establish the structures.

There is still a lack of modifications or new tools that detect those complex core configuration STRs, or more information on these structures for the entire universe of human STRs. The tools (alone or in conjunction with others) are also required to be able to qualify the allelic detections and assignments they achieve. The quality of these assignments will depend on all of the above, in addition to the quality conditions and technologies used in sequencing.

It was possible to check in a simple analysis (and some findings with SNPs) that the working assemblage: HG19 and the current HG38 have annotation errors with respect to the reference allele (the annotated one is not the most frequent) for both SNPs and STRs. of simple structures. This situation is not easily demonstrable with complex STRs.

Once this is overcome, it will be possible to have the best description of all the human STRs. This statement also applies to the rest of the STRs of living beings.

ÍNDICE GENERAL

Índice de Tablas y Figuras

1. Introducción

Los microsatélites, también conocidos como repeticiones de secuencia simple o repeticiones cortas en tándem, son regiones de ADN repetitivas no codificantes compuestas de pequeños motivos de 1 a 10 nucleótidos repetidos en tándem, que están muy extendidos tanto en genomas eucariotas como procariotas (Campo, 1998; Tóth, 2000). Utilizados ampliamente como marcadores genéticos, los microsatélites tienen un atributo particular en el sentido de que sufren mayores tasas de mutación que el resto del genoma (Jarne, 1996) y que van desde 10^{-2} a 10^{-6} nucleótidos por locus por generación (Sia, 2000).

Se han sugerido varios mecanismos para explicar la alta tasa de mutación de los microsatélites, incluidos los errores durante la recombinación, el cruce desigual y el deslizamiento de la polimerasa durante la replicación o reparación del ADN (Strand, 1993).

El alto poder de discriminación de los microsatélites es una característica importante que justifica su uso en estudios genéticos de poblaciones y en ciencias forenses (Rosenberg, 2002; Broman, 1998; Rosenberg, 2005; Bamshad, 1994; Mead, 1994; Van der Velden, 1999; Bar et al, 1997; Gusmao, 2006).

Hasta hace unos años, se pensaba que los microsatélites son marcadores selectivamente neutros y no estaban afectados por presiones selectivas. Sin embargo, ahora es evidente que la expansión del número de repeticiones puede causar enfermedades humanas. Por ejemplo, la enfermedad de Huntington es causada por aumentos en la longitud de una repetición de motivo CAG presente en el gen de la proteína huntingtina en el cromosoma 4 humano (Moxon y Wills, 1999), y un número creciente de trastornos neurodegenerativos se han relacionado con repeticiones de microsatélites expandidos, principalmente en la clase de tri-nucleótido (Goldstein y Schlotterer, 1999; Cummings y Zoghbi, 2000; Everett y Wood, 2004).

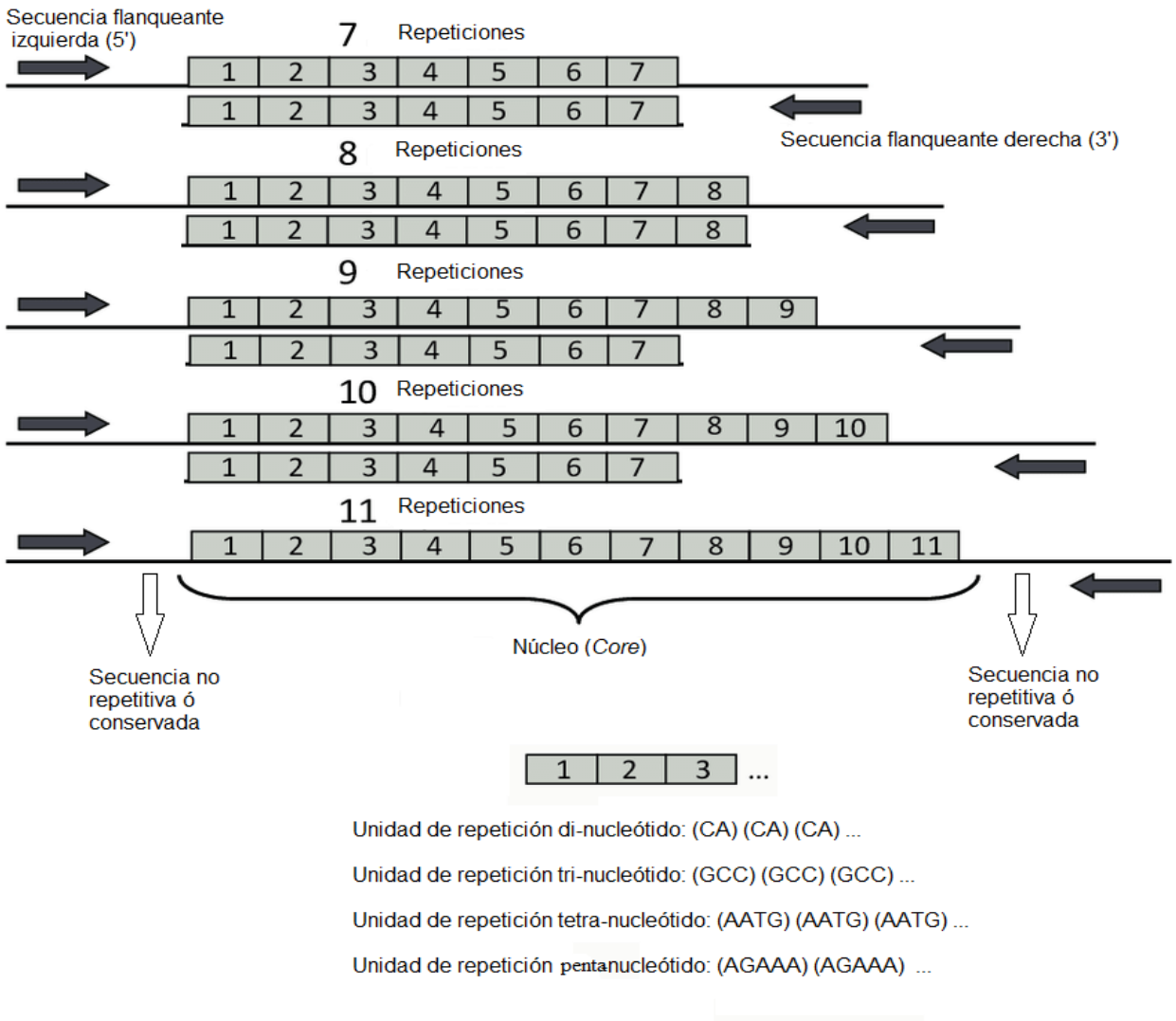
1.1 Criterios de clasificación y abundancia

Existen dos formas de clasificar a los microsatélites, una que indica la longitud de la repetición (mono, di, tri, tetra, etc.) en pares de bases de nucleótidos de la unidad de repetición o motivo, y otra que describe la complejidad del núcleo del STR (se entiende por núcleo al largo completo que involucra todas las repeticiones) y se pueden dividir en perfectos o imperfectos, simples o compuestos e interrumpidos o ininterrumpidos (Oliviera, 2006; Pemberton, 2009; Lareu, 2013), situación que se conoce también como pureza del STR.

Dependiendo de la fuente, el genoma humano posee desde casi un cuarto de millón (Benson, 1999), cerca del millón (Lander, 2001) a un millón y medio de loci STRs (Lareu, 2013). Estas diferencias pueden deberse al uso de distintos algoritmos en las herramientas bioinformáticas para hallarlas, al porcentaje del genoma secuenciado del momento o simplemente el rango en pares de bases de lo que se considera una repetición STR microsatélite (1-6 pb, 2-6 pb, 2-10 pb, etc.).

La propiedad de poseer variantes alélicas según cambios en la longitud le confiere a los sistemas STRs un elevado polimorfismo, por encima de lo que una variante SNP o SNV podría y debido a que la longitud total de los microsatélites promedia las 150 pb, las convierten en candidatas ideales para el uso en análisis de genética de poblaciones y ciencias forenses. La Figura 1 muestra los componentes básicos de esta variante.

Figura 1. Componentes básicos de un STR.



Las secuencias flanqueantes pueden interpretarse como cebadores para aquellas técnicas *in vitro* (PCR) ó secuencias de anclaje para técnicas *in silico* (herramientas bioinformáticas).

1.2 Estructura del núcleo de las variantes STRs

Existe un número reducido de microsatélites humanos usados en ciencias forenses cuyas estructuras del núcleo de las variantes alélicas están ampliamente estudiadas y registradas:

TH01, TPOX, D7S820, SE33, FGA, D12S391, DYS449, etc. se puede acceder al listado completo en STRBase: https://strbase.nist.gov/index.htm (Ruitberg, 2001)

El núcleo de una variante STR puede estar compuesto por repeticiones del mismo tipo o distintas, en secuencia y/o largo, con ausencia o presencia de interrupciones. La manera en que las repeticiones completas, incompletas e interrupciones configuran al núcleo se conoce como estructura. Se generan así distintas combinaciones, desde repeticiones simples no interrumpidas (TH01, TPOX, D7S820, etc.) hasta núcleos heterogéneos, con repeticiones de distinto largo y secuencia, con interrupciones de unas pocas pares de bases (D13S634, D13S305, SE33, FGA, etc). En la Figura 2 se muestran cuatro STRs de distintos grados de complejidad: TH01

considerado simple, D12S391 compuesto ó mixto, DYS449 compuesto con interrupciones y D13S634 complejo.

Figura 2. Estructura del núcleo de los marcadores TH01, D12S391, DYS449 y D13S634.

Dentro de la secuencia se destacan en azul y rojo los cebadores, y en verde, celeste y amarillo las distintas zonas del núcleo.

1.2.1 Representación de la estructura.

La forma habitual de representar la estructura de la secuencia del núcleo STR es encerrar la secuencia de la unidad de repetición entre corchetes y a continuación un subíndice numérico que indica el conteo de repeticiones ó alfabético (letra “n”) para generalizar la estructura a todos los alelos hallados y secuenciados de la población. Por ejemplo:

$[AGAT]_{12}$ para un núcleo con 12 repeticiones AGAT.

[GAAA]$_n$ para un núcleo con n repeticiones GAAA.

[TGAA] $_{3, 4, 5, 6, 7, 8, 9, 9.3, 10, 11, 13.3}$ para un núcleo con un número de repeticiones conocidas de secuencia TGAA del STR TH01.

Aquí se presenta un caso especial de este STR, en los alelos 9.3 y 13.3 el núcleo consta de 9 y 13 repeticiones completas y una secuencia TGA incompleta que se indica .3 (por tres nucleótidos).

Si la estructura del núcleo del STR se hace más compleja pero se conoce más en detalle, se puede generalizar usando un rango de números enteros separados entre guiones para cada repetición distinta, e indicar con 0 si en ocasiones esa secuencia no aparece.

Además se puede incorporar las interrupciones (no entre corchetes) con su secuencia (en minúscula) e incluso indicando con N (mayúscula) si esa secuencia es distinta entre las distintas variantes alélicas, seguido de un número entero que indica la longitud de esa secuencia. De esta manera se puede condensar la secuencia de aquellos STRs complejos. Algunos ejemplos:

D19S433: [CCTT]$_n$ ccta [CCTT]$_n$ cttt [CCTT]$_n$

SE33: [CTTT]$_n$ tt ct [CTTT]$_n$

DYS449: [TTCT]$_n$ N22 [TTCT]$_n$ N12 [TTCT]$_n$

vWA: tga [TCTA] $_{0-4}$ [TCTG] $_{1-5}$ [TCTA] $_{1-20}$ [TCTG] $_{0-5}$ [TCTA] $_{0-5}$ [TCCA] $_{0-1}$ [TCTA] $_{0-4}$ [TCCA] $_{0-1}$ tcc

D21S11: cct [TCTA] $_{4-11}$ [TCTG] $_{3-15}$ [TCTA] $_{3-6}$ [TATCTA] $_{0-1}$ [TCTA] $_{0-6}$ tca [TCTA] $_{2}$ tccata [TCTA] $_{6-15}$ [TCA] $_{0-1}$ [TCTA] $_{0-4}$ [TATCTA] $_{0-1}$ tcg

El sitio web de STRBase: https://strbase.nist.gov/index.htm (Ruitberg, 2001) actualmente es el sitio más completo en informar las estructuras y sus representaciones de todas las variantes alélicas de STRs de uso común en ciencias forenses.

1.3 Asignación alélica

La designación particular de un alelo detectado se conoce como asignación alélica y puede estar basada según la longitud del fragmento amplificado, separado y detectado por electroforesis capilar -*gold standard*- (Wenz, 1998) ó basada en la secuencia determinada por secuenciación Sanger ó NGS.

En la década de los 90's del siglo pasado las tecnologías separativas se hicieron más accesibles para los laboratorios de genética forense, y junto con la técnica de PCR, propició a que la determinación alélica basada en la longitud de los fragmentos se convirtiera en el método estándar preferido por estos laboratorios, desplazando a las técnicas predecesoras RFLP o VNTR y esto obligó a que se establezcan pautas de uso común para la correcta asignación alélica a los fragmentos detectados, la determinación de frecuencias alélicas a los fines de los cálculos estadísticos y que los resultados sean reproducibles en otros laboratorios.

La ISFG (https://www.isfg.org/) ha definido desde entonces la nomenclatura, criterios y demás tópicos relativos al conocimiento de los marcadores STRs y otros marcadores de uso en las ciencias forenses (Bar et al, 1997; Gusmao, 2006; van der Gaag, 2015; Parson, 2016).

La asignación alélica debía ser simple, pero a su vez que informara correctamente el contenido del STR detectado para ese alelo. La solución fue definir un número que indique el número de repeticiones contenidos en el fragmento detectado.

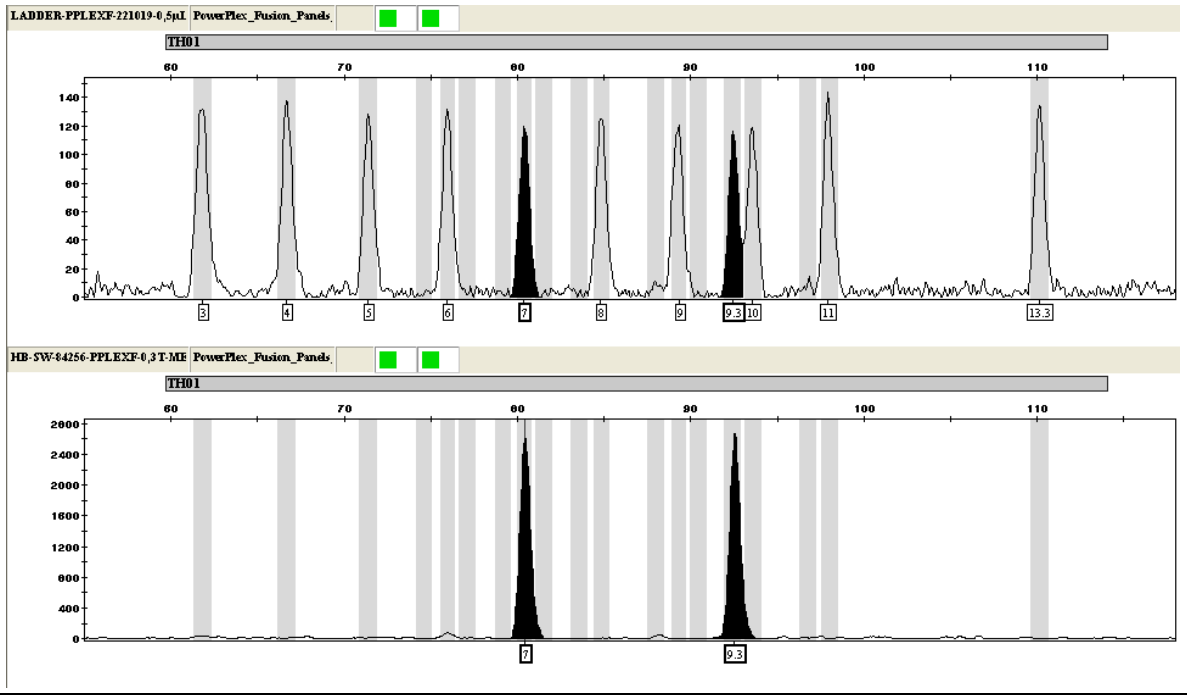
Debido a que secuenciar cada fragmento detectado resultaba costoso y lento, se continuó con la tecnología de fragmentos (electroforesis capilar principalmente) pero con un mejoramiento constante, incorporando estándares de tamaños con pocas fluctuaciones de migración, y luego, cada fabricante de reactivos comerciales de genotipificación debía proveer una escalera alélica que incluyeran todos los alelos conocidos de la población, de manera de asegurar con mucha fiabilidad que el fragmento detectado sea un alelo determinado.

Previo a esto el fabricante debía conocer la secuencia de cada alelo con la que construía la escalera alélica. La asignación alélica entonces resultaba de tres simples pasos:

1) Asignación de tamaño (en pb) al fragmento desconocido.
2) Asignación de tamaños (en pb) a la totalidad de los fragmentos de la escalera alélica.
3) Comparación entre el fragmento desconocido y los fragmentos de la escalera alélica y asignación del número alélico al fragmento desconocido según las coincidencias con la escalera alélica.

Estas tareas se fueron haciendo más eficientes con la evolución de herramientas bioinformáticas como Genescan®, GeneMapper® ó GeneMarker (Liu, 2011), y con la incorporación de más variantes alélicas en la escalera, a medida de que el número de individuos genotipificados fue en aumento. La Figura 3 resume este tipo de asignación alélica.

Figura 3. Asignación alélica a un individuo en estudio para el marcador TH01.



Arriba: Escalera alélica.
Abajo: Alelos 7 y 9.3 asignados al individuo en estudio.

La mayoría de los marcadores usados inicialmente eran simples o compuestos, por lo tanto condensar la secuencia a un número no resultaba en mucha pérdida de información. Además el uso de números enteros (o con puntos y otro número entero para microvariantes, Bar et al, 1997) simplificó los cálculos estadísticos de vínculos biológicos.

Esta forma de denominar a un alelo de acuerdo al número de repeticiones y demás reglas definidas desde la ISFG aún están vigentes en ciencias forenses.

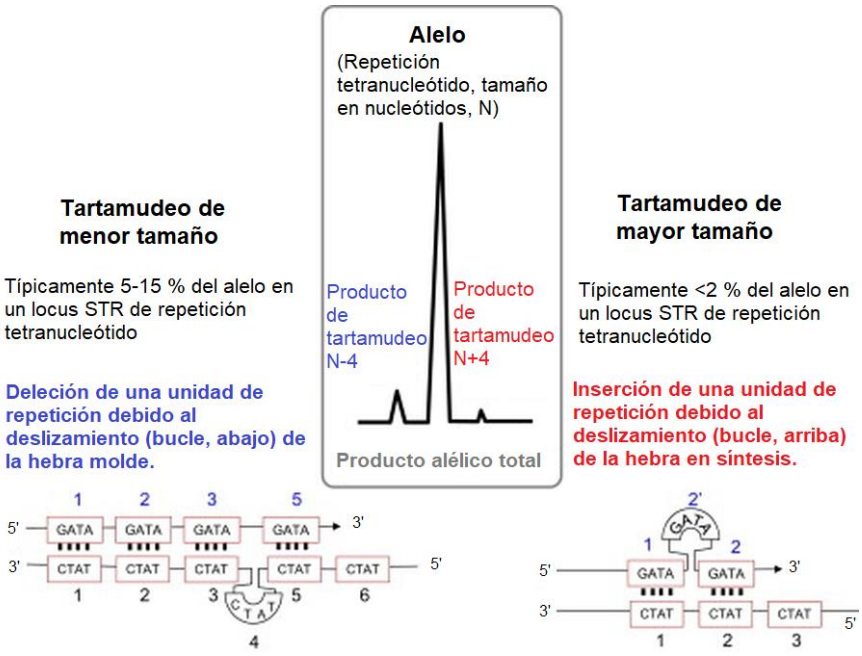
Respetando estos criterios, si se conoce la secuencia, la asignación alélica derivada de esta es directa y será aplicado en la segunda parte de este trabajo, con el uso de herramientas de detección de STRs en archivos de secuenciación de segunda generación.

1.4 Tartamudeo

Un evento difícil de sortear en la amplificación de los microsatélites se conoce como tartamudeo. Es un evento inherente a las ADN polimerasas, sean estas procariotas o eucariotas (Strand ,1993; Fan, 2007).

La hipótesis más aceptada es que se debe a un deslizamiento de una de las dos hebras en el momento de la síntesis debido a un bucle intracatenario en la hebra que se desliza, creando un acortamiento o alargamiento. Si la hebra que sufre el acortamiento es la hebra molde, la hebra que se está sintetizando es más corta que la hebra molde, caso contrario sucede lo inverso, la hebra en síntesis será más larga si esta sufre el bucle (Figura 4). Esto sucede con mayor frecuencia con los microsatélites dinucleótidos, y la mayoría de las veces los acortamientos o alargamientos suceden en números enteros de repeticiones (-1, -2, -3, +1, +2, +3, etc.). Los acortamientos -1 y -2 son más frecuentes. En general, si el largo de la repetición aumenta, la abundancia de tartamudeos disminuye. Para las repeticiones tetra-nucleótido y en condiciones controladas de amplificación los fragmentos amplificados por tartamudeo no superan el 20 % de la población de fragmentos principal.

Figura 4. Hipótesis más aceptada del origen del tartamudeo.



Centro: Representación de un alelo y sus tartamudeos detectados por electroforesis capilar.

Abajo: Esquema de una hebra en síntesis (dirección 5' → 3') sobre una hebra molde (dirección 3' → 5').

1.5 Bases de datos públicas

La recopilación de los primeros datos respecto de microsatélites provino de la recopilación de STSs que fueron suministrados por centros de investigación genómica abocados a la creación de los primeros mapas físicos del genoma humano (Radiation Hybrid Database, Whitehead Institute, Stanford Human Genome Center y Sanger Center). Estos fueron almacenados en sitios como dbSTS y GDB (Letovsky, 1998).

Paralelamente las secuencias de estos marcadores eran cargados en Genbank (NCBI) y posteriormente UniSTS reúne esta información, incluyendo secuencias de cebadores, tamaño del producto de PCR, información de mapeo y referencias cruzadas a LocusLink, dbSNP, RHdb, GDB, MGD y Map Viewer.

Actualmente las bases de datos públicas que albergan información sobre microsatélites humanos podemos accederlos desde sitios tales como Ensembl (Zerbino D., 2018), UCSC (University of California, Santa Cruz) (Kent, 2002), SNPs Database, GeneLoc (Rosen, 2003) que contienen datos de ubicación genómica (cromosoma, coordenadas, cebadores, largo del fragmento de referencia, etc.) y de mapeo (laboratorio que lo realizó, posición en cM –centiMorgan-), desde 5.000 hasta un poco más de 300.000 marcadores, pero poseen poca o nula información sobre frecuencias alélicas, heterocigosidad, o su distribución poblacional.

También existen las bases de datos cuyos datos fueron obtenidos por electroforesis capilar y que contienen información de un número más reducido de variantes STRs acotados a aquellos usados para estudios poblacionales ALFRED (Cheung KH, 2000) o de ciencias forenses STRbase y ALLST*R, y contienen, dependiendo de la base, información más detallada sobre número de alelos, frecuencias alélicas, microvariantes, estructura del STR, separación por población, etc.

Existe también un catálogo de variantes STRs de creación más reciente (Willems, 2014), basado en secuenciación por NGS de individuos del proyecto 1000 Genomas, mucho más completo en cuanto al número de variantes: 668.631 (300.000 MAF superior al 1 %). Desde su sitio *web* (http://strcat.teamerlich.org/ que redirige a: http://webstr.ucsd.edu/) se puede realizar la búsqueda según número de cromosoma, ubicación, secuencia de la repetición, pureza, heterocigosidad por población, entre otros filtros permitiendo hacer una inspección minuciosa, pero es deficiente para aquellas búsquedas que involucren nombres de STRs conocidos a pesar de que este trabajo validó su asignación alélica en concordancia con aquellos STRs que pertenecen al panel forense del CODIS y al panel de ligamiento de Marshfield (Rosenberg, 2002).

Además, buscar un determinado STR implica conocer *a priori* en qué cromosoma se encuentra y su ubicación, y aun sabiendo esto, se debe saber a qué versión del ensamblaje del genoma humano se trata.

1.6 Disponibilidad de genomas.

Desde que se publicó el primer borrador de la secuencia completa del genoma humano en febrero del año 2001 (Lander, 2001; Venter, 2001), vinieron después sucesivos ensamblajes y sus respectivas actualizaciones, las más relevantes son las siguientes:

1.- NCBI34/hg16 julio 2003.
2.- NCBI35/hg17 mayo 2004.
3.- NCBI36/hg18 marzo 2006.
4.- GRCh37/hg19 febrero 2009.
5.- GRCh38/hg38 diciembre 2013.

El acceso a los datos de secuenciación, anotación y demás información es posible desde sitios como UCSC, Ensembl o NCBI.

Además de los ensamblajes del genoma de referencia mencionados, existe un número suficiente de genomas secuenciados disponibles por tecnologías de secuenciación de segunda generación (Proyecto 1000 Genomas, https://www.internationalgenome.org/home) de libre

acceso. Este proyecto ha finalizado la fase 3, donde se incrementó el número de individuos secuenciados y la recopilación de variantes, respecto de la fase 1 y la fase piloto (Tabla 1).

Tabla 1. Datos disponibles del Proyecto 1000 Genomas.

Lanzamiento de 1000 Genomas	Variantes (millones)	Número de individuos	Número de poblaciones
Piloto	14,8	179	4
Fase 1	37,9	1.092	14
Fase 3	84,4	2.504	26

Información obtenida del sitio https://www.internationalgenome.org/data.

1.7 Herramientas bioinformáticas para la detección de STRs

De las herramientas existentes para detectar STRs se pueden distinguir al menos de tres tipos:

1.- Aquellos diseñados para analizar archivos con formato FASTA, por ejemplo: Tandem Repeats Finder (Benson, 1999), RepeatMasker (Smit, 1996), gmato (Wang, 2013), Misa-Web (Beier, 2017), SciRoKO (Kofler, 2007), mreps (Kolpakov, 2003) y STR-FM en Galaxy (Fungtammasan, 2015).

2.- Aquellos diseñados para analizar archivos con formato FASTQ dependientes de alineamiento, por ejemplo: lobSTR 4.0.6 (Gymrek, 2012), RepeatSeq (Highnam, 2013) y STRviper (Cao, 2014).

3.- Y aquellos diseñados para analizar archivos con formato FASTQ independientes de alineamiento, por ejemplo: STRaitRazor v1-v2-v3 (Warshauer, 2013), TSSV (Anvar, 2014), MyFLq (Van Neste, 2014) y SEQ Mapper (Liu, 2018).

Para el presente trabajo se seleccionaron:

a.- Tandem Repeats Finder debido a su sencillez en uso, velocidad y estar presente en el portal de UCSC

b.- lobSTR debido a que es la herramienta usada por el grupo de Erlich (Gymrek, 2012; Willems, 2014) para generar el catálogo de las casi 300.000 variantes que se mencionó en los párrafos anteriores

c.- STRaitRazor 3.0 (Woerner, 2017) que promete tiempos de análisis muchos más rápidos.

En consecuencia, esto permite obtener información nueva de las secuencias de las distintas variantes alélicas. Es necesario entonces tener un buen conocimiento de las características genómicas relevantes de los STRs (ubicación cromosómica, secuencia de repetición y flanqueantes a esta, complejidad del núcleo de repetición, variantes alélicas conocidas, etc.) y que son necesarias para una correcta elaboración del algoritmo y consecuentemente la herramienta informática que deberá encontrar las variantes STR en un genoma secuenciado por NGS, o parte del mismo.

2. Objetivos

En el presente trabajo final, se plantean los siguientes Objetivos:

2.1 Generales

- Recopilación y curado de información de variantes STRs humanas disponibles en bases de datos públicas.
- Corroboración de genotipado de STRs ampliamente utilizados en ciencias forenses y ciencias de la salud sobre genomas secuenciados del proyecto 1000 genomas y otros proyectos.
- Crear una base de datos con funcionamiento local, que integre información proveniente de las bases de datos analizadas y de la información obtenida a partir de las herramientas lobSTR y STRaitRazor 3.0.

Se pretende que esta base de datos contemple algunas de las deficiencias mencionadas anteriormente, permitiendo una búsqueda más intuitiva. La búsqueda para asignación alélica con estas herramientas son obtenidos por NGS para un número reducido de individuos pertenecen al proyecto 1000 genomas. Este proyecto estará enfocado sólo en aquellos STRs de uso común en el campo de la genética forense (FBI CODIS Core STR Loci y European Core Loci) y de diagnóstico prenatal de aneuploidías por QF-PCR (Adinolfi, 2000; Cirigliano, 2009).

2.2 Específicos

- Diseñar *scripts* y *pipelines* para la obtención de la información existente en bases de datos públicas de STRs
- Establecer coincidencias según ubicación cromosómica, según “ID” conforme la técnica de mapeo que se usó para ubicar el marcador (Marshfield, CHCL, Wi YAC, WI RH, Stanford TNG, etc.), nomenclatura previa del marcador (UNISTS, Genbank, Decode, etc.) o alguna otra información unívoca que defina la variante.
- Establecer cuál será la información crítica que poseerá la variante contrastada, que podría contener: nombre trivial, por nomenclatura; longitud, secuencia y complejidad de la repetición; localización en el cromosoma según ensamblaje; número de alelos; longitud o rango alélico; frecuencias alélicas; heterocigosidad; tecnología usada para la determinación de la longitud del STR; etc.
- Evaluar un número reducido de genomas (FASTQ) de aquellas variantes de uso en genética forense **CODIS (US):** CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11 mas D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433, y D22S1045; **Europeos:** SE33, D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433, D22S1045, DYS19, DYS385 a/b, DYS389 /II, DYS390, DYS391, DYS392, DYS393, DYS438, DYS439 entre otros: F13A1, F13B, FES/FPS, HPRTB, LPL, Penta D, Penta E, D6S1043, D14S1434, DYS388, DYS434, DYS437, DYS447, DYS448, DYS456, DYS458, DYS460, DYS464, DYS635, Y-GATA-A4, Y-GATA-A7.1, Y-GATA-A7.2, Y-GATA-A10, Y-GATA-H4 con una herramienta recientemente optimizada para esta tarea: STRait Razor v3.0.
- Se realizará también un análisis similar para aquellos STRs usados en la detección de aneuploidías, a saber: **Aneufast:** TAF9L, X22, DXYS267, DXYS218, DXYS156, HPRT, DXS6803, DXS6809, DXS8377, SBMA, D21S1414, D21S1411, D21S1446, D21S1437, D21S1809, D21S1412, D21S1435, D21S1442, D18S391, D18S390, D18S535, D18S386, D18S858, D18S499, D18S1002, D18S976, D13S631, D13S634, D13S258, D13S305, D13S628, D13S742, D13S797; **Devyser:** D15S643, D15S657, D15S659, D15S822, D15S1513, D16S539, D16S753, D16S2620, D16S3396, D16S2624, D22S1045, D22S683, D22S686, D22S689, GATA198B05, D18S386, D13S742, D13S634, D13S628,

D13S305, D13S1492, D18S978, D18S535, D18S386, D18S976, GATA178F11, D21S1435, D21S11, D21S1411, D21S1444, D21S1442, D21S1437, DXS1187, XHPRT, DXS2390, DXYS267, DXYS218.

- Buscar STRs para cromosomas 9, 14 y 17 donde no mapean las variantes mencionadas en los incisos anteriores (excepto D14S1434), debido a que existe la necesidad diagnóstica de conocer si estos cromosomas están implicados en trisomías o disomías uniparentales en individuos en gestación.

3. Materiales y Métodos

3.1 Datos

Se recopiló la información relativa a microsatélites (STRs, STSs y SNPs) descritos en las Tablas 2 y 3 para la construcción de la base de datos local. Los archivos utilizados en lenguaje Perl se muestran en varios apartados del Anexo III.

Tabla 2. Bases de datos públicas de STRs, STSs y SNPs humanos

Variante principal del sitio	Nombre del sitio	Dirección del sitio web	Número de marcadores/loci	Número de atributos de la variante
STR	A Catalog of Human STR Variation	http://strcat.teamerlich.org/[1]	668.631	16
STR	STRBase (SRD-130)	https://strbase.nist.gov/	56	Hasta 14
STR	Genotyping Y-STR and CODIS markers	http://lobstr.teamerlich.org/ystr-codis.html	110	5
STR	ALLST*R	http://allstr.de/allstr/home.seam	93	18
STR	ALFRED	https://alfred.med.yale.edu/alfred/index.asp	760	20
STS	UCSC - Genome Browser Gateway	https://genome.ucsc.edu/cgi-bin/hgGateway	322.212	52
STS	Ensembl	ensembldb.ensembl.org	328.845	13
STS	UniSTS	https://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring01/unists.html	321.290	8
STS	GeneLoc	https://genecards.weizmann.ac.il/geneloc/index.shtml	405.100	8
SNP	dbNSP	https://www.ncbi.nlm.nih.gov/snp	5.500	26

1 El sitio ha cambiado a http://webstr.ucsd.edu/ en Junio 2020.

3.1.1 Datos de secuenciación de microsatélites por tecnología NGS

Los datos de secuenciación fueron extraídos del archivo "phase_1_final_calls.vcf" proveniente de la anotación realizada con lobSTR (versión 2.0.4), mediante línea de comandos de Linux, filtrando las primeras 8 columnas. Luego, los archivos de texto de salida fueron llevados a una planilla de cálculo de Excel para reorganizar la información de la columna 8 que contenía 7 datos adicionales de modo de expandir los atributos de cada variante. Finalmente, se agregó el valor de heterocigosidad calculado con VCFtools (0.1.15), tal como se muestra en el punto 3.2 (atributo HET en las tablas Fase1 y 319markers).

3.1.2 Datos de marcadores STSs

Desde UCSC: los archivos "STSMap.txt" y "stsInfo2.txt" fueron combinados a partir del atributo "identNo", de modo de unificar los identificadores de los STSs. Se eliminaron las filas con ausencia de datos (ítem UCSC en la Tabla 3).

Desde Ensembl: se realizaron las consultas mediadas por la API de Ensembl (release 94 - October 2018) a la base de datos a través de 30 *scripts* en lenguaje Perl. Se organizaron los archivos de texto de salida (ítem Ensembl en la Tabla 3).

Desde UniSTS: La información disponible en UniSTS no contiene datos de ubicación cromosómica, pero sí de cebadores y de nombres y/o alias. Estos datos no fueron incorporados en la base de datos local, pero fueron luego utilizados para chequear la coincidencia entre las bases de datos UCSC y Ensembl.

3.1.3 Datos de variantes SNPs

Las variantes de nucleótidos simples ó únicos (SNPs ó SNVs) constituyen hasta el 90% de todas las variaciones genómicas humanas, por lo que fueron incluidas para la construcción de la BD local. El número de variantes anotadas en dbSNP asciende a 234.104.110 (SNP150), por lo que sólo se seleccionaron los SNPs contenidos en regiones STRs de los marcadores elegidos para el análisis de las herramientas lobSTR y StraitRazor. La información de los SNPs fueron descargados vía MySQL desde la Tabla "SNP150", base de datos "hg19" disponibles en genome-mysql.soe.ucsc.edu, acotados al cromosoma, comienzo y final de la ubicación de cada STR en estudio. El archivo fue depurado para ser leídos mediante *scripts* en lenguaje Perl, tal como se observa en la Tabla 3 (ítem Snp150common).

3.1.4 Regiones génicas: exónicas e intrónicas

Las variantes microsatélites se encuentran tanto en regiones génicas como en las intergénicas y esta información fue incorporada para la construcción de la BD local.

La anotación de los genes humanos se ha ido refinando y puede variar según la recopilación que se investigue. En un esfuerzo por reunir aquellos que genes que se anotan consistentemente y de buena calidad se ha logrado un conjunto "núcleo" de genes humanos que codifican proteínas en el Proyecto Consensus Coding Sequence (CCDS). Esta información contenida en el archivo "ccdsGene.txt.gz" fue combinada con los nombres disponibles en HGNC, según el identificador de CCDS, y ordenados con *scripts* de Perl (ver ítem "Exon" e "Intron" de la Tabla 3).

3.1.5 Citobandas y segmentos múltiples.

Los datos provenientes de bandeado citogenético correspondientes a regiones visibles en cromosomas obtenidos por técnicas citogenéticas (archivo "cytoBand.txt.gz") y los datos de segmentos múltiples (regiones que se multiplican intra e inter cromosómicamente, con tamaños

que van desde 1.000 pb a 770.000 pb) del archivo “build37.xlsx”, se incluyeron en la BD local (ítem “Cytobands” y “Superdup” en la Tabla 3).

Tabla 3. Datos contenidos en la base de datos local

Tabla	Tuplas	Atributos
Fase 1	668.631	12
UCSC	299.741	22
Ensembl	288.269	10
Snp150common	9.754	26
Exon	302.725	6
Intron	274.089	6
Superdup	51.599	11
Cytobands	862	4
319markers	319	22

3.2 Construcción de la base de datos

Se utilizó el modelo relacional para la confección de la BD local mediante el lenguaje de consulta SQL. SQL usa una combinación de álgebra relacional y construcciones del cálculo relacional. SQL incluye características para definir la estructura de los datos, para la modificación de los datos en la base de datos y para la especificación de restricciones de seguridad. Los esquemas de relación para la construcción de la BD local fueron:

fase1 (NUMCROM, START, END, REF, ALT, QUAL, MOTIF, REFL, RL, RPA, RU, HET)

ucsc (indentno, numcrom, start, end, name, genbank, gdb, othernames, dbstsid, otherdbsts, leftprimer, rigthprimer, distance, mergeucsc, genethonname, marshfieldname, wiyacname, wirhname, gm99gb4Name, gm99g3Name, tngname, decodename)

ensembl (numcrom, start, end)

exon (numcrom, start, end, name, gene, strand)

Snp150common (bin, chrom, chromStart, chromEnd, name, score, strand, refNCBI, refUCSC, observed, molType, class, valid, avHet, avHetSE, func, locType, weight, exceptions, submitterCount, submitters, alleleFreqCount, alleles, alleleNs, alleleFreqs, bitfields)

intron (numcrom, start, end, name, gene, strand)

superdup (chrom, chromStart, chromEnd, size, name, strand, otherchrom, otherstrand, otherend, othersize, uid)

cytobands (NUMCROM, START, END, NAME)

319markers (camp, name, ucsc, ensembl, gene, suggested, chr, start, end, cytoband, segmult, period, period1, purity, seq, structure, structurera, conteos, alleles, anchor5, anchor3, HET)

3.3 Consultas SQL

El resultado de una consulta SQL es una relación.

3.3.1 Criterios para consultas sobre la BD local

Todas la consultas hechas entre las Tablas de la BD local parten de la premisa que la coordenada de inicio siempre es menor que la coordenada fin

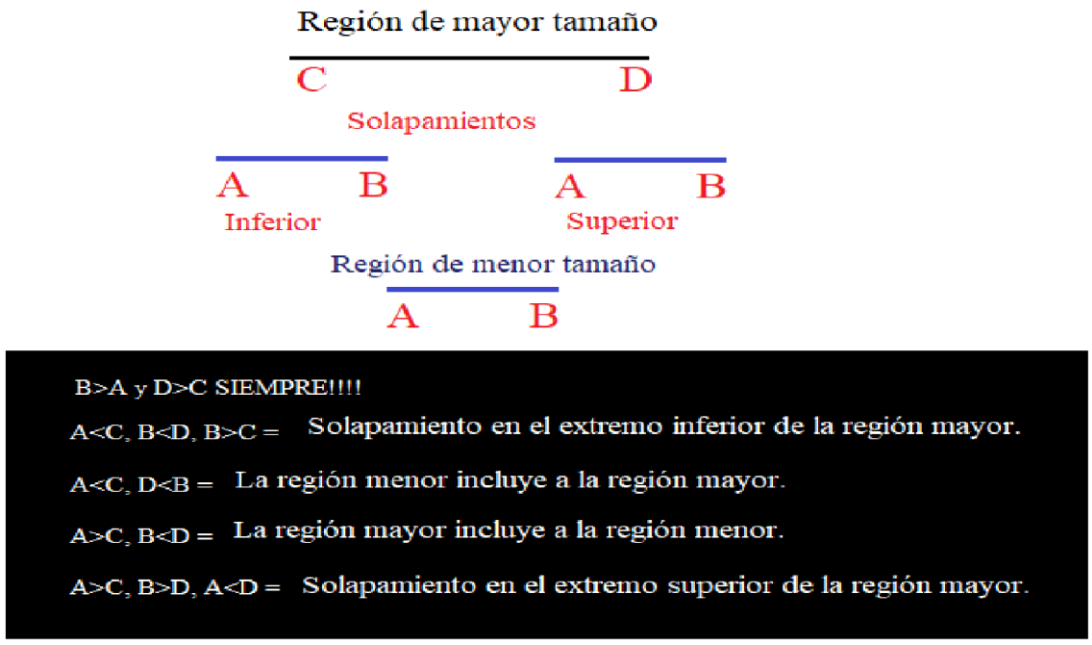
3.2.1 Estructura de los datos

Definir las tres cláusulas: select, from y where.

select *A1 , A2 ,..., An*
from *r1 , r2 ,..., rm*
where *P*

Cada *Ai* representa un atributo, y cada *ri* una relación. *P* es un predicado. La consulta es equivalente a la expresión del álgebra relacionalal para un mismo cromosoma (1 al 22, X e Y), tal como se observa en la Figura 5, siendo A y C coordenadas de inicio, y B y D coordenadas de fin. Se consideró para las consultas que las regiones STSs en general son más grandes (41 – 4.991 pb para UCSC y 30 – 5.025 pb para Ensembl) que las regiones que comprende una variante STRs (8 – 2.546 pb según anotación de Fase 1). Además, los segmentos múltiples son más grandes que los anteriores (1.000 - 770.429 pb). Se observó una alta diversidad de tamaños en las regiones de exones e introns, desde 1 pb hasta 1.160.409 pb.

Figura 5. Criterios para el cruce de datos según número de cromosoma y coordenadas entre tablas.



3.3.2 Cantidad de STRs

Consultas UCSC y Ensembl: se determinó la cantidad de STRs secuenciados contenidos en los marcadores STSs de UCSC y Ensembl, a través de consultas a la BD local (Fase 1 en Tabla 3) siguiendo los criterios expuestos en el punto 3.3.1. Se buscaron las variantes STRs de Fase1 en los marcadores anotados en UCSC y Ensembl. Se consideraron los solapamientos

(inclusión parcial de la variante en un marcador). Asimismo, se consideró la situación inversa: el marcador queda incluido dentro de una región STR.

