Trabajo Fin de Máster

Estudio sobre la asociación entre el índice de conservación y las variantes presentes en las regiones RNR1 y RNR2 del genoma mitocondrial

Borja Freire Castro

Máster en Técnicas Estadísticas

Curso 2020-2021

Propuesta de Trabajo Fin de Máster

Título en galego: Estudo sobre a asociación entre o índice de conservación e as variantes presentes nas rexións RNR1 y RNR2 do xenoma mitocondrial

Título en español: Estudio sobre la asociación entre el índice de conservación y las variantes presentes en las regiones RNR1 y RNR2 del genoma mitocondrial

English title: Study over the associations between the conserved index and mutations in the RNRS-1 and RNRS-2 regions

Modalidad: Modalidad A

Autor/a: Borja Freire Castro, Universidade de A Coruña

Director/a: Jose Antonio Vilar Fernández, Universidade de A Coruña; Antón Vila Sanjurjo, Universidade de A Coruña

Tutor/a: , ; ,

Breve resumen del trabajo:

El alto grado de conservación del rRNA ha provocado que este sea objeto de estudio a la hora de determinar la arquitectura y función del ribosoma. El nivel de conservación de los diferentes residuos o secciones del genoma es considerado como una medida de su importancia funcional y estructural. Dichas y otras regiones sufren cambios esporádicos conocidos como mutaciones. Estas mutaciones pasan totalmente inadvertidas, o juegan un papel vital en la supervivencia y/o desarrollo de las células. Nuestro interés se centra en el estudio de dichas mutaciones en genoma mitocondrial donde estas son más frecuentes. Además, cambios en regiones conservadas no son necesariamente deletéreas debido a la multiplicidad de las moléculas de DNA mitocondrial, mtDNA.

Recomendaciones:

Otras observaciones:

Don/doña Jose Antonio Vilar Fernández, Catedrático de universidad de la Universidade de A Coruña, don/doña Antón Vila Sanjurjo, Profesor contratado interino de sustitución de la Universidade de A Coruña, don/doña , de , y don/doña , de , informan que el Trabajo Fin de Máster titulado

Estudio sobre la asociación entre el índice de conservación y las variantes presentes en las regiones RNR1 y RNR2 del genoma mitocondrial

fue realizado bajo su dirección por don/doña Borja Freire Castro para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En A Corula, a 23 de Diciembre de 2020.

El/la director/a:

VILAR FERNANDEZ JOSE ANTONIO - 32639957J

Firmado digitalmente por VILAR FERNANDEZ JOSE ANTONIO - 32639957J
Fecha: 2021.02.04 13:21:53 +01'00'

Don/doña Jose Antonio Vilar Fernández

El/la director/a:

VILA SANJURJO ANTONIO - 32773313S

Digitally signed by VILA SANJURJO ANTONIO - 32773313S
DN: c=ES, serialNumber=IDCES-32773313S, givenName=ANTONIO, sn=VILA SANJURJO, cn=VILA SANJURJO ANTONIO - 32773313S
Date: 2021.02.04 16:18:55 +01'00'

Don/doña Antón Vila Sanjurjo

El/la tutor/a:

Don/doña

El/la tutor/a:

Don/doña

El/la autor/a:

Firmado por FREIRE CASTRO BORJA -

Don/doña Borja Freire Castro

Agradecimientos

No se me da demasiado bien agradecer las cosas pero en ocasiones es más que necesario hacer el esfuerzo. Me gustaría agradecer a José Antonio, director del trabajo y por aquel entonces coordinador del máster en A Coruña, el haberme motivado a estudiar el máster en primera instancia cuando le consulté hace un par de años. Pese a que como todo tiene cosas mejorables la formación me ha parecido adecuada y espero me sea de utilidad en el presente y futuro. A mayores me llevo un muy buen recuerdo de tanto las clases como de las conversaciones explicativas con José. En las cuales es imposible perder detalle pues José se encarga de que estés siempre al loro: "¿Me sigues?, ¿Vienes conmigo?...".

Por último, me gustaría agradecer a ambos directores del TFM, si a José de nuevo pero en esta ocasión también a Antón, las horas extras que han hecho (que no han sido pocas) en momentos de máxima carga laboral con el único fin de que yo pudiese leer cuando me lo propuse.

Índice general

Resumen

Resumen en español

Esta memoria se centra en el estudio de mutaciones ocurridas en el genoma mitocondrial. En general, estas mutaciones son más frecuentes que en el genoma nuclear debido a la falta de sistemas de corrección eficientes y sus efectos más complejos de estudiar debido a la multiplicidad de dichas moléculas de DNA mitocondrial, mtDNA. Los objetivos finales de este estudio son fundamentalmente dos:

- Estudiar si existe una relación funcional entre el índice de conservación y el número de variaciones registradas en las posiciones del genoma mitocondrial.
- Estudiar si existen variantes coocurrentes fuera de las ya reportadas y, en caso afirmativo, examinar si conllevan algún significado biológico.

Los análisis estadísticos y numéricos realizados parecen indicar que la primera hipótesis es correcta, lo que abre la puerta a la posibilidad de estudios posteriores más complejos asociados con enfermedades como el cáncer. Además, se han desarrollado técnicas basadas en métodos numéricos que parece que permiten asociar variantes entre si capaces de capturar tanto relaciones previamente reportadas como relaciones nunca antes vistas.

English abstract

English abstract of 150 to 250 words.

Prefacio

El nivel de conservación de los diferentes residuos pertenecientes a un gen funcional es normalmente considerado una medida de su importancia funcional y estructural. La secuencia de los genes sufre cambios esporádicos conocidos como mutaciones. Estas mutaciones pueden pasar totalmente inadvertidas, o jugar un papel vital en la supervivencia y/o desarrollo de las células. Nuestro interés se centra en el estudio de dichas mutaciones en el genoma mitocondrial donde estas son más frecuentes, pero mucho más difíciles de analizar. Además, puede darse el caso de que mutaciones en regiones conservadas no sean necesariamente deletereas debido a la multiplicidad de las moleculas de DNA mitocondrial o mtDNA, un fenómeno conocido como heteroplasmia. En concreto nos centraremos en los genes que codifican los ARNs ribosómicos mitocondriales (mt-ARNrs).