Se usó un *script* en lenguaje Perl para generar las consultas SQL por cromosoma (Ver *scripts* en Perl y sentencias SQL generales en el apartado III.1.1).

3.3.3 Localización de STRs en regiones génicas e intergénicas

En regiones génicas (exones e intrones) e intergénicas*:* se obtuvo el número de variantes STRs dentro de exones, intrones y regiones intergénicas para obtener tanto la proporción de STRs en estas regiones, y para asignar un nombre a los microsatélites identificados según la nomenclatura de los genes, mediante el *script* SQL descrito en el apartado III.1.1.2 dentro de "Consultas MySQL con la BD local".

3.3.4 Distancias acumuladas de exones, intrones y variantes STRs de Fase 1 para ser comparadas con el genoma completo.

La consulta SQL obtuvo de las tablas Exon, Intron y variantes STRs de Fase 1, el tamaño en pb (coordenada final menos coordenada inicial) de cada exón e intrón de la base de datos local, luego se sumaron y fueron comparadas con el tamaño del genoma completo de "hg19.fa" (usando la herramienta Bioseq).

3.3.5 STSs y STRs incluidos en segmentos múltiples

Se consultó (MySQL) a la BD local por el número de marcadores y de variantes STRs que están incluidos en segmentos múltiples por cromosoma, siguiendo criterios descritos en 3.3.1. Se calculó el porcentaje del total que están incluidas en estas regiones.

3.4 Análisis genómicos

3.4.1 Análisis de marcadores utilizando Tandem Repeats Finder (TRF)

EL programa TRF es capaz de analizar STRs simples y devuelve la siguiente información:

- índices (comienzo y fin del núcleo),
- largo de la repetición,
- número de repetición,
- secuencia consenso,
- porcentajes de *match* e *indels*,
- una puntuación (*score*),
- datos termodinámicos,
- Salidas de texto plano "*.dat" y o del tipo web "*.html".

Los análisis con TRF se realizaron en:

USCS y Ensembl: las secuencias de marcadores UCSC (299.410 en formato .fasta) se obtuvieron mediante un *script* en lenguaje Perl (detalles en el apartado III.2.2: UCSC) a partir de los cromosomas 1 a 22, X e Y. Las secuencias de marcadores de Ensembl (288.247) se descargaron utilizando la API de Ensembl (detalles en el apartado III.2.2: Ensembl), con archivos de salida convertidos a formato *.fas (fasta). El total de secuencias obtenidas se analizaron con TRF utilizando los parámetros: 2 7 7 80 10 50 10 –d por línea de comandos.

- el genoma humano completo: se detectaron variantes STRs en el archivo "hg19.fa", según los parámetros 2 7 7 80 10 50 10 -d -h –ngs.

- Las secuencias fasta de los marcadores seleccionados para la segunda parte, con los parámetros: 2 7 7 80 10 50 6 –d por línea de comandos.

(Esta información será usada luego para el archivo de configuración de StraitRazor)

3.4.2 Determinación de ubicaciones de marcadores en el genoma

Se asignó la ubicación genómica de marcadores utilizando la herramienta Nucleotide-Nucleotide BLAST 2.7.1+ sobre el archivo "hg19.fa" cuando no se disponía de la ubicación previamente o no se correspondía al ensamblaje de este estudio pero la secuencia estaba disponible.

3.4.3 Coincidencias de secuencias de pares de cebadores obtenidos desde UCSC, Ensembl y UniSTS.

UCSC, Ensembl y UniSTS poseen las secuencias de los cebadores usados para la amplificación de las regiones STSs blancos. Se analizaron las coincidencias en las secuencias de cebadores entre las tres bases de datos (coincidencias estrictas).

Se analizaron los archivos dbID_all_info_1_10K_1-30_markers_GRCh37_v2_reordened_cut1-2-3-4-5-8-9-10.txt (Ensembl), stsInfo2.txt (UCSC) y human-UniSts.txt, mediante scripts en lenguaje Perl (Anexo II: *analyze-primers1.pl, analyze-primers2.pl, analyze-primers3.pl).*

3.4.4 Coincidencias estrictas (strict match) de cebadores de UCSC y Ensembl y dentro de secuencias STSs propias de su base de datos

Las secuencias fueron de dos tipos:

+/- 0 pb: Largo original según sus datos de ubicación.

+/- 250 pb: Largo ampliado en 250 pb hacia ambos extremos 5' y 3'.

Las secuencias STS de Ensembl para este análisis fueron descargadas (+/- 250 pb) y modificadas (+/- 0 pb) y las secuencias UCSC fueron extraídas del genoma "hg19.fa" según su ubicación (ambas +/- 0 pb y +/- 250 pb), y fueron convertidas a formato fasta "*.fas".

La búsqueda de coincidencias estrictas (*strict match*) requiere que ambos cebadores se encuentren incluidos en la secuencia STS a la que están asociados. Asimismo, se consideró que cada cebador pueda estar tanto en orientación "sentido" (*forward*) como "anti-sentido" (*reverse*). Esto propone cuatro posibilidades:

a) El cebador *forward* tiene el mismo sentido que la secuencia STS blanco, y el cebador *reverse* es antisentido e inverso a esa misma secuencia (resultado esperado).
b) El cebador *reverse* tiene el mismo sentido que la secuencia STS blanco, y el cebador *forward* es antisentido e inverso a esa misma secuencia.
c) Ambos cebadores *forward* y *reverse* tienen orientación sentido a la secuencia STS blanco.
d) Ambos cebadores *forward* y *reverse* tienen orientación antisentido e inverso a la secuencia STS blanco.

Estas cuatro situaciones son consideradas condición "MATCH".

Cuando el par completo o uno de sus cebadores no coinciden con la secuencia STS blanco, es considerada como "NO MATCH".

Se contabilizaron todas las situaciones mencionadas en un archivo de Excel, y se hicieron cálculos estadísticos entre los grupos "+/- 0 pb" y "+/- 250 pb". Más detalles en "Conteos y cálculo estadístico de las búsquedas cebadores-vs-STSs*"* en el Anexo III.

3.4.5 Obtención de la frecuencia del alelo minoritario de la anotación de Fase 1

Los archivos "*.frq" generados con VCFtools (ver *III.3.1* dentro de "Sentencias específicas de las herramientas") fueron usados para calcular el valor de heterocigosidad como atributo adicional (HET) en las tablas Fase 1 y 319markers.

Estos mismos archivos son recorridos con el *script* "*extract-frq-for-MAF.pl*" para extraer la frecuencia del alelo minoritario. Se ejecuta dentro del *script* "*for-extract-pl.sh*".

El archivo generado “frequencies-MAF-extracted-1-22-X-Y.txt” es recorrido por cat para eliminar líneas innecesarias y luego con awk y wc para contar aquellas ubicaciones que superan al 0.01 (MAF > 1 %).

3.4.6 Comparación de las longitudes del alelo de referencia y del alelo más frecuente en la anotación de Fase 1.

Con bcftools se extrajo la secuencia del alelo de referencia en la anotación hecha en “phase_1_final_calls.vcf”:

/root/bcftools/./bcftools query -f '%CHROM\t%POS\t%REF\n' phase_1_final_calls.vcf > phase_1_final_calls_ref.txt

El archivo resultante luego fue recorrido con el *script* “*cat-grep-ph1-ref.sh*” para extraer la información por cromosoma, archivos “chr(1 .. 22, X e Y)-ref.txt”

Finalmente, el script “*check-reference.pl*” dentro del *script* “*check-reference.sh*” realizó las comparaciones por cromosoma, entre los archivos generados con bcftools y vcftools (archivos “*.frq” del apartado anterior). Los archivos “ref-new-ref-checked-(1 .. 22, X e Y).txt” luego fueron resumidos con el script “*for-echo-awk-cat1.sh*” al archivo final “ref-new-ref-checked-1-22-X-Y.txt”. Se calculan porcentajes en el archivo “REF-allele_vs_most_frequent.xlsx”.

3.5 Selección de marcadores para análisis con lobSTR y StraitRazor

3.5.1 Recopilación de información para los marcadores propuestos y los agregados posteriormente

De los 93 marcadores propuestos (75 únicos y 18 comunes) se incorporó una serie adicional de marcadores de relevancia en ciencias forenses, ciencias de la salud y otros que fueron considerados de importancia por su elevado valor de heterocigosidad (aquellos STRs cuya proporción de heterocigotas sean mayor al 70 % del total de genotipos).

La información de los marcadores propuestos y de los agregados posteriormente se encontró en distintos sitios web, bases de datos (Tabla 2) e incluso archivos de configuración de las herramientas usadas (StraitRazor y LobSTR).

Se necesitó como mínimo la secuencia ó la ubicación de algún marcador candidato.

En algunos casos, los datos de ubicación en los sitios de la Tabla 2 son vagos ó corresponden a un ensamblaje anterior al de este estudio. Siempre que hubiera referencias a secuencias, se obtuvo la secuencia y fue buscada en el genoma “hg19.fa” con la herramienta Blastn para determinar la ubicación.

Aquellos marcadores cuya ubicación sea conocida se usaron para obtener la secuencia en formato fasta, con *script* de Perl o con la herramienta BioEdit (Hall, 2011).

La base de datos local creada también fue consultada para obtener información de marcadores, en su mayoría los de ciencias de la salud.

La secuencia y la ubicación de cada marcador fueron necesarias para los análisis posteriores con la herramienta TRF y con *scripts* de Perl y MySQL para la configuración de los archivos de StraitRazor y LobSTR.

3.5.2 Nomenclatura de los marcadores utilizados

Se mantuvo la nomenclatura D#S#, donde D# hace referencia al número de cromosoma y S# a un número de segmento definido en los primeros mapas físicos del genoma humano. Se prioriza su utilización debido a que constituye la nomenclatura más antigua sobre anotación de regiones genómicas no asociadas a proteínas.

En el caso de que el marcador seleccionado quedase comprendido sobre una región génica, se utilizó el nombre del gen para nomenclarlo, seguido de guión bajo, una “I” latina mayúscula (por intrón) ó “E” (por exón), seguido de otro guión bajo y el número de intrón ó exón donde se encuentra el STR. Ej.: ANXA11_I_5. Si el marcador no estuviera asociado a ningún gen o nombre D#S#, se buscó otro alias que pudiera asociarse. En caso de ausencia de todas las posibilidades anteriores, se nombró de acuerdo con la citobanda, cromosoma y repetición del STR (Ej.: 7p143AAAG, 9q2113TCTA, 7p121ATCT).

3.5.3 Total de marcadores en estudio

Acorde a los criterios expuestos en el punto 3.5.1, fueron seleccionados 323 marcadores que corresponden a 319 *loci*. Esto se debe a que los *loci* DYS448 y D13S305 fueron incorporados ambos en dos partes (DYS448_1, DYS448_2, D13S305_1 y D13S305_2) y el *locus* DYS389 es un caso especial de tres partes (DYS389I, DYS389II y DYS389-2). Los *loci* Amelogenina y SRY son marcadores de género y no son STRs.

De los 323 marcadores, 129 son de uso en Ciencias Forenses y 47 en Ciencias de la Salud, 5 son comunes a ambas disciplinas y 142 fueron elegidos por tener elevada heterocigosidad (114 STRs de la anotación de Fase 1/lobSTR 2.0.4) o por pertenecer a cromosomas cuyas anomalías tienen significancia clínica y no están contemplados en los grupos anteriores (28 STRs del cromosoma 7 y 9).

La recopilación de información para los marcadores de Ciencias Forenses fue heterogénea en cuanto a los sitios consultados, principalmente los sitios web de STRBase (SRD-130) y “Genotyping Y-STR and CODIS markers” de lobSTR, aunque la base de datos local también fue consultada; en cambio la información de los 47 marcadores exclusivos de las Ciencias de la Salud se obtuvieron de la base de datos local (junto a los 114 STRs de Fase 1) y la información de los 28 marcadores de los cromosomas 7 y 9 fueron obtenidos desde el sitio web “A Catalog of Human STR Variation”.

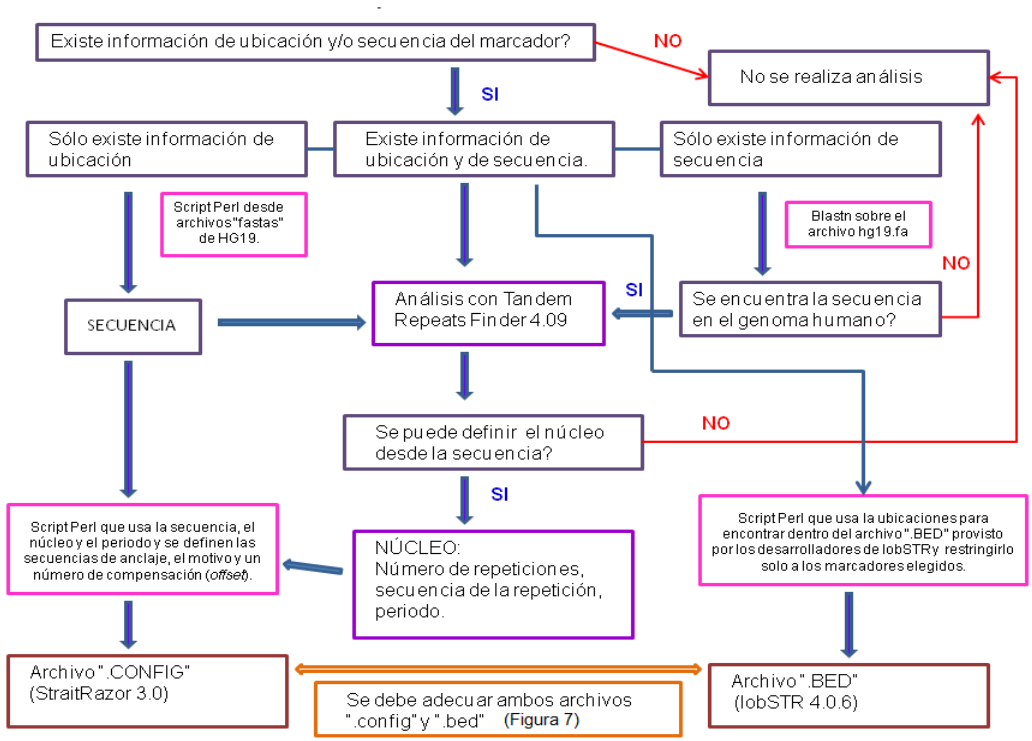
Fueron eliminados durante esta selección los siguientes marcadores de Ciencias Forenses: PentaC, DSX10134, DSX10101 y DSX10146 y los siguientes marcadores de Ciencias de la Salud: ZFYX, TAF9L, X22, DXYS218, SBMA y DXS2390, al no hallarse datos de secuencia o ubicación genómica.

Se incorporaron los marcadores: F13B, FESFPS, F13A01, F13A01, SE33, DXS10146, DXS10079, DXS10148, D18S499, LPL cuya información de ubicación era poco precisa o de ensamblajes anteriores al hg19, sin embargo se obtuvieron sus secuencias mediante sus números de acceso de GenBank, y se determinaron las ubicaciones en genoma hg19.

En el Anexo IV se encuentra el listado completo de marcadores definidos para este estudio.

En la Figura 6 se resumen las distintas instancias para la selección de marcadores.

Figura 6. Selección de marcadores para ser sometidos a los análisis de lobSTR y StraitRazor.



Fueron descartados de los análisis: PentaC, DSX10134 y DXS10101 usados en ciencias forenses y ZFYX, TAF9L, X22, DXYS218, SBMA y DXS2390 de uso en diagnóstico prenatal propuestos en el objetivo.

3.6 Asignación alélica de marcadores

Las herramientas lobSTR y StraitRazor brindan información sobre la detección y/o asignación alélica. Los parámetros a utilizar en los archivos de configuración de ambas herramientas se describen más abajo. La selección de estos parámetros para ambas herramientas requirió del conocimiento de la estructura del núcleo, secuencia de la repetición, periodo y número de repeticiones.

3.6.1 Confección del archivo "*.config" para StraitRazor v3.0

El archivo "*.config" debe contener:

a) nombre del marcador (*Marker*),
b) tipo de marcador (*Type: AUTOSOMAL, X, Y*)
c) dos secuencias de anclaje ó flanqueantes al núcleo del STR (*5'Anchor* y *3'Anchor*, ambas están en la misma hebra, con la orientación 5' -> 3'),
d) una secuencia motivo (*Motif*, que consta de un número de 2, 3 ó 4 repeticiones),
e) largo de la repetición ó periodo (*Period*),
f) un número de compensación (*Offset*) que el programa usa para el cálculo de asignación alélica.

Para generar los parámetros (“a” a la “f”) se usaron los valores obtenidos con TRF (3.4.1): índices, secuencia y largo de la repetición y junto con la secuencia del STR se obtuvieron los anclajes. El *script* de Perl también incorporó datos de SNPs frecuentes (SNP150), de manera tal que los anclajes no tengan estas variantes y minimizar los posibles *mismatchs* de los fragmentos de secuenciación. El cálculo del *offset* por parte del *script* tuvo en cuenta los índices originales de la salida de TRF, pero estos índices encierran un núcleo impuro en la mayoría de los casos, lo que implicó ajustar los índices para que se considere el núcleo puro (ver III.2.5 en el apartado *Archivos de texto plano ejecutables –script-* del ANEXO III)

3.6.2 Confección del archivo “*.bed” para lobSTR v4.0.6

El archivo “.bed” permite crear un índice utilizando el genoma de referencia (hg19). Para la confección de este archivo se deben completar 15 columnas, de las cuales 3 repiten información y 5 no son relevantes, pero no pueden estar vacías (se completaron con puntos).

- Columnas necesarias: 1- número de repeticiones para esa ubicación, 2- secuencia de la/las repeticiones y 3- puntaje: número de repeticiones x largo de la repetición x 2.
- Columnas que se repiten: 1- largo de la repetición; 2- ubicación cromosómica: comienzo y final del núcleo (si es impuro las interrupciones quedan incluidas en esta región); 3- número de cromosoma (dato único),

A partir del archivo de STRs de referencia provisto por los desarrolladores de esta herramienta, se filtraron con *scripts* de Perl a aquellos marcadores seleccionados listados en el Anexo IV. (ver punto III.2.6 en el apartado *Archivos de texto plano ejecutables (scripts)* del ANEXO III)

Se realizaron ajustes a los límites del núcleo, número de repeticiones y se agregaron nuevas líneas a aquellos STRs con más de una secuencia de repetición

3.6.3 Extracción de información de la salida de StraitRazor 3.0

El programa StraiRazor 3.0 devuelve una salida de texto plano con el siguiente contenido por línea:
MARCADOR: ALELO; XXX bases (largo en pares de bases de la secuencia hallada); SECUENCIA (con secuencias de anclaje, opción –i del programa); XXX; YYY (número de lecturas para la secuencia hallada, X: sentido e Y: anti-sentido)

Después de los dos puntos “:” (:ALELO) el programa designa alélicamente la secuencia según los parámetros pasados al archivo config (*offset, period*) considerando el largo completo de la secuencia y las secuencias de anclaje, según la siguiente fórmula:
(Largo secuencia completa – largo *anchor5* – largo *anchor3* – *offset*)/*period*

En caso de la aparición de decimales el programa los convierte a la notación según nomenclatura ISFG.

Esto simplifica todo el manejo posterior de la salida de este programa, dado que sólo requirió reunir en otro archivo las asignaciones alélicas según el marcador para cada fastq de cada individuo ensayado. Las lecturas según asignación alélica coincidente fueron sumadas (ambos XXX e YYY sumados también) para cada marcador. Dos *scripts* en Perl hicieron estas tareas, el primero sólo para análisis exploratorios y para hacer concordancias con lobSTR y el segundo unificando a las salidas de lobSTR para realizar los perfiles finales de los genomas analizados (ver “*Manejo de los archivos de salida de StraitRazor”* en el Anexo III).

Las secuencias obtenidas fueron extraídas con *scripts* en Perl también en dos instancias, una instancia de análisis de las secuencias de marcadores complejos y evaluación del patrón de repeticiones (corrección de parámetros del config) y una segunda para la extracción de todas las

secuencias cuando los perfiles finales fueron corregidos (posterior al contraste con lobSTR) (ver 3.7.4 más adelante).

3.6.4 Extracción de información de la salida de lobSTR 4.0.6

Por cada par de fastq analizado *e*ste programa genera dos archivos de salida: "*.aligned.stats" y "*.aligned.bam". Este último es procesado con el programa Samtools y se generan los archivos "*.sorted.bam", "*.sorted.bam.bai" y "*.vcf" con los comandos –sort, -index y –allelotype respectivamente.

Desde el archivo "*.vcf" y con *scripts* en Perl se extrajo sólo la información necesaria para la asignación de cada alelo detectado. Los campos relevantes dentro del archivo "*.vcf" son los siguientes: CHROM, POS, REF, ALT, INFO (sólo REF y RPA.) y FORMAT (sólo GT y DP). Para más detalle de las especificaciones de anotación VCF y *scripts* vea III.2.8 en el Anexo III.

De igual manera que con StraitRazor, hubieron versiones iniciales de los *scripts* a los fines de contrastar ambas herramientas y corrección de inespecificidades, y una versión definitiva que reunió las salidas en un solo archivo, de texto plano, con las asignaciones alélicas de ambas herramientas y sus respectivas lecturas que la soportaron. Este archivo será usado después en una planilla de cálculo habilitada para macros "*.xlsm" diseñada para los contrastes automáticos entre StraitRazor y LobSTR.

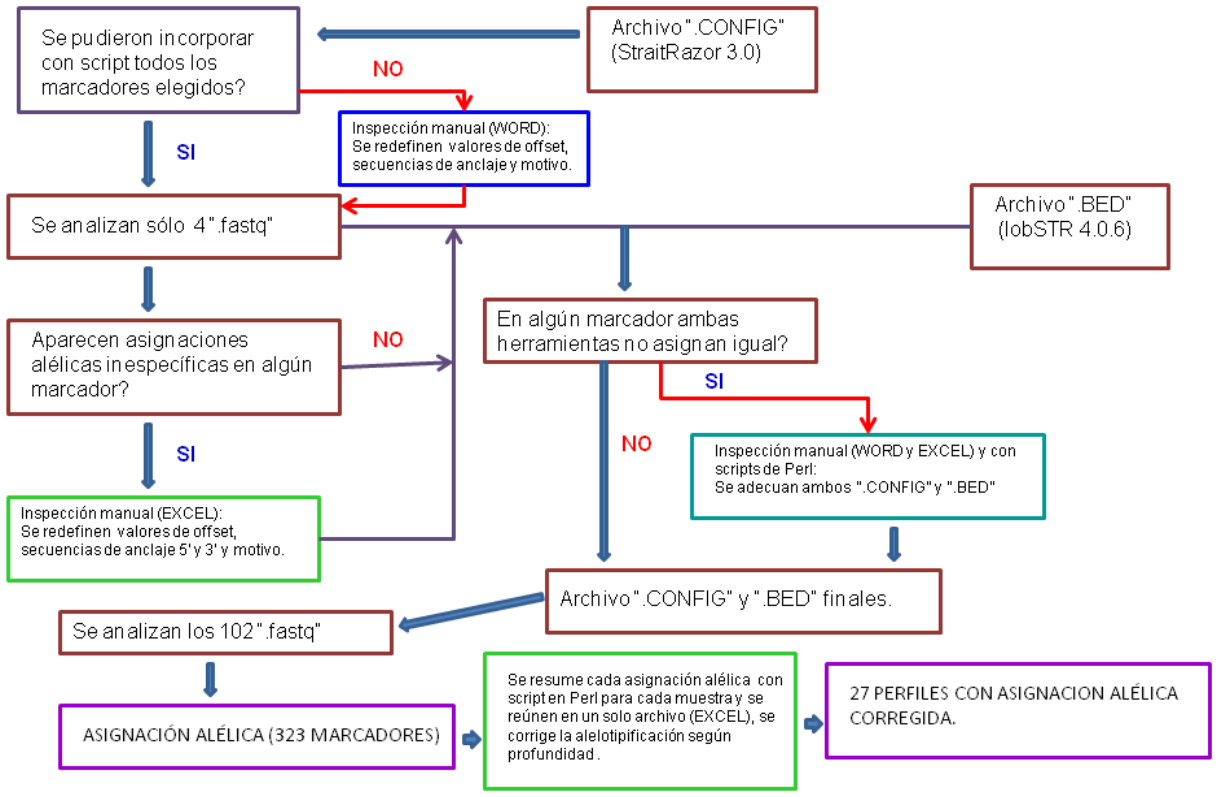
3.6.5 Ajustes de asignación alélica entre ambas herramientas StraitRazor y lobSTR.

Luego de que se establecieran todos los marcadores para el estudio de los 27 genomas, y que se conformaron las primeras versiones de los archivos de configuración "*.config" y "*.bed", fue necesario recorrer la información de ambas para constatar si se obtenía el mismo número de repeticiones para el alelo de referencia de un mismo marcador. Un total de 190 marcadores no coincidían, por lo tanto se hizo una inspección más minuciosa en una plantilla en Word y/ó con *script* en Perl, dependiendo de la causa de la inconsistencia. Ver el listado completo de marcadores inspeccionados de esta manera en el Anexo V.

Se repitió la inspección en algunos casos, con ayuda de las secuencias que se obtenían de unos pocos "fastqs" con StraitRazor (*script* Perl, v17 y v18 del config), para aquellos STRs de estructura compleja, donde fue difícil discernir interrupciones, repeticiones principales y secundarias, repeticiones incompletas, etc.

En la siguiente figura se esquematiza todos los pasos previos a los análisis definitivos con ambas herramientas para el total de los 102 fastqs descargados.

Figura 7. Confección de los archivos “.CONFIG” y “.BED”, corrección de inespecificidades e inconsistencias de asignación alélica entre lobSTR y StraitRazor.



Cada corrección en los archivos “*.config” ó “*.bed” fueron hechos en las planillas de cálculo EXCEL donde estaban alojadas las distintas versiones de estos archivos.
(ver *items* III.2.5 y III.2.6 en el apartado *Archivos de texto plano ejecutables –scripts-* del ANEXO III)

3.6.6 Análisis de los 27 genomas descargados y recopilación de los datos

Un total de 102 archivos “*.fastq” pertenecientes a 27 muestras fueron sometidos a análisis: 3 muestras (6 fastq) pertenecen a un trío (familia 1463) del Centro de Estudios de Polimorfismos Humanos (CEPH, *The Centre d'Etude du Polymorphisme Humain*) y las 24 muestras (96 fastq) restantes corresponden a individuos del Proyecto 1.000 Genomas.

Las muestras son del tipo *paired-end,* o dicho de otra manera una muestra posee cuatro archivos de secuenciación fastq (ver Tabla 4). Esto le otorga a cada marcador por muestra una buena calidad, siempre que el marcador tenga buena cobertura vertical (muchas lecturas para esa región). Esta característica puede no ocurrir en la totalidad de los casos, se discute con más detalle en el apartado 4: “Resultados de utilización y curado de herramientas bioinformáticas”. De las 27 muestras seleccionadas con Alta Cobertura, 16 son femeninas y 11 masculinas.

Tabla 4. Dos ejemplos de disponibilidad de archivos de secuenciación NGS y su uso con lobSTR y StraitRazor.

	Herramientas para la detección de lecturas que contienen STRs	
	StraitRazor	lobSTR
Salida de la herramienta:	txt	aligned.bam (aligned.stats)
Tipo de archivo sometido a analisis:	fastq	fastq
Muestra (género, población)	Análisis ID	Análisis ID
HG00096 (male, GBR*)	SRR1291026_2 SRR1291026_1 SRR1291035_2 SRR1291035_1	SRR1291026 (_1 y _2 juntas) SRR1291035 (_1 y _2 juntas)
NA12878 (female, CEPH*)	SRR622457_1 SRR622457_2	SRR622457 (_1 y _2 juntas)

*GBR: Gran Bretaña y CEU: The Centre d'Etude du Polymorphisme Humain.

El listado completo de las 27 muestras, su respectivo género, población de origen, sus códigos de análisis (análisis ID), rutas para las descargas, datos de secuenciación, etc., se detallan en el archivo “1000genomes.sequence.index” (punto I.1).

Como se indicó en el punto 3.6.4, las salidas de StraitRazor 3.0 (Versión del archivo “*.config”: v183) y de LobSTR (Versión del archivo “*.bed”: v32) fueron reunidas en un único archivo de texto, y este “copiado y pegado” a la primera hoja de una plantilla de Excel, y en la segunda hoja se establecen las asignaciones alélicas (de manera automática) acorde a una serie de reglas condicionales, para unificar lo que detectó cada herramienta por separado.

Estas reglas se discuten con más detalles en la sección “4.8 Asignaciones alélicas” dentro de “4. Resultados de utilización y curado de herramientas bioinformáticas”.

Antes de concluir en la asignación alélica final, en las columnas previas se corrige lo siguiente:

a) Decimales de lobSTR a la notación según nomenclatura de la ISFG.
b) Los valores nulos fueron cambiados a “ND”: No Detectado.
c) Se eliminan espacios que no permiten la comparación.

Además, se establecen cálculos de conteos de lecturas de los dos alelos más detectados y sus respectivas relaciones y se destacan las celdas donde la asignación automática no fue resuelta, entonces el analista interviene manualmente (también se destaca la intervención manual).

Todas las asignaciones alélicas de cada muestra (archivo “PERFIL_v2_(muestra)_lobSTR_v32_str8rzr_v183.xlsm” fueron luego reunidas en el archivo de Excel “perfiles-finales.xlsm”, generando así una plantilla de resumen global de este estudio. También fue generado el archivo “visualizaciones-profundidad-lobSTR-StraitRazor.xlsm” donde fueron reunidos los conteos de lecturas de las herramientas (en hojas separadas).

Estos archivos son la base para la generación de gráficos para mostrar los resultados de la segunda parte. Las asignaciones alélicas finales (ambas herramientas contrastadas e inspeccionadas) fueron usadas para filtrar las secuencias válidas obtenidas con StraitRazor.

Para más detalles de las sentencias en Excel y *scripts* vea el punto III.2.9 dentro de "Archivos de texto plano ejecutables (scripts)" en el Anexo III.

3.6.7 Evaluación de las asignaciones alélicas finales

Se estableció una nueva versión de los archivos "PERFIL_v2_(muestra)_lobSTR_v32_str8rzr_v183.xlsm" del punto anterior a "PERFIL_v3_(muestra)_lobSTR_v32_str8rzr_v183.xlsm" que fueron reunidas en el archivo de Excel "perfiles-finales-v2.xlsx" con la finalidad de evaluar de manera dinámica las reglas condicionales aplicadas en las asignaciones alélicas.

En el archivo "perfiles-finales-v2.xlsx" se obtienen las asignaciones alélicas de cada muestra y se permite variar el número de conteos de lecturas mínimo (establecido en 1 en el archivo "perfiles-finales.xlsm"), además se permite variar las relaciones de conteos de lecturas entre el alelo mayoritario y el siguiente en número de lecturas (establecido en 2.5 para di y trincleotidos y 3 para tetra, penta y hexanucleótidos), esto evalúa tartamudeo versus segundo alelo verdadero.

Se evaluó un mínimo de 1, 5 y 10 de conteos de lecturas mínimas y un rango de 2 a 5 (2, 2.5, 3, 3.5, 4, 4.5 y 5) de relaciones de conteos de lecturas.

Se establecieron 12 niveles de concordancias separados en tres grupos:

a) Las herramientas coinciden completamente.

1: las herramientas coinciden y ambas superan las 10 lecturas (lobSTR y StraitRazor => 10)

2: las herramientas coinciden y sólo una supera las 10 lecturas (lobSTR o StraitRazor => 10 y la otra > 1 y < 10)

3: las herramientas coinciden y ambas NO superan las 10 lecturas (lobSTR y StraitRazor > 1 y < 10)

12: las herramientas coinciden y ambas NO tienen detecciones (lobSTR y StraitRazor = 0) NDs ó No Detectados

b) Las herramientas coinciden parcialmente.

4: las herramientas comparten un alelo, y ambas superan las 10 lecturas (lobSTR y StraitRazor => 10)

5: las herramientas comparten un alelo, y sólo una supera las 10 lecturas (lobSTR o StraitRazor => 10 y la otra > 1 y < 10)

6: las herramientas comparten un alelo, y ambas NO superan las 10 lecturas (lobSTR y StraitRazor > 1 y < 10)

c) Las herramientas no coinciden.

7: las herramientas no coinciden y no comparten alelos y ambas superan las 10 lecturas (lobSTR y StraitRazor => 10)

8: ambas herramientas no coinciden y no comparten alelos y sólo una supera las 10 lecturas (lobSTR o StraitRazor => 10 y la otra > 1 y < 10)

9: ambas herramientas no coinciden y no comparten alelos y sólo una supera las 10 lecturas y la otra no tiene detecciones (lobSTR o StraitRazor => 10 y la otra = 0)

10: ambas herramientas no coinciden y no comparten alelos y ambas NO superan las 10 lecturas (lobSTR y StraitRazor > 1 y < 10)

11: ambas herramientas no coinciden y no comparten alelos y ambas NO superan las 10 lecturas y la otra no tiene detecciones (lobSTR o StraitRazor < 10 y la otra = 0)

3.6.7.1 Puntajes de las asignaciones alélicas finales para marcadores y muestras

Para establecer un puntaje se decidió sumar el valor de la regla aplicada en determinada asignación alélica (1 a 12) a lo largo de las 27 muestras para cada marcador, o lo largo de los 323 para cada muestra. Esto es posible porque desde la regla 1 (asignación alélica óptima) hasta la 12 (peor escenario) se establece una pérdida gradual de performance de la asignación alélica entre lobSTR y StraitRazor.

Es decir, por ejemplo, un marcador con puntaje de 27, significa que las 27 asignaciones alélicas (27 muestras) obtuvieron el resultado óptimo, porque se aplicó la regla número 1. Para marcadores autosómicos, pseudo-autosómicos y exclusivos del cromosoma X (estos últimos sólo en mujeres): valores mayores a 27 significan pérdida de performance entre ambas herramientas y para marcadores exclusivos del cromosoma X e Y (sólo varones) valores mayores a 11 significan pérdida de performance entre ambas herramientas.

3.6.7.2 Archivos derivados de “perfiles-finales-v2.xlsx”.

Del archivo “perfiles-finales-v2.xlsx” se derivan los archivos: “ERRORES_lobSTR_STRAITRAZOR.xlsx”, “combinaciones-de-cortes.xlsx”, “PUNTAJES.xlsx” y “TRIO_CEPH.xlsm”. El contenido de estos archivos soporta lo que se discute en el apartado 4.9.

3.7 Generación de gráficos y otros archivos

3.7.1 Archivos BED (para la visualización de datos de la BD local y marcadores en estudio de la segunda parte) con Integrative Genomics Viewer IGV

La conformación de los archivos “*.BED” para las visualizaciones con esta herrmienta difiere un poco de lo expuesto en el ítem 3.6.2. En esta ocasión el total de columnas son 12, y 3 de ellas son mandatorias ó fijas: *chrom, chromStart* y *chromEnd* y las 9 restantes son opcionales: *name, score, strand, thickStart, thickEnd, itemRgb, blockCount, blockSizes* y *blockStarts*. Y dependiendo del dato genómico para mostrar algunas fueron completadas y otras no, aunque el campo de *score* estuvo vacío en todas las pistas, en cambio los campos *name* (dato principal de la pista)*, itemRgb* (color del datos, ej.: “0,255,0”=”verde”), *thickStart (*idem *chromStart)* y *thickEnd (*idem *chromEnd)* siempre fueron completados. Para más detalles del formato de este archivo ir a http://genome.ucsc.edu/FAQ/FAQformat#format1.

Para la creación de las distintas pistas se partió de los archivos que fueron creados para la BD local, seleccionando y reordenando las columnas con el comando awk y luego *sed*, o a partir planillas Excel previos que ya alojaban estos datos.

Las pistas creadas son: Citobandas (1-citobandas.bed), Segmentos Múltiples (2-segmentos_multiples.bed), Genes (3-intron-exon-gene.bed), STSs de UCSC (6-ucsc2.bed), STSs de Ensembl (7-ensembl2.bed), Marcadores en estudio (5-321markers.bed), Datos de Fase 1 (4a-fase1_cleaned.bed), STRs de referencia (4b-str_reference_9cols.bed), Analisis con TRF

sobre “hg19.fa” (4c-trf-nt10.bed), SNPs versión: 151 (snp151Common_9cols_nt_black_cleaned_chr-1..22 X e Y-).

IGV también admite archivos “*.bam”, por lo tanto todos los archivos “*.sorted.bam” y “*.sorted.bam.bai” (salidas de Samtools) son útiles para las visualizaciones.

3.7.2 Tablas y Figuras

A continuación se lista los archivos que soportan lo expuesto en Tablas y Figuras, y en estos archivos están los datos y/o gráficos:

TABLAS:

Tabla 4: “TABLA-4.xlsx”

Tabla 5: “STRs-dentro-de-STSs-UCSC-Ensembl.xlsx”

Tabla 6: “match segun primers y ubicacion.xlsx”

Tabla 7 y 8: Hoja “SUMAS” en “sumas-introns-exones.xlsx”

Tabla 9: Hoja “cromosomaY-en-mujeres” en “visualizaciones-profundidad-lobSTR-StraitRazor.xlsm”.

Tabla 10: “TABLA_SEGMENTOS_MULTIPLES.docx”

Tabla 11: “compara-final-str8rzr-lobstr.docx”

FIGURAS:

La Figura 3 son dos capturas de pantalla del programa GeneMapper versión 3.2 (Applied Biosystems).

Figura 6: “seleccion-de-marcadores.pptx”

Figura 7: “configuracion-de-herramientas.pptx”

Figura 8: “pipeline.pptx”

Figura 9 a y b: hojas “0pb-graph” y “250pb-graph” en “cebadores-versus-seqs.xlsx”“

Figura 10.: “interseccion-superdup.xlsx”

La Figura 11 son capturas de los archivo “perfiles-finales.xlsm”, “visualizaciones-profundidad-lobSTR-StraitRazor.xlsm” y de archivos generados en el punto 3.7.4.

Figura 12a y 13a: Gráfico en la hoja “varones” del archivo “FIGURA-11a.xlsx”.

Figura 12b y 13b: Gráfico en la hoja “mujeres” del archivo “FIGURA-11b.xlsx”.

Figura 14a y b: Sendos graficos en el archivo “FIGURA-13.xlsx”

Figura 15: Gráfico en la hoja “nd-nc-detected” del archivo “FIGURAS-14y15.xlsx”.

Figura 16: Gráfico en la hoja “automatico-manual” del archivo “FIGURAS-14y15.xlsx”.

Figuras 17, 18 y 19: Capturas de pantalla editadas del programa BioEdit.

Figura 20a: “asignacion-alelica-ambas-herramientas-DYS393-DXYS267.xlsx”.

Figuras 20b y c: Hojas StraitRazor y lobSTR respectivamente del archivo “visualizaciones-profundidad-lobSTR-StraitRazor.xlsm”, filtrado con los marcadores: DXYS267 y DYS393.

Figura 21: Hoja “DYS393-DXYS267-str8rzr” del archivo “DYS393-DXYS267-str8rzr.xlsx”.

Figura 22a y b: Hojas “lobSTR” y “StraitRazor” respectivamente del archivo “asignacion-alelica-profundidad-CromX-ambas-herramientas-11varones.xlsx”

Figura 23: Hoja “lobSTR” del archivo “asignacion-alelica-profundidad-CromX-ambas-herramientas-11varones.xlsx”

Figuras 24a y b: “perfiles-finales-v2.xlsx”, los datos provienen de la hoja “GRAFICOS”.

Figura 25a y b: Hojas “combinaciones-de-cortes” y “idem_idem” respectivamente del archivo “combinaciones-de-cortes.xlsx”.

Figura 26: Archivos “D7S3057.htm” y “NBPF9_I_21.htm”.

Figura 27: “TRIO_CEPH.xlsm”.

Las Figuras 28 a 33 son capturas de pantalla del programa IGV.

Las Figuras 34a a 34d son capturas de pantalla del navegador Firefox que accede a la BD local.

3.7.3 Cálculos estadísticos para los datos de las Figuras 9, 12, 13, 14 y 15

De los datos obtenidos para la Figura 9 se hicieron comparaciones entre las proporciones de los cebadores contenidos en secuencias STS originales +/- 0 pb y 250 +/- pb. Más detalles en el punto 3.4.4.

Para los datos que se muestran en las Figuras 12, 13, 14 y 15 (sólo para los “no detectados”) se realizaron Test. T para comparar los valores medios de “no detecciones” de una herramienta para el grupo de 24 muestras versus las del trío CEPH, y se controló que hubiera homogeneidad de varianzas entre ambos grupos. Ver III.2.4 dentro de “Archivos de texto plano ejecutables –scripts-” en el Anexo III.

3.7.4 Histogramas de frecuencias absolutas y resúmenes

El script de Perl “*resumen-final-v2.pl*” recoge información del archivo “perfiles-finales-depured-v1.txt” (originado a partir de “perfiles-finales.xlsm”) y genera los archivos “report-(marcador).txt” y “plot-(marcador).png” que muestran un resumen y un histograma de frecuencias alélicas absolutas por marcador, respectivamente.

Los archivos de imagen “*.png” fueron incorporados a la rutas para ser visualizados por el navegador (Firefox) en archivos “*.php” (punto 3.7.5).

3.7.5 Secuencias multifasta obtenidas con StraitRazor 3.0

El *script* de Perl “*get-seq-from-results-cv183-by-Sample-by-Marker-v3.pl*” dentro del *script* “*doble-loop-for-get-seq-pl-v2.sh*”, extrajo todas la secuencias contenidas en los archivos “STRs_en_estudio_v183-(acrhivo fastq)_i.txt” (ver *Manejo de los archivos de salida de StraitRazor* en el Anexo III) después de los análisis con esta herramienta. En esta ocasión las secuencias fueron “filtradas” según asignación alélica contrastada entre lobSTR y StraitRazor (archivo “perfiles-finales-depured-v1.txt”).

Se generan así archivos “*.fa”, multifasta para cada muestra según marcador y a su vez según asignación alélica. Cada uno de estos multifasta son convertidos a un archivo simple fasta consenso con los *scripts*: “*consensus-alignio-bySample.pl*” dentro de “*for-consensus-pl.sh*”, se logra así una única secuencia para cada alelo, según marcador por muestra.

Se reúnen aquellas secuencias para un mismo marcador con el comando cat de Linux (*script: for-cat.sh*), se obtienen así todos los alelos consenso de cada marcador, luego ese multifasta será corregido (titulo de la secuencia), re-ordenado por número de alelo (comandos awk y sort en el *script sort-fasta.sh*) y finalmente alineado con el programa MAFFT (*script: mafft-v1.sh*).

Las secuencias multifasta fueron convertidas a archivos “*.html” para poder ser visualizadas en el navegador.

3.7.6 Generación de páginas php y html para mostrar resultados finales

Desde el sitio http://arrobasistemas.com/humstrs2/index.html se puede acceder a la interfaz web para navegar los datos. Los datos más optimizados para la navegación son todos los relacionados con los resultados del estudio de los 323 marcadores.

Se pueden hacer consultas mínimas y de datos de secuenciación de Fase 1 (lobSTR 2.0.4) a las tablas de STS de Ensembl, de USCS y de genes.

Se generaron archivos "*.php" para cada marcador (*script: print-php-strs.pl*) que consulta la base (tablas 319markers y snp150common), alojados en http://arrobasistemas.com/humstrs2/index.html y a su vez esta ubicación contiene las imágenes de histogramas de frecuencias propias (punto 3.7.3) y del reactivo *PowerPlex Fusion* de Promega (/png y /png1), los alineamientos de secuencias del punto 3.7.4 (/fastas), páginas web con la ubicación en el ideograma del cromosoma correspondiente para cada marcador (/loc) y las páginas php que consultan la tabla snp150common (/snp).

Los scripts usados para generar estos archivos fueron: *print-php-strs.pl* que muestra información relevante del marcador, *print-php-snps.pl* para las consultas de SNPs, *print-html-strs.pl* para las ubicaciones del marcador y *print-fastas-html-strs.pl* para poder mostrar los alineamientos en html.

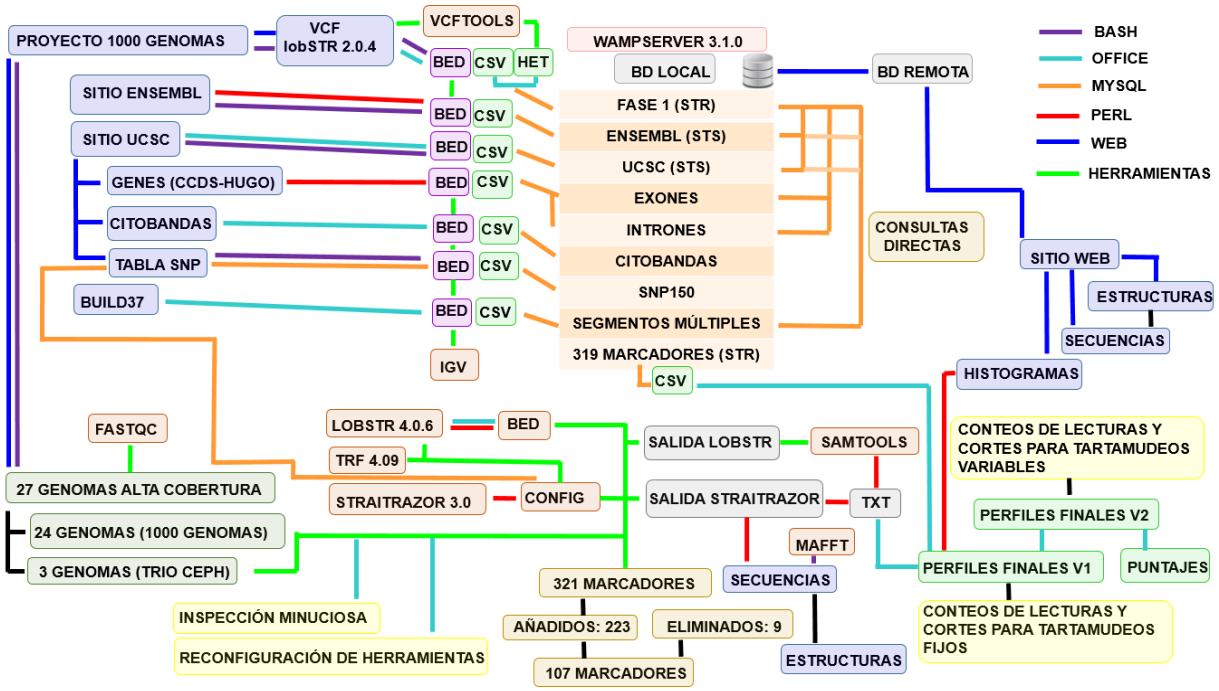
Con la intención de hacer una comparación con frecuencias relativas de aquellos STRs de uso en ciencias forenses, se realizaron histogramas de frecuencias relativas con los *scripts* "*make-script-for-r-strs-freqs-pplexf.pl"* que genera el archivo "*script-for-r-strs-freqs-pplexf.R*" y usa la tabla "frecuencias-strs-forenses.csv" para crear las imágenes "*.png", alojados en /png1.

Los valores de frecuencias alélicas relativas provienen de información suministrada por Promega para su reactivo *PowerPlex Fusion* para muestras de población de EE. UU. de individuos no relacionados (n = 1.036) en 29 loci de autosomas STR.

3.7.7 Pipeline

La siguiente figura resume a modo de *pipeline* el flujo de información descripto en este trabajo.

Figura 8. Pipeline: Flujo de información desde los recursos hasta la construcción de la base de datos y el sitio web.



4. RESULTADOS DE UTILIZACIÓN Y CURADO DE HERRAMIENTAS BIOINFORMÁTICAS

La anotación hecha con lobSTR 2.0.4 para la secuenciación del Proyecto 1.000 Genomas fue llevada a la BD local (ver 3.1.1) y cruzada con los datos de marcadores (STSs) de los sitios UCSC y Ensembl. En la Tabla 5 se destaca lo relevante.

4.1 Cantidad y análisis genómico de STRs provenientes de Ensembl y UCSC

Tabla 5. Cantidad de STRs secuenciados contenidos en los marcadores STSs de UCSC y Ensembl.

Total de variantes STRs anotadas y contenidas en FASE1: 668.631				
BD pública	Número de STSs	TOTAL STRs		
		STRs totalmente incluidos en los STSs	STSs totalmente incluidos en los STRs	Solapamientos
UCSC	299.410	35.712 (11,9 %)	1 (0,0003 %)	2.649 (0,9 %)
		38.362 (12,8 %)		
Ensembl	288.247	29.365 (10,2 %)	2 (0,0007 %)	2.370 (0,8 %)
		31.737 (11,0 %)		

Los marcadores STSs de Ensembl y UCSC (considerando todas las posibilidades, inclusiones y solapamientos) están asociados a 11 y 12,8 % de variantes STRs, respectivamente. Asimismo, si se considera sólo la inclusión de variantes STR dentro de STSs, los marcadores están asociados a un 10,2 % y 11,9 %, respectivamente. Sólo en tres instancias quedaron marcadores incluidos dentro de regiones STRs: REN113313 (UCSC y Ensembl) dentro de una región de repeticiones hexa-nucleótido AGGGGG, y BV728075 (Ensembl) dentro de una región de repeticiones dinucleótido GT.

Por otra parte, el análisis genómico con TRF en secuencias STSs dio como resultado que 7,53 % (22.546 de un total de 299.410) y 6,21 % (17.902 de 288.178) sólo corresponde a variantes STRs en UCSC y Ensembl, respectivamente.

Es un resultado menor a lo observado con las consultas a la BD, TRF es deficiente en encontrar algunos STRs dinucleótidos.

Resulta evidente entonces que no más del 12,8 % de los STRs están incluidos en regiones STSs. Esta situación se venía observando previo a la construcción de la base de datos, con los datos descargados de UCSC y de Fase 1, manejados en archivos de Excel y se hicieron más evidentes ante las consultas a la base de datos (Tabla 5), y esto impactará no sólo en la nomenclatura usada para la segunda parte de este trabajo, sino también en la verdadera utilidad de los STSs para la construcción de una base de datos más completa para STRs.

4.2 Coincidencias entre cebadores y secuencias STRs

Posteriormente, se analizaron las coincidencias estrictas entre el conjunto de cebadores de las bases de datos de UCSC, Ensembl y UniSTS. Paralelamente se analizaron las coincidencias estrictas entre cada par de cebador con su secuencia de STS asociado. En la Tabla 6 y la Figura 9 se resume estos dos tipos de búsquedas realizadas.

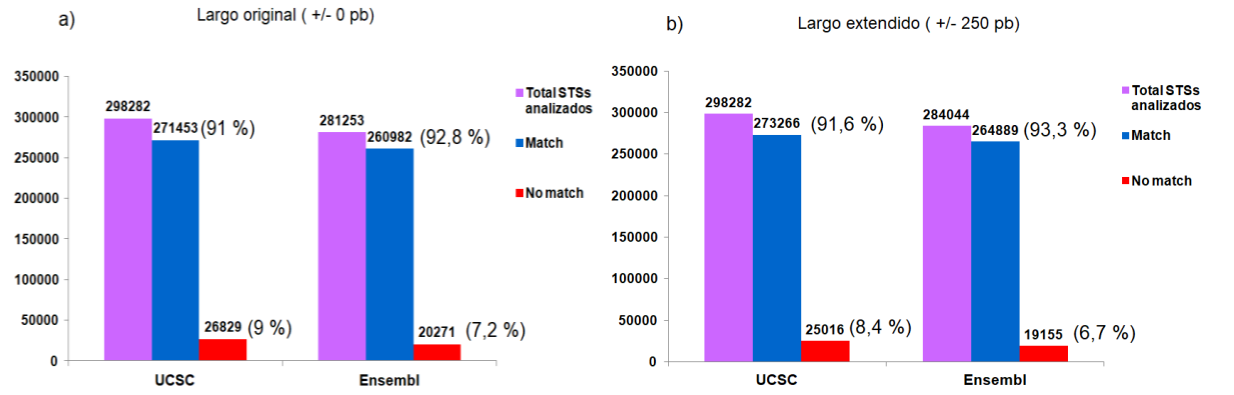
Tabla 6. Coincidencias recíprocas de las secuencias de cebadores de UniSTS, UCSC y Ensembl.

Base de datos		UCSC	Ensembl	UniSTS
Base de datos	Número de marcadores STSs	322.212	298.993	321.290
UCSC	322.212		296.318 (99,1 %)	310.131 (96,5 %)
Ensembl	298.993	296.318 (92 %)		293.993 (91,5 %)
UniSTS	321.290	310.131 (96,3 %)	293.993 (98,3 %)	
Para cada STS existe un par de cebadores asociados.				

Este análisis se hizo con el propósito de corroborar el hecho de que tanto UCSC como Ensembl incorporaron a sus bases de datos las secuencias de los pares de cebadores alojados en UniSTS y la Tabla 6 muestra el grado de coincidencias entre los pares de cada base respecto de las otras, mayores al 90 %.

Desde UCSC se declara que los pares de cebadores fueron usados para localizar en el genoma humano las regiones que amplifican (herramienta ePCR: https://www.ncbi.nlm.nih.gov/tools/epcr/) pero esos mismos cebadores no parecen estar actualizados según el ensamblaje que se investigue en el mismo portal. Esto se discute a continuación.

Figura 9. Búsquedas de pares de cebadores de UCSC y Ensembl dentro de secuencias propias de la base.



Condición “MATCH”: Todas las combinaciones mencionadas en el punto 3.4.4 (ítems a, b, c y d).
Condición “NO MATCH”: Se consideró como “*No match*” cuando el par de cebadores ó al menos uno, no fueron encontrados en la secuencia STS asociada.

Se observa en la Figura 9 que hay un porcentaje elevado (> 5%) de secuencias de cebadores (el par ó al menos uno) que no coinciden con la secuencia STS blanco (*No match*).

Además, se observaron diferencias significativas entre búsquedas de secuencias con sus longitudes originales (0+/- pb) y aquellas ampliadas en +/- 250 pb (*P* <.001) lo que indicaría que con la ampliación +/- 250 pb se encontraron más ubicaciones de cebadores que flanquean la región STS.

Otra situación encontrada en estas búsquedas tanto en UCSC como en Ensembl, dentro de los resultados considerados de "MATCH" hay aproximadamente un 18 a 18,5 % de coincidencias que responden a la situación "b" del punto 3.4.4.

Es decir, los cebadores están acordes a una secuencia STSs reversa complementaria.

Finalmente, las coincidencias entre pares de cebadores orientados de la manera esperada (situación "a" del punto 3.4.4) dan cuenta de un 72,5 y 74,8 % del total de los archivos "fasta" analizados de UCSC y Ensembl, respectivamente.

Cuando uno o ambos cebadores no coinciden con su secuencia STS asociada, hay uno o más SNPs no contemplados en la secuencia del cebador ó de la secuencia STS, ó están por fuera de la región STS analizada.

Resulta necesario entonces la generación de nuevas secuencias de cebadores para las regiones STSs que incluyen STRs, según el ensamblaje que corresponda, y existen muchas herramientas bioinformáticas para realizarlo (primer3 por ejemplo), pero como se discutirá en la segunda parte de este trabajo, no existen herramientas destinadas a encontrar secuencias flanqueantes a los fines bioinformáticos.

Por todo esto, se planteó incorporar otros eventos genómicos como SNPs, genes y segmentos múltiples a la BD local.

4.3 Localización de variantes STR en regiones génicas e intergénicas

Se analizó la cantidad de variantes STRs de Fase 1 que están incluidos en regiones génicas (exónicas, intrónicas) y extragénicas. Los resultados se resumen en la Tabla 7.

Tabla 7. Número de variantes STRs contenidos en regiones génicas y extragénicas.

Total de variantes de Fase 1	Regiones génicas		Regiones extragénicas
	Exones	Intrones	
668.631	2.237 (0,6 %)	373.017 (99,4 %)	293.377
	375.274 (56,12 %)		43,88 %

El 56,12% de las STRs se encuentran en regiones génicas, y de ellas el 99,4% corresponde a regiones intrónicas. Estos valores podrían cambiar si se considera el corte de MAF > 1 %, dado que el total de variantes de Fase 1 desciende a 320.479, ver punto 3.4.5.

Para este número, la distribución entre exones, intrones y regiones intergénicas no fueron obtenidas dado que las frecuencias alélicas no fueron incorporadas a la BD local.

Tabla 8. Distancias acumuladas de variantes STRs de Fase 1, exones e intrones obtenidas desde la base de datos local.

		Fase 1	Exón	Intrón	Exón+Intrón
	Mpb*	17,33	49,53	1.603,78	1.653,30
HG19	3.137,16	0,6 %	1,6 %	51,1 %	52,7 %
Exón+Intrón:	1.653,30	1,0 %	3,0 %	97,0 %	

*Mpb: tamaño en mega pares de bases.

Aproximadamente la mitad de las variantes STRs están comprendidos en regiones intrónicas, lo que era esperable debido que el 51,1 % del genoma humano corresponde a regiones intrónicas, mientras que sólo un 3 % a exónicas, y confirma el hecho de que los STRs están distribuidos de manera homogénea.

Esta condición fue la causa de usar la nomenclatura de regiones génicas para asociarlos a STRs que no están incluidos a STSs y que tampoco poseen ningún nombre heredado (ver punto 3.5.2).

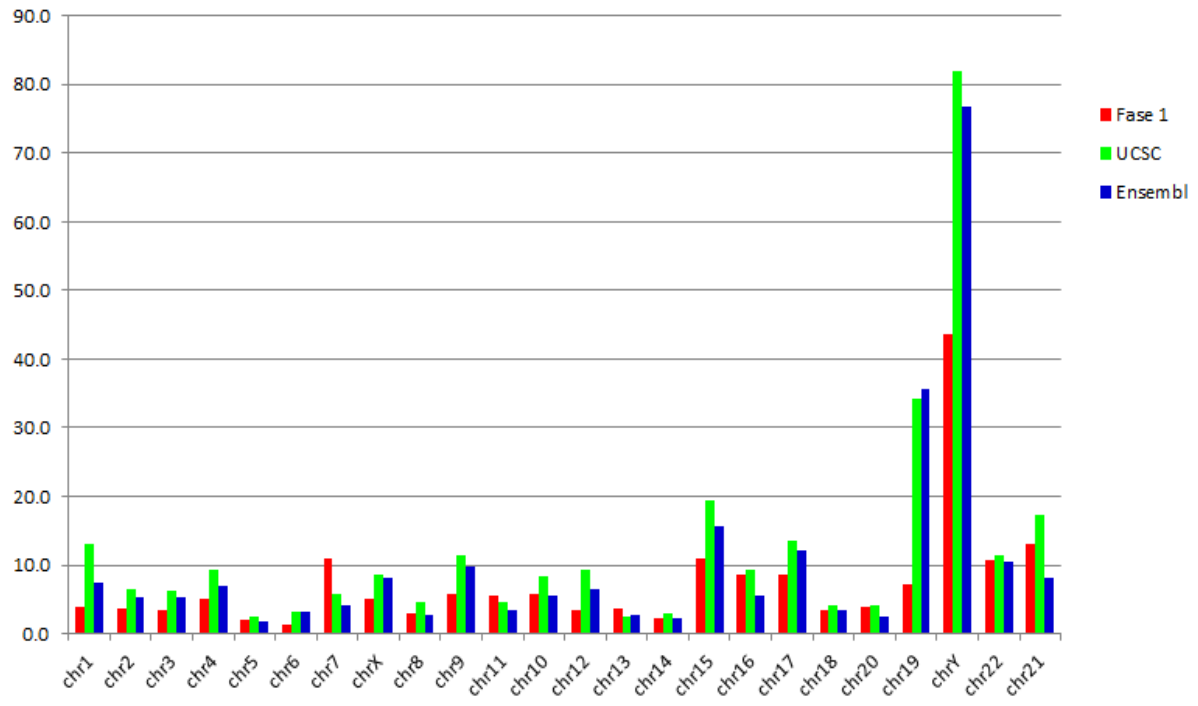
4.4 Localización de variantes STR en segmentos múltiples

Los segmentos múltiples son regiones del genoma que poseen una elevada identidad (> 90 %) con otras porciones del mismo, poseen tamaños desde 1.000 pb hasta casi 800.000 pb, siendo estos eventos intra e inter-cromosómicos, y provienen de duplicaciones, triplicaciones, etc. Los marcadores STSs y las variantes STRs pueden estar comprendidos dentro de estos segmentos múltiples.

Se analizó la presencia de variantes STRs anotados en Fase 1, y de STSs de UCSC y de Ensembl, que podrían estar en los segmentos múltiples.

La Figura 10 muestra una cuantificación de STRs expresados en valores porcentuales para Fase 1 ó STSs (UCSC ó Ensembl).

Figura 10. Porcentajes de STRs (Fase 1) y STSs (UCSC y Ensembl) incluidos en segmentos múltiples del Genoma Humano.



Los 24 cromosomas humanos están ordenados sobre la abscisa según sus tamaños de manera decreciente.

Se observó que a medida que la longitud del cromosoma se acorta, se incrementa el número de variantes STRs ó secuencias STSs incluidas en segmentos múltiples. Sin embargo, el cromosoma Y muestra un porcentaje muy elevado (> 50%), tanto de STRs como de STSs incluidos en segmentos múltiples. No se observó lo mismo para los cromosomas 18 y 20, cercanos en tamaño.

El cromosoma 19 posee un porcentaje también elevado, esta vez sólo de secuencias STSs (UCSC y Ensembl) incluidas en estos segmentos.

Por último, en menor proporción, están los cromosomas 15, 17, 9 y 22 (tanto STRs como STSs) y los cromosomas 21 y 7 (sólo STRs).

Dentro del análisis de STRs con las herramientas propuestas lobSTR y StraitRazor para la segunda parte de este trabajo (secciones 4.7 en adelante) se discutirá el impacto de contemplar aquellos STRs incluidos en segmentos múltiples, no sólo por cuestiones de especificidad de las herramientas sino también, habiendo establecido adecuadamente las diferencias entre segmentos, como marcadores útiles que diferencien dos cromosomas por ejemplo.

4.5 Una nomenclatura completa pendiente

Resulta evidente que la totalidad de marcadores incorporados en bases de datos públicas como Ensembl, Probe (anteriormente UniSTS) o del sitio Genome Browser Gateway de la Universidad de California (Santa Cruz) son insuficientes para asociar sus datos con los datos de secuenciación de nueva generación de proyectos de grandes consorcios como el de *1.000 Genomes Proyect*, que con la herramienta LobSTR creada por Gymrek et al., se generó una recopilación ambiciosa de variantes microsatélites (700.000) usando sólo los genomas secuenciados de la fase 1 de este proyecto (Williams et al).

Asociar el identificador #rs del catálogo de variantes SNVs (SNPs) que en la versión actual (SNP151, 09 Diciembre 2018) contiene 14.831.956 variantes es una alternativa usada por algunos autores ante la falta de nombres o alias para muchos microsatélites, y hacer esto responde al hecho de que es casi seguro que habrá una variante SNP dentro o en una íntima cercanía (10 nt). Esto no sólo es incorrecto debido a que son variantes que se originan por eventos biológicos distintos, sino que anotar variantes STR con nomenclatura de SNPs crea confusión en cuanto a identificar la variabilidad de esa región. Aun así la dbSNP posee registradas unas 5.500 variantes microsatélites manteniendo el nombre de GDB original (nomenclatura DS), en lo que parece un intento de identificar regiones que poseen inserciones y deleciones (INDELS) muy largas respectos de los SNPs.

El número registrado de 5.500 está muy lejos de contemplar la totalidad de variantes STRs humanas, que con un corte de MAF > 1 % para una longitud de la repetición de 2 nt a 6 nt, supera las 300.000 variantes.

Otra alternativa para nombrar variantes podrían ser sus datos de ubicación cromosómica (número de cromosoma y coordenadas) pero quedaría sujeta a la versión (y sus correcciones) del ensamblaje estudiado, y estaría afectado de desactualizaciones y no sería fácil de trazar con otras versiones, previas o posteriores.

La alternativa propuesta en este trabajo y que se usó para nombrar aquellos marcadores estudiados que carecían de alias, fue la de asociarlas con nombres de genes, acorde a las reglas que la ISFG establece para STRs contenidos en regiones génicas. Lo destacable de este enfoque es que casi la mitad de variantes STRs se encuentran en regiones génicas, y de estas, las regiones intrónicas contienen casi la totalidad. A su vez un intrón puede contener varios STRs. Lo que obliga no sólo a usar el nombre del gen (de HUGO) y el número de intrón y exón, sino que se requiere de un orden dentro de estos. Y eso fue lo que se hizo para nombrar a 98 marcadores en este estudio.

Los nombres para microsatélites de las regiones intergénicas se asociarían a marcadores previos respetando su nomenclatura, pero estos cubrirían un número menor a 70.099 (38.362 de UCSC y 31.737 de Ensembl, porque hay mucho solapamiento entre estas bases). Respetando el enfoque propuesto, aquellos STRs que quedan comprendidos entre regiones intergénicas llevarían el nombre de ambos genes, o entre marcadores, entre genes y marcadores, etc. y su respectivo orden, incluso podrían usarse también otros eventos genómicos, tales como otras regiones repetitivas: LINEs, SINEs, LTRs, etc. (no fueron incorporados a BD local ni fueron tenidas en cuenta en este estudio).

Está pendiente entonces una nomenclatura completa para una gran mayoría de microsatélites, y como se discute en 4.10 también de las variantes alélicas de estos microsatélites.

4.6 Cebadores y/o secuencias de anclaje o secuencias flanqueantes.

Aunque *a priori* se podría pensar que cebadores y secuencias flanqueantes son lo mismo, los cebadores son secuencias de importancia estrictamente biológicas-bioquímicas (*in vitro*) para la amplificación fragmentos de interés y es por eso que en la recopilación original de STSs era importante informar los cebadores usados, en cambio el concepto de secuencias de anclajes o flanqueantes son de importancia bioinformática (*in silico*).

En esencia comparten determinadas características, de poseer tamaños de entre 18 y 30 pb, y que deben ser específicos de la región diana.

Sin embargo, los cebadores están regidos por requisitos termodinámicos para su unión a la secuencia blanco, en cambio las secuencias flanqueantes son cadenas de texto que deben coincidir de manera más o menos estricta con la cadena blanco, y están regidos por principios informáticos y probabilísticos.

Y como ya se mencionó en el punto 4.2 muchos de los cebadores contenidos en Ensembl, UCSC y UniSTS no coinciden exactamente con su secuencia STS blanco asociada ó no están orientados de la manera adecuada a esa secuencia, además de que contienen un número bajo de STRs.

No existe al momento de la realización de este trabajo una recopilación de secuencias flanqueantes para microsatélites a los fines de uso con herramientas bioinformáticas, a pesar de la creciente aparición de nuevas herramientas con algoritmos mejorados o nuevas estrategias de detección.

4.7 Asignaciones alélicas

4.7.1 Reglas implementadas para adecuar las asignaciones alélicas entre lobSTR y StraitRazor

Dentro de la hoja “re-perfilado” de los archivos “PERFIL_v2_PLANTILLA_lobSTR_v32_str8rzr_v183.xlsm” de cada muestra, se suceden de manera automática antes de la asignación alélica final una serie de correcciones (ver punto 3.6.6. y relacionados).

Se utilizaron reglas condicionales para valorar cada asignación alélica, teniendo en cuenta coincidencias y profundidad. Se definieron las siguientes instancias:

- Se descartaron asignaciones alélicas donde la lectura que lo respaldó sólo apareció una sola vez.
- En caso de que ambas herramientas no detectaran alelos, se definió como marcador “No detectado”.
- Si ambas herramientas coincidieron en la asignación alélica se decidió por cualquiera de ambas.
- Si una herramienta no tuvo detección del sistema en cuestión, pero la otra sí, con más de una lectura de respaldo, se consideró entonces lo detectado por la mencionada

herramienta. Caso contrario, sin más de una lectura, se consideró como sistema “No detectado”.

- En caso de que ambas herramientas detectaran alelos, pero no coincidieran en la asignación, se definió aquella asignación alélica con mayor respaldo en lecturas.

Se definieron **reglas especiales** para los Tartamudeos.

- Todas las asignaciones alélicas fueron corregidas si alguna asignación alélica fuera un potencial tartamudeo (+/- 1 repetición del alelo principal), para esto se aplicó un criterio de abundancia entre ambos alelos:

(*) Para repeticiones de 2pb y 3pb, si los conteos del alelo principal fueron mayores a 2,5 veces los conteos del potencial tartamudeo, se consideró tartamudeo y se descartó, caso contrario fue considerado como segundo alelo.

(**) Para repeticiones de 4pb, 5pb y 6pb, si los conteos del alelo principal fueron mayores a 3 veces los conteos del potencial tartamudeo, se consideró tartamudeo y se descartó, caso contrario fue considerado como segundo alelo.

Si de todas las reglas aplicadas, algún sistema quedara sin asignación alélica o con más de dos alelos asignados, se corrigió manualmente.

- Se definieron como **“No concordantes”** los alelos detectados pero no coincidentes entre las herramientas y que a su vez, el análisis de profundidad o de tartamudeo no lo hubiera resuelto.
- Se definió como **“No detectado”** a los casos en que una herramienta no obtuviese alelos y la otra no fuese lo suficientemente coherente para ser aceptada.

4.7.2 Visualizaciones de los resultados de asignación alélica

En el archivo “perfiles-finales.xlsm” (Anexo VI) se pueden observar cada una de las 8.721 asignaciones alélicas (323 marcadores para las 27 muestras) donde se muestran las asignaciones automáticas (celdas verdes y grises “NDs”), intervenciones manuales (celdas amarillas) y no concordantes “NCs” (celdas rojas), se muestra una captura en la Figura 11a.

Como se detalla en el punto 3.7.2, del mismo archivo anterior se derivan una serie de otros archivos en Excel que dan origen a las Figuras 16 a 19.

Las Figuras 12 a 15 derivan de la hoja “Stats” de cada perfil obtenido para cada individuo ensayado.

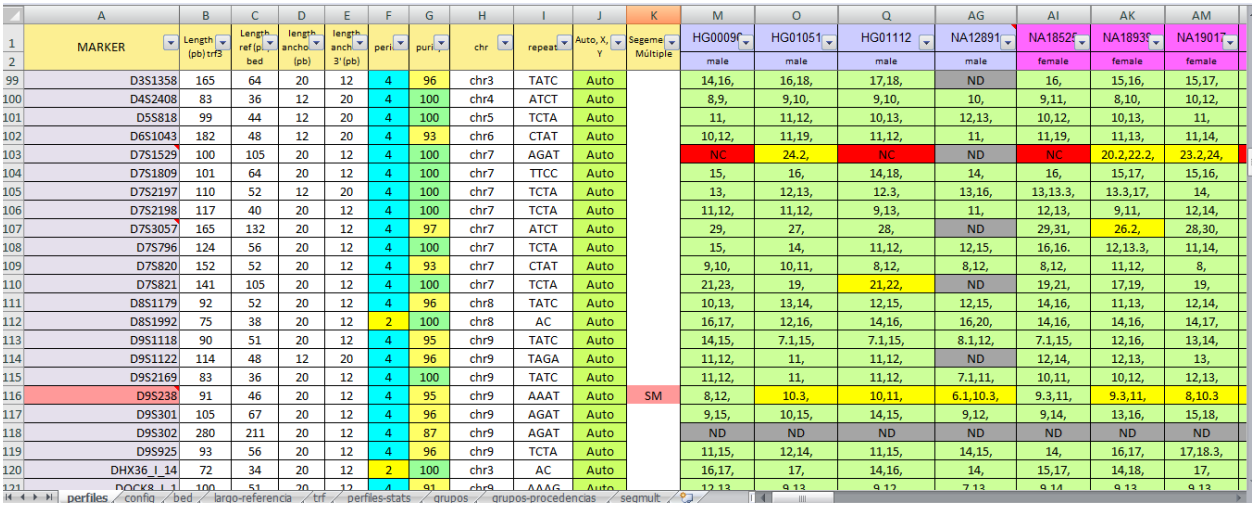
En las Figuras 11b y 11c se muestran capturas del archivo “visualizaciones-profundidad-lobSTR-StraitRazor.xlsm” donde se pueden ver los conteos de lectura para cada marcador y muestra, para ambas herramientas (punto 3.6.6 y Anexo VI).

Además se hicieron resúmenes con datos relevantes de cada marcador y la distribución alélica en histogramas para los 323 marcadores (vea el punto 3.7.4 y Anexo VI). En las capturas de la Figura 11d se muestran miniaturas como ejemplo.

Estas imágenes y datos más completos de los 319 marcadores (no se incluyen SRY y AMELOGENINA, y se reúnen los datos de los marcadores D13S305 y DYS448) también son accesibles desde la BD local vía interfaz web (ver punto 6.2)

Figura 11. Capturas de los archivos “perfiles-finales.xlsm” y “visualizaciones-profundidad-lobSTR-StraitRazor.xlsm” (a, b, c) y ejemplos de los reportes mínimos e histogramas (d).

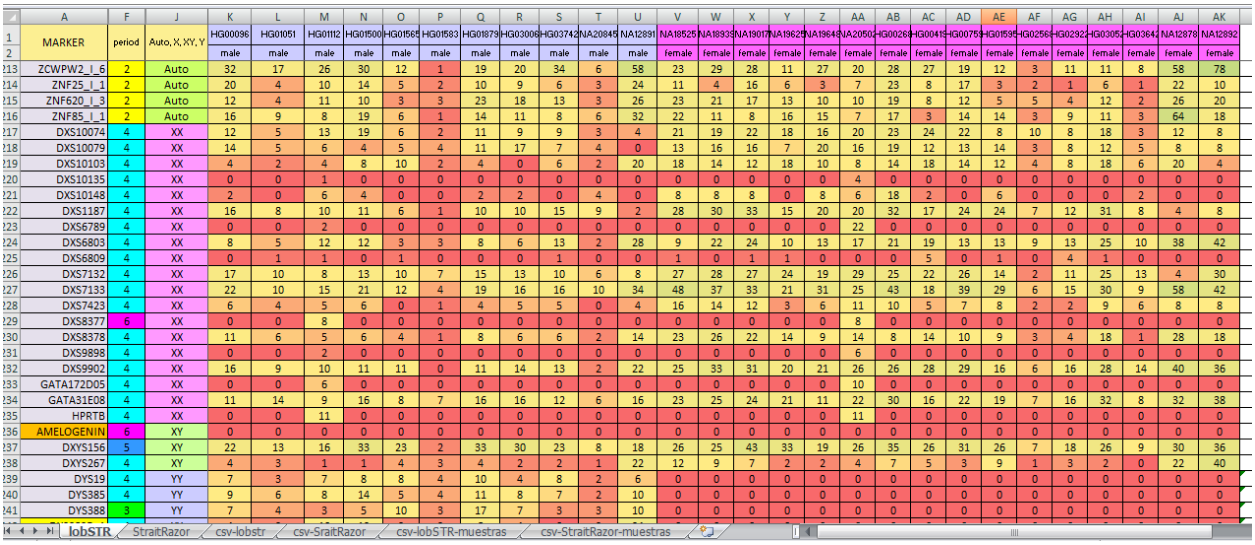
a)



	A	B	C	D	E	F	G	H	I	J	K	M	O	Q	AG	AI	AK	AM
1	MARKER	Length (pb) trf3	Length ref (p… bed	length ancho (pb)	length anchi 3' (pb)	peri…	puri…	chr	repeat…	Auto, X, Y	Segme… Múltiple	HG0009…	HG01051…	HG01112	NA12891…	NA1852…	NA1893…	NA1901…
2												male	male	male	male	female	female	female
99	D3S1358	165	64	20	12	4	96	chr3	TATC	Auto		14,16,	16,18,	17,18,	ND	16,	15,16,	15,17,
100	D4S2408	83	36	12	20	4	100	chr4	ATCT	Auto		8,9,	9,10,	9,10,	10,	9,11,	8,10,	10,12,
101	D5S818	99	44	12	20	4	100	chr5	TCTA	Auto		11,	11,12,	10,13,	12,13,	10,12,	10,13,	11,
102	D6S1043	182	48	12	20	4	93	chr6	CTAT	Auto		10,12,	11,19,	11,12,	11,	11,19,	11,13,	11,14,
103	D7S1529	100	105	20	12	4	100	chr7	AGAT	Auto		NC	24.2,	NC	ND	NC	20.2,22.2,	23.2,24,
104	D7S1809	101	64	20	12	4	100	chr7	TTCC	Auto		15,	16,	14,18,	14,	16,	15,17,	15,16,
105	D7S2197	110	52	12	20	4	100	chr7	TCTA	Auto		13,	12,13,	12.3,	13,16,	13,13.3,	13.3,17,	14,
106	D7S2198	117	40	20	12	4	100	chr7	TCTA	Auto		11,12,	11,12,	9,13,	11,	12,13,	9,11,	12,14,
107	D7S3057	165	132	20	12	4	97	chr7	ATCT	Auto		29,	27,	28,	ND	29,31,	26.2,	28,30,
108	D7S796	124	56	20	12	4	100	chr7	TCTA	Auto		15,	14,	11,12,	12,15,	16,16.	12,13.3,	11,14,
109	D7S820	152	52	20	12	4	93	chr7	CTAT	Auto		9,10,	10,11,	8,12,	8,12,	8,12,	11,12,	8,
110	D7S821	141	105	20	12	4	100	chr7	TCTA	Auto		21,23,	19,	21,22,	ND	19,21,	17,19,	19,
111	D8S1179	92	52	20	12	4	96	chr8	TATC	Auto		10,13,	13,14,	12,15,	12,15,	14,16,	11,13,	12,14,
112	D8S1992	75	38	20	12	2	100	chr8	AC	Auto		16,17,	12,16,	14,16,	16,20,	14,16,	14,16,	14,17,
113	D9S1118	90	51	20	12	4	95	chr9	TATC	Auto		14,15,	7.1,15,	7.1,15,	8.1,12,	7.1,15,	12,16,	13,14,
114	D9S1122	114	48	12	20	4	96	chr9	TAGA	Auto		11,12,	11,	11,12,	ND	12,14,	12,13,	13,
115	D9S2169	83	36	20	12	4	100	chr9	TATC	Auto		11,12,	11,	11,12,	7.1,11,	10,11,	10,12,	12,13,
116	D9S238	91	46	20	12	4	95	chr9	AAAT	Auto	SM	8,12,	10.3,	10,11,	6.1,10.3,	9.3,11,	9.3,11,	8,10.3
117	D9S301	105	67	20	12	4	96	chr9	AGAT	Auto		9,15,	10,15,	14,15,	9,12,	9,14,	13,16,	15,18,
118	D9S302	280	211	20	12	4	87	chr9	AGAT	Auto		ND	ND	ND	ND	ND	ND	ND
119	D9S925	93	56	20	12	4	96	chr9	TCTA	Auto		11,15,	12,14,	11,15,	14,15,	14,	16,17,	17,18.3,
120	DHX36_I_14	72	34	20	12	2	100	chr3	AC	Auto		16,17,	17,	14,16,	14,	15,17,	14,18,	17,
121	DOCK8_I_1	100	51	20	12	4	91	chr9	AAAG	Auto		12,13	9,13	9,12	7,13	9,14	9,13	9,13

perfiles | config | bed | largo-referencia | trf | perfiles-stats | grupos | grupos-procedencias | seqmult

b)



	A	F	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
1	MARKER	period	Auto, X, XY, Y	HG00096	HG01051	HG01112	HG01500	HG01565	HG01583	HG01879	HG03006	HG03742	NA20845	NA12891	NA18525	NA18939	NA19017	NA19625	NA19648	NA20502	HG00269	HG00419	HG00759	HG01595	HG02568	HG02922	HG03052	HG03642	NA12878	NA12892
2				male	male	male	male	male	male	male	male	male	male	male	female	female	female	female	female	female	female	female	female	female	female	female	female	female	female	female
213	ZCWPW2_I_6	2	Auto	32	17	26	30	12	1	19	20	34	6	58	23	29	28	11	27	20	28	27	19	12	3	11	11	8	58	78
214	ZNF25_I_1	2	Auto	20	4	10	14	5	2	10	9	6	3	24	11	4	16	6	3	7	23	8	17	3	2	1	6	1	22	10
215	ZNF620_I_3	2	Auto	12	4	11	10	3	3	23	18	13	3	26	23	21	17	13	10	10	19	8	12	5	5	4	12	2	26	20
216	ZNF85_I_1	2	Auto	16	9	8	19	6	1	14	11	8	6	32	22	11	8	16	15	7	17	3	14	14	3	9	11	3	64	18
217	DXS10074	4	XX	12	5	13	19	6	2	11	9	9	3	4	21	19	22	18	16	20	23	24	22	8	10	8	18	3	12	8
218	DXS10079	4	XX	14	5	6	4	5	4	11	17	7	4	0	13	16	16	7	20	16	19	12	13	14	3	8	12	5	8	8
219	DXS10103	4	XX	4	2	4	8	10	2	4	0	6	2	20	18	14	12	18	10	8	14	18	14	12	4	8	18	6	20	4
220	DXS10135	4	XX	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
221	DXS10148	4	XX	2	0	6	4	0	0	2	2	0	4	0	8	8	8	0	8	6	18	2	0	6	0	0	0	2	0	0
222	DXS1187	4	XX	16	8	10	11	6	1	10	10	15	9	2	28	30	33	15	20	20	32	17	24	24	7	12	31	8	4	8
223	DXS6789	4	XX	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0
224	DXS6803	4	XX	8	5	12	12	3	3	8	6	13	2	28	9	22	24	10	13	17	21	19	13	13	9	13	25	10	38	42
225	DXS6809	4	XX	0	1	1	0	1	0	0	0	1	0	0	1	0	1	1	0	0	0	5	0	1	0	4	1	0	0	0
226	DXS7132	4	XX	17	10	8	13	10	7	15	13	10	6	8	27	28	27	24	19	29	25	22	26	14	2	11	25	13	4	30
227	DXS7133	4	XX	22	10	15	21	12	4	19	16	16	10	34	48	37	33	21	31	25	43	18	39	29	6	15	30	9	58	42
228	DXS7423	4	XX	6	4	5	6	0	1	4	5	5	0	4	16	14	12	3	6	11	10	5	7	8	2	2	9	6	8	8
229	DXS8377	6	XX	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0
230	DXS8378	4	XX	11	6	5	6	4	1	8	6	6	2	14	23	26	22	14	9	14	8	14	10	9	3	4	18	1	28	18
231	DXS9898	4	XX	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
232	DXS9902	4	XX	16	9	10	11	11	0	11	14	13	2	22	25	33	31	20	21	26	26	28	29	16	6	16	28	14	40	36
233	GATA172D05	4	XX	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
234	GATA31E08	4	XX	11	14	9	16	8	7	16	16	12	6	16	23	25	24	21	11	22	30	16	22	19	7	16	32	8	32	38
235	HPRTB	4	XX	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0
236	AMELOGENIN	6	XY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
237	DXYS156	5	XY	22	13	16	33	23	2	33	30	23	8	18	26	25	43	33	19	26	35	26	31	26	7	18	26	9	30	36
238	DXYS267	4	XY	4	3	1	1	4	3	4	2	2	1	22	12	9	7	2	2	4	7	5	3	9	1	3	2	0	22	40
239	DYS19	4	YY	7	3	7	8	8	4	10	4	8	2	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
240	DYS385	4	YY	9	6	8	14	5	4	11	8	7	2	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
241	DYS388	3	YY	7	4	3	5	10	3	17	7	3	3	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

lobSTR | StraitRazor | csv-lobstr | csv-StraitRazor | csv-lobSTR-muestras | csv-StraitRazor-muestras

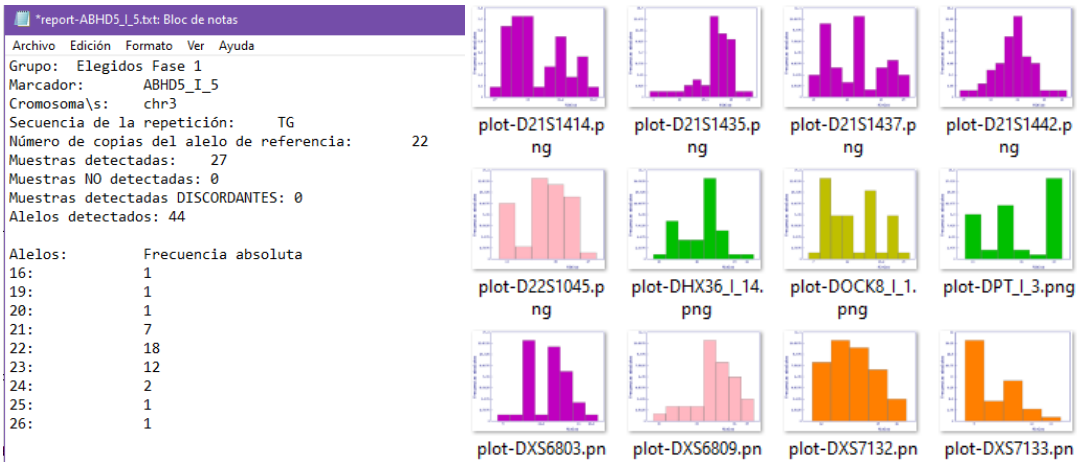
c) siguiente página.

c) (Figura 11 continuación)

	A	F	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
1	MARKER	period	Auto, X, XY, Y	HG00096	HG01051	HG01112	HG01500	HG01565	HG01583	HG01879	HG03006	HG03742	NA20845	NA12891	NA18525	NA18939	NA19017	NA19625	NA19648	NA20502	HG00268	HG00419	HG00753	HG01595	HG02568	HG02922	HG03052	HG03642	NA12878	NA12892
2				male	male	male	male	male	male	male	male	male	male	male	female	female	female	female	female	female	female	female	female	female	female	female	female	female	female	female
213	ZCWPW2_I_6	2	Auto	37	25	36	36	25	8	28	23	38	18	24	19	28	35	18	29	24	29	28	25	21	11	25	16	17	8	22
214	ZNF25_I_1	2	Auto	30	15	15	18	10	9	12	16	7	13	14	16	11	22	9	13	13	26	14	30	11	5	10	10	8	30	18
215	ZNF620_I_3	2	Auto	23	18	16	19	11	8	30	23	19	12	36	31	28	23	26	19	17	30	23	17	17	10	21	20	14	26	24
216	ZNF85_I_1	2	Auto	17	19	16	20	11	6	30	17	23	17	20	35	16	14	21	27	9	28	14	17	27	7	23	16	15	52	12
217	DXS10074	4	XX	17	6	17	17	12	7	11	6	10	7	0	15	19	17	16	18	17	26	20	25	7	10	14	19	13	0	0
218	DXS10079	4	XX	19	9	4	4	7	11	14	20	10	5	0	18	23	22	13	23	26	22	18	15	18	11	17	13	14	0	0
219	DXS10103	4	XX	8	12	12	17	16	7	14	10	13	6	0	19	29	20	20	18	20	17	21	20	30	8	14	25	23	0	0
220	DXS10135	4	XX	5	5	7	6	5	2	3	1	2	3	0	9	15	9	6	6	9	10	5	10	2	5	11	12	1	0	0
221	DXS10148	4	XX	1	0	0	1	0	0	1	2	0	0	0	4	1	1	0	1	0	2	1	4	1	0	0	3	0	0	0
222	DXS1187	4	XX	20	24	25	11	13	6	16	13	23	16	0	31	39	31	26	23	22	28	20	28	43	14	21	34	21	0	0
223	DXS6789	4	XX	10	13	4	15	14	6	19	13	16	7	0	19	22	36	16	14	21	46	29	36	24	11	17	28	15	0	0
224	DXS6803	4	XX	16	12	25	23	18	21	29	17	27	17	22	40	43	39	38	24	35	47	48	51	46	23	39	33	37	30	20
225	DXS6809	4	XX	1	7	4	3	3	6	8	5	8	3	0	14	11	10	6	9	10	13	21	6	17	6	7	7	3	0	0
226	DXS7132	4	XX	10	9	2	9	7	5	10	7	12	4	0	15	16	15	16	17	18	9	18	19	12	1	19	7	13	0	0
227	DXS7133	4	XX	29	16	27	21	23	18	28	26	23	17	24	62	53	39	37	38	31	50	28	58	41	15	42	51	31	38	24
228	DXS7423	4	XX	3	5	3	2	0	4	4	5	5	2	0	6	14	10	5	9	10	12	3	7	11	5	3	17	11	0	0
229	DXS8377	6	XX	4	4	6	1	3	0	2	0	4	0	0	8	5	1	3	4	4	1	4	3	3	1	6	3	0	0	0
230	DXS8378	4	XX	9	8	7	5	12	8	13	11	8	5	0	18	26	25	12	12	11	12	21	6	18	9	11	23	9	0	0
231	DXS9898	4	XX	13	12	6	8	4	6	11	12	9	6	0	17	23	19	13	11	10	20	25	12	7	9	17	18	18	0	0
232	DXS9902	4	XX	22	18	19	11	20	10	14	20	23	14	22	32	40	43	32	31	39	34	38	52	30	12	30	28	34	10	8
233	GATA172D05	4	XX	15	14	13	11	17	13	28	22	36	19	18	34	35	38	34	34	36	30	28	43	41	11	27	26	30	40	68
234	GATA31E08	4	XX	14	21	18	22	14	19	22	23	19	12	10	30	36	35	36	26	26	39	20	40	37	12	41	45	33	42	38
235	HPRTB	4	XX	5	6	7	4	6	6	6	6	4	3	0	16	15	10	9	12	12	21	12	13	13	11	15	18	12	0	0
236	AMELOGENIN	6	XY	23	40	32	38	29	15	26	13	26	18	8	24	29	26	12	40	23	32	25	31	23	18	19	23	13	4	10
237	DXYS156	5	XY	30	23	33	52	37	9	44	33	37	23	0	24	27	50	37	22	23	40	34	39	36	9	38	31	28	10	18
238	DXYS267	4	XY	22	26	24	37	23	15	28	21	21	21	24	20	33	31	16	12	23	22	18	20	21	12	8	26	18	16	20
239	DYS19	4	YY	6	6	11	6	13	6	10	7	16	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
240	DYS385	4	YY	3	5	4	4	3	9	3	7	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
241	DYS388	3	YY	8	10	11	10	18	8	19	8	6	13	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

lobSTR | StraitRazor | csv-lobstr | csv-SraitRazor | csv-lobSTR-muestras | csv-StraitRazor-muestras

d)



Los colores en los histogramas responden a distintos grupos de los marcadores sometidos a estudio: Naranja: Ciencias Forenses, Lila: Ciencias de la Salud, Rosa: comunes a Ciencias Forenses y de la Salud, Verde: Elegidos de Fase1/lobSTR y Verde-Oliva: Elegidos del cromosoma 7 y 9.