A pesar de toda esta controversia, nuestro análisis permite confirmar la existencia de una clara relación inversa entre el número de variaciones del mt-rRNA y los niveles de conservación de las posiciones donde suceden dichas variaciones en el mt-rRNA humano [26]. Esta conclusión se ha obtenido tras aplicar análisis de regresión para estudiar el vínculo entre mutaciones e índices de conservación en el mt-rRNA de los humanos. Dada la naturaleza complicada de los datos ha sido necesario considerar un abanico de modelos de regresión de complejidad variable, incluyendo modelos paramétricos y también modelos no paramétricos que otorgan mayor flexibilidad para describir la relación funcional subyacente. Los resultados de ajustar estos modelos confirman que los procesos adaptativos constituyen una fuerza importante en la configuración de la evolución de los mt-rRNA.

A medida que se iba avanzando en el análisis acerca del índice de conservación y su relación con las variantes mitocondriales, surgió la duda acerca de si existía la posibilidad de desarrollar alguna metodología capaz de capturar coocurrencia significativa de variantes. Para dar respuesta a este reto, se exploró la posibilidad de utilizar métodos estadísticos y/o numéricos capaces de reconstruir los perfiles haplotípicos conocidos a día de hoy. Los resultados obtenidos son prometedores y muestran una alta tasa de acierto en los métodos desarrollados siendo en la mayoría de casos capaces de reconstruir las variantes asociadas a los haplogrupos y adicionalmente ofrece nuevas variantes, que han de ser confirmadas como de interés, no relacionadas previamente.

El presente trabajo se divide en dos partes claramente diferenciadas:

- Presentación del análisis desarrollado para modelizar la relación entre el índice de conservación y las variaciones asociadas con dicho índice. Se describe y discute el aparato metodológico empleado, enfatizando la idoneidad de los ajustes realizados así como su alcance en términos predictivos. Se hará también especial hincapié en el interés biológico y la importancia que supone dicha relación de cara a estudios posteriores más complejos.

- Descripción detallada de los mecanismos propuestos para capturar la coocurrencia de variantes así como las técnicas de evaluación empleadas para verificar el correcto funcionamiento de las mismas. Es valioso hacer notar que, hasta donde sabemos, este problema no ha sido abordado en la literatura especializada en el área.

Capítulo 1

Presentación del problema

Este capítulo proporciona una breve introducción a los fundamentos biológicos necesarios para la correcta compresión de la memoria. Además, presenta de manera formal las dos problemáticas específicas abordadas. El objetivo último del capítulo es facilitar al lector una mayor comprensión tanto de las técnicas empleadas como de la finalidad y objetivo de las mismas.

1.1. Fundamentos biológicos

Pese a no ser necesario tener un conocimiento exhaustivo del dominio biológico, sí que es necesario familiarizarse con algunos conceptos antes de abordar los problemas que se tratan a lo largo del trabajo [29].

1.1.1. Genoma

El genoma es el conjunto de genes contenidos en cromosomas, lo que puede interpretarse como la totalidad del material genético que posee un organismo o una especie en particular. El material genético se transcribe para producir ARNs, que son de tres tipos ARNs de transferencia o ARNts, ribosómicos o ARNrs y ARNs mensajeros o ARNms. Estos últimos se traducen a proteínas en el ribosoma, que se compone de los ARNrs y un conjunto proteínas ribosómicas, y que requiere a los ARNts para completar su función.

El genoma se replica en el proceso de duplicación celular para producir nuevas células. Estos procesos de replicación del material genético son muy precisos pero no son perfectos, por lo que pueden producirse errores que generan cambios. Estos cambios, tradicionalmente llamados mutaciones, son variaciones espontáneas o inducidas del genoma que producen un cambio permanente y heredable en la secuencia del ADN. Las consecuencias de dichas mutaciones pueden ser desde nulas, mutaciones silenciosas, a dañinas, mutaciones deletéreas. Afortunadamente todos los seres vivos cuentan con sistemas de reparación de ADN, pero en ocasiones las lesiones del ADN no se pueden revertir a la secuencia original dando lugar a mutaciones permanentes. Es importante destacar que la distribución de las tasas de mutación a lo largo del genoma, lejos de ser uniforme, es extremadamente irregular. Existiendo desde regiones conocidas como *hotspots*, puntos con tasas de mutación mucho más elevadas de lo normal, a regiones altamente conservadas, como el RNA ribosomal que es generalmente empleado en la reconstrucción de filogenias.

El genoma en los seres eucariotas, organismos con núcleo diferenciado, comprende el ADN contenido en el núcleo, organizado en cromosomas, y el genoma de orgánulos celulares, como las mitocondrias y cloroplastos (en vegetales y algas). Es por tanto interesante recalcar que en eucariotas pueden existir diversos genomas dentro de la misma célula y cuyo ADN, pese a compartir el objetivo básico de ser transcrito a ARN de cualesquiera de los tres posibles tipos: mensajero, transferencia y ribosomal, da lugar a conjuntos de genes diferentes. En este trabajo nos centraremos en al ADN mitocondrial

o ADNmt. El ADNmt de humanos codifica 13 proteínas 22 ARNts y 2 ARNrs (codificados por los genes RNR1 y RNR2), que se usan para construir las subunidades pequeña (SSU, del inglés "small subunit") y grande (LSU, del inglés "large subunit") del ribosoma mitocondrial o mitoribosoma. Debemos introducir los conceptos de homoplasmia y heteroplasmia para designar la presencia en una misma célula de uno o más tipos de ADNmt en base a la presencia (heteroplasmia) o no (homoplasmia) de variaciones en alguna de sus posiciones o residuos. Así pues el genoma de una célula puede no solo estar compuesto por diferentes tipos de ADN sino que además el ADNmt, pueden estar compuesto de moléculas diferentes de ADN.