4.7.3 Resultados globales relevantes

Sólo dos marcadores no fueron encontrados en estas 27 muestras, uno autosómico D9S302 (crom. 9) y otro exclusivo del crom. Y, DYS389B_2, ambos exceden los 250 pb del largo del fragmento probable de lectura en la secuenciación (ver hoja “largo-referencia” dentro del archivo “perfiles-finales.xlsm”).

Los tres marcadores pseudo-autosómicos: Amelogenina (sólo StraitRazor), DXYS267 y DXYS156 fueron detectados en las 27 muestras.

El marcador SRY (sólo StraitRazor) fue detectado en los 11 varones y no fue detectado en las 16 mujeres, evidenciando la especificidad.

Ambos marcadores Amelogenina y SRY arrojaron resultados acorde con lo declarado en el género para cada muestra.

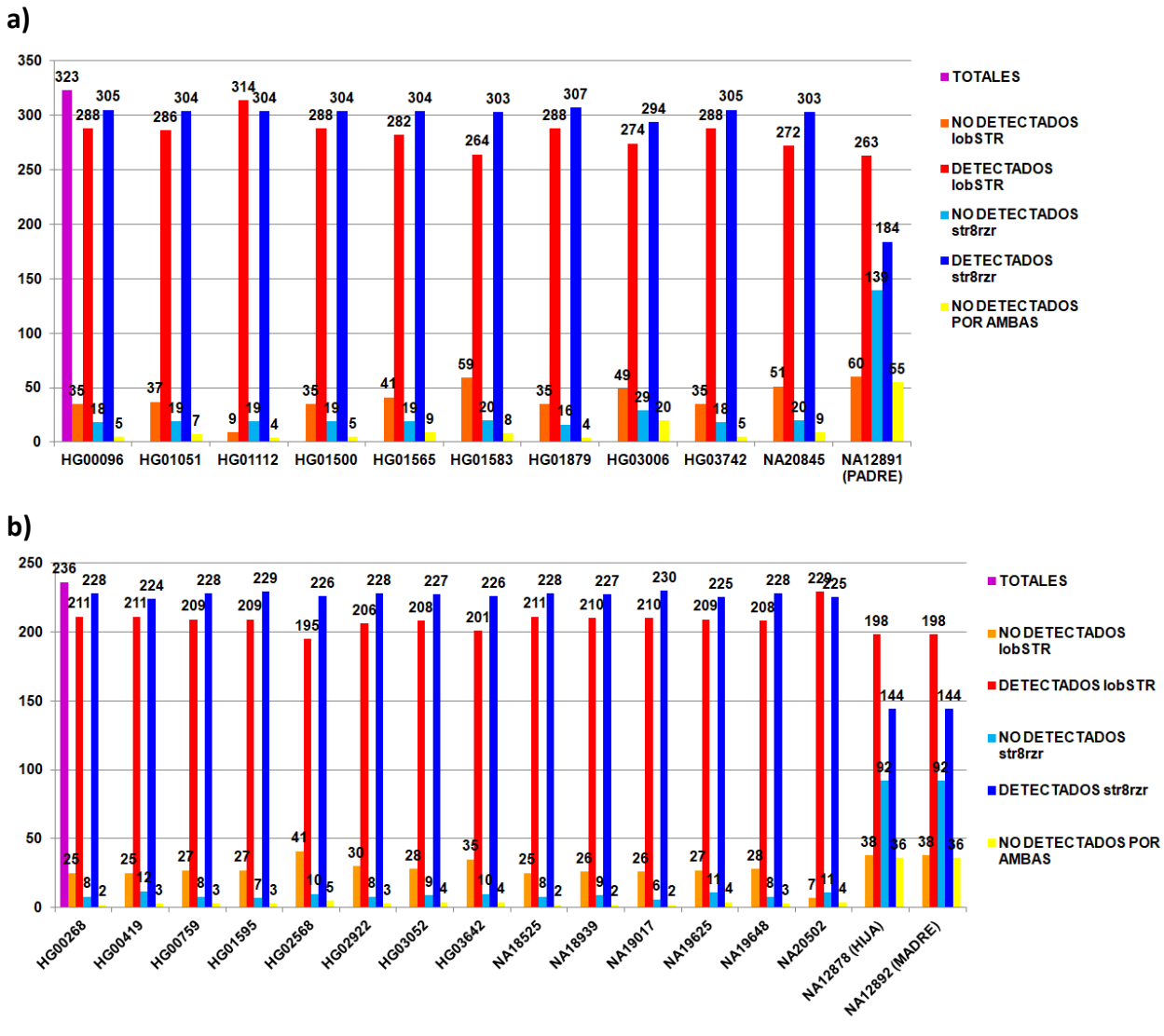
De los marcadores exclusivos del cromosoma X (hemicigota para varones) 10 de 19 (53 %) fueron detectados en las 27 muestras.

Los marcadores SRY y Amelogenina fueron incorporados al archivo BED (versión 32) para lobSTR, pero no hubo detecciones con esta herramienta.

4.7.4 Detecciones de las herramientas de forma unilateral

En la Figura 12 se muestran la cantidad de marcadores detectados y no detectados por cada herramienta por separado y se observa que en general el número de marcadores detectados por cada herramienta es similar, con cierta ventaja de StraitRazor (barras azules). Los totales de marcadores analizados para varones y mujeres fueron de 321 y 234 (se excluyeron los 87 del cromosoma Y) respectivamente.

Figura 12. Marcadores detectados por ambas herramientas para 11 varones (a) y 16 mujeres (b).



En las muestras HG01112 (varón) y NA20502 (mujer) se observa un número de detecciones de marcadores de lobSTR levemente superior a lo conseguido con StraitRazor.

En las tres muestras del trío CEPH (NA12878, NA12891, NA12892) se observa un número de detecciones de marcadores de lobSTR considerablemente superior a StraitRazor, aunque el número de marcadores no detectados para ambas herramientas en estas tres muestras se incrementan considerablemente 36-55 (versus 2-20 del resto, P = .02611, α = 0.05) y 92-139 (versus 6-29 del resto, P = .02594, α = 0.05) para StraiRazor -barras celestes-, en cambio lobSTR se mantiene sin demasiados cambios en el número de detecciones para estas muestras, 38-60 (versus 7-59 del resto, P = .07019, α = 0.05).

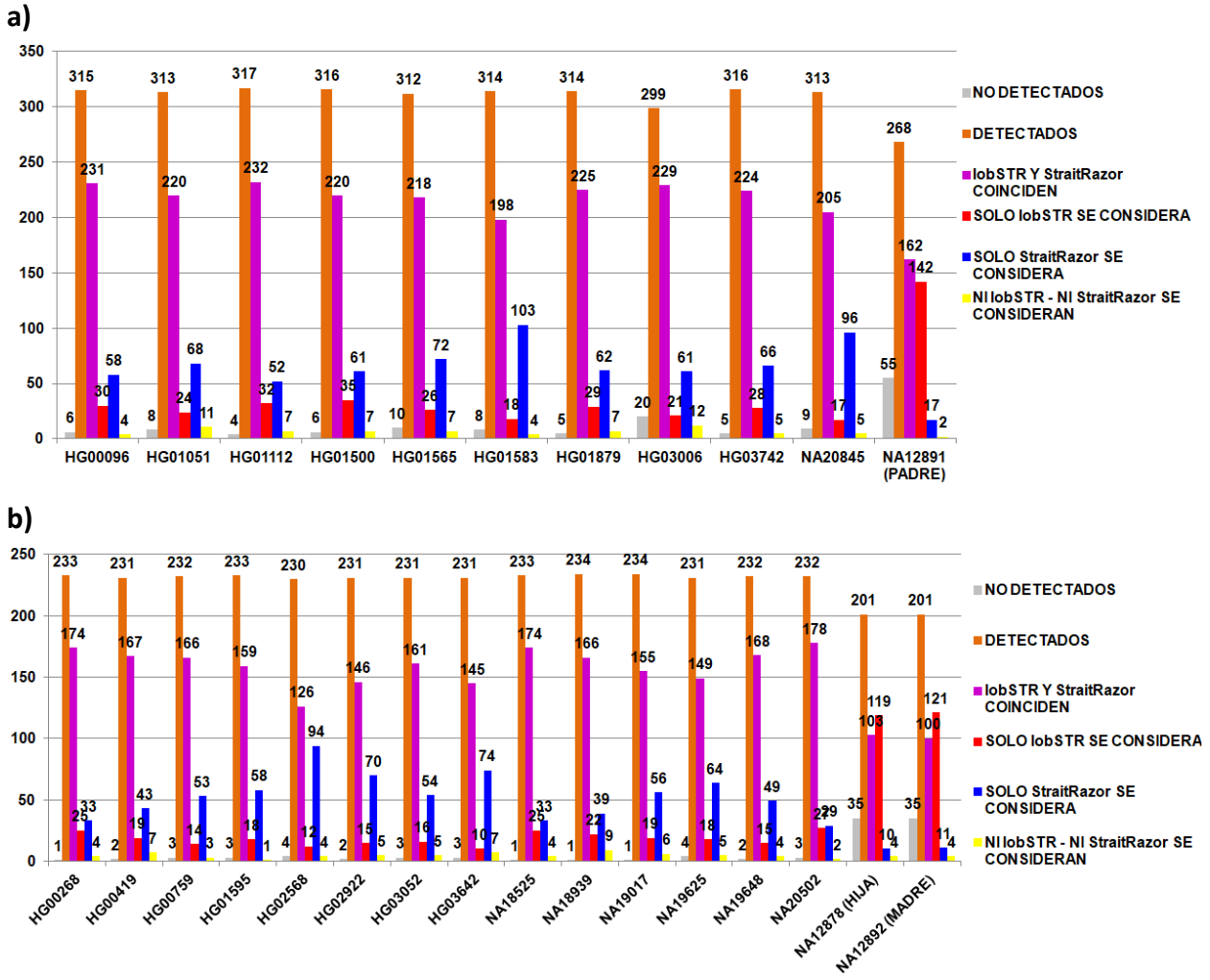
Para las muestras de CEPH se usó una librería distinta de la usada para las 24 muestras restantes: Illumina en vez de Solexa. (Ver Anexo I).

Esto sugiere que StraitRazor es susceptible para otras condiciones de secuenciación. La longitud de las lecturas de la librería producida por Illumina (101 pb) son menores que la producida por Solexa (250 pb). Datos provenientes de los análisis de calidad a los 102 fastqs con FastQC (detalles en “Desde sitio web” del Anexo II)

4.7.5 Detecciones de las herramientas de manera conjunta

En la Figura 13 se muestran la cantidad de marcadores detectados y no detectados por ambas herramientas cuando fueron coincidentes (barras púrpuras) o si alguna asignación fue exclusiva para alguna de las dos herramientas (lobSTR: rojo, StraitRazor: azul).

Figura 13. Marcadores detectados por ambas herramientas, juntas y por separado para 11 varones (a) y 16 mujeres (b).



En las Figuras 13a y 13b se evidencian como ambas herramientas hicieron asignaciones alélicas (con posteriores correcciones) que fueron concordantes. También se evidencia que StraitRazor definió unilateralmente algún sistema cuando la asignación no fue coincidente con lobSTR y esto se debe principalmente a que StraitRazor entregó en la salida mayores conteos de lecturas que soportaron los alelos designados, y el conteo de lecturas formó parte de las decisiones para definir algunas asignaciones alélicas.

De nuevo, y de manera más evidente, lobSTR fue la mejor opción frente a StraitRazor con las muestras del trío CEPH, incluso las asignaciones de lobSTR fueron superiores a las asignaciones en conjunto de ambas herramientas.

Pero en estas tres muestras también se observa el mayor número de marcadores no detectados, por una u otra herramienta, o ambas: 35-55 versus 1-20 del resto, P = .0294, α = 0.05 (barras grises).

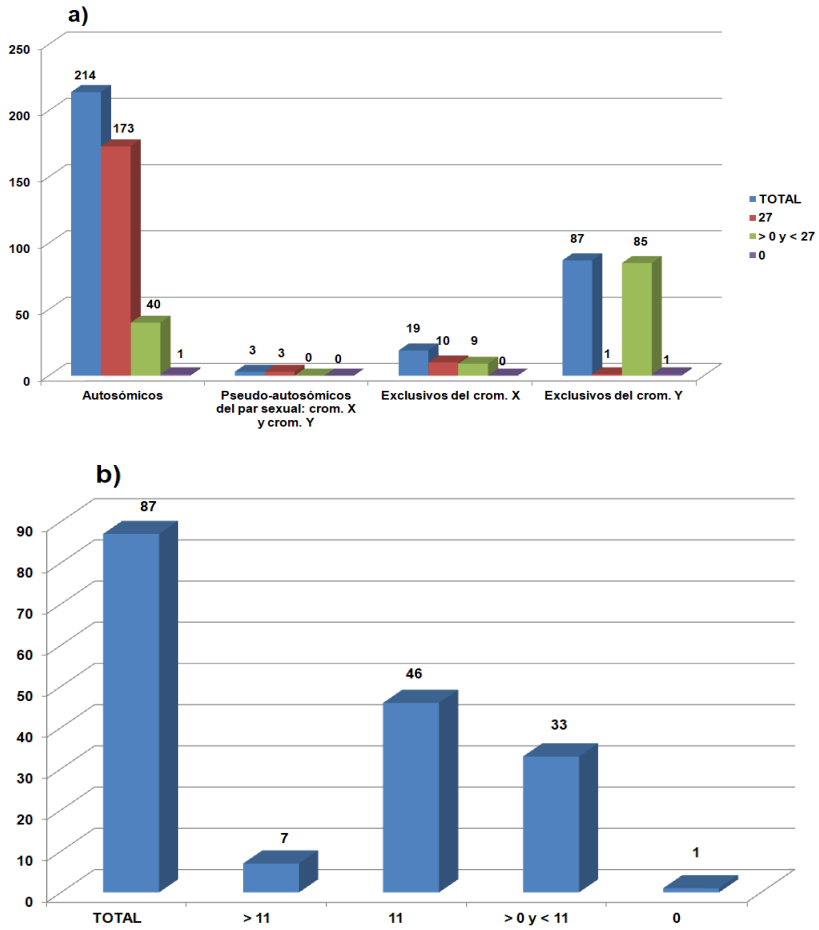
4.7.6 Detecciones de los marcadores según número de muestras y procedencia cromosómica.

La procedencia cromosómica indica si el marcador proviene de un cromosoma autosómico (1 al 22), pseudo-autosómico del par sexual (el marcador existe en ambos cromosomas X e Y), exclusivos del cromosoma X y exclusivos del cromosoma Y (sólo existen en los cromosomas mencionados).

Entonces, excluyendo a los marcadores exclusivos del cromosoma Y, si un marcador fue encontrado en las 27 muestras analizadas la detección fue del 100 %, 186 marcadores obtuvieron ese número (87 % de un total de 236 marcadores).

En cambio para los marcadores exclusivos del cromosoma Y, 11 es el número máximo de muestras, de 87 marcadores totales, 46 obtuvieron la máxima detección (53%), ver Figura 14b.

Figura 14. Muestras versus marcadores detectados según su procedencia cromosómica.



a) Se espera que para los marcadores autosómicos, pseudo-autosómicos y exclusivos del crom. X se detecten como máximo en las 27 muestras ensayadas (16 mujeres y 11 varones), en cambio,

b) para los marcadores exclusivos del crom. Y ese máximo es 11 (11 varones).

Aquellos marcadores del cromosoma Y que obtuvieron un número mayor a 11 se debió a inespecificidad de las herramientas.

Los siete marcadores fueron DYS393, DYS413, DYS446, DYS490, DYS572, DYS636 y DYS640.

Se discute en el apartado 4.8.1.1 el caso especial: DYS393 vs DXYS267, cuando se analice más en profundidad las inespecificidades de las herramientas.

Los 6 restantes marcadores exclusivos del crom. Y (ver barra "> 11" en la Figura 14) que se sospecha inespecificidad por tener detecciones en mujeres son: DYS413, DYS446, DYS490, DYS572, DYS636 y DYS640. Excepto DYS636, todos están incluidos en segmentos múltiples. DYS636 tiene sólo una detección (asignado como "0.2,") en el individuo de género femenino "NA12892" del trío CEPH. De hecho ambas mujeres del trío CEPH presentan para estos marcadores un porcentaje alto de detecciones (9 de 12).

4.7.7 Intervención del analista en las asignaciones alélicas

En la Figura 15 se puede observar del total de muestras sometidas a análisis (27) para los 323 marcadores, la proporción de marcadores detectados, no detectados o con asignaciones alélicas discordantes, y en la Figura 15 la proporciones de asignaciones automáticas y de intervención manual.

Figura 15. Muestras detectadas, no detectadas y con asignaciones discordantes.

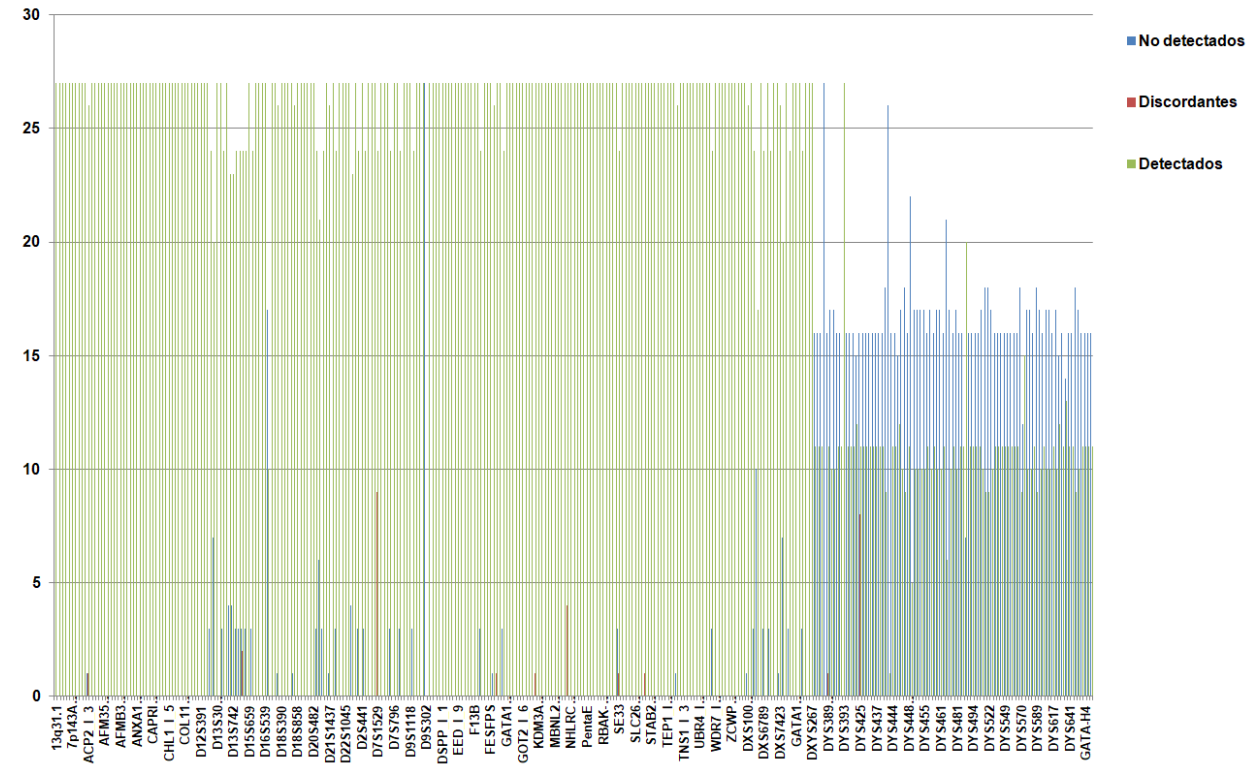
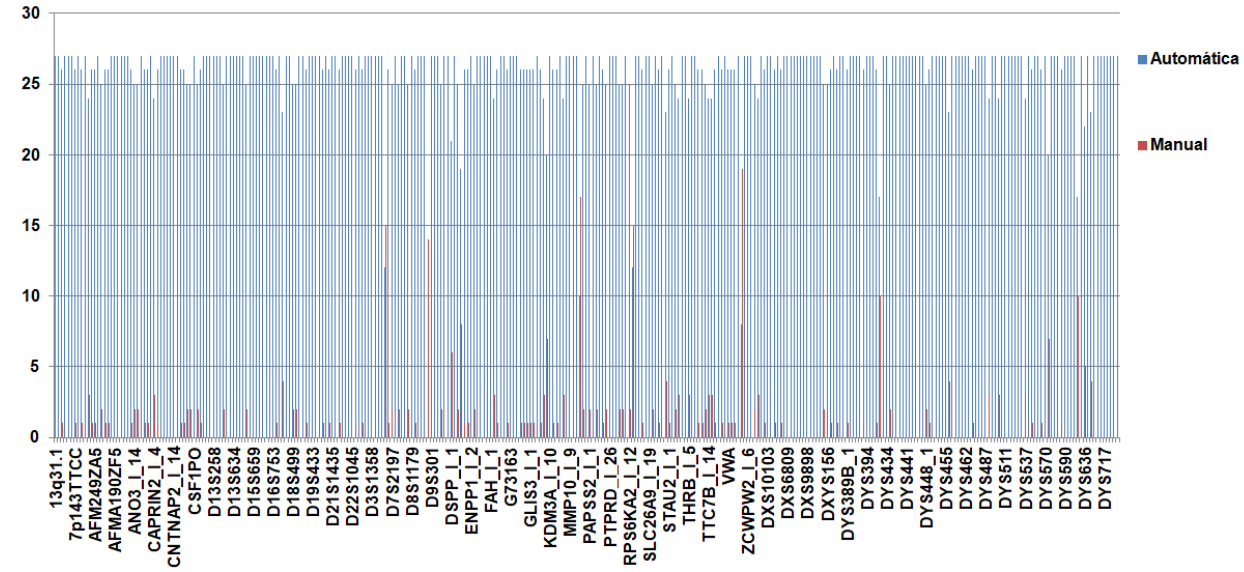


Figura 16. Proporciones de asignaciones alélicas automáticas e intervención manual



Se destacan los marcadores D7S1529, DYS425, NBPF9_I_21 y YY1AP1_I_7 con muchas asignaciones alélicas discordantes

Para el marcador D7S1529 se encontró un error en el archivo BED (lobSTR), donde el largo de una de las repeticiones no coincide con el largo de la secuencia de repetición (2 para una secuencia ACAC, debería haber sido 4). En cambio DYS425 cae dentro de un segmento múltiple (ver todos los marcadores en esta situación en la Tabla 10), el marcador YY1AP1_I_7 no fue revisado minuciosamente (aunque si hubieron modificaciones en sus secuencias de anclaje y el motivo) la inespecificidad no se evidenció en los análisis exploratorios. Por último NBPF9_I_21, a diferencia del STR anterior; si se hizo inspección minuciosa y correcciones, sin embargo aparecieron inespecificidades. A estos últimos dos marcadores no se le encontró una razón obvia de su comportamiento en las asignaciones alélicas.

Incluso, con el análisis hecho con Blastn, YY1AP1_I_7 y NBPF9_I_21 no fueron secuencias de aparición frecuente en el genoma: YY1AP1_I_7 sin *hits* y NBPF9_I_21 un *hit*, *score* 213 y e-value: 1e-53 ó 1 x 10^{-53}, que se corresponde con la ubicación esperada de este marcador.

Usando esta misma herramienta, para los marcadores D19S433 y G73163, por mencionar sólo dos ejemplos con mayor cantidad de *hits*, se obtuvieron 44 y 51 *hits* en el genoma (en ambos casos el *hit* con *score* más alto tiene un 100 % de identidad con el largo total de la secuencia, los siguientes *hits* sólo de manera parcial y con identidad menor al 100%).

Sin embargo, D19S433 no tuvo intervenciones manuales y G73163 sólo dos.

Se observó que encontrar las secuencias de anclaje y la secuencia motivo adecuada para ser usados con StraitRazor no sólo se consigue aislándolos de las inmediaciones del núcleo, esquivando SNPs de aparición frecuente, si no que requiere un enfoque que contemple la probabilidad de captura de lecturas de manera inespecífica, previo al uso de la herramienta. Blastn no resolvió esta necesidad.

En este trabajo, la evaluación de inespecificidad se hizo *a posteriori* del uso de StraitRazor con un número reducido de fastqs (hasta 56, v17 del config), e implicó redefinir los parámetros en el config, inspeccionando de manera minuciosa cada marcador (190 de 323 marcadores)

Los marcadores D7S1529, DYS425, NBPF9_I_21 y YY1AP1_I_7, D9S238, DYS617 y SE33 tuvieron muchas intervenciones manuales (de 10 a 19 de las 27).

4.8 Especificidad de las herramientas

La manera correcta de evaluar la especificidad de estas herramientas bioinformáticas es contrastarlas con otro método que haya sido extensamente probado, por ejemplo: para los marcadores STRs usados en ciencias forenses la asignación alélica se corresponde con la migración electroforética de un fragmento amplificado por PCR, junto con un estándar de tamaños, y por último se define el alelo según una escalera alélica (todos los alelos conocidos de la población) A su vez, los fabricantes de los reactivos de detección de STRs deben secuenciar las variantes alélicas que componen la escalera alélica. Para más detalles vea el punto 1.3.

Entonces, existe un gran impedimento operativo y costoso para evaluar la asignación alélica de una herramienta bioinformática en contra del método *gold standard*. Y si a eso le sumamos más individuos y pretendemos ampliar el número de marcadores, resulta imposible.

La aproximación propuesta en este trabajo es la de examinar las asignaciones alélicas devueltas por lobSTR y StraitRazor de manera recíproca, y existen cinco escenarios posibles:

a) El uso de marcadores autosómicos (cromosomas 1 al 22) y pseudo-autosómicos (cromosomas X e Y), donde se espera a lo sumo dos asignaciones alélicas.
b) El uso de marcadores cromosoma X en mujeres, y también se espera a lo sumo dos asignaciones alélicas.
c) El uso de marcadores del cromosoma Y en varones, donde se espera a lo sumo una asignación alélica.
d) El uso de marcadores del cromosoma X en varones, y también se espera a lo sumo una asignación alélica.
e) El uso de marcadores del cromosoma Y en mujeres, donde no se espera asignaciones alélicas.

Un número mayor de lo esperado en cualquiera de estos escenarios incurren en inespecificidad.

Los escenarios “a” y “b” no resultan útiles, dado que las reglas aplicadas en los perfiles escogen los dos alelos con mayor conteos de lecturas, y a su vez estas están sometidas a las reglas donde se filtran asignaciones alélicas según el cociente entre esos conteos (reglas para tartamudeos).

En una primera inspección de las asignaciones alélicas puras, lobSTR casi siempre devolvió uno o dos alelos, en cambio StraitRazor fue más frecuente en devolver tres o más alelos, y debido a este hecho es que se aplicaron estas reglas, y poder emparejar las asignaciones alélicas entre ambas herramientas de manera automática.

Sin embargo, los escenarios que se discuten a continuación brindan información directa sobre la inespecificidad (sea esta un tartamudeo u otro evento).

Se trabajó de manera más fina con los marcadores del cromosoma Y (escenario “e”) y los marcadores del cromosoma X (escenario “d”), en los apartados siguientes.

El escenario “c” es similar al “d”.

4.8.1 Evaluación de especificidad mediante los marcadores del cromosoma Y

En la sección 4.7 se analizaron las asignaciones alélicas de ambas herramientas, de manera global y conjunta, y esas asignaciones alélicas son producto de las reglas aplicadas en cada planilla de Excel de los perfiles de cada individuo. Para el análisis de especificidad con los marcadores del cromosoma Y en mujeres (escenario "e"), prescindimos de las asignaciones alélicas y nos detenemos en el conteo de lecturas, dado que cualquier resultado mayor a 0 es indicativo de inespecificidad. En el archivo "visualizaciones-profundidad-lobSTR-StraitRazor.xlsm" se encuentran estas lecturas, y en la Tabla 9 lo que se desprende del manejo de esos datos.

Tabla 9. Conteos de lecturas para los marcadores del cromosoma Y en las 16 mujeres (incluyendo y excluyendo marcadores que caen dentro de segmentos múltiples).

	lobSTR		StraitRazor	
	Sin segmentos múltiples	Con segmentos múltiples Sin DYS393, DYS490, DYS446 y DYS572[1]	Sin segmentos múltiples	Con segmentos múltiples Sin DYS393, DYS490, DYS446 y DYS572
Detecciones posibles	960 (16 x 60) [2]	1.328 (16 x 83)	960 (16 x 60)	1.328 (16 x 83)
Inespecíficas	10 (1,04 %)[3]	16 (1,20 %)[4]	7 (0,73 %)[3]	13 (0,98 %)[4]

1. Excluidos en este análisis debido a errores en los archivos de configuración de las herramientas (DYS393, se discute en 4.8.1.1) ó por que fueron eliminados del archivo "*.config" (v183) de StraitRazor.
2. Número de mujeres x número de marcadores
3. P = .6261, α = 0.05
4. P = .7088, α = 0.05

El cuadro anterior sugiere que ambas herramientas se asemejan respecto a la cantidad de detecciones inespecíficas: 10 de 960 (1,04 %) de lobSTR contra 7 de 960 (0,73 %) de StraitRazor (P = .6261, α = 0.05) considerando aquellos marcadores que no están dentro de segmentos múltiples, y 16 de 1328 (1,20 %) de lobSTR contra 13 de 1328 (0,98 %) de StraitRazor (P = .7088, α = 0.05) considerando marcadores que están contenidos en segmentos múltiples (se excluyen los marcadores DYS393, DYS490, DYS446 y DYS5721 que tuvieron errores en los archivos de configuración de las herramientas o simplemente no fueron incluidos en los análisis de una de las herramientas).

Estos mismos marcadores junto con DYS640 explicarían la totalidad de detecciones inespecíficas con lobSTR, y todos pertenecen a segmentos múltiples (duplicaciones inter-cromosómicas entre el Cromosoma X e Y, ver el listado completo en la Tabla 10 en el apartado 4.8.2), sin embargo otros ocho marcadores (DYS391, DYS393 –sólo lobSTR-, DYS434, DYS435,

DYS437, DYS456, DYS505, DYS522) en idénticas condiciones de duplicación entre el Cromosoma X e Y, no tienen detecciones inespecíficas para esta herramienta.

En cambio, para StraitRazor se encontraron detecciones inespecíficas en los marcadores DYS393 (16, se discute en el próximo apartado), DYS435 (2), DYS456 (3) y DYS640 (1).

En el apartado 4.8.2 se discute en profundidad sobre segmentos múltiples.

.

4.8.1.1 Caso especial: DYS393 vs DXYS267

Un error cometido en la configuración de las herramientas (StraitRazor en este caso) llevó a la detección del marcador DYS393 en las 16 mujeres. Esta situación también deja evidencia de cómo impacta el largo de las secuencias de anclaje en la detección de lecturas para esta herramienta.

El marcador DYS393 forma parte del panel de marcadores del *European core Y-STR* (recomendado por SWGDAM) y es sumamente usado en ciencias forenses integrando reactivos multiplex para la determinación del haplotipo del cromosoma Y, por su parte el marcador DXYS267 es ampliamente usado en ciencias de la salud como marcador para el diagnóstico de aneuploidías originadas en el par sexual.

Pero, en definitiva, los marcadores DYS393 y DXYS267 pertenecen a una duplicación de segmentos compartidos entre los cromosomas X e Y, UID-25539:

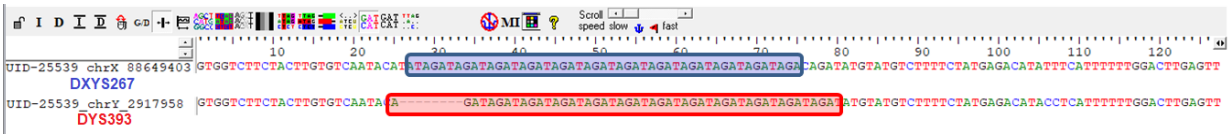
Segmento del cromosoma X: chrX: 88649403 - 88987546 (largo 338.143 pb).

Segmento del cromosoma Y: chrY: 2917958 - 3258597 (largo 340.639 pb).

Identidad: 98,12 %.

Ambas regiones circunscritas a los marcadores tienen “casi” la misma estructura del núcleo: $[ATAG]_n$ para DXYS267 y $[AGAT]_n$ para DYS393 y “casi” la misma secuencia, ver Figura 17.

Figura 17. Alineamiento acotado a la región del segmento UID-25539 que contienen los marcadores DXYS287 y DYS393.



Es importante aclarar que cuando se menciona al marcador DYS393 sólo nos referimos a aquella región que sólo pertenece al segmento del cromosoma Y (de ahí el sufijo “DYS”), en cambio cuando mencionamos al marcador DXYS267 nos referimos a ambas regiones de los segmentos de los cromosomas X e Y (de ahí el sufijo “DXYS”).

Para evaluar la razón por la cual el marcador DYS393 fue detectado en las 16 mujeres, se inspeccionaron con cuidado las secuencias de anclaje para ambos marcadores (config v183 para StraitRazor 3.0), junto con las secuencias blanco (HG19), además de incorporar en esta inspección a los cebadores descritos en STRBase (SRD-130) para el marcador DYS393. Figuras 21 y 22.

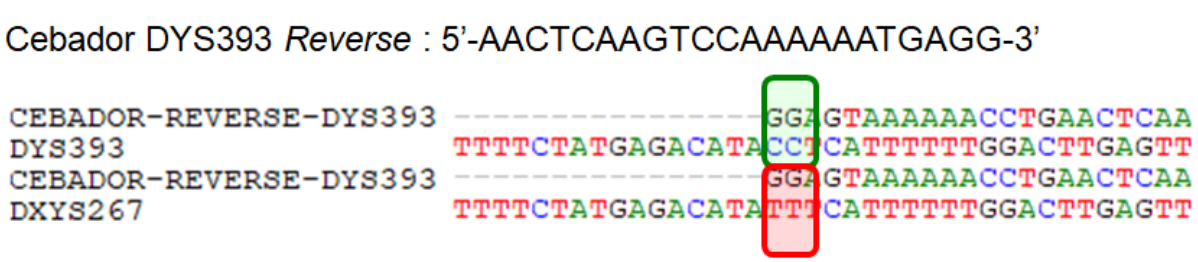
Figura 18. Ubicación de los cebadores de uso en ciencias forenses para el marcador DYS393 y de las secuencias de anclaje para la herramienta StraitRazor 3.0 para ambos marcadores sobre sus respectivas secuencias del alelo de referencia HG19.

Los cuadros en la figura anterior encierran las secuencias flanqueantes, en verde ambos cebadores del marcador DYS393 (para PCR), en azul las secuencias de anclaje 5' de StraitRazor para ambos marcadores (son idénticas), en amarillo y naranja las secuencias del anclaje 3', no son idénticas pero la secuencia de DXYS267 (12 nt) está contenida en la DYS393 (20 nt), y en rojo dos nucleótidos que al cebador reverse del marcador DYS393 le confiere especificidad.

Este comportamiento que fue considerado para el diseño de cebadores específicos para el marcador DYS393 no fue tenido en cuenta en el diseño de las secuencias de anclaje para StraitRazor, por lo tanto las secuencias de anclaje de DYS393 hacen *match* también en el otro segmento DXYS267, o lo que es lo mismo, las secuencias de anclaje de DYS393 detectaron lecturas de los cromosomas X e Y, en varones, y detectaron lecturas de los cromosomas X en mujeres.

Se detalla en la Figura 19 como el cebador reverse del marcador DYS393 se uniría específicamente al segmento del cromosoma Y. Las dos bases GG en el extremo 3' del cebador es termodinámicamente suficiente para la unión a la secuencia blanco, y el *mismatch* GG/TT con la región del segmento del cromosoma X evita la unión, por consiguiente no amplificaría esa región.

Figura 19. Especificidad del cebador reverse para DYS393, que no hace *match* con la secuencia del marcador DXYS267 debido a un cambio TT/CC entre ambos segmentos.



Sería esperable ante la inespecificidad de DYS393 que las asignaciones alélicas fueran idénticas a las de DXYS267, pero no es así, por dos razones, los valores de *offset* para el cálculo de asignación alélica en ambos marcadores fueron distintos, 3 para DYS393 y 11 para DYS267 (diferencia: 8), lo que explica que muchas de las asignaciones alélicas hechas por DYS393 son 2 repeticiones (8 pb) mayor a las de DXYS267 (ver Figura 23a), la otra razón es que las secuencias de anclaje no fueron del todo idénticas, ambos marcadores comparten la misma secuencia de anclaje 5', pero la secuencia 3' de DYS393 fue de 20 nt y la de DXYS267 fue de 12 nt, que casualmente, la última está incluida en la primera (justo los primeros 12 nt), esto permite evaluar el impacto del largo de las secuencias de anclaje.

Respecto de la herramienta StraitRazor, es importante aclarar entonces que cuando se menciona a DYS393 nos referimos DXYS267 (20 nt - 20 nt) y a DXYS267 nos referimos a DXYS267 (20 nt - 12 nt).