La principal finalidad del genoma, como se comenta en el anterior párrafo, es contener el material genético de un individuo. Este material se transcribe para producir proteínas, t-RNAs o r-RNAs, pero además se replica en el proceso de duplicación celular para producir nuevas células. Estos procesos de replicación del material genético son muy precisos pero no son perfectos, por lo que pueden producirse errores que generan cambios. Estos cambios, tradicionalmente llamados mutaciones, son variaciones espontáneas o inducidas del genoma, pudiendo estos ser permanentes y heredables en la secuencia del ADN, en nucleótidos, o bien en la disposición del ADN en el genoma. Las consecuencias de dichas mutaciones dependen entonces del nivel de afectación: desde cambios en la secuencia nucleotídica, en los genes o los productos génicos, hasta en los tipos celulares involucrados. Afortunadamente todos los seres vivos cuentan con sistemas de verificación y reparación de los procesos celulares. En el caso de la replicación del ADN existen sistemas de reparación específicos y bien coordinados, pero en ocasiones las lesiones del ADN que escapan a esta verificación serán las causantes de ocasionar mutaciones. Por lo tanto, la efectividad de la lesión dependerá tanto de la tasa de división celular como de la eficiencia de los mecanismos de reparación contra el aumento en las lesiones al ADN. Cuando ambos factores están incrementados, el resultado será una mayor generación de mutaciones o mutagénesis (y carcinogénesis). Afortunadamente, en estos niveles de variación están medidos y calculados para los genomas nucleares más representativos y es tradicionalmente conocido como *índice de conservación*, el concepto es introducido en detalle en la siguiente sección.

1.1.2. Índice de conservación

Para entender si existe o no una relación entre el número de variaciones y la conservación de un residuo dado, decidimos usar una medida de conservación universal, CVs [2]. Estos Cv son calculados a través de secuencias de rRNA de *Escherichia coli*, E. coli, como referencia y otras 8513 ARNrs SSU y 1045 ARNrs LSU secuencias, correspondientes a los tres dominios filogeneticos, Eubacteria, Archaea, y Eucarya (incluyendo en este último dominio tanto secuencias nucleares, mitocondriales y cloroplásticas). En su estudio, Cannone et al. (Cannone et al., 2002; Ortoleva-Donnelly et al., 1998) definieron los siguientes intervalos de conservación:

- 100 % de conservación corresponde con un Cv de 2.
- 90 % de conservación sería un Cv de alrededor de 1.5.
- 60-80 % de conservación corresponde con un 1.0.
- 40-50 % de conservación serían valores cercanos al 0.5.
- total falta de conservación serían valores por debajo de 0.

. Las Figuras 1.1 y 1.2 muestran la distribución de las variantes registradas tanto para la LSU como para la SSU en la plataforma GenBank para los distintos Cv. De cada región se proporciona:

- Figura superior izquierda: Histograma del número total de variantes con el objetivo de visualizar la cantidad de variaciones que se tienen y cómo se distribuyen.
- Figura superior central: Histograma del logaritmo del total de variantes con la idea de tratar de suavizar el impacto de las variantes más repetidas.

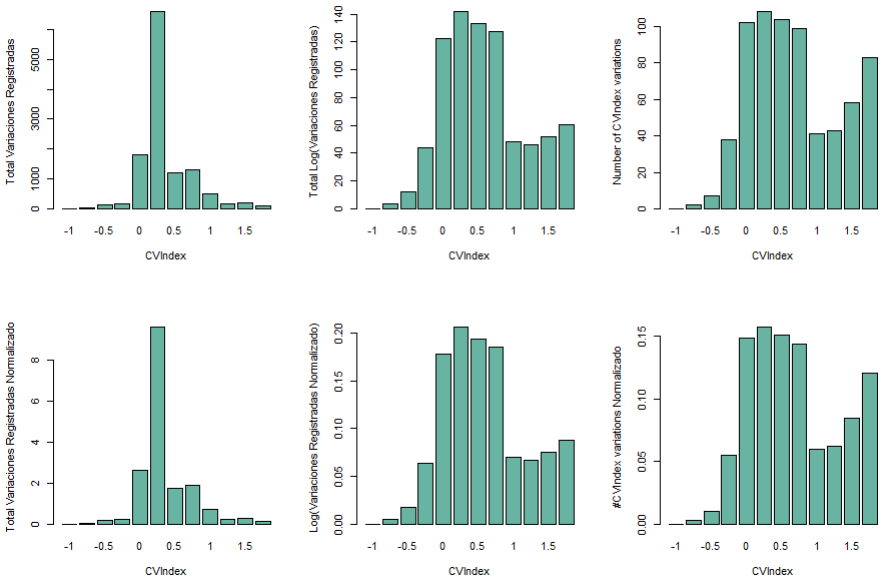

Figura 1.1: Distribución de los residuos de acuerdo a su CV para el LSU

- Figura superior derecha: Caso extremo del histograma mencionado previamente donde cada variante cuenta una sola vez.

- Figuras paneles inferiores: Versiones normalizadas de los histogramas de los paneles superiores.

Se observa como existe una acumulación mayoritaria entorno al intervalo $[0, 0.5]$ para la LSU, mientras que la mayor concentración en el caso de la SSU se extiende desde $[0, 2]$. Además, existen claramente variantes altamente repetidas en ciertas posiciones genómicas esto pudo ser devido deberse tanto a una variante común a diversos haplogrupos o a una variante de un haplogrupo secuenciado en multitud de ocasiones.

1.2. El problema abordado: Motivación e interés

La alta conservación del ARNr en evolución ha sido una piedra angular para los estudios que tienen como objetivo comprender la arquitectura y la función del ribosoma. Para los residuos altamente conservados, su grado de conservación se ha tomado como una medida de su importancia funcional y/o estructural [3]. Mutaciones encontradas en residuos conservados a menudo condujeron a fenotipos deletéreos, cuya gravedad coincidía con su grado de conservación [7, 10, 11, 12, 13]. Un subconjunto mucho más grande de residuos menos conservados se estimó que co-variaban durante la evolución en patrones que podrían usarse para predecir interacciones secundarias y terciarias [3]. A menudo, las mutaciones disruptivas dirigidas podrían rescatarse mediante cambios de base compensatorios en el sitio complementario [5, 6, 7, 8, 9]. Finalmente, existe un subconjunto importante de residuos escasamente conservados que a menudo se agrupan en regiones periféricas del ribosoma pero que presentan escasa participación en la función ribosómica [3, 14]. Todas estas observaciones llevaron al desarrollo de algoritmos de plegamiento basados en filogenia capaces de predecir la estructura secundaria (y una parte importante de la estructura terciaria) de los rRNA [2]. Sorprendentemente, la calidad de estas predicciones ha sido confirmada por una amplia gama de estructuras tridimensionales derivadas biofísicamente [4, 15, 16, 17, 18, 19, 20].

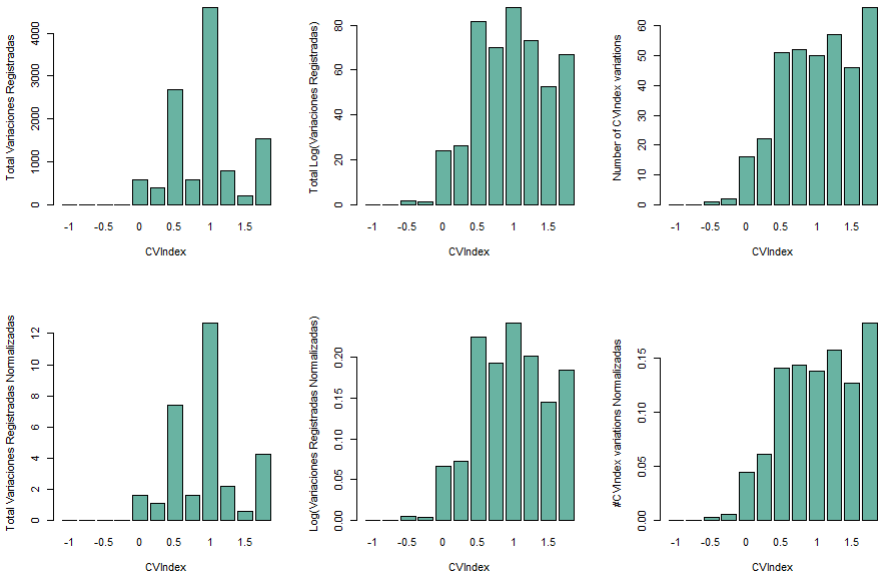

Figura 1.2: Distribución de los residuos de acuerdo a su CV para el SSU

A pesar del origen monofilético de las mitocondrias, la alta diversidad de arreglos organizativos del mtDNA encontrados en el espectro filogenético de los individuos eucarióticos apunta a la existencia de restricciones evolutivas muy diferentes para estos orgánulos en los diferentes linajes eucariotas [21, 22]. Por lo tanto, no fue sorprendente encontrar que los modelos de evolución no darwinianos se usaron desde el principio para describir el origen de los eucariotas por endosimbiosis [22]. La genética mitocondrial es marcadamente diferente de la genética mendeliana que impulsa la herencia de genes nucleares [23]. La multiplicidad de moléculas de mtDNA en una sola célula conduce a la acumulación gradual de mutaciones sin un impacto deletéreo inmediato, un fenómeno que explica en parte la evolución más rápida del mtDNA en relación con el ADN nuclear [21]. De hecho, se ha estimado que el ADNmt tiene una tasa de evolución de 9 a 25 veces más rápida que la del ADN nuclear [23, 24]. Además del efecto de multiplicidad mencionado anteriormente, se ha propuesto que otros factores, como la producción de radicales libres de oxígeno dañinos, la replicación continua de ADNmt por polimerasas de baja fidelidad y la falta de una maquinaria de reparación de ADN eficiente en el orgánulo, podrían estar produciendo esta rápida evolución [23, 24]. También se ha planteado la hipótesis de que el mecanismo inusual de herencia del ADNmt, junto con su haploidía, debe reducir la eficacia de la selección natural para dirigir su evolución, en relación con la del ADN nuclear, mientras que el papel de la deriva genética debería amplificarse [23]. De hecho, el mtDNA está predispuesto a la acumulación no adaptativa de mutaciones deletéreas a través del trinquete de Muller [23] . Finalmente, el hecho de que todos los genes involucrados en el ensamblaje y función del mitoribosoma estén codificados en el núcleo, complica aún más la conformación de la secuencia de mt-rRNA durante la evolución. Por lo tanto, no sería una sorpresa que la acción de todos estos factores pudiera borrar efectivamente la firma de conservación filogenética del espectro mutacional de los mt-rRNA humanos. En el caso específico de las mitocondrias de mamíferos, incluso se ha dicho que "los conceptos clásicos de asociaciones genotipo-fenotipo y su interacción con la selección son violados por la genética del mtDNA" [25]. Todo esto pone en tela de juicio el papel de la selección adaptativa en la evolución del ADNmt de mamíferos.

Hace unos años se comenzaron análisis de las mutaciones que mapean el ARNr de las mitocondrias humanas con el objetivo de comprender su potencial papel en las enfermedades humanas [26]. En el momento inicial existía poca información estructural, bioquímica y mutacional sobre el mitoribosoma

de los mamíferos. Por esta razón, los análisis se basaron principalmente en el poder predictivo proporcionado por la conservación universal del pliegue del rRNA. Como resultado, se pudieron hacer predicciones estructurales y funcionales sobre el papel disruptivo de las mutaciones del mt-rRNA humano mediante el uso de fuentes ribosómicas heterólogas para las que se disponía de dicha información. Una fuente importante de los datos mutacionales utilizados en estos estudios fue la base de datos cada vez más creciente de secuencias de ADNmt de GenBank, información que ahora mismo se encuentra recogida en la base de MitoMap. La validez de esta información para estudios inferenciales claramente requería la existencia de una fuerte relación entre las mutaciones y las conservación de los mt-rRNA humanos. Si bien se han encontrado pruebas de que la selección adaptativa contribuye a la evolución del mtDNA en todo el reino animal, si el genoma mitocondrial evoluciona o no de forma adaptativa o no adaptativa sigue siendo un tema de debate [23]. En principio, la relación entre las mutaciones y los niveles conservación de los mt-rRNA humanos podría verse debilitada por la rápida evolución del mtDNA animal. La mayor tasa de evolución del genoma mitocondrial en relación con la de su equivalente nuclear (9-25 veces más rápida) es causada por varios factores, como la producción orgánular de radicales de oxígeno libres dañinos, la replicación continua del mtDNA por polimerasas de baja fidelidad. y la falta de una maquinaria de reparación de ADN eficiente en el orgánulo [23, 24]. También se ha planteado la hipótesis de que el mecanismo inusual de herencia del ADNmt, junto con su haploidía, debe reducir la eficacia de la selección natural para dirigir su evolución, en relación con la del ADN nuclear, mientras que el papel de la deriva genética debería amplificarse [23]. Finalmente, el hecho de que todos los genes involucrados en el ensamblaje y función del mitoribosoma estén codificados en el núcleo, complica aún más la conformación de la secuencia de mt-rRNA durante la evolución. Por lo tanto, no sería una sorpresa que la acción de todos estos factores pudiera borrar efectivamente la firma de conservación filogenética del espectro mutacional de los mt-rRNA humanos.