Las siguientes Figuras 20a, 20b y 20c muestran las asignaciones alélicas, conteos de lecturas para StraiRazor y LobSTR respectivamente, restringidos a los dos marcadores.

Figura 20. Conteo de lecturas (profundidad) para los marcadores DYS393 y DXYS267, según ambas herramientas (Cabeceras de las columnas en fucsia las 16 mujeres y en celeste los 11 varones)

a)

MARKER	HG00096	HG01051	HG01112	HG01500	HG01565	HG01583	HG01879	HG03006	HG03742	NA20845	NA12891	NA18525	NA18939	NA19017	NA19625	NA19648	NA20502	HG00268	HG00419	HG00759	HG01595	HG02568	HG02922	HG03052	HG03642	NA12878	NA12892
DXYS267	11,12,	11,	12,	11,	10,11,	10,11,	11,	10,	10,11,	12,13,	11,	10,11,	13,	11,13,	14,	11,13,	13,15,	11,12,	11,13,	10,11,	11,	12,13,	11,12,	11,12,	10,11,	11,12,	12,13,
DYS393	13,14,	13,	13,14,	13,	12,13,	12,13,	13,	12,	12,13,	14,15,	13,	12,13,	15,	13,15,	13,16,	13,15,	15,17,	13,14,	13,15,	12,13,	13,	14,15,	13,14,	13,14,	12,13,	13,14,	14,

b)

MARKER	HG00096	HG01051	HG01112	HG01500	HG01565	HG01583	HG01879	HG03006	HG03742	NA20845	NA12891	NA18525	NA18939	NA19017	NA19625	NA19648	NA20502	HG00268	HG00419	HG00759	HG01595	HG02568	HG02922	HG03052	HG03642	NA12878	NA12892
DXYS267	22	26	24	37	23	15	28	21	21	21	24	20	33	31	16	12	23	22	18	20	21	12	8	26	18	16	20
DYS393	31	38	32	45	30	25	38	33	30	25	8	31	40	41	19	18	28	31	25	29	28	15	15	36	30	6	4

c)

MARKER	HG00096	HG01051	HG01112	HG01500	HG01565	HG01583	HG01879	HG03006	HG03742	NA20845	NA12891	NA18525	NA18939	NA19017	NA19625	NA19648	NA20502	HG00268	HG00419	HG00759	HG01595	HG02568	HG02922	HG03052	HG03642	NA12878	NA12892
DXYS267	4	3	1	1	4	3	4	2	2	1	22	12	9	7	2	2	4	7	5	3	9	1	3	2	0	22	40
DYS393	4	7	5	8	2	2	5	12	6	4	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

La casilla "DYS393" (en rojo) alerta que las asignaciones alélicas para este marcador son incorrectas, serían asignaciones de DXYS267 con la secuencia de anclaje 3' de 20 nt (en vez de

12 nt). Como se mencionó antes, las asignaciones alélicas (Figura 20a) difieren en 2 repeticiones debido a diferencias de parámetros en el archivo config (*offset*). Sólo las muestras HG01112 y NA19625 no coinciden. En ambas aparece un alelo (11) extra, HG01112: 12 **->** 11, 12 (*offset* corregido) y NA19625: 14 **->** 11, 14. Este cambio obedece a conteos diferentes de lecturas, lo que cambia las relaciones del alelo mayoritario vs alelo minoritario con el parámetro de corte: 3 (regla de tartamudeo para una repetición tetra-nucleótido).

Siempre que las herramientas coincidieron, fue indistinto asignar de acuerdo a una u otra herramienta, sin embargo, cuando estas no coinciden, se decidió por la que entregó mayor conteos de lecturas. Esa es la razón por la que hay asignaciones alélicas en DYS393, cuando por ejemplo LobSTR no tuvo detecciones en mujeres, Figura 20c: conteos igual a 0 (comportamiento específico esperable); y se puede apreciar en 19b y 19c que StraitRazor tiene mayores conteos, para ambos DXYS267 y DYS393 (erróneo).

En la Figura 21 se evalúa el impacto en los conteos de acuerdo a diferencias en los parámetros de anclaje: 5’ de 20 nt para ambos (secuencia idéntica) y 3’ de 12 y 20 nt (DXYS267 y DYS393 erróneo, respectivamente).

Figura 21. Conteo de lecturas (profundidad) para los marcadores DYS393 y DXYS267, según ambas herramientas (Cabeceras de las columnas en fucsia las 16 mujeres y en celeste los 11 varones)

MARKER	Auto, X, XY, Y	length anchor 5' (pb)	length anchor 3' (pb)	HG00096	HG01051	HG01112	HG01500	HG01565	HG01583	HG01879	HG03006	HG03742	NA20845	NA12891	NA18525	NA18939	NA19017	NA19625	NA19648	NA20502	HG00268	HG00419	HG00759	HG01535	HG02568	HG02922	HG03052	HG03642	NA12878	NA12892
DXYS267	XY	20	12	22	26	24	37	23	15	28	21	21	21	24	20	33	31	16	12	23	22	18	20	21	12	8	26	18	16	20
DYS393	YY	20	20	31	38	32	45	30	25	38	33	30	25	8	31	40	41	19	18	28	31	25	29	28	15	15	36	30	6	4
Diferencia en conteos de lecuturas				9	12	8	8	7	10	10	12	9	4	-16	11	7	10	3	6	5	9	7	9	7	3	7	10	12	-10	-16
%				40,9	46,2	33,3	21,6	30,4	66,7	35,7	57,1	42,9	19,0	-66,7	55,0	21,2	32,3	18,8	50,0	21,7	40,9	38,9	45,0	33,3	25,0	87,5	38,5	66,7	-62,5	-80,0

Celdas en verde: ganancia de lecturas, celdas en rojo pérdida de lecturas.

Se observa que hay incrementos de conteos de DXYS267 (20 nt - 20 nt) en un % promedio (rango: 18,8 % – 87,5 %) respecto de DXYS267 (20 nt - 12 nt), excepto en aquellas muestras del trío CEPH, que sucede lo contrario, se pierden en promedio un 69,7 % de lecturas (rango: 62,5 - 80,0 %).

Las correcciones finales en estas asignaciones alélicas implicó sólo considerar lo devuelto por lobSTR para el marcador DYS393, en cambio, para el marcador DXYS267 se consideraron ambas herramientas, teniendo en cuenta las asignaciones alélicas hechas por StraitRazor con los parámetros de 5’ = 20 nt y 3’: 20 nt.

4.8.2 Segmentos múltiples

En los apartados anteriores se pudo observar que para aquellos marcadores que pertenecen a segmentos múltiples se obtienen detecciones inespecíficas con ambas herramientas. Los mejores ejemplos son los marcadores DYS490, DYS446 y DYS572 (eliminados originalmente en el config de StraitRazor) pero que también tuvieron detecciones inespecíficas en lobSTR.

En la Tabla 10 se destacan las principales características de todos los marcadores en estudio comprendidos en segmentos múltiples.

Tabla 10. Detalle de los marcadores de este estudio que pertenecen a segmentos múltiples.

Marcador	UID	Cromosomas	Intra ó inter cromosómico	Cantidad de segmentos extras [1]
DYS19	25602	Y	Intra	1
HPRTB	11012	X, 15	Inter	6
DYS434	25491	Y, X	Inter	1
DXYS156	25539	Y, X	Inter	1
DYS393	25539	Y, X	Inter	1
DYS572 [2]	25541	Y, X	Inter	1
7p143AAAG	22463	7	Intra	1
DYF387S1	25774	Y	Intra	1
D9S238 [2]	8925	9, 14, 2, 15, 18, 21	Intra e Inter	6 (1,5) [3]
DYS505	25541	Y, X	Inter	1
DYS425	25684	Y	Intra	3
DYS435	25489	Y, X	Inter	1
DYS522	25498	Y, X	Inter	1
DYS406S1	25634	Y	Intra	1
DYS389II	25484	Y, X	Intra e Inter	2 (1,1)
DYS456	25542	Y, X	Inter	1
TYW1_I_13 [2]	22922	7	Intra	1
DYS395S1	6940	Y, 12	Intra e Inter	2 (1,1)
DYS568	25630	Y	Intra	3
RBAK-RBAKDN_I_2 [2]	22346	7	Intra	18
DYS459	25774	Y	Intra	1
DYS487	25633	Y	Intra	1
DYS413	25515	Y, X	Intra e Inter	5 (1,4)
DYS385	25710	Y	Intra	1
DYS446 [2]	25539	Y, X	Inter	1
DYS389I	25484	Y, X	Intra e Inter	2 (1,1)
DYS389-2	25484	Y, X	Intra e Inter	2 (1,1)
DYS490 [2]	25540	Y, X	Inter	1
DYS394	25602	Y	Intra	1
DYS464	20140	Y, 3	Intra e Inter	4 (3,1)
DYS437	25491	Y, X	Inter	1
DYS640	25540	Y, X	Inter	1
DXYS267	25539	Y, X	Inter	1
DYS391	25496	Y, X	Inter	1
YCAII	6936	Y, 12, chrUn_gl000232	Intra e Inter	4 (1,3)

1 El valor expresa el número de segmentos extras (1: duplicación, 2: triplicación, etc.)
2 No incluidos en el archivo config v183 de StraitRazor.
3 Entre paréntesis la cantidad de segmentos intra e inter cromosómicos, en ese orden.

De estos marcadores, se puede apreciar que la mayoría pertenecen al cromosoma Y, y estos a su vez comparten segmentos con el cromosoma X. También se observa que en su mayoría son duplicaciones (valor 1 en la última columna). El evento de segmento múltiple está repartido entre inter e intra-cromosómicas para estos marcadores. La inespecificidad evidenciada en los análisis exploratorios es evidente para los marcadores D9S238 (6 segmentos) y RBAK-RBAKDN_I_2 (18 segmentos), eliminados del archivo de configuración de StraitRazor, pero no es evidente para los marcadores DYS572, TYW1_I_13, DYS446 y DYS490 (también eliminados del config) con solo un segmento duplicado.

Además, de un total de 35 marcadores de la Tabla 10, sólo los marcadores mencionados, e incluyendo a DYS425, DYS413, y DYS490 (9 marcadores en total) tuvieron distintas instancias de asignación alélica problemática. Lo que no ocurrió con los restantes 25 marcadores, lo que sugiere que un marcador incluido en segmentos múltiples no siempre incurre en asignaciones alélicas inespecíficas. En ese sentido, de estos 25 marcadores DYS391, 7p143AAAG, DYS19, DYS391, DYS395S1, DYS406S1, DYS435, DYS437, DYS456, DYS459 y DYS640 tuvieron los mejores comportamientos, comparables a aquellos marcadores que no pertenecen a segmentos múltiples. En el apartado 4.9 se analizan estos comportamientos de los 323 marcadores estudiados, donde se deja evidencia de las decisiones tomadas para las asignaciones alélicas automáticas.

4.8.3 Evaluación de especificidad mediante los marcadores del cromosoma X en varones

Para la evaluación de los STRs del cromosoma X se usaron 18 marcadores detectados en los 11 varones (escenario “d”), se excluyó HPRTB por pertenecer a un segmento múltiple (UIDs: 11012, 11016, 11018, 11019, 11020 y 11021). En la Figura 22 se muestran las 198 asignaciones alélicas posibles (18 x 11 varones) y la profundidad que sustenta esa detección, según lobSTR y StraitRazor.

Figura 22. Asignación alélica y profundidad para los marcadores del cromosoma X en los 11 varones analizados: a) lobSTR y b) StraitRazor.

a)

MARKER	purity	period	Auto, X, XY, Y	HG00096 male	HG01051 male	HG01112 male	HG01500 male	HG01565 male	HG01583 male	HG01879 male	HG03006 male	HG03742 male	NA20845 male	NA12891 male
DXS10074	94	4	XX	9, (12)	17, (5)	8.2, 9, (3.5-9.5)	9, (19)	19, (6)	16, (2)	9, (11)	17, (9)	16, (9)	17, (3)	18, (4)
DXS10079	95	4	XX	20, (14)	19, (5)	20, (6)	17, 18, (1 - 3)	22, (5)	19, (4)	19, (11)	16, (17)	21, (7)	17, (4)	ND (0)
DXS10103	90	4	XX	18, (4)	20, (2)	19, (4)	20, (8)	16, (10)	16, (2)	20, (4)	ND (0)	16, (6)	20, (2)	18, (20)
DXS10135	83	4	XX	ND (0)	ND (0)	19.1, (1)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)
DXS10148	88	4	XX	22, (2)	ND (0)	27.1, (6)	26.1, (4)	ND (0)	ND (0)	32.2, (2)	30.1, (2)	ND (0)	24, (4)	ND (0)
DXS1187	94	4	XX	16, (16)	16, (8)	16, (10)	19, (11)	17, (6)	19, (1)	17, (10)	18, (10)	17, (15)	18, (9)	17, (2)
DXS6789	92	4	XX	ND (0)	ND (0)	20, (2)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)
DXS6803	100	4	XX	13, (8)	10, 11, (1-4)	11, (12)	11, (12)	13.3, (3)	12, (3)	12.3, (8)	12, (6)	12, (13)	12, (2)	11, (28)
DXS6809	91	4	XX	ND (0)	32, (1)	35, (1)	ND (0)	34, (1)	ND (0)	ND (0)	ND (0)	37, (1)	ND (0)	ND (0)
DXS7132	91	4	XX	12, (17)	13, (10)	12, (8)	14, (13)	13, (10)	14, (7)	13, (15)	14, (13)	15, (10)	13, (6)	15, (8)
DXS7133	100	4	XX	9, (22)	9, (10)	9, (15)	11, (21)	9, (12)	9, (4)	9, (19)	9, (16)	11, (16)	9, (10)	9, (34)
DXS7423	92	4	XX	15, (6)	15, (4)	15, (5)	15, (6)	ND (0)	16, (1)	14, (4)	16, (5)	14, (5)	ND (0)	16, (4)
DXS8377	95	6	XX	ND (0)	ND (0)	19.3, (8)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)
DXS8378	100	4	XX	11, (11)	12, (6)	12, (5)	10, (6)	11, (4)	13, (1)	11, (8)	11, (6)	11, (6)	11, (2)	9, (14)
DXS9898	78	4	XX	ND (0)	ND (0)	11, (2)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)
DXS9902	100	4	XX	11, (16)	11, (9)	11, (10)	11, (11)	12, (11)	ND (0)	13, (11)	11, (14)	12, (13)	13, (2)	10, (22)
GATA172D05	94	4	XX	ND (0)	ND (0)	6, (6)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)	ND (0)
GATA31E08	100	4	XX	7, (11)	10 (14)	9, (9)	7, (16)	10, (8)	6, (7)	8, (16)	7, (16)	7, (12)	11, (6)	10, (16)

b)

MARKER	purity	period	Auto, X, XY, Y	HG00096 male	HG01051 male	HG01112 male	HG01500 male	HG01565 male	HG01583 male	HG01879 male	HG03006 male	HG03742 male	NA20845 male	NA12891 male
DXS10074	94	4	XX	9, (17 - 0)	17, (6 - 0)	8.2, 9, (1 - 16)	9, (17 - 0)	19, (12 - 0)	16, (7 - 0)	9, (11 - 0)	17, (6 - 0)	16, (10 - 0)	17, (7 - 0)	ND (0 - 0)
DXS10079	95	4	XX	20, (19 - 0)	19, (9 - 0)	20, (4 - 0)	17, 18, (1 - 3)	22, (7 - 0)	1.2, 19, (1 - 10	19, (14 - 0)	16, (20 - 0)	21, (10 - 0)	17, (5 - 0)	ND (0 - 0)
DXS10103	90	4	XX	17, 18, (1 - 7)	20, (12 - 0)	19, (12 - 0)	19, 20, (1 - 16)	16, (16 - 0)	16, (7 - 0)	20, (14 - 0)	16, (10 - 0)	16, (13 - 0)	20, (6 - 0)	ND (0 - 0)
DXS10135	83	4	XX	24, (5 - 0)	20, 21, (3 - 2)	24, (7 - 0)	18, (6 - 0)	24, (5 - 0)	28, (2 - 0)	27, (3 - 0)	19, (1 - 0)	24, (2 - 0)	25, (3 - 0)	ND (0 - 0)
DXS10148	88	4	XX	22, (1 - 0)	ND (0 - 0)	ND (0 - 0)	26.2, (1 - 0)	ND (0 - 0)	ND (0 - 0)	32.3, (1 - 0)	30, 30.2, (1 - 1)	ND (0 - 0)	ND (0 - 0)	ND (0 - 0)
DXS1187	94	4	XX	16, (20 - 0)	16, (24 - 0)	16, (25 - 0)	19, (11 - 0)	17, (13 - 0)	19, (6 - 0)	17, (16 - 0)	18, (13 - 0)	17, (23 - 0)	18, (16 - 0)	ND (0 - 0)
DXS6789	92	4	XX	22, (10 - 0)	20, 21.2, (12 - 1)	20, (4 - 0)	19, (15 - 0)	21, (14 - 0)	19, (6 - 0)	20, (19 - 0)	19, (13 - 0)	19, (16 - 0)	20, (7 - 0)	ND (0 - 0)
DXS6803	100	4	XX	13, (16 - 0)	10, 11, (1-11)	11, (25 - 0)	11, (23 - 0)	13.3, (18 - 0)	12, (21 - 0)	12.3, (29 - 0)	12, (17 - 0)	12, (27 - 0)	12, (17 - 0)	11, (22 - 0)
DXS6809	91	4	XX	37, (1 - 0)	32, (7 - 0)	35, (4 - 0)	36, (3 - 0)	34, (3 - 0)	36, (6 - 0)	34, (8 - 0)	35, (5 - 0)	36, 37, (2 - 6)	35, (3 - 0)	ND (0 - 0)
DXS7132	91	4	XX	12, (10 - 0)	13, (9 - 0)	12, (2 - 0)	14, (9 - 0)	13, (7 - 0)	14, (5 - 0)	13, (10 - 0)	14, (7 - 0)	15, (12 - 0)	13, (4 - 0)	ND (0 - 0)
DXS7133	100	4	XX	9, (29 - 0)	9, (16 - 0)	9, (27 - 0)	11, (21 - 0)	9, (23 - 0)	9, (18 - 0)	9, (28 - 0)	9, (26 - 0)	11, (23 - 0)	9, (17 - 0)	9, (24 - 0)
DXS7423	92	4	XX	15, (3 - 0)	15, (5 - 0)	15, (3 - 0)	15, (2 - 0)	ND (0 - 0)	16, (4 - 0)	14, (4 - 0)	16, (5 - 0)	14, (5 - 0)	16, (2 - 0)	ND (0 - 0)
DXS8377	95	6	XX	21, (4 - 0)	19.3, (4 - 0)	19.3, (6 - 0)	20, (1 - 0)	20, (3 - 0)	ND (0 - 0)	21, (2 - 0)	ND (0 - 0)	20.3, (4 - 0)	ND (0 - 0)	ND (0 - 0)
DXS8378	100	4	XX	11, (9 - 0)	12, (8 - 0)	12, (7 - 0)	10, (5 - 0)	11, (12 - 0)	13, (8 - 0)	11, (13 - 0)	11, (11 - 0)	11, (8 - 0)	11, (5 - 0)	ND (0 - 0)
DXS9898	78	4	XX	11, (13 - 0)	10, (12 - 0)	11, (6 - 0)	6.3, (8 - 0)	10, (4 - 0)	6.3, (6 - 0)	10, (11 - 0)	11, (12 - 0)	9, (9 - 0)	6.3, (6 - 0)	ND (0 - 0)
DXS9902	100	4	XX	11, (22 - 0)	11, (18 - 0)	11, (19 - 0)	11, (11 - 0)	12, (20 - 0)	11, (10 - 0)	13, (14 - 0)	11, (20 - 0)	12, (23 - 0)	13, (14 - 0)	10, (22 - 0)
GATA172D05	94	4	XX	8, (15 - 0)	11, (14-0)	6, (13-0)	12, (11-0)	6, (17-0)	10, (13-0)	10, (28-0)	6, (22-0)	10, (36-0)	10, (19-0)	10, (18-0)
GATA31E08	100	4	XX	7, (14 - 0)	9,10, (1-20)	9, (18-0)	4.1,7, (1-21)	10, (14-0)	6, (19-0)	8, (22-0)	7, (23-0)	7, (19-0)	11, (12-0)	10, (10-0)

Se destacan en las celdas las siguientes situaciones:
Verde: ambas herramientas coinciden y no tienen más de una asignación alélica.
Amarillo: hay más de una asignación alélica para esta herramienta.
Naranja: hay una asignación alélica para esta herramienta y sin detecciones en la otra.
Blanco: hay una asignación alélica para esta herramienta pero no coincide con la otra.
Rojo: sin detecciones en esta herramienta.

La Figura 23 resume lo que se observa en la Figura 22:

Figura 23. Resumen del análisis hecho con marcadores del cromosoma X en 11 varones.

HERRAMIENTA	StraiRazor	LobSTR
TOTAL	198	
NO DETECTADOS	23	67
DETECTADOS	175	131
MÁS DE UN ALELO ASIGNADO	12	3
% VERSUS DETECTADOS	6,9	2,3
UN ALELO ASIGNADO Y AMBAS HERRAMIENTAS COINCIDEN	110	110
% VERSUS DETECTADOS	62,9	84,0
UN ALELO ASIGNADO PERO NO COINCIDE CON LA OTRA HERRAMIENTA	3	10
% VERSUS DETECTADOS	1,7	7,6
UN ALELO ASIGNADO PERO SIN DETECCIONES CON LA OTRA HERRAMIENTA	50	8
% VERSUS DETECTADOS	28,6	6,1

Del cuadro anterior se puede afirmar que ambas herramientas detectan alelos (y los asignan) erróneamente de igual manera: 3 de 131 (2,3 %) de lobSTR versus 12 de 175 (6,9 %) de StraitRazor, P=.118, α = 0.05; en cambio, StraitRazor logra mayores detecciones: 175 de 198 (88,4%) contra 131 de 198 (66,2%) de lobSTR, P < .001, α = 0.05.

Este análisis deja evidencia del funcionamiento de las herramientas por separado, pero muestra también que hay distintos grados de concordancias entre las asignaciones alélicas de ambas herramientas. Estos marcadores analizados son tetra-nucleótidos (excepto DXS8377 que es hexa-nucleótido) y fueron usados debido a dos situaciones relevantes:

a) los tartamudeos son menos frecuentes con este largo de repeticiones (4 nt, 5 nt y 6 nt).

b) las diferencias de asignación alélica no es atribuible al largo de la repetición.

A continuación se detalla entre paréntesis el largo en pb del alelo de referencia (hg19) y la pureza en porcentaje calculada con TRF:

Se observó que los marcadores DXS1187 (128 pb, 94%), DXS6803 (85 pb, 100%), DXS7132 (203 pb, 91%), DXS7133 (93 pb, 100%) y DXS9902 (92 pb, 100%) tuvieron óptimas detecciones concordantes entre LobSTR y StraitRazor, por lo contrario, los marcadores DXS10135 (251 pb, 83%), DXS10148 (202 pb, 88%), DXS6789 (103 pb, 92%), DXS6809 (212 pb, 91%) y DXS9898 (224 pb, 78%) fueron las menos concordantes.

Si bien no es concluyente, porque hay otros factores intervinientes, los datos sugieren que a menor longitud del alelo a detectar y mayor pureza de la región STR es esperable un comportamiento más concordante entre ambas herramientas.

Entonces, habiendo visto tasas de error comparables en ambos escenarios, se deduce que las diferencias de asignación alélica provienen de otros factores, por lo tanto surgió la

necesidad de realizar una revisión de las asignaciones alélicas hechas en “perfiles-finales.xlsm” (punto 3.6.6 y Anexo VI) en un nuevo archivo “perfiles-finales-v2.xlsx” (punto 3.6.7 y Anexo VI), para evaluar las 8.721 asignaciones alélicas. Se discute a continuación.

4.9 Revisión de las asignaciones alélicas.

Las asignaciones alélicas automáticas y correcciones manuales realizadas y resumidas dentro del archivo “perfiles-finales.xlsm” permitió un primer pantallazo, y todo lo derivado de este archivo fue agregado a la base de datos local. Sin embargo, por todo lo expuesto en los apartados anteriores fue necesario establecer las causas que llevó a las herramientas a no coincidir, principalmente en aquellas asignaciones alélicas con buenos conteos de lecturas (> 10).

El archivo “perfiles-finales-v2.xlsx” fue diseñado para responder lo siguiente:

a) Impacto del conteo de lecturas (sólo se consideran los dos mayoritarios).
b) Impacto de los cortes para tartamudeos (ver reglas especiales en el punto 4.7.1 y reglas para la evaluación de las asignaciones alélicas, ver 3.6.7).
c) Comportamiento global mediante puntajes de las asignaciones alélicas según marcador o muestra (ver punto 3.6.7.1).
d) Eventos que produjeron que determinadas asignaciones alélicas incurrieran en reglas con criterios de menor performance (reglas 4 a la 12)

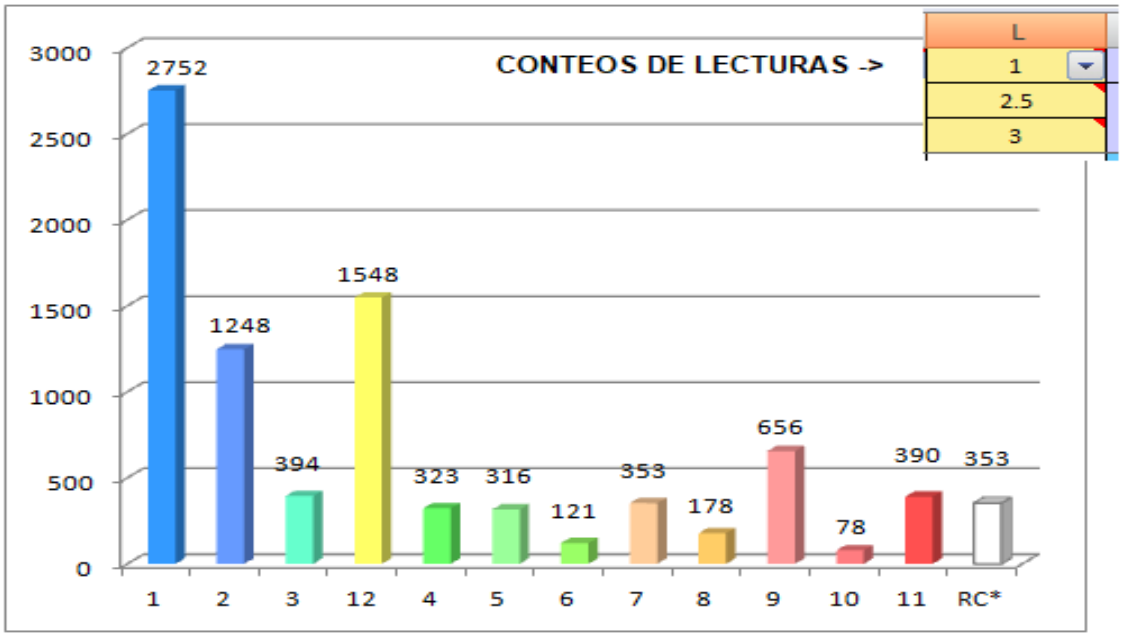
4.9.1 Impacto del conteo mínimo de lecturas

Las asignaciones alélicas hechas en el archivo “perfiles-finales.xlsm” tienen un corte de conteos de lecturas de 1 a partir de la cual puede ser considerada, para ambas herramientas.

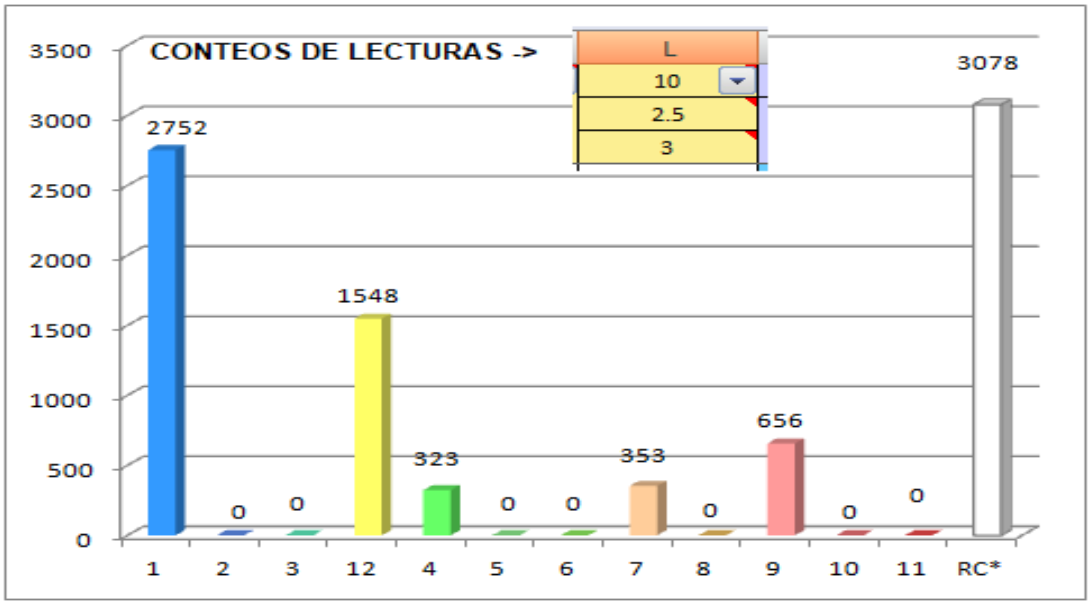
Se decide evaluar entonces, cuántas asignaciones alélicas se pierden si ese corte se hace más elevado. En el archivo “perfiles-finales-v2.xlsx” aquellas asignaciones alélicas que respondieron a las reglas 1, 4, 7, 9 y 12 no son sensibles de la modificación del mínimo de conteos de lecturas, debido a que estas reglas establecen que los conteos de lecturas sean iguales o mayores a 10, en cambio, las reglas restantes 2, 3, 5, 6, 8, 10 y 11 si están afectadas. En la Figura 24 se puede apreciar el cambio en número de asignaciones alélicas afectadas ante la modificación del mínimo de conteos de lecturas.

Figura 24. Impacto del conteo mínimo de lecturas.

a)



b)



RC*: Requiere corrección.

En la ordenada el número de asignaciones alélicas y sobre la abscisa las reglas 1 a 12 y RC*.

La barra “Requiere corrección” son aquellas asignaciones alélicas que no respondieron a ninguna de las 12 reglas (353, 4,05 % del total). Entonces, cuando se incrementó el mínimo de conteos de lecturas de 1 a 10, 2.725 asignaciones alélicas realizadas bajo las reglas afectadas

se incorporaron a los datos de “Requiere corrección” (31,25 % del total). De estas 2.725 asignaciones alélicas perdidas, 1.642 (60,26 %) corresponden a las reglas 2 y 3, consideradas aun de buena performance (ambas herramientas coinciden, pero con conteos de lecturas menores a 10).

En la confección inicial de los primeros perfiles de las muestras (analysis_v2_lobSTR_str8rzr_[SRR129…..]_v183_v32.xlsx, archivos no descritos en Materiales y Métodos) se hizo evidente que los conteos de lecturas eran bajos: lobSTR 11,74 ± 13,85 y StraitRazor 15,37 ± 13,64. Por ese motivo se decidió que el mínimo de conteos de lecturas empiecen desde el valor 1. Las 1.642 asignaciones alélicas donde las herramientas coinciden justifican este valor. Está claro que esta decisión es global, eventualmente el valor de corte puede variar según el marcador o la muestra, no se hizo esa distinción, dado que *a priori* se desconoce cómo respondería una, otra o ambas herramientas, con las muestras y marcadores elegidos.

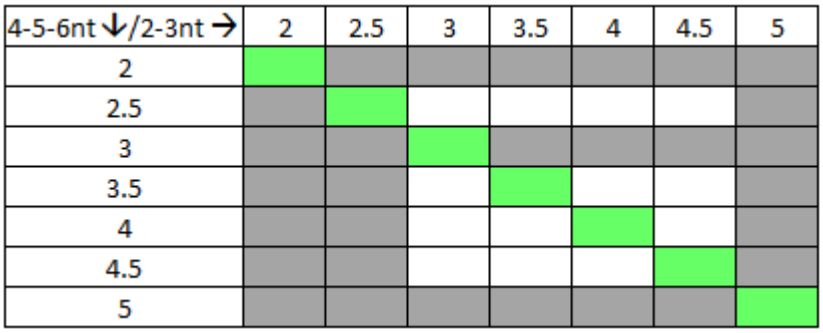
4.9.2 Impacto de la reglas para tartamudeos

Cuando se decidió incorporar lobSTR como segunda herramienta para efectuar las asignaciones alélicas, se hizo debido a que StraitRazor entregaba en muchos marcadores (análisis exploratorio) más de dos asignaciones alélicas (para marcadores de cromosomas autosómicos, pseudo-autosómicos o del cromosoma X en mujeres) ó más de una (para marcadores del cromosoma X e Y en varones). Incluso corrigiendo los valores del archivo config (*offset*, secuencia *5'Anchor*, secuencia *3'Anchor* y *Motif*) muchos marcadores analizados con StraitRazor continuaban con el mismo comportamiento. Por lo tanto fue necesario incluir en los análisis los conteos de lecturas de los alelos detectados. Se decidió establecer un corte, un tanto arbitrario, pero que respondió a lo observado en los primeros perfiles obtenidos con las primeras correcciones, y al hecho de que mientras más larga es la secuencia de la repetición, menor es el porcentaje de tartamudeo, se sabe que para STRs cuya repetición es de 4 nt, este porcentaje no supera el 20 %. Es decir, una relación 5 a 1 del alelo con mayores conteos respecto del que le sigue en conteos. El corte definido para STRs de 4 nt, 5 nt y 6 nt fue de 3, es decir, se considera que los tartamudeos están por debajo del 33 % del mayoritario. Superando este umbral, se considera segundo alelo. Para aquellos STRs de 2 nt y 3 nt el corte fijado fue de 2.5, un umbral de 40 %, más exigente por ser repeticiones cortas.

En el archivo “perfiles-finales-v2.xlsx” ese corte se hizo variable, asociando los 27 perfiles de manera dinámica (los 27 archivos Excel están vinculados a “perfiles-finales-v2.xlsx”). De esta manera, las celdas L2 y L3 albergan los valores de corte para STRs de 2nt-3nt y 4nt-5nt-6nt respectivamente. Los rangos de valores evaluados son de 2 a 5. La Figura 25 resume lo observado.

Figura 25. Número de asignaciones alélicas según las 12 reglas y bajo distintas combinaciones de rangos de cortes.

a)



4-5-6nt ↓/2-3nt →	2	2.5	3	3.5	4	4.5	5
2							
2.5							
3							
3.5							
4							
4.5							
5							

b)

	Combinaciones de cortes								
2-3nt →	2	2.5	3	3.5	4	4.5	5	Diferencia: cortes 5 y 2	Asign. alélicas
4-5-6nt →	2	2.5	3	3.5	4	4.5	5		
Condiciones ↓									
1	2503	2721	2799	2844	2868	2873	2881	378.0	671
2	1120	1234	1276	1312	1322	1338	1340	220.0	
3	365	389	404	430	436	438	438	73.0	
12	1548	1548	1548	1548	1548	1548	1548	0.0	
4	500	344	283	256	239	232	229	-271.0	-507.0
5	424	327	291	261	255	241	239	-185.0	
6	140	124	111	95	90	89	89	-51.0	
7	425	363	346	328	321	323	318	-107.0	-156.0
8	198	181	175	169	165	163	163	-35.0	
9	656	656	656	656	656	656	656	0.0	
10	83	79	78	71	71	69	69	-14.0	
11	390	390	390	390	390	390	390	0.0	
RC*	358	354	353	350	349	350	350	-8.0	-8.0

a) En gris todas combinaciones de cortes para ambos grupos de STRs (2nt-3nt y 4nt-5nt-6nt) y verde un grupo de combinaciones elegidos.

b) Variaciones del número de asignaciones alélicas según combinaciones de cortes 2-2, 2.5-2.5, 3-3, 3.5-3.5, 4-4, 4.5-4.4 y 5-5.

La ganancia de 671 asignaciones alélicas para las reglas 1, 2 y 3 cuando el corte se pasó de 2 a 5 (para ambos grupos de STRs) se debió principalmente a una pérdida de asignaciones alélicas de las reglas 4, 5 y 6 (507). Estas combinaciones de cortes tuvieron los cambios más pronunciados. Pero en los otros grupos de combinaciones de cortes se observó un comportamiento similar. De esto se desprende que, cuando se pasó de un corte estricto a uno más tolerante respecto del porcentaje de aparición de tartamudeos (de 50 % a 20 %), muchas asignaciones alélicas pasaron de estar en el grupo “ambas herramientas comparten un alelo” al grupo de “ambas herramientas” coinciden. Esta mejora puede tener un costo, y es que ambas herramientas podrían estar detectando segundos alelos que en realidad son tartamudeos, pero se desconoce en qué grado se comete este error. Eventualmente los cortes de 2.5 o 3 para STRs

de 2 nt y 3nt, por un lado, y de 3 o 3.5 para STRs de 4 nt, 5 nt y 6 nt, por el otro, sean más preferibles, y tal vez dependa de cada situación modificar estos cortes.

De nuevo, estos cortes fueron manejados de manera global, eventualmente cada STR puede tener un comportamiento diferente ante la posibilidad de tartamudeos. Por ejemplo, para STRs tetra-nucléotidos usados en ciencias forenses, los porcentajes de tartamudeos varían de 7 a 18 %. Estos valores no son fácilmente aplicables, la cantidad de moléculas obtenidas por PCR son absolutamente superiores a las pocas moléculas detectadas por NGS.

Es oportuno aclarar que ambas herramientas fueron diseñadas para estrategias de secuenciación muy diferentes, StraitRazor apunta a secuenciaciones de profundidades elevadas y acotado a marcadores de ciencias forenses (cuyas librerías amplifican sólo las regiones de interés) y lobSTR puede funcionar con secuenciaciones de menor profundidad, con mayor cobertura horizontal, y como se discutió en el apartado 4.7.4, el tipo de librería impacta en la respuesta de la herramienta. Veremos más fortalezas y desventajas en la Tabla 11.

4.9.3 Puntajes

Aprovechando que para cada asignación alélica hubo una regla aplicada, se usó ese valor para establecer un puntaje para los marcadores y las muestras. Esto se hizo sólo a los fines de establecer qué marcadores o muestras respondieron mejor a la doble asignación alélica, y cuales estuvieron en la peor situación.

Los primeros 10 marcadores con la mejor puntuación son:

D4S2408 (29), SLC7A7_I_6 (30), D18S391 (33), D9S301 (33), ZCCHC7_I_5 (33), D21S1809 (35), TPOX (35), D9S1118 (36), D18S858 (38) y AFMB313ZH5 (39).

Los últimos 10 marcadores con la peor puntuación son:

D3S1358 (249), DXS8377 (251), D13S258 (252), DXS10135 (252), D7S1529 (253), DXS9898 (253), D18S386 (258), FAH_I_1 (263), D2S1338 (269) y D9S302 (324).

Para los marcadores del cromosoma Y, el mejor puntaje esperado es 11, y son considerados aparte de los marcadores autosómicos, pseudo-autosómicos y exclusivos del cromosoma X, debido a que se excluyen las asignaciones alélicas hechas en mujeres (donde predomina la regla 12, y así debe ser).

Los primeros 10 marcadores del cromosoma Y con la mejor puntuación son:

DYS459***** (13), DYS556 (15), DYS395S1***** (16), DYS537 (16), DYS578 (16), DYS594 (16), DYS638 (16), DYS445 (18), DYS472 (18) y DYS485 (19). (Identificados con asterisco "*****" son marcadores que pertenecen a segmentos múltiples, ver punto 4.8.2).

Los últimos 10 marcadores del cromosoma Y con la peor puntuación son:

DYS441 (72), GATA-A10 (87), DYS612 (92), DYS461 (104), DYS449 (105), DYS511 (106), DYS520 (109), DYS460 (111), DYS389-I (116) y DYS442 (129).

Para la puntuación de las muestras se descartaron las asignaciones alélicas de los marcadores del cromosoma Y debido a que sesgan los valores a favor o en contra dependiendo de si la muestra pertenece a un hombre o a una mujer, respectivamente.

Las 10 primeras muestras con la mejor puntuación son (puntaje óptimo: 323):

NA20502 (mujer, 644), HG00268 (mujer, 645), HG01112 (hombre, 666), NA18525 (mujer, 667), HG00096 (hombre, 681), NA19017 (mujer, 695), NA18939 (mujer, 699), HG00759 (mujer, 705), HG01500 (hombre, 714) y NA19648 (mujer, 731)

Las 10 últimas muestras con la peor puntuación son:

HG01051 (hombre, 818), NA19625 (mujer, 875), HG02922 (mujer, 886), HG03642 (mujer, 894), NA20845 (hombre, 915), HG02568 (mujer, 990), HG01583 (hombre, 1028), NA12891 (hombre, 1511), NA12892 (mujer, 1570) y NA12878 (mujer, 1576). Las tres últimas del trío CEPH.

4.9.4 Principales eventos que condujeron a inconsistencias entre lobSTR y StraitRazor

En los apartados 4.7.4 y 4.7.5 se discutió el comportamiento de las herramientas cuando las asignaciones alélicas fueron unilaterales ó de manera conjunta. Y en los apartados 4.7.6 y 4.7.7 se establecieron aquellos marcadores con muchas asignaciones alélicas discordantes ó con muchas intervenciones manuales por parte del analista.

Con el archivo “perfiles-finales-v2.xlsx” y aplicando las 12 reglas se pudo discriminar a qué regla estaba respondiendo cada asignación alélica, y eventualmente observar por marcador a lo largo de las 27 muestras algún predominio de determinada regla.

Por ejemplo: el marcador D16S2624 posee 24 muestras cuyas asignaciones alélicas responden a la regla 7 y 3 muestras a la regla 9. Es decir, en todas las asignaciones alélicas las herramientas no coinciden y no comparten alelos y con buen conteo de lecturas (regla 7) y similar situación para la regla 9 (pero en este caso las herramientas no coinciden porque una de las herramientas no tiene detecciones).

Concretamente lobSTR estaba asignando con una fracción de repetición menos (-0.3), esto significa que tres nucleótidos están generando la diferencia, y precisamente existe un SNP (rs67810177: -/CTA) dentro del núcleo de este STR que lobSTR lo consideró para el cálculo de asignación alélica, pero esta secuencia fue descartada cuando se corrigió el valor de *offset* del config versión v183 de StraitRazor. Este error no es precisamente de la herramienta lobSTR, sino del alelo de referencia del ensamblaje HG19. Y la corrección con StraitRazor se hizo con las primeras secuencias obtenidas (análisis exploratorio). El alelo de referencia tiene la secuencia CTA, siendo que por la frecuencia informada no debería estar (0,988618 para “-“ y 0,011382 para “CTA”). Esto cambiaría el estado de deleción a inserción (atributo “CLASS” en snp150 y snp151).