1.3. Presentación de los datos

En esta sección se hace una breve presentación de los datos que se manejaron a lo largo de todo el proyecto. Dicha información ha sido extraída de la plataforma `mitomap.org`. Esta plataforma busca reunir toda la información publicada acerca de variaciones en el DNA mitocondrial. Desafortunadamente, la plataforma está pensada para agrupar la información pero no para suministrarla de manera sencilla. De hecho, a la hora de acceder a cualquier recurso(i.e variantes en una región, información de individuos que poseen una variante determinada, entre otros), es necesario implementar software propio para la descarga y parseo de toda la información, añadiendo así cierto tedio al trabajo con la plataforma.

En el Cuadro 1.1 se muestra de manera resumida y agregada la información contenida a día de hoy (31/01/2021) en MitoMap para las regiones RNRS1 (SSU) y RNRS2 (LSU). Lo más destacable de la tabla es la enorme cantidad de variaciones que presenta la subunidad pequeña (LSU) frente a la subunidad grande cuando esta es 600 pares de bases más pequeña. Por otro lado, también impacta la concentración de valores en el intervalo $[1, 1.5]$ en la subunidad pequeña que recoge un $25.05\,\%$ de los residuos correctamente emparejados, frente al $9.81\,\%$ de la misma región en el LSU.

El cuadro 1.2 y la Figura 1.3 resumen la información más relevante en el estudio de la coocurrencia de variantes y su asociación a los haplogrupos. En primer lugar, el Cuadro 1.2 agrupa la información acerca de los haplogrupos, componiendose esta de la información recabada por MitoMap para haplogrupos y para los individuos:

- Para cada haplogrupo se tiene las variantes asociadas al mismo: `https://www.mitomap.org/foswiki/bin/view/MITOMAP/HaplogroupMarkers`
- Para cada individuo la información completa no es inmediatamente accesible pero a través de múltiples peticiones se puede acceder a ella.
    - `https://www.mitomap.org/foswiki/bin/view/MITOMAP/PolymorphismsCoding`

	SSU-rRNA			LSU-rRNA		
		Número de sitios de variación	Número de variaciones		Número de sitios de variación	Número de variaciones
Longitud	954			1559		
Residuos con heterólogo	792 (83.01 %)			1050 (67.35 %)		
Residuos sin heterólogo	162 (16.99 %)			509 (32.65 %)		
Residuos con variaciones	558 (58.49 %)			964 (61.83 %)		
Número estimado de variaciones	138053			87136		
Variaciones/Residuo	144.71			55.89		
Máximo número de variaciones	3290 (709)			40793 (2706)		
Cv <0	162 (16.99 %)	127	66814	509 (32.65 %)	567	33283
1 >Cv >= 0	277 (29.03 %)	178	55000	547 (35.09 %)	329	52884
1.5 >Cv >= 1	239 (25.05 %)	140	14345	154 (9.81 %)	73	654
1.9 >Cv >= 1	148 (15.51 %)	73	550	213 (13.66 %)	102	280
Cv >1.9	114 (11.94 %)	40	1344	79 (5.07 %)	24	28

Cuadro 1.1: Tabla de datos para SSU y LSU

- `https://www.mitomap.org/cgi-bin/index_mitomap.cgi?title=Coding+Polymorphism+A-G+at+rCRS+position+648&pos=648&ref=A&alt=G&purge_type=`, query construida a partir de la variación A → G en la posición 648.
- `https://www.ncbi.nlm.nih.gov/nuccore/DQ341065.1`

El Cuadro 1.2 agrupa y sintetiza parte de la información una vez ha sido recabada:

- Las variantes asociadas a cada haplogrupo. Claramente destacan los haplogrupos $L*$, que mantienen mayores niveles de asociación.
- El número de variantes únicas por haplogrupo, destacando L6 con 20 variantes únicas de 51.
- La media de haplogrupos a los que pertenece cada una de las variantes dentro de cada haplogrupo.

Adicionalmente, tendríamos también los individuos secuenciados pero esta información surge de lo que podríamos denominar muestra de entrenamiento[1] [37]. De esta muestra sale también la información de la Figura 1.3 donde podemos comprobar los niveles de dispersión en número de variantes en cada uno de los individuos en cada haplogrupo. Esto es interesante pues se ve claramente que dentro de cada haplogrupo el número de variantes varía. Por lo tanto, la asociación de una variante a un haplogrupo si y solo si todos los individuos del haplogrupo la contienen puede verse como algo demasiado estricto[2]. Siendo necesario un estudio a través de una técnica más permisiva que permita evaluar todas esas variaciones asociadas o no asociadas y ver si existen relaciones no vistas ni reportadas hasta el momento.

Al ajustar los residuos de los ARNr-mts a los índices de conservación [2] surge el problema de que 160 residuos de los 954 residuos del RNRS1 y 509 residuos de los 1559 del RNRS2 quedan sin

[1] Entendiendo esta como la información que se va a emplear para construir los predictores posteriormente. En el mismo hilo añadir que la muestra de *Test* sería la información recabada para cada uno de los haplogrupos mostrada en el Cuadro 1.2

[2] Hacer énfasis aquí en el haplogrupo L1 donde puede verse que la dispersión obtenida en la muestra de entrenamiento va desde **ligeramente** superior a 40 hasta poco más de 10. Siendo esto curioso pues las variantes asociadas son 47 con lo que esto evidentemente implica.

Haplogrupo	Variantes Asociadas	Individuos Secuenciados	#Variantes Únicas	Media Haplogrupo/Variante
L0	50	1669	15	8.32
L1	47	825	5	11.02
L2	42	1283	7	12.5
L3	36	2106	6	14.25
L4	18	109	0	25.33
L5	23	37	1	20.91
L6	51	12	20	10.56
C	34	1774	5	14.88
D	28	2373	5	17.71
E	33	460	8	14.63
G	27	509	3	18.44
M	22	5290	0	22.09
Q	36	422	11	13.66
Z	34	198	5	14.94
A	23	1365	8	17.39
B	15	4346	2	23.93
F	16	1765	5	21.81
H	6	9607	0	32.5
HV	9	778	1	28.77
I	29	754	9	15.10
J	21	2484	4	18.23
K	27	1934	9	14.62
N	13	843	0	29.61
O	16	8	3	24.25
P	12	411	0	28.83
R	9	1128	0	31.88
S	13	49	1	28.23
T	26	2351	8	14.46
U	14	4531	0	25
V	12	740	3	22
W	30	579	11	13.93
X	23	508	5	17.95
Y	19	185	4	20.68

Cuadro 1.2: Tabla resumen de variantes y haplogrupos

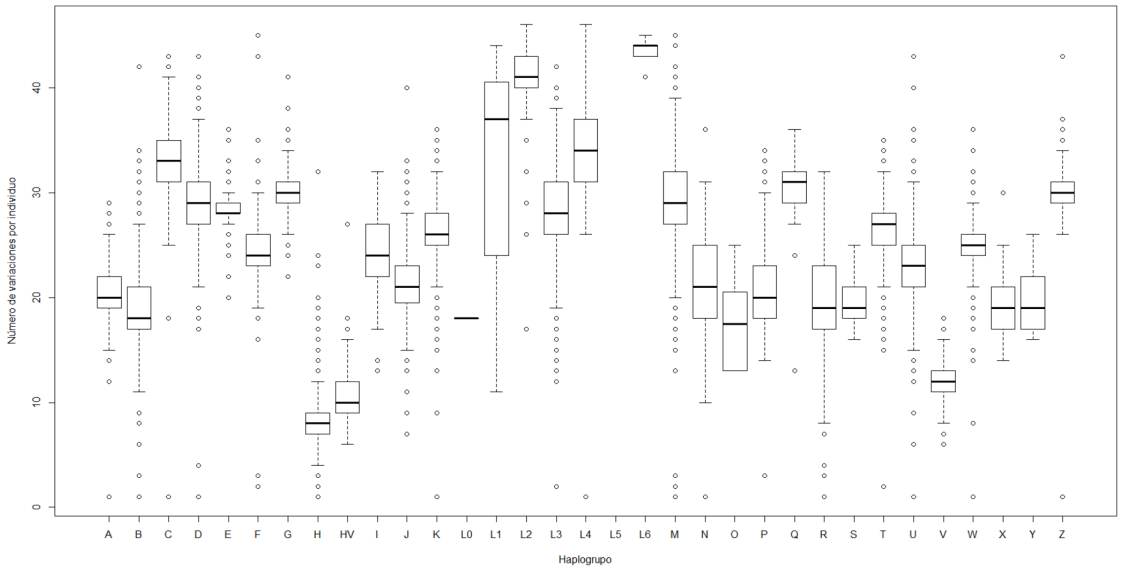

Figura 1.3: Boxplot de variantes por individuo por haplogrupo

Cv debido a la falta de homología entre las estructuras secundarias de la referencia, *E. coli*, y de los ARNr-mts [1]. Esta falta de homología motivó que en primera instancia incluyésemos estos residuos en el grupo con menor índice de conservación (-1), aunque finalmente debido a inconsistencias a nivel estadístico[3] optamos por obviar dichos resultados.

[3]Dar valor de conservación -1 a todos los residuos con falta de homología a nivel estadístico provoca sesgos claros. La falta de información no puede ser asumida como la similitud de la misma.

Capítulo 2

Estudio convergencia RNRS-1 y RNRS-2

2.1. Introducción

Este capítulo expone los análisis que realizamos sobre los datos anteriormente presentados en busca de verificar estadísticamente la hipótesis de que el índice de conservación y el número de variaciones reportadas presentan correlación negativa. La Sección 2.2 describe brevemente los métodos empleados y los resultados obtenidos se recogen en la Sección 2.3. Conviene enfatizar que originariamente manejamos datos erróneos y nuestros estudios exploratorios preliminares no permitieron concluir la existencia de la mencionada relación que, por otro lado, es una hipótesis a priori muy clara desde un punto de vista biológico. Tras el procesado y análisis pormenorizado de los datos, optamos por considerar un amplio abanico de técnicas de regresión con la intención de obtener el refrendo de esta hipótesis con todas ellas.

2.2. Modelos de regresión

Una curva de regresión $m(\cdot)$ describe una relación general entre una variable explicativa X (aleatoria o no) y una variable aleatoria respuesta Y [32]. Específicamente, modeliza el comportamiento en media de Y como función de X:

$$m(x) = E(Y|X = x)$$

Dada una muestra $\{(X_i, Y_i), i = 1, \ldots, n\}$, se ha de satisfacer:

$$Y_i = m(X_i) + \epsilon_i,$$

donde se asume que ϵ_i, $i = 1 \ldots, n$, son realizaciones de una variable aleatoria ϵ tal que $E(\epsilon|X = x) = 0$ y $Var(\epsilon|X = x) = \sigma^2(x)$.

El interés radica en emplear la muestra $\{(X_i, Y_i), i = 1, \ldots, n\}$ para obtener una estimación de $m(\cdot)$, digamos $\hat{m}(\cdot)$. En aras del rigor, procede comenzar distinguiendo dos posibles escenarios para la toma de los datos muestrales:

- Diseño Fijo. La variable explicativa X (también llamada regresora) no es aleatoria, es decir los valores de X_i han sido prefijados. Esta situación es típica en el contexto de un diseño experimental donde se desea evaluar la respuesta Y para valores concretos de X.