Otro ejemplo y que también obedece a asignaciones alélicas bajo la regla 7 (19 muestras y 8 muestras bajo la regla 8) es el marcador FGD6_I_13, en esta caso fue un error de configuración del archivo BED de lobSTR, donde la región del genoma “hg19.fa” considerada fue: chr12: chr12:95500670-95500699 y debió haber sido: chr12:95500670-95500711, por un lado, y en la configuración del archivo config de StraitRazor no se consideraron 2 repeticiones dentro de esa región, además el alelo de referencia contiene un SNP (rs72166124: -/AA) que,

por lo observado en las secuencias (finales) es de baja frecuencia, sólo aparece en el alelo 6 (alelo 8 corregido), esta inserción es considerada como interrupción.

Para indagar sobre estos errores se usó todo lo realizado para este trabajo:

a) Scripts de Perl: get-anchors-considering-snp150f-argv-todos2.pl, chek-lobSTR-str8rzr-by-sample-by-marker6.pl, chek-lobSTR-str8rzr-by-analisisID-by-marker.pl, generate_planilla_structure_argv.pl.
b) Comandos de Linux: cat STRs_en_estudio_v183*.txt | grep (marcador) | grep -W (algún valor de alelo, o todos) guardados en un archivo y convertido en formato fasta, alineados eventualmente con BioEdit (ClustalW),
c) Comandos de Linux: cat (análisis ID de alguna muestra).vcf | grep (posición de comienzo del núcleo para algún marcador).
d) Comandos de Linux: mysql --user=genome --host=genome-mysql.soe.ucsc.edu -A -P 3306, use hg19, SELECT * FROM SNP150 ó SNP151 WHERE name="(nombre #rs de algún SNP)".
e) Carga del archivo BAM de los análisis ID (ambos) de determinada muestra, y la pista "5-321markers.bed" en IGV.
f) Archivos de Word donde se hizo la inspección para confeccionar el archivo config v183 ([marcador]_structure_2.docx)
g) Archivos en Excel derivados de "perfiles-finales-v2.xlsx": "ERRORES_lobSTR_STRAITRAZOR.xlsx" y "PUNTAJES.xlsx".
h) Por último y no menos importante, a la BD local mediante FIREFOX (ver apartado 6.12), accediendo principalmente a la tabla "319markers".

En el archivo "ERRORES_lobSTR_STRAITRAZOR.xlsx" dentro de la hoja "PUNTAJE_MARCADORES" se puede acceder al listado completo de los marcadores y bajo qué reglas se hicieron las asignaciones alélicas, algunos ejemplos (entre paréntesis el número de muestras involucradas):

Regla 4: NBPF9_I_21 (13), YY1AP1_I_7 (10), DYS425 (10), DUSP22_I_5 (9) y WDR7_I_22 (6).

Regla 5: YY1AP1_I_7 (11), ITSN2_I_18 (10), USP37_I_19 (8), DXYS267 (8) y FAN1_I_9 (7).

Regla 6: PPM1L_I_1 (7), RH61194 (4), SHKBP1_I_4 (4), 7p143TTCC (4), KDM3A_I_10 (4), D21S1414 (4).

Regla 7: D21S1446 (25), D16S2624 (24), FGD6_I_13 (19), TTC7B_I_14 (11) y D7S821 (11).

Regla 8: D12S391 (10), D13S305_2 (10), D13S742 (9), UBR4_I_74 (9) y D22S683 (8).

Regla 9: GATA172D05 (25), AMELOGENIN (25), D18S535 (23), D15S657 (22), D9S1122 (22) y D15S822 (22).

4.10 Estructura y nomenclatura de las variantes alélicas

Como se mencionó en los apartados 1.2 y 1.2.1, la estructura de un STR es un atributo propio de esta variante, y definirla correctamente implica necesariamente conocer las secuencias de las variantes alélicas. En este trabajo se obtuvo en algunos marcadores una cantidad suficiente de secuencias de las variantes alélicas detectadas por StraitRazor

En la Figura 26 se muestra dos ejemplos de las estructuras de dos STRs: NBPF9_I_21 (repetición dinucleótido, Figura 26a) y D7S3057 (repetición tetranucleótido, Figura 26b).

Figura 26. Estructuras obtenidas con secuencias provistas por StraitRazor de NBPF9_I_21 y D7S3057.

a)

NBPF9_I_21

STR dinucleótido GT con una interrupción GA

Alelos:
14, 16, 17, 18, 19, 20, 21, 27.

Estructura de las variantes alélicas (número de secuencias):

14: [GT]$_6$ GA [GT]$_8$ (7)
16: [GT]$_6$ GA [GT]$_{10}$ (2)
17: [GT]$_6$ GA [GT]$_{11}$ (8)
17: [GT]$_5$ GA [GT]$_{12}$ (5)
18: [GT]$_6$ GA [GT]$_{12}$ (1)
19: [GT]$_6$ GA [GT]$_{13}$ (2)
20: [GT]$_6$ GA [GT]$_{14}$ (8)
21: [GT]$_6$ GA [GT]$_{15}$ (1)
27: [GT]$_6$ GA [GT]$_{21}$ (1)

Estructura general:

[GT]$_{5\text{-}6}$ GA [GT]$_{8\text{-}10\text{-}11\text{-}12\text{-}13\text{-}14\text{-}15\text{-}21}$

b)

D7S3057

STR tetranucleótido [ATCT] y [TCCA] con interrupciones ATC, CTTTG y CCCA
(este STR requiere corrección en las asignaciones alélicas)

Alelos:
22, 24, 26, 26.2, 27, 28, 29, 30, 31, 32

Estructura de las variantes alélicas (número de secuencias):

Antes 22, ahora: 21: [ATCT]$_2$ ATC [ATCT]$_7$ CTTTG [TCCA]$_{11}$ CCCA [TCCA]$_1$ (2)
Antes 24, ahora: 23: [ATCT]$_2$ ATC [ATCT]$_{12}$ CTTTG [TCCA]$_8$ CCCA [TCCA]$_1$ (1)
Antes 26, ahora: 25: [ATCT]$_2$ ATC [ATCT]$_{11}$ CTTTG [TCCA]$_{11}$ CCCA [TCCA]$_1$ (1)
Antes 26.2, ahora: 25.2: [ATCT]$_2$ ATC [ATCT]$_{12}$ TTG [TCCA]$_{11}$ CCCA [TCCA]$_1$ (2)
Antes 27, ahora: 26: [ATCT]$_2$ ATC [ATCT]$_{12}$ CTTTG [TCCA]$_{11}$ CCCA [TCCA]$_1$ (1)
Antes 28, ahora: 27: [ATCT]$_2$ ATC [ATCT]$_{13}$ CTTTG [TCCA]$_{11}$ CCCA [TCCA]$_1$ (7)
[ATCT]$_2$ ATC [ATCT]$_{12}$ CTTTG [TCCA]$_{12}$ CCCA [TCCA]$_1$ (1)
[ATCT]$_2$ ATC [ATCT]$_{14}$ CTTTG [TCCA]$_{10}$ CCCA [TCCA]$_1$ (2)
Antes 29, ahora: 28: [ATCT]$_2$ ATC [ATCT]$_{14}$ CTTTG [TCCA]$_{11}$ CCCA [TCCA]$_1$ (9)
[ATCT]$_2$ ATC [ATCT]$_{13}$ CTTTG [TCCA]$_{12}$ CCCA [TCCA]$_1$ (1)
Antes 30, ahora: 29: [ATCT]$_2$ ATC [ATCT]$_{15}$ CTTTG [TCCA]$_{11}$ CCCA [TCCA]$_1$ (7)
Antes 31, ahora: 30: [ATCT]$_2$ ATC [ATCT]$_{16}$ CTTTG [TCCA]$_{11}$ CCCA [TCCA]$_1$ (3)
Antes 32, ahora: 31: [ATCT]$_2$ ATC [ATCT]$_{17}$ CTTTG [TCCA]$_{11}$ CCCA [TCCA]$_1$ (1)

Estructura general:
[ATCT]$_2$ ATC [ATCT]$_{7,11,12,13,14,15,16,17}$ CT$_{0\text{-}1}$TTG [TCCA]$_{8,10,11,12}$ CCCA [TCCA]$_1$

Secuencias en verde y celeste son repeticiones principales del núcleo y las secuencias en amarillo son consideradas interrupciones.
Lo destacado en rojo significa que las asignaciones alélicas hechas para este STR requieren corrección.

Sólo considerando el STR D7S3057 (marcador revisado de manera minuciosa, ver Anexo V), se presentaron inquietudes al momento de definir repeticiones principales, repeticiones incompletas e interrupciones. Estas inquietudes fueron resueltas a medida que se obtuvieron las primeras secuencias. Por ejemplo: es evidente que ambas repeticiones ATCT y TCCA a ambos lados de la secuencia CTTTG, son las repeticiones que le otorgan variabilidad al núcleo. Lo que implica que deben ser consideradas en la asignación alélica. Sin embargo, en ambos extremos están esas mismas secuencias [ATCT]$_2$ y [TCCA]$_1$, sin variar, pero por ser idénticas a las repeticiones principales, fueron también consideradas en la asignación alélica. La secuencia ATC, podría considerarse una repetición incompleta de ATCT, pero al aparecer sistemáticamente en todas las secuencias, sin variar, fue considerada interrupción. La secuencia CCCA fue considerada interrupción, debido a que tampoco varió en todas las secuencias halladas en estas variantes alélicas. Por último, se observó una deleción CT (rs775414159: -/TC, SNP150 y

SNP151) en la secuencia CTTTG (interrupción) en el alelo 25.2, única variante alélica con este evento, correctamente asignada por ambas herramientas.

En concreto, para este STR, la estructura general del núcleo quedó definida de la siguiente manera:

$[ATCT]_{2}$ ATC $[ATCT]_{7,11,12,13,14,15,16,17}$ $CT_{0\text{-}1}$TTG $[TCCA]_{8,10,11,12}$ CCCA $[TCCA]_{1}$

Esta representación (acorde a lo expuesto en el punto 1.2.1) es la información más detallada del núcleo del STR, que eventualmente está sujeta a cambios en caso de presentarse secuencias con un patrón diferente.

La estructura es un atributo que debe considerarse en la anotación de los STRs, lo que plantea no sólo la necesidad de las secuencias de las variantes alélicas de la población sino también del algoritmo apropiado para extraerlas de esa secuencias. Sé probó el programa GMATO (ver apartado 1.7) para realizar esta tarea, con éxito parcial, por lo que la inspección manual de las secuencias sigue siendo lo más eficiente.

En general, los criterios a continuación fueron los adoptados, no sólo en la inspección minuciosa de los 193 marcadores del Anexo V, sino también de los marcadores que aún faltan obtener su estructura:

a) Establecer repeticiones (principales y secundarias).
b) Establecer repeticiones incompletas (debe coincidir en parte con alguna repetición principal o secundaria, y aparecer esporádicamente).
c) Establecer interrupciones (debe ser una secuencia que claramente no se repite, aunque sea pasible de inserciones o deleciones).
d) Las secuencias que no varían pero coinciden con repeticiones principales o secundarias deben ser consideradas, aunque no pertenezcan al intervalo de las repeticiones principales o secundarias.
e) Establecer la generalización según las estructuras de las distintas variantes alélicas.
f) La asignación alélica resultante dependerá de todas de las repeticiones principales y secundarias, completas e incompletas, incluyendo secuencias del punto d.

La nomenclatura de los distintos alelos hallados en la población, planteados de acuerdo a una estructura definida es un hecho que requiere actualmente una revisión más profunda, no sólo en muchos STRs complejos de ciencias forenses, sino obviamente en el universo de STRs humanos.

La anotación encontrada en algunos STRs simples incluidos en bases de datos públicas sigue respetando el criterio del número de repeticiones, pero al término de este trabajo, no se ha encontrado anotación de variantes alélicas de STRs complejos.

Se ha planteado que las variantes alélicas sean descritas según la variabilidad del STR (estructura) y de los SNPs contenidos en la región de amplificación, además del cromosoma y ubicación genómica, según ensamblaje (van der Gaag, 2015).

4.11 Trío CEPH

Habiendo realizado los perfiles de los tres individuos que pertenecen al panel CEPH (NA12891, NA12892 y NA12878) y conociendo que son tres individuos relacionados (Padre, Madre e Hija, respectivamente), se indagó la aparición de mutaciones en líneas germinales de ambos progenitores, y que deberán hallarse en el descendiente. De los 323 marcadores en estudio se excluyeron todos los marcadores del cromosoma Y (87, que no son útiles dado que la descendiente es mujer) y aquellos marcadores incluidos en segmentos múltiples (7) y el marcador amelogenina. De un total de 228 marcadores, sólo 171 tuvieron asignaciones alélicas confiables (mínimo de conteos de lecturas en 5, y cortes de tartamudeos en 2.5 para di-tri-nucleótidos y 3 para tetra-penta-hexa-nucleótidos) y que su vez permiten el análisis del vínculo biológico. En la Figura 27 se detalla estos hallazgos.

Figura 27. Mutaciones en líneas germinales de los progenitores NA12891 (padre) y NA12892 (madre).

		2	3	4	5	6
		101	3	66	1	0
CORRECTO	121	77	2	41	1	0
CORRECTO (SIN PADRE)	15	4	1	10	0	0
CORRECTO (SIN MADRE)	12	5	0	7	0	0
MUTACIÓN MADRE	5	4	0	1	0	0
MUTACIÓN PADRE	16	10	0	6	0	0
MUTACIÓN MADRE (SIN PADRE)	1	1	0	0	0	0
MUTACIÓN PADRE (SIN MADRE)	1	0	0	1	0	0
MARCADOR NO INFORMATIVO	57	10	0	43	3	1

Las columnas 2, 3, 4, 5 y 6 indican el largo (nt) de la repetición del marcador.
En las filas, las distintas situaciones del vínculo Padre-Hija-Madre, evaluados en el archivo "TRIO_CEPH.xlsm".

Se evidenciaron 23 eventos atribuibles a mutación (zona roja de la Figura 27) de los 171 marcadores (13,45 %). De estos 23 eventos, 16 (recuadro rojo) están asociados al progenitor paterno, lo que está vinculado con la gametogénesis masculina. No hay diferencias significativas entre el número de mutaciones de marcadores dinucleótidos: 15 de 101, de los tetra-nucleótidos: 8 de 66 (P = 0.7865).

Este análisis si bien no evalúa a las herramientas usadas, muestra el comportamiento de los STRs *in vivo*.

En ese sentido, no fueron contemplados en estos análisis, la aparición esporádica en los STRs de un tercer alelo (Error Tipo I y II), y que son responsables de la aparición de trialélicos verdaderos, que eventualmente son eliminados por las herramientas.

5. Abordaje de la doble asignación alélica

Por último, y habiendo evaluado ambas herramientas, se detalla en la Tabla 11 las características relevantes.

Tabla 11. Comparación entre LobSTR y StraitRazor.

	StraitRazor 3.0	LobSTR 4.0.6
Plataforma	Linux, Mac y Windows.	Linux, Mac y Curoverse (Online)
Archivos de entrada que admite	FASTQ/FASTA	FASTQ/FASTA/BAM (Mac)
Librería óptima	Solexa	Illumina
Tecnología de secuenciación	Illumina	Illumina
Archivos de salida	Texto plano	aligned.bam (aligned.stats)
¿Admite la ejecución de muchos núcleos (CPUs)?	Si (opción –p)	NO
Tiempo de ejecución	40 min a 1 hr	6 - 8 hrs
Formato/s del archivo de configuración	Config (texto con tabulaciones)	BED (ver 3.7.1)
Creación del archivo de configuración	Requiere secuencias flanqueantes, secuencia motivo, valor de *offset* y libre de SNPs comunes.	Requiere definir ubicación genómica, número de copias del alelo de referencia, secuencia de ó de las repeticiones, cálculo de entropía (TRF) y valor de *score*.
Refinamiento o maduración del archivo de configuración	Requiere refinación. Demandó más tiempo para crearlo.	Sólo se necesitó filtrar el archivo provisto por los desarrolladores los marcadores de interés. Refinamiento mínimo.
Manejo de los archivos de salida	Pocos pasos	Muchos pasos
¿Se pueden analizar archivos *paired-end* juntos?	Si (no usado en este trabajo)	Si
Notación ISFG	Si	No
¿Se obtienen secuencias?	Si (opción –i para obtener la región flanqueada)	Sólo el núcleo (y sólo acorde a la repetición configurada en el archivo BED)
Complejidad del STR	Todas (aunque en STRs complejos requiere más detalle de las estructuras del núcleo de las variantes alélicas)	Simples (se puede agregar repeticiones, ajustar las coordenadas y redefinir el alelo de referencia)
Impacto de SNPs en la detección	Más de dos nt distintos en la secuencias de anclaje evitan la captura de las lecturas	No hay pérdida de lecturas por cambios en la composición de nt en la región blanco
Impacto de SNPs en la asignación alélica conjunta	Presencia de INDELS provocarán asignaciones alélicas distintas respecto de lobSTR.	Cualquier cambio que altere el núcleo puede provocar pérdida o ganancia de repeticiones.

6. VISUALIZACIONES

6.1 Visualizaciones con IGV (Integrative Genomic Viewer)

Esta herramienta (basada en JAVA) es muy útil para mostrar todos los datos recopilados en este trabajo. Acepta archivos BED, BAM, BigBED, etc. En las Figuras 28, 29 y 30 se muestran 3 capturas de pantalla en tres niveles de acercamiento: Todos los cromosomas del genoma humano, el cromosoma 7 y la región que incluye 6 marcadores del cromosoma 7 de este estudio. Y en orden descendente debajo del ideograma del cromosoma las siguientes pistas: Refseq Genes (por defecto en IGV, exones en cajas e intrones en líneas, en azul), bandeo citogenético (grises, negro y beige), los segmentos múltiples (marrón), genes (construido para la BD local, exones en fucsia e intrones en verde), los datos de secuenciación anotados con lobSTR de la Fase 1 (salmón), el catálogo de microsatélites usados por lobSTR (en rojo), la salida de Tandem Repeats Finder de 1 nt a 10 nt (fucsia), los 321 marcadores estudiados (colores varios), los marcadores de UCSC (fucsia), los marcadores de Ensembl (naranja), y por último el catálogo de variantes SNP151 comunes (en negro).

Figura 28. Captura de pantalla de la visualización del genoma humano completo (ensamblaje GRCh37/hg19). Se observan todas las pistas creadas (ver 3.7.1)

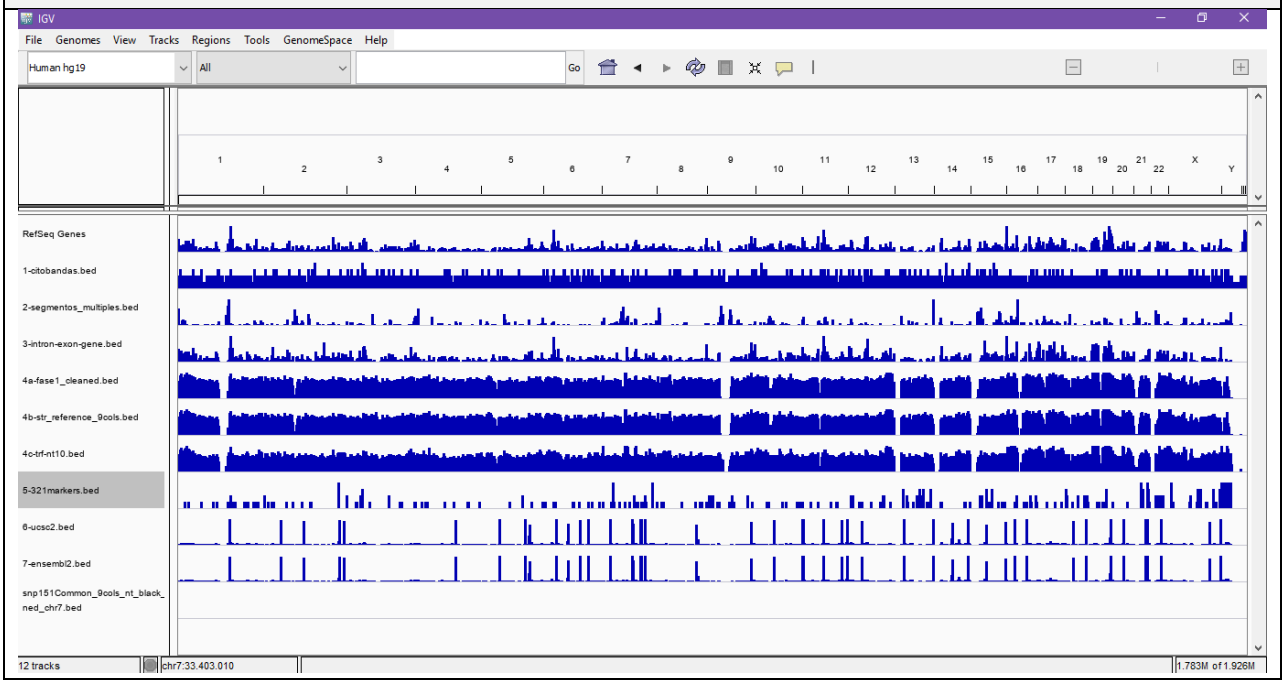


Figura 29. Captura de pantalla de la visualización del cromosoma 7 con IGV.

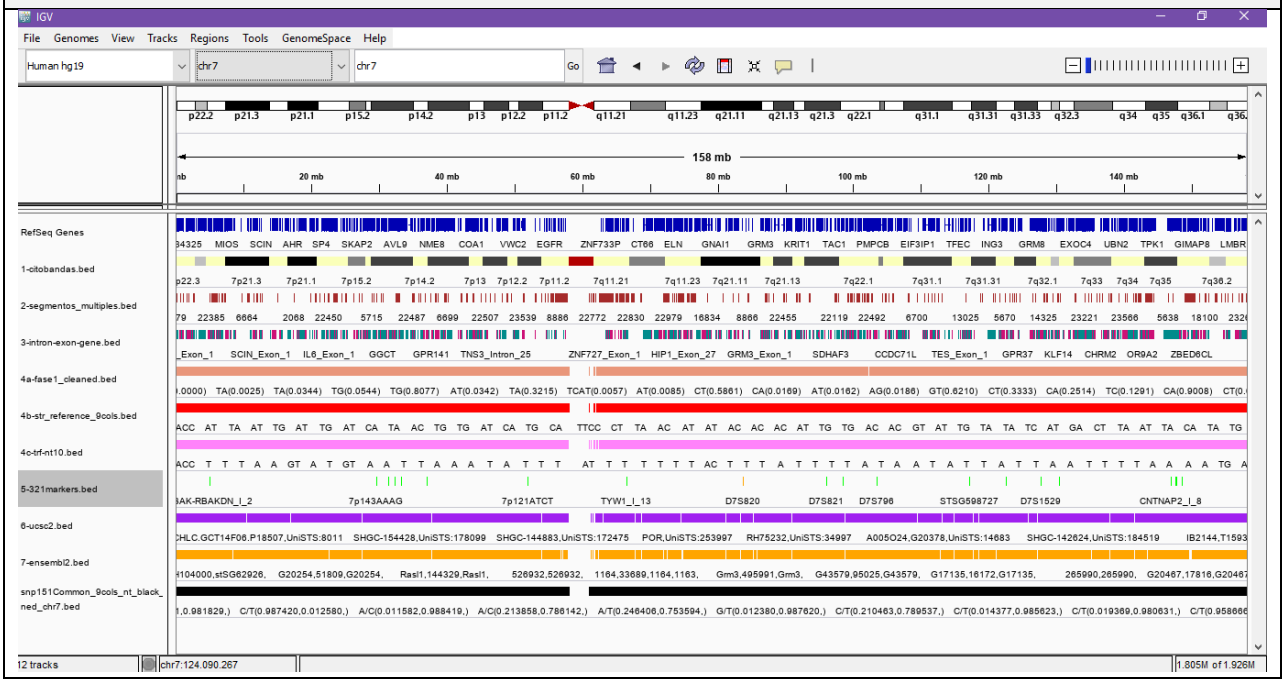
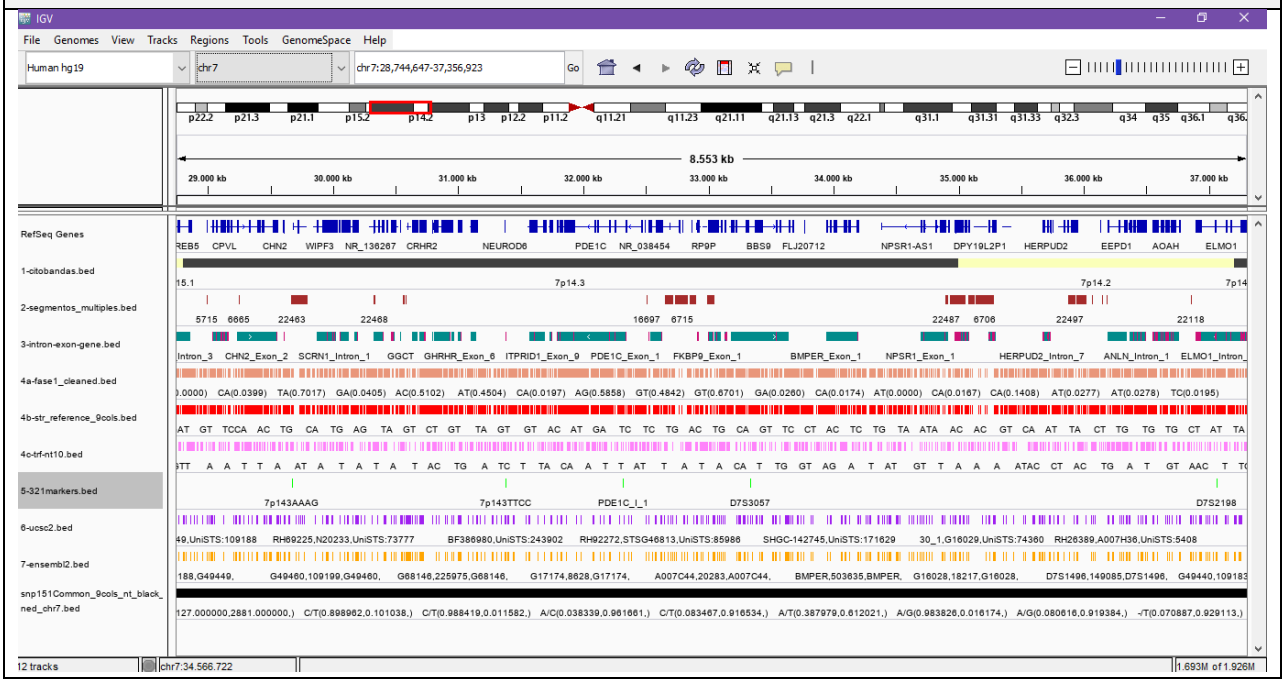


Figura 30. Captura de pantalla de la visualización de las bandas p14.2 y p14.3 del cromosoma 7.



Y en las Figuras 31 y 32 se muestran los alineamientos hechos con lobSTR para las muestras HG00096 (Figura 31) y HG01051 (Figura 32) para el marcador D16S2624. Este

marcador tiene un evento de deleción (inserción es lo correcto: rs67810177, se discutió en 4.9.4) que generó discordancias entre las herramientas, se puede apreciar el *gap* de 3 nt a lo largo de las lecturas (recuadro rojo) y los *gaps* de 12 y 8 nt (alelos 15 y 16 respectivamente, recuadro verde). En ambos ejemplos los alelos detectados son menores al alelo de referencia: 18 (eso explica los *gaps*)

Figura 31. Alineamientos hechos con lobSTR para la muestra HG00096 (Varón, archivos sorted.bam y sorted.bam.bai: SRR1291026 y SRR1291035 en la región del marcador D16S2624).

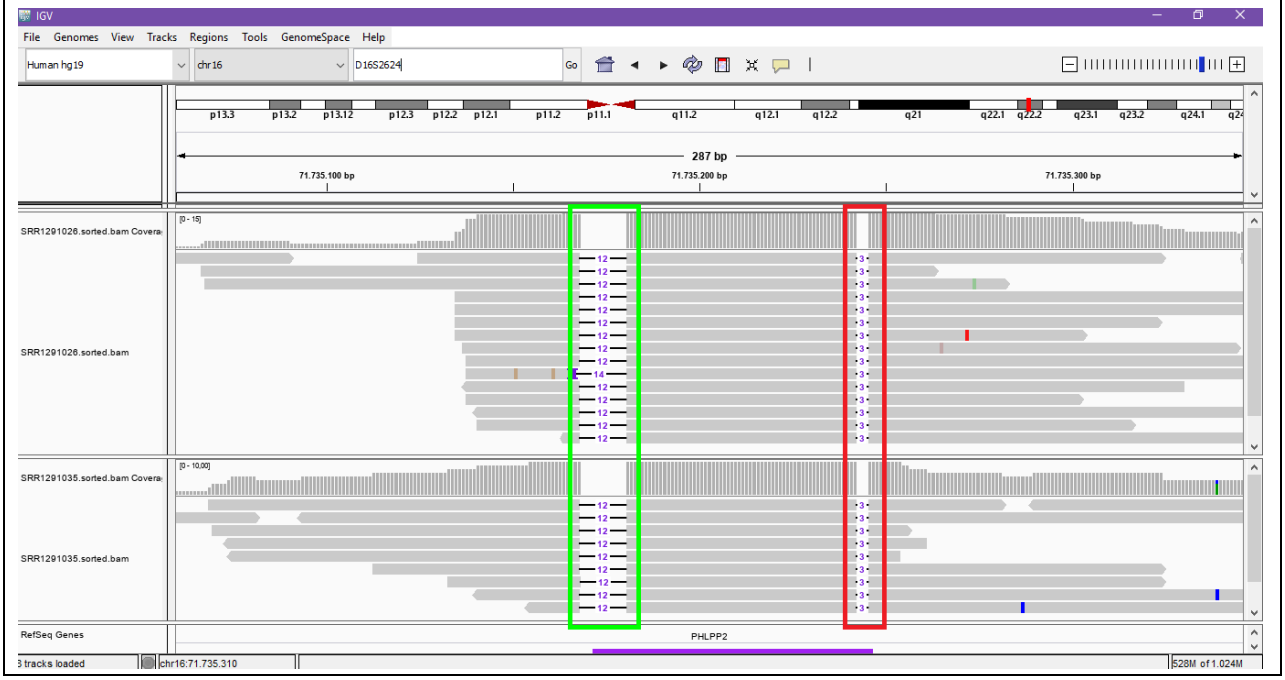
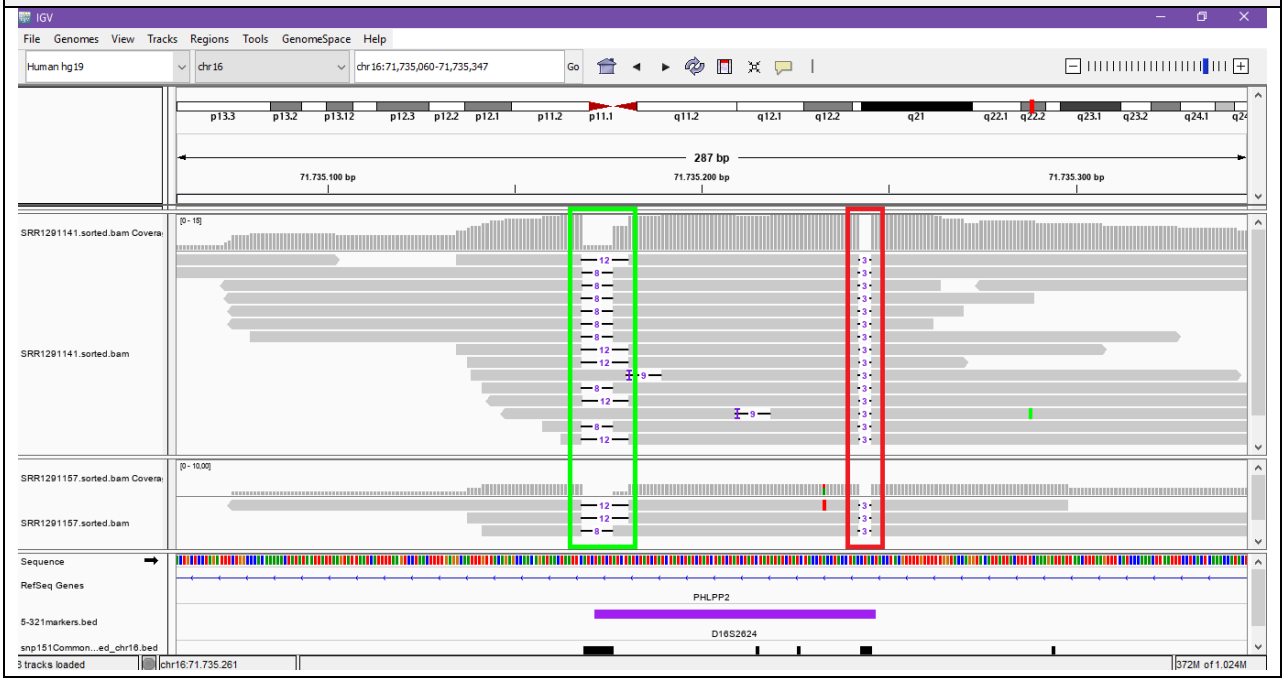
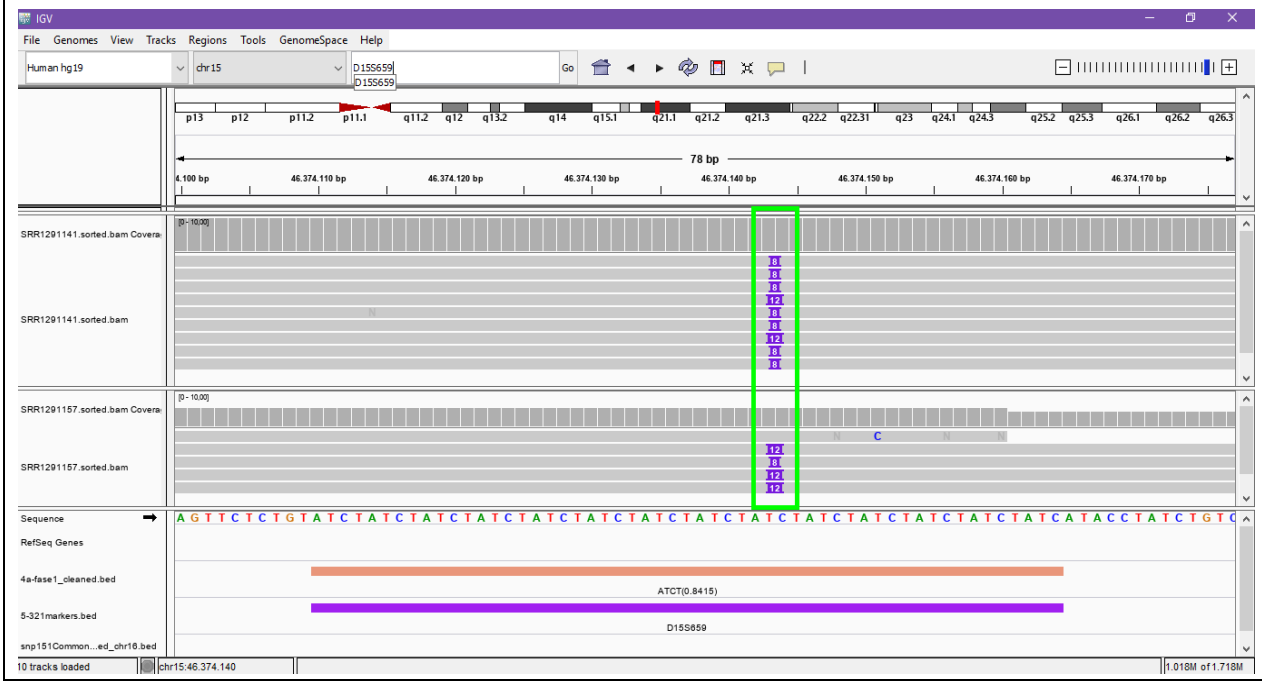


Figura 32. Alineamientos hechos con lobSTR para la muestra HG01051 (Varón, archivos sorted.bam y sorted.bam.bai: SRR1291141 y SRR1291157 en la región del marcador D16S2624).



Y en la Figura 33 se muestra los alineamientos hechos con lobSTR para la muestra HG01051 para el marcador D15S659. El propósito de esta figura es mostrar el alineamiento de las lecturas cuando estas superan en repeticiones al alelo de referencia. En este ejemplo el alelo de referencia es 14 y los alelos detectados son 16 y 17.Se observa los números 8 y 12 nt destacados en púrpura (recuadro en verde).

Figura 33. Alineamientos hechos por lobSTR para la muestra HG01051 (Varón, archivos sorted.bam y sorted.bam.bai: SRR1291141 y SRR1291157 en la región del marcador D15S659).



6.2 Visualizaciones con el navegador.

El sitio http://arrobasistemas.com/humstrs2/index.html permite inspeccionar todos los STRs de este estudio, además permite la búsqueda de STSs (Ensembl y UCSC, requiere conocer nombres o alias de los STSs). También se puede inspeccionar STRs contenidos en genes (intrones y exones, requiere conocer nombres de acuerdo a HUGO).

En la página de inicio del sitio hay una breve descripción de la variante en estudio, además de proporcionar los vínculos a: STRs de este estudio, STRs asociados a STSs, STRs asociados a genes y Búsquedas avanzadas.

Dentro de STRs en estudio podemos acceder a los 319 STRs (se excluyen Amelogenina, SRY, DYS389-2 y D13S305_1 fue unificado con D13S305_2), separados en los siguientes grupos: Ciencias Forenses, Ciencias de la Salud, Usados en ambos Ciencias Forenses y de la Salud, Elegidos del Cromosoma 7 y 9 y Elegidos de Fase 1 / LobSTR v2.0.4.

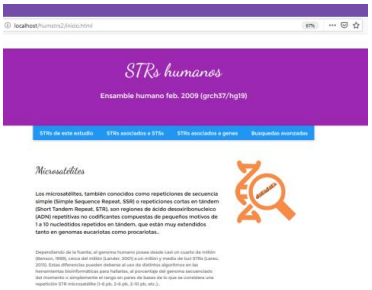
Dentro de STRs asociados a STSs hay dos vínculos: a STSs de UCSC y de Ensembl, a su vez, ambos vínculos llevan a la página para las búsquedas de marcadores (por Nombre ó por Otros Nombres), con opción a datos mínimos del marcador o con datos de secuenciación (Fase 1 / LobSTR 2.0.4).

Dentro de STRs asociados a genes, se puede buscar datos de secuenciación de Fase 1 / LobSTR 2.0.4 tanto en exones como en intrones (la búsqueda debe ser exacta y es sensible de mayúsculas y minúsculas: por ejemplo DPT Intron 3).

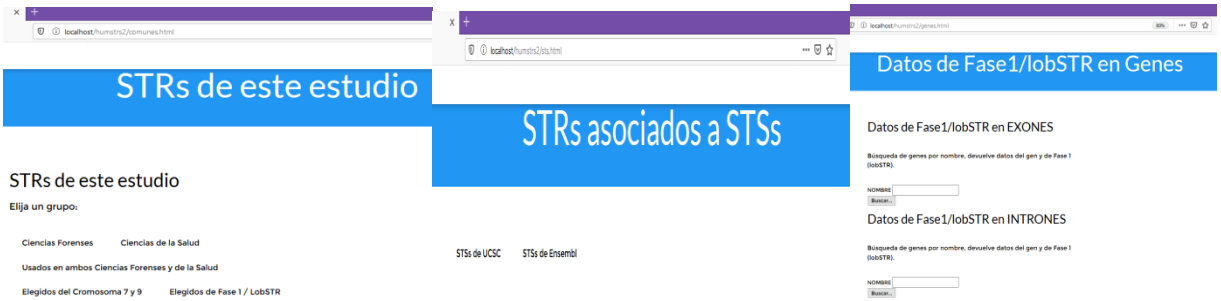
Por último, el vínculo de Búsquedas avanzadas permite inspeccionar los datos de secuenciación de Fase 1 / LobSTR 2.0.4 filtrando según ubicación genómica, largo de la repetición, secuencia de la repetición o valor de heterocigosidad.

Figura 34. Interfaz web para acceder a la BD local.

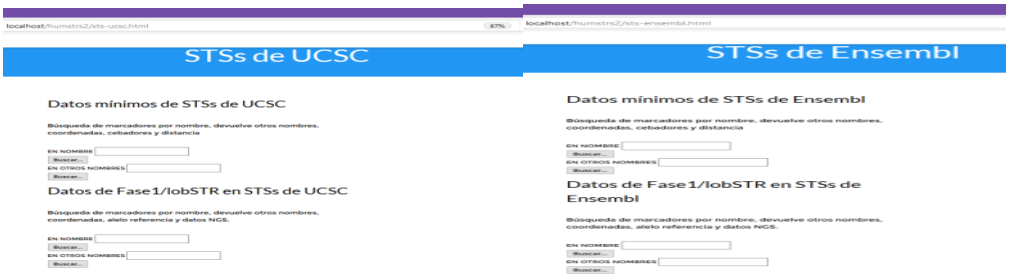
a)



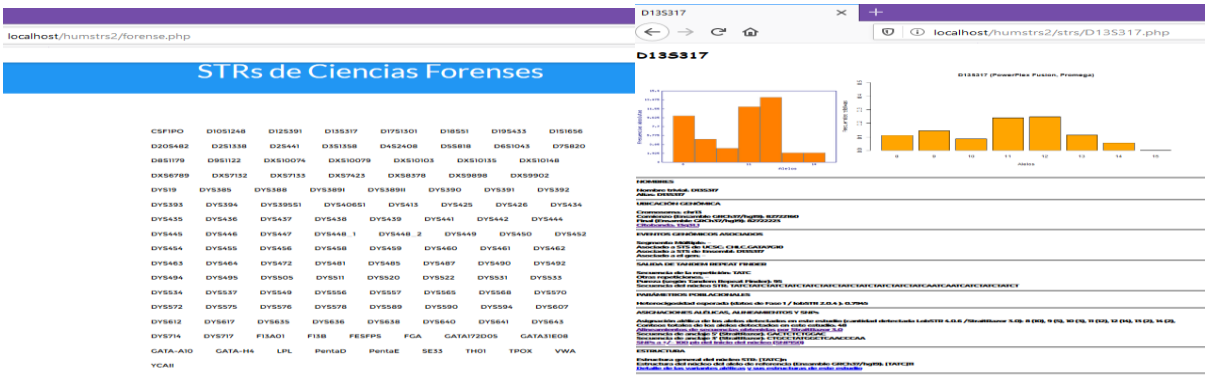
b)



c)



d)



7. Conclusiones

RESPECTO DE LA BASE DE DATOS:

Los datos de secuenciación y posterior anotación de 1.092 individuos de Fase 1 del Proyecto 1000 Genomas (herramienta lobSTR 2.0.4) incorporados a la BD local permitió encontrar aquellos STRs más polimórficos, por lo tanto los más útiles, perteneciente a cualquier cromosoma de interés, para cualquiera de las disciplinas que los utilizan.