- Diseño Aleatorio. Los pares $(X_i, Y_i), i = 1, ..., n$ son realizaciones de un vector aleatorio bidimensional (X, Y), con fdp conjunta $f(x, y)$. En tal caso, $m(x)$ puede expresarse como:

$$m(x) = E(Y|X = x) = \int_R yf(y|x)dy = \int_R y\frac{f(x,y)}{f(x)}dy = \frac{1}{f(x)}\int_R yf(x,y)dy$$

  donde $f(x) = \int_R f(x, y)dy$ denota la densidad marginal de X.

Nuestro caso de estudio es claramente un diseño aleatorio por lo que no incidiremos más en diseños fijos.

La estimación de la función de regresión puede enfocarse vía paramétrica o vía no paramétrica:

- Enfoque paramétrico. Se asume que $m(\cdot)$ pertenece a una familia indexada por un vector de parámetros (por ejemplo la familia de polinomios de grado r) y el objetivo se centra en estimar los valores de los parámetros que conducen a explicar mejor las respuestas observadas, i.e. a un mejor "ajuste".

- Enfoque no paramétrico. No se asume estructura paramétrica alguna para $m(\cdot)$ y se deriva su forma funcional a partir únicamente de la información que dan los datos muestrales bajo hipótesis generales de regularidad tales como continuidad o diferenciabilidad.

Ambas vías de análisis son interesantes, presentan ventajas y desventajas y, en cualquier caso, deben ser consideradas como herramientas complementarias. Así, mientras la vía paramétrica puede resultar muy restrictiva para explicar cambios inesperados o poco regulares (generando estimaciones muy sesgadas), la vía no paramétrica ofrece mayor versatilidad para modelizar la relación de regresión desconocida. Como contrapartida, esta flexibilidad tiene un coste en términos de elevada variabilidad en las predicciones y una tasa de convergencia más lenta. En este proyecto hemos optado por considerar algoritmos de ambas vías de análisis al objeto de corroborar que en todos los casos se obtienen conclusiones análogas.

2.2.1. Modelos lineales generalizados

Entre los modelos paramétricos, además de los modelos lineales, cabe reseñar a los modelos lineales generalizados [34, 42]. Los modelos lineales generalizados [30] son una extensión de los lineales para el caso de que la distribución condicional de la variable respuesta no sea normal, introduciendo una función de enlace (o link) g tal que:

$$g(E(Y|X)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

En la práctica la estimación de los parámetros se realiza empleando el método de máxima verosimilitud. Además, la función link debe ser invertible, de forma que se pueda volver a transformar el modelo ajustado (en la escala lineal de las puntuaciones) a la escala original.

Regresión logística

Un primer modelo de interés entre los GLM es el modelo de regresión logística [31]. Se fundamenta en modelar una variable indicadora, con distribución de Bernouilli donde $E(Y|X) = p(X)$ es la probabilidad de éxito, mediante el empleo de la función logit:

$$logit(p(X)) = log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

La transformación inversa se conoce como función logística y viene dada por:

$$p(X) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}}$$

Sus coeficientes β_j, $j = 0, 1, \ldots, p$, se estiman por máxima verosimilitud, con función $L(\vec{\beta})$:

$$L(\vec{\beta}) = \prod_i p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

Regresión de Poisson

El modelo de regresión de Poisson surge cuando la variable respuesta es una cantidad discreta, con valores en $0, 1, 2, ...$ Este tipo de variable respuesta suele representar el recuento de sucesos o hechos (número de individuos de un grupo, **número de mutaciones por CV**), y nos podemos plantear si ciertas variables explicativas influyen en la variable respuesta y cómo lo hacen.

A diferencia de los modelos lineales de regresión donde se asumen hipótesis de normalidad, homoscedasticidad y linealidad, en la regresión logística y de Poisson estas hipótesis de regularidad se ven claramente vulneradas por la naturaleza de la respuesta. En el caso de la regresión de Poisson:

- **Linealidad**, la naturaleza positiva de los recuentos evita que esta condición se cumpla pues valores negativos son alcanzables a través de una recta y esto no es admisible.
- **Homocedasticidad**, el valor esperado de una Poisson coincide con su varianza, luego si la media condicional crece también lo hará la dispersión y ese es precisamente uno de los fundamentos de la regresión de Poisson, valores elevados de recuentos se acompañan de mayor dispersión.
- **Normalidad**, el carácter discreto y la asimetría de la respuesta impiden asumir un modelo de distribución normal.

De forma análoga a la regresión logística, cuando la respuesta Y condicionada a X se distribuye de acuerdo a una Poisson, cabe modelizar la esperanza condicionada $E(Y|X) = \lambda(X)$, i.e. la función de regresión, mediante una transformación de un predictor lineal sin más que encontrar una función de enlace adecuada. Debido a la naturaleza positiva de la variable respuesta se usa la función de enlace *log*:

$$log(\lambda(x, \beta)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

Quedando por tanto la función de regresión del modelo de Poisson como:

$$\lambda(x, \beta) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p}$$

cuyos coeficientes β se estimaran a partir de máxima verosimilitud, con función $L(\beta)$:

$$L(\beta) = \prod_i^n [e^{-\lambda(x_i,\beta)} \frac{\lambda(x_i, \beta)^{y_i}}{y_i!}]$$

Métodos de tratamiento de la sobredispersión

El modelo de regresión de Poisson presentado es algo restrictivo. Lo más destacable es que impone que la varianza de la variable respuesta coincida con la media, lo cual es inherente a la distribución de Poisson, pero puede no corresponderse con la realidad en muchas circunstancias reales.