La incorporación del dato de heterocigosidad, de las tablas de segmentos múltiples y de SNPs permitió no sólo filtrar datos, sino que también permitió responder anomalías en el uso de las herramientas lobSTR y StraitRazor.

La información de STSs de Ensembl y de UCSC fue útil para definir nombres para los marcadores en estudio que no están dentro de regiones génicas, además de proveer las coordenadas para extraer las secuencias (*scripts* de Perl para los STSs de UCSC sobre "hg19.fa").

Heredar nombres provenientes de STSs para asignarlos a STRs es claramente insuficiente (11 y 12,8 % de STRs dentro de STSs de Ensembl y UCSC respectivamente), incluso heredarlos desde nomenclatura para genes también lo es, aunque son más abarcativos en el genoma (56,12 %, principalmente intrones, 99,4%). La nomenclatura de SNPs (rs#) suele ser usada actualmente para nombrar STRs huérfanos, pero se debe destacar que STRs y SNPs responden a eventos biológicos de distinto origen y comportamientos diferentes.

Las consultas a la base también sirvieron para confirmar lo observado en otros análisis genómicos:

a) Las secuencias de los cebadores alojados en UCSC y Ensembl están desactualizados respecto de la secuencia de referencia.
b) Se confirmó la distribución homogénea de los STRs en el genoma humano.
c) En concordancia con la base de datos de segmentos múltiples usado en este trabajo (Bailey et al. 2001, archivo "build37.xlsx") se corroboró el predominio de segmentos múltiples inter e intra cromosómicos del cromosoma Y. Evento que explica la aparición de más de un alelo para muchos STRs en este cromosoma (haploide).
d) Tandem Repeat Finder es deficiente en detectar STRs dinucleótidos (al menos con los parámetros usados). Es recomendable usar otras herramientas, como RepeatMasker, Gmato, STR-FM en Galaxy o similares (punto 1.7).

Los archivos generados para la construcción de la BD local sirvieron de sustrato para nuevos archivos, por ejemplo, para las visualizaciones en IGV.

Se pudieron definir las siguientes características para que un STR quede debidamente anotado:

Ubicación genómica: cromosoma, coordenadas de comienzo y fin del núcleo y citobanda. Estos datos son susceptibles de cambio de acuerdo a la versión de ensamblaje del genoma y sus actualizaciones.

Nombres: Actualmente no existe una nomenclatura que identifique la totalidad de STRs en el genoma humano ó de cualquier otro genoma, por esa razón las asociaciones a otros eventos genómicos siguen siendo bastante utilizadas.

Eventos genómicos asociados: STSs, genes, segmentos múltiples y SNPs deben acompañar y complementar la información de cada STR.

Núcleo y repeticiones: existen numerosas herramientas que puede determinar la zona de repeticiones dentro de una región STR. Esto aplica a la variante más frecuente (alelo de referencia) y a cada una de las variantes alélicas existentes.

Estructura: como ya se discutió, la estructura de las variantes alélicas es un atributo a considerar, y es de mucha ayuda para comprender el comportamiento de las repeticiones en el núcleo, y esto también incluye a los SNPs.

Asignación alélica: el uso de números enteros es útil para nombrar las variantes alélicas, sin embargo debido a la pérdida de información (especialmente en STR complejos) es probable que esto deba cambiar en el futuro. Se ha propuesto que la asignación alélica en su versión más completa incorpore SNPs, ubicación genómica y ensamblaje y la estructura (van der Gaag, 2015)

Parámetros poblacionales: las frecuencias de las variantes alélicas de acuerdo a la población en estudio deberán ser agregadas, y todos los parámetros que deriven de ellas (heterocigosidad, contenido de información polimórfica, poder de discriminación, etc.)

Secuencias de anclaje (opcional): eventualmente las secuencias usadas con herramientas bioinformáticas podrían agregarse. Hay numerosas herramientas para obtener secuencias de cebadores de una determinada región, pero no existe al momento de finalizar este trabajo software, API o herramienta que permite obtener secuencias flanqueantes con fines bioinformáticos.

Es importante destacar que el ensamblaje GRCh37/hg19 posee alelos de referencia que no son consistentes con sus frecuencias alélicas (Willems, 2014), situación corroborada en este trabajo (ver punto 3.4.6) donde 12,45 % ± 1,48 % no coincide el alelo de referencia incorporado en el ensamblaje con el alelo más frecuente, lo que plantea la necesidad de correcciones.

RESPECTO DEL USO DE LAS HERRAMIENTAS

Se analizaron 24 genomas humanos de individuos no relacionados y 3 individuos relacionados del panel de CEPH (Madre, Hija y Padre) con dos herramientas bioinformáticas que admiten archivos "*.fastq": StraitRazor v3.0 y LobSTR v4.0.6. Estas herramientas tienen un funcionamiento distinto para detectar variantes STRs, siendo lobSTR dependiente de alineamiento y StraitRazor independiente de alineamiento.

Del panel de STRs propuesto inicialmente, se amplió a 323: 127 de uso en Ciencias Forenses, 47 de uso en Ciencias de la Salud, 5 comunes a ambas disciplinas, 28 elegidos de los cromosomas 7 y 9, y 114 elegidos de los datos de secuenciación y anotación de Fase 1 sobre 1.009 individuos, que superan el valor de heterocigosidad mayor a 0,7, repartidos en todos los cromosomas.

Sólo dos marcadores no fueron detectados en las 27 muestras por ambas herramientas: D9S302 y DYS389-2. Para D9S302 la razón es fácilmente atribuible a lo extenso del núcleo: 211 pb, sin embargo no hay una razón concluyente para DYS389-2, siendo que este marcador junto con DYS389I forman DYS389II, ambos DYS389I y DYS389II si fueron detectados.

Un total de 232 marcadores (71,8 %) fueron detectados completamente: 186 (78,8 %) marcadores autosómicos, pseudo-autosómicos y del cromosoma X en las 27 muestras y 46 (52,9 %) marcadores del cromosoma Y en 11 varones.

22 marcadores no fueron detectados en los individuos del trío CEPH aunque no hay explicación para este hallazgo.

Se evidenció que lobSTR funciona mejor con secuenciaciones cuyas lecturas son cortas (100 pb en promedio), esto depende de la librería usada para generar los fragmentos del genoma,

en cambio StraitRazor pudo entregar resultados con un rango mayor de la longitud de la lectura de secuenciación.

También se observó que si bien ambas herramientas tienen tasas de error comparables respecto de la detección de lecturas (lobSTR 1,20 % y StraitRazor 0,98 %, punto 4.8.1), al momento de la asignación alélica, StraitRazor devuelve 6,9 % asignaciones alélicas inespecíficas, contra un 2,3 % de lobSTR (punto 4.8.3). Los cocientes aplicados para los tartamudeos solucionan este problema en la mayoría de los casos (cuando los conteos de lecturas son mayores a 10).

Con la aplicación de reglas para la asignación alélica automática (tanto para el conteo mínimo de lecturas como para los cocientes para tartamudeos) se logró un total de 4.394 asignaciones alélicas consideradas confiables (reglas 1, 2 y 3, conteo mínimo de lecturas en 1 y cocientes de tartamudeo en 5), que sumado a las “No detecciones” y las asignaciones alélicas de las reglas 4, 5 y 6 se logró un total de 6.702 (76,9 % de las 8.710 asignaciones alélicas posibles), en condiciones un poco más estrictas (conteo mínimo de lecturas en 5 y cocientes de tartamudeo en 2) ese porcentaje cae al 65,4 % (5.696 asignaciones alélicas confiables). Ciertamente es un valor lo suficientemente alto considerando que ambas herramientas funcionan de manera distinta, diseñadas no solo con algoritmos diferentes, sino también para distintas estrategias de secuenciación. Además, las 2.008 asignaciones alélicas restantes (ó 3.014 dependiendo si se hace estricto el corte) son pasibles de correcciones.

Toda la información recopilada de la segunda parte de este trabajo fue incorporada a la BD local ó al sitio http://arrobasistemas.com/humstrs2/index.html: asignaciones alélicas, composición del núcleo del STR y estructura, alineamientos de secuencias obtenidas con StraitRazor, histogramas de frecuencias alélicas, SNPs hallados en las inmediaciones del núcleo, por mencionar los más importantes

8. DISCUSIÓN

En este trabajo se deja evidencia que las variantes microsatélites humanas (y que es extensible al resto de los seres vivos) necesitan un abordaje más específico para definir la nomenclatura de las regiones STRs y de sus variantes alélicas, y si se consideran las razones biológicas de cómo evoluciona en una población, y a su vez la información que brinda para diferentes disciplinas, este tipo de variante tiene identidad propia suficiente para tener un tratamiento único y que la diferencie del resto de eventos genómicos que otorgan variabilidad a un individuo. En ese sentido, la tarea no sólo atañe a la nomenclatura, si no también, y de manera muy minuciosa, a las estructuras del núcleo de una región STR, que al término de este trabajo (con poco más de 190 estructuras) se pudo observar una diversidad elevada de patrones, sobre todo en aquellos núcleos de elevada complejidad. No fueron pocos los casos en donde un SNP hallado dentro del núcleo estaba sucediendo en una determinada variante alélica, lo que refuerza esta propuesta de tratamiento especial. Es decir, observar un SNP de manera aislada dentro de la región núcleo del STR, no brinda más información sobre el entorno donde se encuentra.

Sin embargo, para muchos STRs cuyo núcleo es de elevada complejidad, obtener un patrón claro de su variabilidad resulta no sólo engorroso para quien analiza las secuencias, sino también resulta complicado para los algoritmos de las herramientas que detectan esa región. Hasta ahora, y habiendo usado dos herramientas de detección de STRs en archivos de secuenciación de segunda generación con funcionamientos diferentes (con o sin alineamiento), y una herramienta que detecta STRs en archivos con formato FASTA, se pudo evidenciar que los algoritmos para detectar STRs complejos requieren una mejora para que la asignación alélica sea correcta, y no sólo esto, la aparición de SNPs dentro del núcleo (y en las secuencias no repetitivas incluidas en la región detectada) alteran el patrón de detección, cualquiera sea el camino para detectarlos y asignarle nombre al alelo. Además, sería importante incorporar en la salida de las herramientas un puntaje (*score)* que evalúe la asignación alélica (principalmente para aplicaciones en ciencias forenses) que pueda considerar no solo posibles tartamudeos, sino también otros factores que afectan la calidad de la asignación (ver más abajo HipSTR).

El conocimiento de las estructuras de las distintas variantes alélicas de una región STR determinada, soportada por un número elevado de secuencias podrá dar una solución y otorgar una mejora en los algoritmos, ya sean previos (archivos de configuración de las herramientas), o posterior a la detección (intervención del analista para decidir sobre probables tartamudeos). Y sobre esto último, es importante mencionar que los desarrolladores de lobSTR posteriormente lanzaron una nueva herramienta: HipSTR (Willems et al, 2017), donde incluyen parámetros relacionados con los tartamudeos, además de considerar la calidad y profundidad de la secuenciación. Sin embargo, las herramientas aun apuntan a la detección de STRs simples (o con poca complejidad) y que tampoco consideran las regiones no repetitivas por fuera del núcleo, ni tampoco consideran SNPs dentro y fuera del núcleo. El detalle de las estructuras puede dar solución a esta situación, y la información proveniente de SNPs es necesaria también. Como ya se dijo, muchas veces un determinado SNP pertenece a una determinada variante alélica STR y no aparece en las otras, lo que da una idea del grado de separación que debe establecerse entre ambas variantes.

Antes de trabajar con StraitRazor, se probó otra herramienta (solo de manera exploratoria): TSSV, que considera la estructura de la secuencia hallada y las devuelve con la

nomenclatura convencional, pero hubo dos motivos para ser descartada en este trabajo: el archivo de configuración requiere proporcionar una estructura general que al momento de iniciar este estudio con todos los marcadores se desconocían la mayoría, y además esta herramienta devolvía resultados con mucha demora (posiblemente debido a que solo usa un núcleo de procesamiento) respecto de StraitRazor.

En este sentido, luego de finalizado este trabajo, en el sitio web creado están alojados todos los datos obtenidos, de los STRs seleccionados, cuyos atributos (incluyendo estructuras) para anotarlos debidamente están para la mayoría (de algunos marcadores no se obtuvieron las secuencias y por ende tampoco las estructuras).

Obviamente, el sitio puede alojar en el futuro más datos poblacionales, lo que implica una mejora en la base de datos. De todas formas, el sitio actualmente posee información relevante, y precisa de los STRs más usados en distintas disciplinas, de manera ordenada y clara para su inspección.

Como se dijo, actualmente hay en el sitio web 190 STRs cuyas estructuras del núcleo están descriptas, de acuerdo a los criterios del apartado 4.10.

Un aspecto no menos importante respecto de la anotación de STRs, es el hecho de que las frecuencias alélicas de muchos STRs detectados en los trabajos que fueron de sustento de este, y que también se hizo aquí esa corroboración, es aquel donde el alelo de referencia no coincidió con el alelo más frecuente hallado, imprecisión que no solo afecta al ensamblaje usado GRCh37/hg19 (feb. 2009) sino también al ensamblaje de referencia vigente GRCh38/hg38 (dic. 2013). La corrección necesariamente estará ligada a definiciones que aún se requieren para esta variante.

A pesar del advenimiento de nuevas tecnologías de secuenciación y del desarrollo de herramientas para detectar STRs, el uso de tecnologías separativas para determinarlos sigue siendo ampliamente usadas, debido al menor costo y también al hecho de que los STRs son altamente informativos, con un panel de unos pocos marcadores (de 21 a 25 en Ciencias Forenses y de 20 a 40 en Ciencias de la Salud) se puede establecer con un margen de error muy bajo un vínculo biológico o un diagnóstico, respectivamente.

Podría pensarse que no hay razones para describir las más de 320.000 STRs del genoma humano si con unos pocos marcadores hemos resuelto necesidades analíticas de las dos disciplinas. Y si, de alguna manera, hasta ahora es así, sin embargo, hay algunas razones para proponer mejoras en la anotación de esta variante:

a.- En Ciencias de la Salud, se usan marcadores de los cromosomas 21, 18, 13, 15, 16, 22, X e Y, pero cuando se quiere investigar algún otro cromosoma, sin dudas que se empieza a hacer búsquedas y esa información es un tanto escasa y a veces poco precisa o desactualizada. Puntualmente, hay otros cromosomas involucrados en aneuploidías (mixoploidías) donde pueden estar afectados los cromosomas 7, 9, 14 ó 17, de la línea celular que genera la placenta, provocando un desarrollo anormal del individuo en gestación (que por lo general no está afectado por la aneuploidía).

b.- Se ha demostrado recientemente (Bagshaw, 2017; Gymrek, 2016) que los microsatélites tienen impacto a nivel de expresión génica, y se proponen diversos mecanismos para explicar su influencia (modulación de la unión del factor de transcripción, espaciado entre elementos promotores, potenciadores, metilación de citosina, corte y empalme alternativo, estabilidad del ARNm, selección de sitios de inicio y terminación de la transcripción,

conformaciones estructurales inusuales, posicionamiento y modificación de nucleosomas, estructura de cromatina de orden superior, ARN no codificante y recombinación meiótica en puntos calientes) brindando más fuerza al planteo de una mejor anotación de los STRs, esta vez involucrando regiones génicas.

c.- La aparición de las distintas variantes alélicas obedece principalmente a lo explicado en la introducción (punto 1.4, Figura 4) y que difiere del mecanismo biológico propuesto para los SNPs, lo que implica que esta variante requiera un tratamiento específico, y no debe ser incluida en otras bases de datos (al menos no como SNP). En este estudio han aparecido algunos STRs cuya anotación en la base de datos de SNPs están más detalladas, lo que indica que habría intenciones de una mejor anotación:

STR **AFM249ZA5** (rs3220775): (**CA**) 16/17/19/20/21/22/23/25.

STR **AFM343VF1** (rs3221549): (**CA**) 16/17/18/19/20/21/22.

STR **AFM359TB5** (rs3221630): (**CA**) 14/15/16/17.

Por mencionar algunos ejemplos, pero la gran mayoría tienen una descripción muy pobre: "*lengthTooLong*" dice como cambio observado.

d.- La versión del ensamblaje actual del genoma de referencia GRCh38/hg38 ha recibido al momento de finalización de esta redacción un nuevo parche (*patch*): GRCh38.p14 (03-02-2022), en este caso debido a que se ha logrado secuenciar exitosamente regiones que permanecían desconocidas (Nurk, 2022). La secuenciación aleatoria de lectura larga ha superado las limitaciones del ensamblaje basado en BAC. Entonces, poder incorporar un mejor detalle de las variaciones en los microsatélites es un objetivo a conseguir.

De esta manera queda reflejado en este trabajo toda la información que será necesario recopilar para lograr una mejor descripción del universo completo de STRs humanos.

9. BIBLIOGRAFÍA

Adinolfi M., Sherlock J., Cirigliano V., Pertl B. (2000). *Prenatal screening of aneuploidies by quantitative fluorescent PCR*. Community Genet. 3: 50-60.

Anvar S.Y., Van Der Gaag K.J.,Van Der Heijden J.W.F., Veltrop M.H.A.M, Vossen R.H.A.M., De Leeuw R.H., Breukel C., Buermans H.P.J., Verbeek J.S., De Knijff P., Den Dunnen J.T., Laros J.F.J. (2014). *TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes*. Bioinformatics 30 1651–1659, http://dx.doi.org/10.1093/bioinformatics/btu068.

Bagshaw A. (2017). *Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes*. Genome biology and evolution, 9(9), 2428–2443. https://doi.org/10.1093/gbe/evx164

Bailey J.A., Church D.M., Ventura M., Rocchi M., and Eichler E.E. (2004). *Analysis of Segmental Duplications and Genome Assembly in the Mouse*. Genome Res.; 14(5): 789–801. doi: 10.1101/gr.2238404

Bamshad M., Watkins W.S., Zenger R.K., Bohnsack J.F., Carey J.C., Otterud B., Krakowiak P.A., Robertson M., Jorde L.B. (1994). *A Gene for Distal Arthrogryposis Type I Maps to the Pericentromeric Region of Chromosome 9.* Am.J. Hum. Genet. 55:1153-1158

Beier S., Thiel T., Münch T., Scholz U., Mascher M. (2017). *MISA-web: a web server for microsatellite prediction*. Bioinformatics. 15;33 (16):2583-2585.

Benson G. (1999). *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Research Vol. 27, No. 2, pp. 573-580

Broman K., Murray J., Sheffield V., White R., Weber J. (1998). *Comprehensive Human Genetic Maps Individual and Sex-Specific Variation in Recombination.* Am. J. Hum. Genet. 63:861–869,

Bär W., (1997). *DNA recommendations - further report of the DNA Comission of the ISFH regarding the use of short tandem repeat systems*. International. Society for Forensic Haemogenetics 87 179-184

Cao M.D.,Tasker E., Willadsen K., Imelfort M., Vishwanathan S., Sureshkumar S., Balasubramanian S., Boden M. (2014). *Inferring short tandem repeat variation from paired-end short reads.* Nucleic Acids Res. 42, http://dx.doi.org/10.1093/nar/gkt1313.

Cheung K.H., Miller P.L., Kidd J.R., Kidd K.K., Osier M.V., Pakstis A.J. (2000). *ALFRED: a Web-accessible allele frequency database*. Pac Symp Biocomput 639-50.

Cirigliano V., Voglino G., Ordoñez E., Marongiu A., Cañadas M.P., Ejarque M., Rueda L., Lloveras E., Fuster C., Adinolfi M. (2009). *Rapid prenatal diagnosis of common chromosome aneuploidies by QF-PCR, results of 9 years of clinical experience*. Prenat Diagn; 29: 40–49. DOI: 10.1002/pd.2192

Cummings C.J. and Zoghbi H.Y. (2000). *Trinucleotide repeats: Mechanisms and pathophysiology*. Annual Review of Genomics and Human Genetics 1:281-328

Everett C.M. and Wood N.W. (2004). *Trinucleotide repeats and neurodegenerative disease.* Brain 127:2385-2405

Fan H., and Chu J.Y. (2007). *A Brief Review of Short Tandem Repeat Mutation.* Geno. Prot. Bioinfo. Vol. 5 No. 1, 7-14

Field D. and Wills C. (1998). *Long polymorphic microsatellites in simple organisms*. Proceeding of the Royal Society of London, Series B: Biological Sciences 263:209-215

Fungtammasan A., Ananda G., Hile S., Su M., Sun C., Harris R., Medvedev P., Eckert K., Makova K. (2015). *Accurate typing of short tandem repeats from genome-wide sequencing data and its applications*. Genome Res. 2015;25(5):736–49

GeneMapper™ ID Software Version 3.2 Human Identification Analysis User Guide. Part Number 4357520 Rev. A 11/2004 .© Copyright 2004, Applied Biosystems. All rights reserved.

GeneScan® Analysis Software Version 3.1 User's Manual. Copyright 2000, Applied Biosystems

Goldstein D.B. and Schlotterer C. (1999). *Microsatellites: Evolution and Applications*. Oxford University Press, New York, 343 pp.

Gusmao L., Butler J., Carracedo A., Gill P., Kayser M., Mayr W., Morling N., Prinz M., Roewer L., Tyler-Smith C., Schneider P. (2006). *DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic análisis.* Forensic Science International 157 187–197

Gymrek M., Golan D., Rosset S. y Erlich Y. (2012). *lobSTR: A short tandem repeat profiler for personal genomes*. Genome Research.

Gymrek M., Willems T., Guilmatre A., Zeng H., Markus B., Georgiev S., Daly M.J., Price A.L., Pritchard J.K., Sharp A.J. y Erlich Y. (2016) *Abundant contribution of short tandem repeats to gene expression variation in humans*. Nat Genet 48, 22–29.

Hall T. (2011). *BioEdit: An important software for molecular biology*. GERF Bulletin of Biosciences. June 2011, 2(1):60-61

Highnam G., Franck C., Martin A., Stephens C., Puthige A., Mittelman D. (2013). *Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles*. Nucleic Acids Res 41: e32.

Jarne P. and Lagoda P.J.L. (1996). *Microsatellites, from molecules to populations and back*. Trends in Ecology and Evolution 11:424-429

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. (2002). *The human genome browser at UCSC*. Genome Res. 12(6):996-1006.

King J. L., Peter de Knijff, Morling N., Prinzo M., Schneider P. M., Van Neste C., Willuweit S., Phillips C. (2016). *Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements.* Forensic Science International: Genetics 22 54–63.

Kofler R., Schlötterer C. and Lelley T.(2007). *SciRoKo: a new tool for whole genome microsatellite search and investigation*. Bioinformatics, Volume 23, Issue 13, Pages 1683–1685.

Kolpakov R., Bana G., Kucherov G. (2003). *mreps: efficient and flexible detection of tandem repeats in DNA*. Nucleic Acids Res. 31:3672-3678.

Lander E. (2001). *Initial sequencing and analysis of the human genome*.Nature 409: 860–921

Lareu M. (2013). *Short Tandem Repeats.* Encyclopedia of Forensic Sciences, 2nd Edition doi: 10.1016/B978-0-12-382165-2.00040-4 219-226

Liu C.S.J., Hulce D., Li X., Snyder-Leiby T. (2011). *GeneMarker® Genotyping Software: Tools to Increase the Statistical Power of DNA Fragment Analysis*. SoftGenetics, LLC, State College, PA, United States. J Biomol Tech. 22(Suppl): S35–S36.

Liu Y., Harbison S.A. (2018). *A review of bioinformatic methods for forensic DNA analyses*. Forensic Science International: Genetics 33 117–128

Mead L.J., Gillespie M.T., Irving L.B., Campbell L.J. (1994). *Homozygous and Hemizygous Deletions of 9p Centromeric to the Interferon Genes in Lung Cancer* CANCER RESEARCH 54, 2307-2309.

Morgante M., Hanafey M. and Powell W. (2002). *Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes*. Nature Genetics 30:194-200

Moxon R and Willis C (1999). *DNA microsatellites: Agents of evolution?.* Scientific American 280:94-99

Nurk S., Koren S., Rhie A., Rautiainen M., Bzikadze A. V., Mikheenko A., Vollger M. R., Altemose N., Uralsky L., Gershman A., Aganezov S., Hoyt S. J., Diekhans M., Logsdon G. A., Alonge M., Antonarakis S. E., Borchers M., Bouffard G. G., Brooks S. Y., Caldas G. V.,… Phillippy A. M. (2022). *The complete sequence of a human genome*. Science. 44-53:376 D.O.I.: https://doi.org/10.1126/science.abj6987

Oliveira E., Pádua J., Zucchi M., Vencovsky R., Vieira M. (2006). *Origin, evolution and genome distribution of microsatellites.* Genetics and Molecular Biology, 29, 2, 294-307

Parson W., Ballard D., Budowle B. , Butler J. M., Gettings K. B., Gill P., Gusmão L., Hares D. R., Irwin J. A., Pemberton T., Sandefur C., Jakobsson M., Rosenberg N. (2009). *Sequence determinants of human microsatellite variability*. BMC Genomics, 10:612 doi:10.1186/1471-2164-10-612

Rosen, N., V. Chalifa-Caspi, O. Shmueli, A. Adato, M. Lapidot, J. Stampnitzky, M. Safran, and D. Lancet (2003). *GeneLoc: Exon-based integration of human genome maps*. Bioinformatics 19(S1). URL: http://genecards.weizmann.ac.il/geneloc

Rosenberg N., Pritchard J., Weber J., Cann H., Kidd K., Zhivotovsky L., Feldman M. (2002). *Genetic Structure of Human Populations*. Science Vol. 298 20 2381-2385

Rosenberg N., Mahajan S., Ramachandran S., Zhao C., Pritchard J., Feldman M., (2005). *Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure*. Vol. 1 (6) 660-671

Ruitberg Ch., Reeder D., Butler J. (2001). *STRBase: a short tandem repeat DNA database for the human identity testing community*. Nucleic Acids Research, , Vol. 29, No. 1. 320–322.

Sia E.A., Butler C.A., Dominska M., Greenwell P., Fox T.D., and Petes T.D. (2000). *Analysis of microsatellite mutations in the mitochondrial DNA of Saccharomyces cerevisiae*. Proceedings of the National Academy of Sciences 97:250-255

Smit A.F.A., Hubley R., Green P. (1996–2013). *RepeatMasker Open-4.0*. URL http://www.repeatmasker.org.

Strand M., Prolla T.A., Liskay R.M. and Petes T.D. (1993). *Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair*. Nature 365:274-276

Toth G., Gaspari Z., Jurka J. (2000). *Microsatellites in different eukaryotic genomes: survey and analysis*. Genome Res,10(7):967-981.

Van der Gaag K.J., de Knijff P. (2015). *Forensic nomenclature for short tandem repeats updated for Sequencing.* Forensic Science International: Genetics Supplement Series 5 e542–e544

Van der Velden P., Sandkuijl L.A., Bergman W., Hille E.T., Frants R.R., Gruis N.A.. (1999). *A Locus Linked to p16 Modifies Melanoma Risk in Dutch Familial Atypical Multiple Mole Melanoma (FAMMM) Syndrome Families.* Genome Res. 9:575-580

Van Neste C., Gansemans Y., De Coninck D., Van Hoofstat D., Van Criekinge W., Deforce D. (2015). *Forensic massively parallel sequencing data analysis tool: implementation of MyFLq as a standalone web- and illumina BaseSpace (1)-application.* Forensic Sci. Int. Genet. 15 2–7, doi:http://dx.doi.org/10.1016/j.fsigen.2014.10.006.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., Kodira C.D., Zheng X.H., Chen L.,… Zhu X. (2001). *The sequence of the human genome.* Science. 291 (5507): 1304–1351. Bibcode:2001Sci...291.1304V. doi:10.1126/science.1058040. PMID 11181995.

Wang X., Lu P., Luo Z. (2013). *GMATo: A novel tool for the identification and analysis of microsatellites in large genomes*. Bioinformation. 9(10):541-4.

Warshauer D., Lin D., Hari K., Jain R., Davis C., LaRue B., King J., Budowle B. (2013). *STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data*. Forensic Science International: Genetics 7 409–417

Wenz H., Robertson J., Menchen S., Oaks F., Demorest D., Scheibler D., Rosenblum B., Wike C., Gilbert D., Efcavitch J. (1998). *High-Precision Genotyping by Denaturing Capillary Electrophoresis.* Genome Research 8:69–80

Willems T., Gymrek M., Highnam G. The 1000 Genomes Project Consortium. Mittelman D., and Erlich Y. (2014). *The landscape of human STR variation.* Genome Res Nov;24(11):1894-904. doi: 10.1101/gr.177774.114.

Willems T., Zielinski D., Yuan J., Gordon A., Gymrek M., Erlich Y. (2017). *Genome-wide profiling of heritable and de novo STR variations.* Nature Methods, 14:590–592

Woerner A., King J., Budowle B. (2017). *Fast STR allele identification with STRait Razor 3.0.* Forensic Science International: Genetics http://dx.doi.org/10.1016/j.fsigen.2017.05.008

Zerbino D., Achuthan P., Akanni W., Amode M., Barrell D., Bhai J., Billis K., Cummins C., Gall A., Giro´n C., Gil L., Gordon L., Haggerty L., Haskell E., Hourlier T., Izuogu O., Janacek S., Juettemann T., To J., … Flicek P. (2018). *Ensembl 2018* . PubMed PMID: 29155950. doi:10.1093/nar/gkx1098

ANEXO I. DESCARGA DE ARCHIVOS DESDE BASES DE DATOS PÚBLICAS

I.1 Desde sitios web

"stsMap.txt.gz", "stsInfo2.txt.gz", "stsAlias.txt.gz", "cytoBand.txt.gz" y "ccdsGene.txt.gz" desde http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/
y posterior descompresión por línea de comandos (*bash, gunzip –k*)

“phase_1_final_calls.vcf.gz” desde http://strcat.teamerlich.org/download/
y posterior descompresión por línea de comandos (*bash, gunzip –k*)

“str_catalog_scripts.tar.gz”, “hg19_ref.tar.gz” y “expanded_hg19_ref.tar.gz” desde http://strcat.teamerlich.org/download/ y posterior con *tar zxf*

“build37.xlsx” desde http://humanparalogy.gs.washington.edu/build37/data/
“1000genomes.sequence.index” desde ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/

Información de genes (Id, CCDS id, sinónimos, etc.) descargados desde el sitio https://www.genenames.org/download/custom/.

I.2 Desde línea de comandos (*bash, wget –c*)

Archivos en formato fasta comprimido (extensión "*.fa.gz") de los 24 cromosomas humanos del ensamblaje GRCh37/hg19 (Feb. 2009). Desde http://hgdownload.soe.ucsc.edu/goldenPath/hg19/chromosomes/ y posterior descompresión con *gunzip -k*

Archivo único en formato fasta (extensión "*.2bit") del ensamblaje del genoma humano GRCh37/hg19 (Feb. 2009) desde http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/ convertido luego en formato fasta (extensión "*.fa") por la herramienta twoBitToFa (ver más adelante).

Archivo UniSTS_human.sts desde ftp://ftp.ncbi.nih.gov/pub/ProbeDB/legacy_unists/

Archivos en formato fastq comprimido (extensión "*.fastq.gz") de los 24 genomas humanos secuenciados con tecnología Illumina HiSeq 2000 a partir de librerías hechas con Solexa desde ftp://ftp.sra.ebi.ac.uk/vol1/fastq/~/~/~/ (la ruta varía dependiendo del archivo). Este es el listado:
HG00096 (masculino): SRR1291026, SRR1291035; HG00268 (femenino): SRR1293236, SRR1293262; HG00419 (femenino): SRR1295433, SRR1295554; HG00759 (femenino): SRR1293295, SRR1293326; HG01051 (masculino): SRR1291141, SRR1291157; HG01112 (masculino): SRR1291024, SRR1291070; HG01500 (masculino): SRR1295423, SRR1298981; HG01565 (masculino): SRR1295426, SRR1298989; HG01583 (masculino): SRR1291030, SRR1291036; HG01595 (femenino): SRR1295536, SRR1295552; HG01879 (masculino): SRR1295533, SRR1295534; HG02568 (femenino): SRR1295425, SRR1298980; HG02922 (femenino): SRR1295543, SRR1295553; HG03006 (masculino): SRR1295568, SRR1295570; HG03052 (femenino): SRR1295432, SRR1295535; HG03642 (femenino): SRR1295466, SRR1295515; HG03742 (masculino): SRR1293251, SRR1293283; NA18525 (femenino): SRR1295532, SRR1295539; NA18939 (femenino): SRR1295537, SRR1295540; NA19017 (femenino): SRR1295544, SRR1295546; NA19625 (femenino): SRR1295538, SRR1295545; NA19648 (femenino): SRR1291041, SRR1291138; NA20502 (femenino): SRR1295424, SRR1298988; NA20845 (masculino): SRR1295465, SRR1295542.
Nombre del estudio: High Coverage PCR Free sequencing of 1000 Genomes Samples.

Archivos en formato fastq comprimido (extensión "*.fastq.gz") de los 3 genomas humanos secuenciados con tecnología Illumina HiSeq 2000 a partir de librerías hechas con Illumina desde ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR622//~/ (la ruta varía dependiendo del archivo). Este es el listado:
NA12878 (femenino): SRR622457; NA12891 (masculino): SRR622458;
NA12892 (femenino): SRR622459.
Nombre del estudio: 1000 Genomes CEU high coverage sequencing

Todos las secuenciaciones se hicieron del tipo apareados o *paired-end* por lo tanto se descargaron ambos archivos identificados como "_1.fastq.gz" y "_2.fastq.gz", que oportunamente fueron descomprimidos con *gunzip –k* para los análisis.

Se hicieron análisis de calidad a todos los fastqs descargados (ver FastQC en el Anexo siguiente, en el apartado "Desde sitio web").

ANEXO II. DESCARGA E INSTALACIÓN DE HERRAMIENTAS

Algunas herramientas no necesitaron instalación, otras sus ubicaciones requieren ser incorporadas a las variables del entorno del sistema o reubicación de los binarios y otras necesitaron ser compiladas desde el código fuente. Algunas necesitaron una combinación de las anteriores, siempre siguiendo las recomendaciones de instalación por parte de los desarrolladores y/o autores. En caso de errores, se consultaron foros específicos dedicados a soluciones en línea.

II.1 Desde sitio web

Notepad++ v7.5.9 (32 bit) desde https://notepad-plus-plus.org/download/v7.5.9.html

WampServer
"wampserver3.1.0_x64.exe" desde
https://sourceforge.net/projects/wampserver/files/WampServer%203/WampServer%203.0.0/wampserver3.1.0_x64.exe/download (plataforma que incluye phpMyAdmin 4.7.4 y MySQL 5.7.19)

Tandem Repeats Finder 4.09 for 64 bit Linux ("trf409.linux64") desde
http://tandem.bu.edu/trf/trf409.linux64.download.html

STRaitRazor desde https://github.com/Ahhgust/STRaitRazor/archive/master.zip, luego descomprimido con *winrar.*

Bioedit Sequence Alignement Editor (versión 7.2.5)
Desde http://www.mbio.ncsu.edu/BioEdit/bioedit.html y descomprimido con *winrar*

Word y Excel (Microsoft Office 2007)

IGV, Integrative Genomics Viewer
Se descargó desde:
https://data.broadinstitute.org/igv/projects/downloads/2.7/IGV_Win_2.7.2-installer.exe
y se ejecutó el instalador. Este programa requiere Java SE Runtime Environment (JRE ™) para su funcionamiento.
El sitio de descarga de JRE es:
https://www.oracle.com/technetwork/es/java/javase/downloads/jre8-downloads-2133155.html?printOnly=1

FastQC v0.11.8 (Win/Linux zip file)
Se descargó desde:
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.8.zip
luego descomprimido con *unzip*.
Este programa se puede ejecutar desde el binario "fastqc" en la carpeta "FastQC", en Ubuntu usa J.R.E para abrir la aplicación. Desde ahí se hicieron todos los análisis de calidad de los 102 fastq descargados.

II.2 Desde línea de comandos (*bash, wget ó wget –c desde http o usando Github, git clone*)

Instalación de API (del inglés *application program interface*, interfaz del programa de aplicación) de Ensembl (release 94 - October 2018):
Bioperl desde https://github.com/bioperl/bioperl-live/archive/bioperl-release-1-2-3.zip (*wget*), posterior descompresión con *unzip*.
Ensembl desde ftp://ftp.ensembl.org/pub/ensembl-api.tar.gz (*wget*) y posterior descompresion con *tar -zxf*

Bioseq, version 1.12
Instalado según instrucciones de los desarrolladores desde https://github.com/bioperl/p5-bpwrapper (*git clone*)

VCFtools, version 0.1.15 desde https://github.com/vcftools/vcftools.git (*git clone*)

Nucleotide-Nucleotide BLAST 2.7.1+
"ncbi-blast-2.7.1+-x64-linux.tar.gz" desde
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ (*wget –c*), luego descomprimido con *tar zxvpf*

TSSV desde https://git.lumc.nl/j.f.j.laros/tssv.git (*git clone*)

lobSTR
"lobSTR-bin-Linux-x86_64-4.0.6.tar.gz" desde https://github.com/mgymrek/lobstr-code/releases/download/v4.0.6/, luego descomprimido con *tar zxf.*

Samtools (Versión: 0.1.19-96b5f2294a)
Desde https://github.com/samtools/samtools.git (*git clone*)

Bedtools
versión 2.17.0-1 0 usando el comando *apt-get install bedtools*

Bcftools
Desde https://github.com/samtools/bcftools.git (*git clone*)

CrossMap
Usando el comando *pip2 install CrossMap –upgrade*
Y el archivo hg19ToHg38.over.chain.gz desde
http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/
Descompresión con *gunzip –k*

MAFFT
versión: 7.407-1 usando el comando *apt-get install mafft.*

Tablas principales de la base de datos ("ucsc", "fase1", "cytobands", "intron", "exon", "superdup", "ensembl", "snp150common" y "319markers"):

Las Tablas fueron realizadas en Microsoft Excel inicialmente, para ser exportadas en formato en "*.csv", para combinar columnas de otras Tablas {funciones indice () y coincidir ()} y para hacer los primeros cruces de datos: ubicación de repeticiones secuenciadas versus ubicación de marcadores UCSC (repeticiones dentro de marcadores, marcadores dentro de repeticiones, ubicaciones que se solapan y ubicaciones en las cercanías) idem funciones

funciones indice () y coincidir (). Como resultado de este procesamiento, las nuevas Tablas obtenidas modificadas a partir de las originales se continuaron editando en este programa, o bien fueron manejadas por sentencias en línea de comandos de Linux (Ubuntu).

ANEXO III. SENTENCIAS Y SCRIPTS

En este apartado se presenta un resumen sobre la utilización de ejecutables pre-instalados o pertenecientes a la distribución de Linux: Ubuntu 14.04.5 LTS (Release: 14.04, Codename: trusty) tales como *bash*, *sh*, *ls*, *awk*, *cat, wc, cut, tr* y *grep* que junto con Perl v5.18.2 (y Bioperl 1.007002 para la ejecución de algunas tareas) y Python v2.7.6 fueron sumamente usados para el recorrido de archivos y generación de nuevos, con distintos propósitos, desde listados del contenido de directorios, generación de archivos en formato fasta (“*.fasta”, “*.fas”, “*.fa” usando módulo Bio::SeqIO de Bioperl), archivos ejecutables (scripts *.sh, *.pl, *.sql y *.py), limpieza de contenidos en archivos “*.txt”, generación de Tablas en formato “*.csv” o cualquier salida de texto plano (“*.bed”, “*.config”, etc.) de los programas mencionados en el apartado anterior.

Todos los scripts y sentencias fueron generados y editados con Notepad ++.
La totalidad de sentencias (historial de línea de comandos, función *history*) y *scripts* usados en este trabajo se adjuntan en las rutas provistas en el Anexo VI.

III.1 Misceláneos por líneas de comandos (bash, cmd)

Extracción de columnas 1 a 9 del archivo de anotación “vcf” de 1000 genomas:

```
cat phase_1_final_calls.vcf | grep "chr1" | cut -f 1-(8,9) > phase_1_final_calls_chr1-9cols.txt

cat phase_1_final_calls_chr1yotros-9cols.txt | grep -w "chr(1,10..19)" | cut -f 1-9 > phase_1_final_calls_chr1-9cols.txt

cat phase_1_final_calls.vcf | grep "chr2" | cut -f 1-8 > phase_1_final_calls_chr2yotros-8cols.txt

cat phase_1_final_calls_chr2yotros-8cols.txt | grep -w "chr(2, 20..22)" | cut -f 1-8 > phase_1_final_calls_chr2-8cols.txt

cat phase_1_final_calls.vcf | grep "chr(3..9, X e Y)" | cut -f 1-8 > phase_1_final_calls_chr3-8cols.txt
```

Limpieza del archivo “csv” de la Tabla principal de marcadores de UCSC en la base de datos realizada con Excel:

```
awk -F, 'length($3) && length($4) ' t3_markers.csv > t3_markers_filtered.csv
```

Extracción de datos de las salidas del programa TRF sobre las secuencias obtenidas de la base de datos de Ensembl:

```
awk 'FNR==(16..29) {print FILENAME,"\t" $1 "\t" $2 "\t" $3 "\t" $4 "\t" $5 "\t" $6 "\t" $7 "\t" $8 "\t" $9 "\t" $10 "\t" $11 "\t" $12 "\t" $13 "\t" $14 "\t" $15}' *.dat > fila-16-todo-ensembl-part1-2000.txt

cat fila-(16..29)-todo-ensembl-part1-2000.txt > fila-16al29-todo-ensembl-part1-2000.txt
```

III.1.1 Consultas MySQL con la BD local

III.1.1.1 Variantes STRs dentro de regiones STSs.

Se generaron las consultas con el script de Perl: "*gen-sql-strin-mkin-solmin-solmax.pl*", para cada cromosoma: "*strin-mkin-solmax-solmin-ensembl_v1-chr(1..22,X,Y).sql",* y se ejecutaron con el script: "*strin-mkin-solmax-solmin-ensembl_v1_all.sql*"
De manera genérica se detalla la consulta a continuación:

Inclusión (STRIN):

SELECT COUNT(*) FROM fase1,(ensembl ó UCSC) WHERE fase1.numcrom='chr(1..22,X,Y)' AND fase1.numcrom=(ensembl ó UCSC).numcrom AND fase1.start >= (ensembl ó UCSC).start AND (ensembl ó UCSC).end >= fase1.end; (ubicaciones de repeticiones secuenciadas dentro de ubicaciones de marcadores). Guardados en un archivo, función TEE (TEE c:/wamp64/tmp/strin_mkin_solmax_solmin_chr(1..22,X,Y)_(ensembl ó UCSC)_v1.txt, NOTEE).

Inclusión (MKIN):

SELECT COUNT(*) FROM fase1,(ensembl ó UCSC) WHERE fase1.numcrom='chr(1 .. 22, X, Y)' AND fase1.numcrom=(ensembl ó UCSC).numcrom AND (ensembl ó UCSC).start >= fase1.start AND fase1.end >= (ensembl ó UCSC).end; (ubicaciones de repeticiones secuenciadas dentro de ubicaciones de marcadores). Guardados en un archivo, función TEE (TEE c:/wamp64/tmp/ strin_mkin_solmax_solmin_chr(1 .. 22, X, Y)_(ensembl ó UCSC)_v1.txt, NOTEE).

Solapamientos:
Inferior (SOLMAX):
SELECT COUNT(*) FROM fase1,(ensembl ó UCSC) WHERE fase1.numcrom='chr1' AND fase1.numcrom=(ensembl ó UCSC).numcrom AND fase1.start >= (ensembl ó UCSC).start AND (ensembl ó UCSC).end <= fase1.end AND (ensembl ó UCSC).end >= fase1.start; Guardados en un archivo, función TEE (TEE c:/wamp64/tmp/ strin_mkin_solmax_solmin_chr(1..22,X,Y)_(ensembl ó UCSC)_v1.txt, NOTEE).

Superior (SOLMIN):
SELECT COUNT(*) FROM fase1,(ensembl ó UCSC) WHERE fase1.numcrom='chr1' AND fase1.numcrom=(ensembl ó UCSC).numcrom AND (ensembl ó UCSC).start >= fase1.start AND (ensembl ó UCSC).end >= fase1.end AND fase1.end >= (ensembl ó UCSC).start; Guardados en un archivo, función TEE (TEE c:/wamp64/tmp/ strin_mkin_solmax_solmin_chr(1..22,X,Y)_(ensembl ó UCSC)_v1.txt, NOTEE).

III.1.1.2 Variantes STRs dentro de regiones génicas e intergénicas.

El script de Perl *"fase1_vs_(exon ó intrón)_v2_all.pl"* generó las consultas por cromosoma: *"fase1_vs_(exon ó intrón)_chr(1..22,X,Y).sql",* reunidos en un único script *"fase1_vs_(exon ó intrón)_v2_all.sql"* que las ejecutó.

III.1.1.3 Distancias acumuladas en pb de variantes STRs de Fase 1, exones e intrones obtenidas desde la base de datos local.

Se calcularon los tamaños en pb de las variantes STRs de Fase1, los exones e intrones según su inicio y final, y se sumaron estos valores:
SELECT END-START FROM intron;
SELECT END-START FROM exon WHERE strand='+';
SELECT START-END FROM exon WHERE strand='-';

SELECT END-START FROM fase1;

III.1.1.4 Marcadores STSs de UCSC y de Ensembl y las variantes STRs dentro de segmentos múltiples.

Se utilizó un criterio de inclusión de la variante STR ó la secuencia STS (según su ubicación en cada cromosoma) dentro de los segmentos incluidos en la base de datos local.

Los scripts de Perl: *"fase1_vs_superdup_v1_all.pl", "UCSC_vs_superdup_v1_all.pl"* y *"ensembl_vs_superdup_v1_all.pl"* generaron los scripts MySQL: *"fase1_vs_superdup_chr(1..22,X,Y).sql", "UCSC_vs_superdup_chr(1..22,X,Y).sql"* y *"Ensembl_vs_superdup_chr(1..22,X,Y).sql"* respectivamente, y fueron ejecutados con los scripts: *"fase1_vs_superdup_v1_all.sql", "UCSC_vs_superdup_v1_all.sql "* y *"Ensembl_vs_superdup_v1_all.sql"*

De manera genérica se detalla la consulta a continuación:

SELECT COUNT(*) FROM (fase1, ensembl ó UCSC),superdup WHERE (fase1, ensembl ó UCSC).numcrom='chr(1..22,X,Y) AND (fase1, ensembl ó UCSC).numcrom=superdup.chrom AND (fase1, ensembl ó UCSC).start >= superdup.chromStart AND superdup.chromEnd >= (fase1, ensembl ó UCSC).end; Guardados en un archivo, función TEE (TEE c:/wamp64/tmp/ (fase1, ensembl ó UCSC)_vs_superdup_chr(1..22,X,Y).txt, NOTEE).

III.1.2 Consultas MySQL desde la base de datos de UCSC

Habiendo definido los marcadores del punto 3.5, se usaron las ubicaciones para obtener de esta base las variantes SNPs en esas regiones.

Ingreso por línea de comandos:

mysql --user=genome --host=genome-mysql.soe.ucsc.edu -N -A -P 3306

use hg19; (consulta a Tabla SNP150)

De manera genérica la consulta fue:

*SELECT * FROM snp150 WHERE chrom='chr*(1..22,X,Y)*' AND chromStart >= (comienzo STR) AND chromEnd <= (final STR);*

Guardados en un archivo, función TEE (TEE c:/wamp64/tmp/ result_query-mysql_chr(1 .. 22, X, Y)_ *(comienzo STR) _(final STR)*.txt, NOTEE).

Las consultas "sql" fueron generadas con scripts de Perl: *"print-sql-sentences-search-by-location.pl",* y las salidas depuradas con *"clean-line-sql-result.pl".*

III.1.2.1 Reordenamiento de la salida de las 30 partes de los scripts de perl de la API de Ensembl "dbID_all_info_(1..30)_10K_2_markers_GRCh37_v2.txt":

awk '{if ($2 ~ /^[0-9]/ && $2 <= 22) {print $1, "\t", $2, "\t", $3, "\t", $4, "\t", $5, "\t", $6,"\t", $7,"\t", $8, "\t", $9, "\t", $10,"\t", $11,"\t", $12, "\t", $13, "\t", $14, "\t", $15, "\t", $16, "\t", $17, "\t", $18, "\t", $19, "\t", $20, "\t", $21, "\t", $22, "\t", $23, "\t", $24, "\t", $25, "\t", $26, "\t", $27, "\t", $28, "\t", $29 } else if ($2 !~ /^[0-9]/ && ($2 ~ "X" || $2 ~ "Y")) {print $1, "\t", $2, "\t", $3, "\t", $4, "\t", $5, "\t", $6,"\t", $7,"\t", $8, "\t", $9, "\t", $10,"\t", $11,"\t", $12, "\t", $13, "\t", $14, "\t", $15, "\t", $16, "\t", $17, "\t", $18, "\t", $19, "\t", $20, "\t", $21, "\t", $22, "\t", $23, "\t", $24, "\t", $25, "\t", $26, "\t", $27, "\t", $28, "\t", $29, "\t", $30 } else if ($2 ~ /^[0-9]/ && $2 > 22) {print $1, "\t", "\t", "\t", "\t", "\t","\t","\t", $2, "\t", $3, "\t", $4, "\t",$5, "\t", $6,"\t", $7,"\t", $8, "\t", $9, "\t", $10,"\t", $11,"\t", $12, "\t", $13, "\t", $14, "\t", $15, "\t", $16, "\t", $17, "\t", $18, "\t", $19, "\t", $20, "\t", $21, "\t", $22,"\t", $23}}' dbID_all_info_(1..30)_10K_2_markers_GRCh37_v2.txt > dbID_all_info_(1..30)_10K_2_markers_GRCh37_v2_reordened.txt

y posterior concatenación:

cat dbID_all_info_(1..30)_10K_1_markers_GRCh37_v2_reordened.txt > dbID_all_info_1_10K_1-30_markers_GRCh37_v2_reordened.txt

III.1.2.2 Extracción de algunas columnas de los archivos anteriores para la comparación estricta de la secuencia de cebadores de Ensembl con los de UCSC y UniSTS:

cat dbID_all_info_(1..30)_10K_1_markers_GRCh37_v2_reordened.txt | cut -f 1-5,8-10 > dbID_all_info_1_10K_1-30_markers_GRCh37_v2_reordened_cut1-2-3-4-5-8-9-10.txt

III.2 Archivos de texto plano ejecutables (scripts)

Usando la API de ensembl se obtuvo los 299.818 identificadores (IDs) de los marcadores que contiene esta base de datos (*script perl,* módulo: Bio::EnsEMBL::Registry, Puerto: 3337 – ensamblaje GRCh37-, **adaptador: 'Marker'**: fetch_all).

Dado que la conexión a la base era inestable con la totalidad de los identificadores para traer información de cada marcador (**Adaptador 'Marker'** : fetch_by_dbID: *dbID, left_primer, right_primer, min_primer_dist;* get_all_MarkerFeatures: *seq_region_name, display_id, start, end*; get_all_MarkerSynonyms: source, name; get_all_MapLocations: *map_name, position*; **Adaptador 'Slice'**: fetch_by_region: *seq, strand*.) Se probó un número menor de IDs: 100.000, 50.000 y 10.000 siendo este último el óptimo para la consulta. Se particionó entonces el *script* de perl original en 30 *scripts* y se ejecutaron dentro de un *script* de bash (*.sh).

Scripts: "*splicing-script-splice-array-by-10K-markers-dbDIs-GRCh37.pl"* que genera los 30 *scripts* para la consulta: *"fetch_all_info_1_markers_by_10K_(1 al 30)_GRCh37_v2.pl* (dentro del *script for-perl1.sh* que los ejecuta)"

III.2.1 Conformación de las Tablas para EXONES e INTRONES

Se combinan los nombres de HUGO con demás datos de CCDS:

Se parte de los archivos "ccdsGene.txt" y "HGNC-gene-names.txt" y se generan otros con los siguientes *scripts*:

symbol-ccds.pl, change-CCDS_ID-by-symbol.pl, gene-ccds-intron-y-numero-posiciones-totales.pl, gene-ccds-intron-y-numero-posiciones-totales.pl

Para generar las Tablas "exon-gene.csv" y "intron-gene.csv".

III.2.2 Generación de archivos fasta de las secuencias STS Ensembl:

Las secuencias contenidas en el archivo "dbID_all_info_1_10K_1-30_markers_GRCh37_v2_reordened.txt" fueron extraídas con el script de Perl "*export-fasta-from-ensembl-seqs-1-30.pl*", se genera un archivo multifasta, posteriormente fue dividido con el comando:

Bioseq - - break dbID_all_info_1_10K_1-30_markers_GRCh37_v2_reordened.fa

Generando así archivos individuales de cada secuencia STS extraída.

Estas secuencias son 500 pb más grandes que el largo original contenido en esta base, por lo tanto fueron recortadas con el *script* "*shortening-seqs.pl*" para generar los archivos de secuencias del largo original (+/- 0 pb).

UCSC:

A partir de los datos de coordenadas y cromosoma del archivo "t3_markers_filtered_id_chr_start_end_cleaned.txt", el archivo del genoma "hg19.fa" y el script "*gen-seq.pl*" se generaron los archivos de secuencias STS de largo original de esta base (+/- 0 pb) y con el script "*gen-seq-mm250pb.pl*" aquellas secuencias de largo extendido de 250 pb a ambos lados del largo original.

III.2.3 Coincidencias de las secuencias de pares de cebadores dentro de las secuencias STSs blanco.

Los archivos "human-UniSts-primers-2p-replaced.txt", "ucsc_name_r_l_primers_awk-cleaned-2p-replaced.txt" y "primer-ensembl_awk-cleaned.txt" fueron preparados a partir de archivos originales de descarga a los fines de ser recorridos junto con las secuencias del paso anterior con 12 scripts de Perl y obtener las coincidencias.

Script: "*analyze-seq-primers-(DB1)-into-(DB2)-seqs-mm(n)pb.pl*" donde DB1 son UniSTS, Ensembl y UCSC e indica el listado de cebadores de esas bases, DB2 es Ensembl o UCSC e indica las secuencias STS de esas bases y n es 0 ó 250, largo original o extendido respectivamente.

Las salidas de estos *scripts* proporcionan los conteos de las búsquedas generando de manera genérica los archivos "primers_(DB1)_into_(DB2)_seqs_mm(n)pb.txt".

III.2.4 Archivos que alojan cálculos estadísticos

En el archivo "cebadores-versus-seqs.xlsx" se reúnen los conteos de los archivos del apartado anterior y los datos son usados en las hojas "0pb-graph" y "250pb-graph" para generar los gráficos de la Figura 9 y llevados al *script* "*figura9.R*" para realizar el test de homogeneidad X^2 [Chi cuadrado, función *chisq.test()]*

De los archivos "fig11.xlsx" y "fig12.xlsx" se obtuvieron aquellos valores donde las herramientas lobSTR y StraitRazor ó ambas no tuvieron detecciones y fueron reunidas en el archivo "datos-fig-11-12-for-stats.xlsx", y desde la hoja "csv" se exportó a un archivo que es leído por el *script* "*datos-fig-11-12-for-stats.R*", que ejecutado genera un gráfico en cajas [*boxplot*()] y los resultados de la Prueba T [*t.test*()] y Prueba F [*var.test*()].

/II.2.5 Manejo de los archivos de configuración de StraitRazor

Los *scripts* de Perl *"get-anchors-considering-snp150f-todos3.pl"* y *"get-anchors-considering-snp150f-argv-todos2.pl"* se usaron para generar el archivo "*.config" y consultar datos de cada marcador de manera individual respectivamente.

Hubieron distintas versiones del archivo "*.config" cuya maduración se hizo en la hoja de cálculo: "*STRs_en_estudio_vXX_config.xlsx*". Las correcciones en cada versión se debió a las siguientes cuatro situaciones:

a) aparición de inespecificidades evidentes,
b) desaparición de asignación alélica cuando una versión anterior la tenía,
c) agregado de marcadores encontrados con posterioridad,
d) falta de concordancias de asignación alélica con la herramienta lobSTR.

Después de algunos análisis exploratorios con 4 a 56 fastqs seleccionados se llegó a una versión final "*STRs_en_estudio_v183.config*".

III.2.6 Manejo de los archivos de configuración de lobSTR

A partir del archivos "lobstr_v3.0.2_hg19_ref.bed" y con varios *scripts* de Perl *"analyze-location(2 a 5).pl"* se generaron varios archivos "*.bed" que fueron reunidos en la hoja de cálculo: "*CTRL-STRS-FOR-CUSTOMIZED-BED.xlsx*" para una depuración. Se hicieron correcciones debido a las siguientes situaciones:

a) aparición de inespecificidades evidentes,

b) agregado de marcadores encontrados con posterioridad,
c) falta de concordancias de asignación alélica con la herramienta StraitRazor.

Después de algunos análisis exploratorios con los mismos 4 a 56 fastqs seleccionados se llegó a una versión final "*depured_bed_data_v32.bed*".

Este archivo es usado por los *scripts* de Python "*lobstr_index.py*" y "*GetSTRInfo.py*" provistos por el desarrollador para generar la carpeta index_custom_(v1, v2) y el archivo "strinfo_custom_(v1, v2).tab" respectivamente, que usará la herramienta en los análisis.

III.2.7 Manejo de los archivos de salida de StraitRazor

Todas las salidas de este programa tuvieron de forma genérica el siguiente nombre de archivo:

STRs_en_estudio_v(XX)-(acrhivo fastq)_i.txt

Donde v(XX) se refiere a la versión del archivo "*.config" (versiones: v1, v14, v15, v16, v161, v17, v18, v181, v182 y v183).

Y (archivo fastq) es el nombre (sin extensión) del archivo fastq analizado. Para las versiones v1 a la v161 sólo se analizaron dos archivos fastq: SRR1291024_1.fastq y SRR1295424_1.fastq para la evaluación y corrección de inespecificidades. Con las versiones v17 y v18 se continuaron esas mismas inspecciones, pero con un número mayor de "fastqs", sumando ahora las secuencias obtenidas con el *script* "get-*seq-from-results.pl*". Esta necesidad surgió debido a la complejidad de algunos STRs para pasar los parámetros correctos al "*.config".

En esta instancia de análisis se hizo evidente incorporar una segunda opción de análisis, y lobSTR fue la segunda opinión para contrastarlo con StraitRazor.

El *script* "*analyze-output-str8rzr2.pl*" simplificó los resultados extrayendo las asignaciones alélicas por marcador de todas las salidas, incluyendo las de la versión v183 (final) para los 102 fastqs totales.

Hasta aquí no se consideraba los conteos de lecturas. Con la incorporación de la herramienta lobSTR, y haciendo los primeros contrastes entre ambas herramientas (con las primeras versiones de planillas de Excel para los contrastes) y debido a que las intervenciones manuales eran predominantes, fue necesario incorporar el valor de conteo de lecturas y establecer así decisiones automáticas. Ver más adelante en III.2.9.

III.2.8 Manejo de los archivos de salida de lobSTR

Los *scripts "samtools-allelotype-awk-script-perl(0, 1, 2, 3, 4, 5, 6 y 7).sh"* ejecutó Samtools (comandos -sort, -index y –allelotype) generando el archivo "(fastq1y2).vcf" a partir del achivo "(fastq1y2). aligned.bam" realizado por lobSTR. (ver Tabla 4 donde se detalla el uso de los fastq_1 y fastq_2 por Análisis ID)

Luego los *scripts* "*make-genotype.pl*" y "*analyze-lobSTR-vcf-results3.pl*" dentro de los *scripts* "*analyze-lobSTR-vcf-results(3 y 4).sh*" generó los archivos "genotype-v2-(fastq1y2).txt" y "genotype-v2-(fastq1y2)-short.txt" Siendo ambos archivos extractos ampliados y resumidos de la información anotada en el archivo VCF. Estos archivos reunieron información de conteos de lecturas, y fueron usados para todos los contrastes exploratorios y finales, pero estos últimos con el script "*chek-lobSTR-str8rzr-by-sample-by-marker7.pl"* que se explica en el siguiente apartado.

La anotación VCF (versión 4.1 para la versión 4.0.6 de lobSTR) contiene los siguientes 8 campos mandatorios:

CHROM, POS, ID, REF, ALT, QUAL, FILTER e INFO.

Y el noveno FORMAT opcional, si hay datos de genotipos en la línea.

Los campos INFO y FORMAT para la anotación con lobSTR son los siguientes:

INFO (datos separados por punto y coma):

RPA ="*Repeats per allele*"
END ="*End position of variant*"
MOTIF ="*Canonical repeat motif*"

NS ="*Number of samples with data*"
REF ="*Reference copy number*"
RL ="*Reference STR track length in bp*"
RU ="*Repeat motif*"
VT ="*Variant type*"
De estos sólo fueron extraídos RPA y REF.

FORMAT: (datos separados por dos puntos):

ALLREADS="*All reads aligned to locus*"
AML="*Allele marginal likelihood ratio scores*"
DISTENDS="*Average difference between distance of STR to read ends*"
DP="*Read Depth*"
DPA="*Read Depth, including filtered reads*"
GB="*Genotype given in bp difference from reference*"
PL ="*Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification*"
PQ="*-log10(1-Q), where Q is as reported in the Q field*"
Q="*Likelihood ratio score of allelotype call*"
GT="*Genotype*"
SB="*Strand bias*"
STITCH="*Number of stitched reads*"
De estos sólo fueron extraídos DP y GT.

El campo GT de FORMAT contiene el dato de genotipo y fue el más conveniente para determinar la asignación alélica (se podría haber usado GB), y fue manejado dentro de los *scripts* según los siguientes criterios:

0/0 = asignación alélica igual al alelo de referencia.

0/1 ó 1/0 = un alelo con asignación alélica igual al alelo de referencia y se debe calcular la asignación del alelo alternativo.

1/1 = sólo se debe calcular la asignación del alelo alternativo.

1/2 = se debe calcular la asignación de ambos alelos alternativos.

Los cálculos se hicieron con las longitudes de las secuencias de REF y ALT de los campos mandatorios y la longitud de la repetición (del archivo *.bed). Se podrían haber usado otros métodos para obtener las asignaciones alélicas.

Hasta aquí las asignaciones alélicas contienen decimales no acordes a la nomenclatura de ISFG. En el apartado siguiente se detallan las conversiones efectuadas en las planillas de Excel.

III.2.9 Reunión de las salidas de ambas herramientas para cada muestra

Los archivos “bed-info-for-analysis-lobSTR-v2.txt”, “genotype-v2-(fastq1y2).txt” de lobSTR y “STRs_en_estudio_v183-(acrhivo fastq)_(1, 2)_i.txt” de StraitRazor fueron examinados con el *script* “*chek-lobSTR-str8rzr-by-sample-by-marker7.pl”* para reunir en un único archivo de texto plano las asignaciones alélicas de ambas herramientas, por muestra (1 ó 2 Análisis ID, 2 ó 4 fastqs, ver Tabla 4)

Este *script* de Perl fue ejecutado dentro del *script* “*doble-loop-for-check-pl.sh*” y se obtuvieron los 27 perfiles para los 323 marcadores STR elegidos, en 27 archivos de texto plano (tabulado).

Los distintos archivos de la muestras (nombre: “lobSTR-STR8Rzr-(muestra).txt”) fueron copiados y pegados posteriormente a la planilla de cálculo habilitada para macros “PERFIL_v2_PLANTILLA_lobSTR_v32_str8rzr_v183.xlsm” en la hoja “salida-script”, en la hoja

siguiente “re-perfilado” se realizan las comparaciones automáticas y/o intervenciones manuales entre ambas herramientas.

En esta hoja se hacen una serie de correcciones de los datos ingresados en la primera hoja y cálculos de conteos de lecturas. Los más relevantes son:

a) Notación decimal de lobSTR a nomenclatura de ISFG
Largo de la repetición 2 nt: .5 a .1
Largo de la repetición 3 nt: .333 a .1 y 0.666 a .2
Largo de la repetición 4 nt: .25 a .1; .5 a .2 y .75 a .3
Largo de la repetición 5 nt: .2 a .1; .4 a .2; .6 a .3 y .8 a .4
Largo de la repetición 6 nt: .167 a .1; .333 a .2; .5 a .3; .666 a .4 y .833 a .5
b) Los valores nulos fueron cambiados a “ND”: No Detectado.
c) Se eliminan espacios que no permiten la comparación.
d) Se obtienen los valores de los alelos con mayores conteos de lecturas y se realiza el cociente entre estos dos valores, para aplicar las reglas que filtran los tartamudeos. Estas reglas se discuten con más detalles en la sección “4.7.1 Reglas implementadas para adecuar las asignaciones alélicas entre lobSTR y StraitRazor” dentro de “4 Resultados de utilización y curado de herramientas bioinformáticas”.

Las columnas AG: “FINAL (str8rzr-lobSTR merged)” y AH: “FINAL (*merged* + intervención manual)” fueron copiados y incorporados en la planilla “perfiles-finales.xlsm” de todos los perfiles. Este archivo centraliza todas las asignaciones alélicas de las 27 muestras y 323 marcadores. Además, este archivo fue base del archivo “perfiles-finales-depured-v1.txt” que sirvió para generar todos los resúmenes e histogramas de frecuencias absolutas por marcador, secuencias fastas finales e incorporación de datos en la Tabla “319markers” de la BD local.

Las columnas AC: “suma 1er y/o 2do mayoritarios de lobSTR” y AE: “suma 1er y/o 2do mayoritarios de StraitRazor” también fueron reunidos en un único archivo “visualizaciones-profundidad-lobSTR-StraitRazor.xlsm” que muestra los conteos de lecturas de ambas herramientas, y de donde se hace gráficos en R (archivos “profundidad-lobSTR.csv”, “profundidad-StraitRazor.csv” y *script* “*profundidades.R*”)

Por último, la hoja “stats” de cada perfil toma datos de la hoja “re-perfilado” para realizar cálculos, que reunidos en el archivo “perfiles-finales-v1-estadisticas.xlsm” se realizan resúmenes globales de las asignaciones alélicas y gráficos.

En el Anexo VI se comparte todas las sentencias de reglas condicionales usados para el manejo de estos datos, también se comparten todos los perfiles y la plantilla “molde”, donde se observan no sólo el funcionamiento de las reglas en las celdas, sino también formatos condicionales de las columnas AG y AH para las asignaciones automáticas y manuales.

III.3 Sentencias específicas de las herramientas

III.3.1 Extracción de frecuencias de cada repetición del archivo de anotación “vcf” de 1000 genomas:

vcftools --vcf phase_1_final_calls.vcf --freq --chr (1..22, X e Y) --out chr(1..22, X e Y)_analysis

III.3.2 Recodificación del archivo "phase_1_final_calls.vcf" por cromosoma y extrayendo aquellas anotaciones que tengan como mínimo 2 alelos:

vcftools --vcf phase_1_final_calls.vcf --chr chr(1..22, X, Y) --min-alleles 2 --out chr(1..22, X, Y) --recode

1238q-++,- 4, 5 se corresponden con datos de secuenciación de las muestras en las columnas 24, 79, 128, 129, 235, 343, 405, 526, 732, 782, 901 elegidas por que se trabajarán con ellas más adelante para los análisis de lobSTR y Straitrazor. En este filtro fueron elegidas 11 pero se trabajarán con 27 en total.

cat chr(1..22, X, Y).recode.vcf | cut -f 1,2,4,5,24,79,128,129,235,343,405,526,732,782,901 | grep -v ./.:.:.:.:.:.:. | grep -v \## > chr(1..22, X, Y) -recode-min2alleles-high-coverage-only-detected.txt

III.3.3 Obtención de zonas repetitivas de secuencias con el programa TRF (binario trf409.linux64) sobre secuencias en formato fasta (*.fa, *.fasta ó *.fas):

./trf409.linux64 (*.fa, *.fas) 2 7 7 80 10 50 (6-10) (-f -d –h –ngs)

Los argumentos varían dependiendo de la secuencia blanco y de lo que se quiere obtener como salida. Las secuencias diana fueron secuencias de marcadores de la base de datos, marcadores específicos de uso en ciencias forenses o para detección de aneuploidías, de cromosomas individuales del genoma humano e incluso del genoma completo (ensamblajes hg19 y hg38). Luego del nombre completo del archivo fasta, vienen seis parámetros: 2, 7, 7, 80, 10 y 50 y son valores de configuración recomendados por el autor, equivalen a una ponderación por emparejamiento, por no emparejamiento, por inserción-deleción, probabilidad de emparejamiento, probabilidad de inserción-deleción (ambos en números enteros), y puntuación de alineamiento mínimo que se reporta, respectivamente. El último parámetro obligatorio corresponde al número de periodo (largo de la repetición) máximo que se reporta (6 ó 10). Dentro de los argumentos opcionales, -d genera un reporte de texto plano de salida. Se usaron según el caso: –h (supresión de la salida html), -f (secuencias flanqueantes del núcleo de la repetición) y –ngs (salida más compacta).

III.3.4 Ubicación de marcadores usando la herramienta Nucleotide-Nucleotide BLAST 2.7.1+, creando primero la base de datos con el ejecutable makeblastdb:

makeblastdb -in /mnt/d/Varioma/Datos/HG19/hg19.fa -input_type fasta -dbtype nucl -title hg19

Luego se usó cada secuencia disponible de los marcadores y fueron ubicados con el ejecutable blastn. Las sentencias fueron de manera genérica como sigue:

blastn -query /path-to-fasta-file/file.fas -db /mnt/d/Varioma/Datos/HG19/hg19.fa -out /path-to-output-destination/file-hg19-blastn-0.html -outfmt 0 -html -evalue 0.01

Se consideró aquella ubicación cuyo puntaje (*score*) sea el más elevado y que sea coherente con el cromosoma y citobanda donde debe encontrarse el marcador.

Con una sentencia similar a la anterior también se evaluaron todos los marcadores en estudio, pero acotados a las secuencias de anclaje del archivo config versión 183 (*scripts get-sequence-config183.pl* y *blastn1.sh*), con la intención de indagar cuántas veces aparece (con

100 % de identidad o menor, e-value: 0.01) una determinada secuencia dentro del genoma, y tener estos datos como criterio de especificidad.

III.3.5 Análisis de archivos fastq con StraitRazor 3.0

Se muestra la sentencia más usada de forma genérica:

./str8rzr –i -p 8 -c (archivo).config /ruta/(archivo).fastq > (archivo config)-(archivo fastq).txt
Donde:
-i: incluye las secuencias de anclaje dentro de la secuencia hallada
-p: número de núcleos del procesador
-c: archivo de configuración
Más la ubicación del archivo fastq
“>” y la salida se guarda en un archivo que contiene ambos nombre del archivo “config” y del fastq.
Se ejecutó la herramienta dentro de un script de bash, junto a otros comandos (*gunzip –k, rm*)

III.3.6 Análisis de archivos fastq con lobSTR 4.0.6

Se muestra la sentencia más usada de forma genérica:
/(ruta al binario)/./lobSTR --p1 /(ruta a fastq 1)/’(archivo)_1.fastq' --p2 /(ruta a fastq 2)/’(archivo)_2.fastq' -q --index-prefix /(ruta al índice)/index_custom_v2/lobSTR_ -o /(ruta de la salida)/lobSTR/(archivo.aligned.bam) --rg-sample (nombre) --rg-lib Solexa-206009

Cuando se definieron los archivos de configuración finales se ejecutaron ambas herramientas StraitRazor y lobSTR dentro de varios *scripts* de bash, junto a otros comandos (*gunzip –k, rm*)

III.3.7 Alineamientos de secuencias fasta con la herramienta MAFFT

La sentencia usada fue:
mafft *.fa > *_aligned_mafft.fa

III.3.8 Sentencias con bioseq

bioseq -l hg19.fa > length-hg19.txt (determina el largo en pb de cada cromosoma)
bioseq - - break dbID_all_info_1_10K_1-30_markers_GRCh37_v2_reordened.fa
(divide un multifasta y genera archivos fasta individuales)

ANEXO IV TOTALIDAD DE MARCADORES DEFINIDOS PARA ESTE ESTUDIO

Listado completo de los 323 marcadores discriminados en 4 grupos: Ciencias Forenses, Ciencias de la Salud, Comunes a Cs. Forenses y de la Salud, Elegidos de Fase 1 y Elegidos del cromosoma 7 y 9.

Entre paréntesis la procedencia de la información necesaria para incorporarlos en los archivos de configuración de las herramientas:
1 - STRbase (SRD-130),
2 - Genotyping Y-STR and CODIS markers,
3 - A Catalog of Human STR Variation,

4 - Base de datos local,
5 - Archivos “Forenseq.config” y “Powerseq.config” de StraiRazor 3.0

Ciencias Forenses
DYS389-2 (2), DYS394 (2), DYS395S1 (2), DYS406S1 (2), DYS413 (2), DYS426 (2), DYS434 (2), DYS435 (2), DYS436 (2), DYS441 (2), DYS442 (2), DYS444 (2), DYS445 (2), DYS446 (2), DYS447 (2), DYS448_1 (2), DYS448_2 (2), DYS449 (2), DYS450 (2), DYS452 (2), DYS454 (2), DYS455 (2), DYS459 (2), DYS462 (2), DYS463 (2), DYS464 (2), DYS472 (2), DYS485 (2), DYS487 (2), DYS490 (2), DYS492 (2), DYS494 (2), DYS495 (2), DYS511 (2), DYS520 (2), DYS531 (2), DYS534 (2), DYS537 (2), DYS556 (2), DYS557 (2), DYS565 (2), DYS568 (2), DYS572 (2), DYS575 (2), DYS578 (2), DYS589 (2), DYS590 (2), DYS594 (2), DYS607 (2), DYS617 (2), DYS636 (2), DYS638 (2), DYS640 (2), DYS641 (2), DYS714 (2), DYS717 (2), GATA-A10 (2), GATA-H4 (2), YCAII (2), DYS505 (2,5), DYS522 (2,5), D17S1301 (5), D20S482 (5), D4S2408 (5), D9S1122 (5), DYS612 (5), DYS425 (4), DXS10079 (1), DXS10148 (1), F13A01 (1), F13B (1), FESFPS (1), SE33 (1), DXS10074 (1,5), DXS10135 (1,5), DXS6789 (1,4), DXS7133 (1,4), DXS9898 (1,4), DXS9902 (1,4), GATA172D05 (1,4), GATA31E08 (1,4), LPL (1,4), D10S1248 (1,4,5), D12S391 (1,4,5), D19S433 (1,4,5), D1S1656 (1,4,5), D2S1338 (1,4,5), D2S441 (1,4,5), D6S1043 (1,4,5), DXS10103 (1,4,5), DXS7132 (1,4,5), DXS7423 (1,4,5), DXS8378 (1,4,5), VWA (1,4,5), DYS389II (1,2,5), DYS389I (1,2,5), DYS437 (1,2,5), DYS438 (1,2,5), DYS439 (1,2,5), DYS456 (1,2,5), DYS458 (1,2,5), DYS460 (1,2,5), DYS461 (1,2,5), DYS481 (1,2,5), DYS533 (1,2,5), DYS549 (1,2,5), DYS570 (1,2,5), DYS576 (1,2,5), DYS635 (1,2,5), DYS643 (1,2,5), PentaD (1,2,5), PentaE (1,2,5), DYS388 (1,2,4), CSF1PO (1,2,4,5), D13S317 (1,2,4,5), D18S51 (1,2,4,5), D3S1358 (1,2,4,5), D5S818 (1,2,4,5), D7S820 (1,2,4,5), D8S1179 (1,2,4,5), DYS19 (1,2,4,5), DYS385 (1,2,4,5), DYS390 (1,2,4,5), DYS391 (1,2,4,5), DYS392 (1,2,4,5), DYS393 (1,2,4,5), FGA (1,2,4,5), TH01 (1,2,4,5), TPOX (1,2,4,5).

Ciencias de la Salud
D13S1492 (4), D13S258 (4), D13S305_1 (4), D13S305_2 (4), D13S628 (4), D13S631 (4), D13S634 (4), D13S742 (4), D13S797 (4), D15S1513 (4), D15S643 (4), D15S657 (4), D15S659 (4), D15S822 (4), D16S2620 (4), D16S2624 (4), D16S3396 (4), D16S753 (4), D18S499 (4), D18S1002 (4), D18S386 (4), D18S390 (4), D18S391 (4), D18S535 (4), D18S858 (4), D18S976 (4), D18S978 (4), D21S1411 (4), D21S1412 (4), D21S1414 (4), D21S1435 (4), D21S1437 (4), D21S1442 (4), D21S1444 (4), D21S1446 (4), D21S1809 (4), D22S683 (4), D22S686 (4), D22S689 (4), DXS1187 (4), DXS6803 (4), DXS8377 (4), DXYS156 (4), DXYS267 (4), GATA178F11 (4), GATA198B05 (4), SRY (4).

Comunes a Cs. Forenses y de la Salud
AMELOGENIN (4,5), HPRTB (1,5), DXS6809 (1,4), D22S1045 (1,4,5), D16S539 (1,2,4,5).

Elegidos de Fase 1
13q311CA (4), 16p121AC (4), 19p1312AC (4), 2QTEL47 (4), ABHD5_I_5 (4), ACOXL_I_11 (4), ACP2_I_3 (4), AF009206 (4), AFM249ZA5 (4), AFM311ZB1 (4), AFM343VF1 (4), AFM359TB5 (4), AFMA046XH1 (4), AFMA133YE5 (4), AFMA190ZF5 (4), AFMB002YA5 (4), AFMB307WA5 (4), AFMB313ZH5 (4), AFMC020WE9 (4), ALPK1_I_11 (4), ANO3_I_14 (4), ANXA1_I_10 (4), ANXA11_I_5 (4), ARAP2_I_20 (4), BV166198 (4), BV208754 (4), CAPRIN2_I_4 (4), CASP4_I_3 (4), CCDC13_I_1 (4), CEACAM7_I_1 (4), CHL1_I_5 (4), CNTNAP2_I_8 (4), COA6_I_1 (4), COL11A1_I_57 (4), CRYBG1_I_1 (4), CSMD3_I_63 (4), D12S69 (4), D8S1992 (4), DHX36_I_14 (4), DPT_I_3 (4), DSPP_I_1 (4), DUOX1_I_13 (4), DUOX1_I_9 (4), DUSP22_I_5 (4), EDIL3_I_4 (4), EED_I_9 (4), ENPP1_I_2 (4), EPRS_I_20 (4), FABP3_I_3 (4), FAH_I_1 (4), FAN1_I_9 (4), FBN1_I_42 (4), FGD6_I_13 (4), GATAD2B_I_1 (4), GDB_180250 (4), GDB_190840 (4), GLIS3_I_1 (4), GOT2_I_6 (4), HSD17B8_I_6 (4), HSPBAP1_I_4 (4), IQCB1_I_8 (4), ITSN2_I_18

(4), KDM3A_I_10 (4), KIF5B_I_2 (4), LCT_I_2 (4), LRIG1_I_14 (4), MBNL2_I_1 (4), MMP10_I_9 (4), MORN3_I_2 (4), MYBPC1_I_19 (4), NBPF9_I_21 (4), NHLRC3_I_4 (4), OTUD7A_I_6 (4), PAPSS2_I_1 (4), PHTF1_I_17 (4), PPM1L_I_1 (4), PTPRD_I_26 (4), REN24986 (4), RH61194 (4), RH70304 (4), RPS6KA2_I_12 (4), SFXN4_I_12 (4), SH3D19_I_3 (4), SHKBP1_I_4 (4), SLC13A3_I_9 (4), SLC26A9_I_19 (4), SLC41A2_I_6 (4), SLC7A7_I_6 (4), SLIT2_I_18 (4), SPOPL_I_2 (4), STAB2_I_54 (4), STAU2_I_1 (4), STSG598727 (4), STXBP4_I_10 (4), TDRD9_I_25 (4), TEP1_I_53 (4), THRB_I_5 (4), TMEM175_I_6 (4), TMEM71_I_4 (4), TNS1_I_3 (4), TSBP1_I_15 (4), TTC7B_I_14 (4), UBR4_I_74 (4), USP37_I_19 (4), UTRN_I_47 (4), VPS53_I_4 (4), WDR7_I_22 (4), WDR72_I_5 (4), WIF1_I_4 (4), YY1AP1_I_7 (4), ZCWPW2_I_6 (4), ZNF25_I_1 (4), ZNF620_I_3 (4), ZNF85_I_1 (4).

Elegidos del cromosoma 7 y 9

7p121ATCT (3), 7p143AAAG (3), 7p143TTCC (3), 9q2113TCTA (3), CCDC171_I_24 (3), CNTNAP2_I_14 (3), CNTNAP2_I_16 (3), D7S1529 (3), D7S1809 (3), D7S2197 (3), D7S2198 (3), D7S3057 (3), D7S796 (3), D7S821 (3), D9S1118 (3), D9S2169 (3), D9S238 (3), D9S301 (3), D9S302 (3), D9S925 (3), DOCK8_I_1 (3), EZH2_I_3 (3), G73163 (3), PDE1C_I_1 (3), PTPRD_I_3 (3), RBAK-RBAKDN_I_2 (3), TYW1_I_13 (3), ZCCHC7_I_5 (3).

ANEXO V TOTALIDAD DE MARCADORES INSPECCIONADOS DE MANERA MINUCIOSA.

13q311CA, 16p121AC, 2QTEL47, 7p121ATCT, 7p143AAAG, 7p143TTCC, 9q2113TCTA, ABHD5_I_5, ACOXL_I_11, ACP2_I_3, AFM311ZB1, AFMA133YE5, AFMB307WA5, ANO3_I_14, ARAP2_I_20, BV166198, BV208754, CCDC13_I_1, CCDC171_I_24, CNTNAP2_I_14, CNTNAP2_I_16, COL11A1_I_57, CRYBG1_I_1, CSMD3_I_63, D12S391, D12S69, D13S258, D13S305, D13S628, D13S631, D13S634, D13S742, D13S797, D15S1513, D15S643, D15S657, D15S659, D15S822, D16S2620, D16S2624, D16S753, D18S1002, D18S386, D18S391, D18S499, D18S535, D18S858, D18S976, D1S1656, D20S482, D21S11-D21S1414, D21S1411, D21S1412, D21S1435, D21S1437, D21S1442, D21S1444, D21S1446, D22S1045, D22S683, D22S686, D22S689, D2S1338, D3S1358, D5S818, D6S1043, D7S1529, D7S1809, D7S2197, D7S2198, D7S3057, D7S796, D7S821, D8S1992, D9S1118, D9S1122, D9S2169, D9S238, D9S301, D9S302, D9S925, DOCK8_I_1, DPT_I_3, DUOX1_I_9, DUSP22_I_5, DXS10074, DXS10079, DXS10135, DXS10146, DXS10148, DXS1187, DXS6789, DXS6803, DXS6809, DXS7132, DXS8377, DXS8378, DXS9898, DXYS156, DXYS267, DYF387S1, DYS19-DYS394, DYS385, DYS388, DYS389II, DYS393, DYS395S1, DYS406S1, DYS434, DYS437, DYS439, DYS442, DYS444, DYS446, DYS447, DYS448, DYS449, DYS450, DYS452, DYS458, DYS460, DYS461, DYS463, DYS464, DYS490, DYS505, DYS511, DYS520, DYS522, DYS533, DYS534, DYS557, DYS572, DYS575, DYS576, DYS589, DYS607, DYS612, DYS636, DYS640, DYS714, DYS717, EDIL3_I_4, EED_I_9, ENPP1_I_2, EZH2_I_3, F13A01, F13A1, F13B, FAH_I_1, FES-FPS, FGA, FGD6_I_13, G73163, GATA172D05, GATA178F11, GATA198B05, GATAD2B_I_1, GOT2_I_6, HPRTB, HSD17B8_I_6, ITSN2_I_18, LPL, MBNL2_I_1, NBPF9_I_21, OTUD7A_I_6, PDE1C_I_1, PTPRD_I_3, RBAK-RBAKDN_I_2, SE33, SH3D19_I_3, SLC13A3_I_9, SLC26A9_I_19, SPOPL_I_2, STSG598727, TH01, TMEM175_I_6, TMEM71_I_4, TTC7B_I_14, TYW1_I_13, UBR4_I_74, UTRN_I_47, vWA, WDR72_I_5, YCAII, Y-GATA-A10, Y-GATA-H4, ZCCHC7_I_5, ZCWPW2_I_6, ZNF85_I_1.