Los modelos de Poisson se aplican generalmente donde se realiza un recuento de individuos o sucesos en el tiempo o en el espacio. Aplicar el modelo de Poisson equivale a suponer que los individuos se distribuyen al azar en el territorio, sin atraerse ni repelerse. Si se repelieran, los individuos tenderían a distribuirse de manera uniforme en el territorio, y de esta manera el recuento de individuos en cada parcela sería muy similar, lo cual produciría un fenómeno de infra-dispersión. Por el contrario, si los individuos se atraen, formarán grupos de individuos de cierto tamaño y la variabilidad de recuentos por parcelas será más grande (sobre-dispersión).

Una de las maneras tradicionales de abordar el problema de la sobredispersión es usar como modelo una Binomial Negativa pues este tipo de distribuciones presentan inherentemente una varianza mayor

a la media. Como recordatorio, una variable Binomial Negativa de parámetros r y p, $BN(r,p)$, mide el número de éxitos que se producen hasta el r-ésimo fracaso, siendo p la probabilidad de éxito. Su función de masa de probabilidad viene dada por:

$$P(Y=f)=\binom{f+r-1}{f}(1-p)^r p^f$$

Con medias y varianza respectivamente:

$$E(Y)=\frac{pr}{1-p}=\mu \qquad\qquad Var(Y)=\frac{pr}{(1-p)^2}=\theta$$

Se sigue que siempre se verifica: $E(Y)\geq Var(Y)$. Como detalle adicional indicar que para el uso práctico de este modelo se suele extender a un contexto continuo a traves del cambio en su función de masa de probabilidad por:

$$f_Y(y;\mu;\theta)=\frac{\Gamma(\theta+y)}{\Gamma(\theta)y!}\frac{\mu^y\theta^\theta}{(\mu+\theta)^{\mu+y}}$$

Una manera alternativa de tratar la sobredispersión es a través de la estimación ad-hoc de la sobredispersión. Sin embargo, esta metodología ofrece la misma estimación en los parámetros que la regresión de Poisson estándar y no nos pararemos en ella.

2.2.2. Modelos inflados con ceros

Una de las situaciones más frecuentes cuando se hacen estudios biológicos/mutacionales es que no solo aparecen ceros, sino que lo hacen de manera muy superior a lo que correspondería para modelos de Poisson o Binomial Negativo.

Los modelos con ceros inflados consisten en suponer que hay dos tipos de ceros: algunos ceros proceden de la distribución de Poisson y Binomial Negativa de manera natural, mientras que hay otros ceros, que se han añadido, y que generalmente tienen otra interpretación.

Entonces un modelo con ceros inflados asume que se están mezclando dos fuentes de datos: con cierta probabilidad (π) se obtiene un cero (llamémosle falso cero o cero estructural) y con la probabilidad complementaria ($1-\pi$) se toma una observación de la distribución de Poisson o Binomial Negativa. Así, por ejemplo, si $Z\in Poisson(\lambda)$, la observación Y con ceros inflados tendría distribución:

$$P(Y=0)=\pi+(1-\pi)P(Z=0)=\pi+(1-\pi)e^{-\lambda}$$

$$P(Y=y)=(1-\pi)P(Z=y)=(1-\pi)e^{-\lambda}\frac{\lambda^y}{y!}, si\ y\in 1,2,...$$

De igual modo se puede definir la distribución Binomial Negativa inflada en el cero, cambiando las expresiones de las probabilidades $P(Z=0)$ y $P(Z=y)$.

Las medias también se ven modificadas como consecuencia del cero inflado. Así, la media de la distribución inflada en el cero verifica:

$$E(Y)=E(Z)(1-\pi)$$

de modo que para la Poisson inflada en el cero será $E(Y)=\lambda(1-\pi)$, mientras que para la Binomial Negativa inflada en el cero resulta $E(Y)=\mu(1-\pi)$.

Si disponemos de variables explicativas, entonces tanto μ como π se pueden expresar en función de las variables explicativas mediante un modelo de regresión. En el caso de μ sería un modelo log-lineal, como los que hemos visto anteriormente, mientras que para π se construiría un modelo de regresión logística, u otro modelo adecuado para una probabilidad de éxito (en este caso la probabilidad de falso cero).

2.2.3. Estimación via splines

Hasta aquí se han mencionado modelos de regresión paramétricos. Como ya se ha indicado una vía alternativa consiste en dejar que sean los datos los que conduzcan la estimación de la regresión subyacente sin previa especificación de un modelo paramétrico en que confinar al modelo. Una técnica habitual de realizar esto es aproximar la regresión subyacente mediante una combinación lineal de funciones de una base de splines [38]. De forma resumida, los splines son funciones polinómicas definidas a trozos sobre las que se imponen restricciones en los puntos de unión llamados nodos (knots). Estos nodos dividen el rango de la variable explicativa en subintervalos. Así por ejemplo, si el rango es un intervalo $[a, b]$, un spline de grado p con n knots tales que $a < t1 < t2 < ... < tn < b$ definiendo una partición del intervalo es una función polinómica $s(\cdot)$ tal que:

- $s(\cdot)$ es un polinomio de grado menor o igual que p sobre cada subintervalo $[t_{i-1}, t_i]$.
- la derivada de orden $p-1$ de $s(\cdot)$ es continua sobre todo el rango.

El atractivo de este enfoque es que puede verse como una extensión relativamente sencilla de la regresión lineal.

Regression Splines

Los *regression splines* o los splines de regresión consisten en escoger una base de funciones de tipo spline, y proyectar los datos sobre la base por mínimos cuadrados.

Para ello es preciso determinar:

- el grado (*grado*) de los splines (i.e. del polinomio que los define),
- el número n de nodos y su localización,
- la base específica de splines sobre la que proyectar.

Seleccionados estos hiperparámetros se procede a la estimación que conduce a estimar los $n+grado+1$ parámetros que definen la combinación lineal de elementos de la base y que conforman por tanto sus grados de libertad.

En definitiva, el spline por regresión toma la forma:

$$S(t) = \sum_{k=1}^{L+grado+1} c_k F_k(t, \tau)$$

donde:

- F_k, es la k-ésima función de la base de splines F.
- τ, es el conjunto de nodos sobre el que se define la base F.
- t, es el punto donde se evalúa la función spline S.
- c_k, son los coeficientes en la base de la función spline S.

Pese a existir infinitas bases splines, las más usadas son:

- Base de polinomios truncados
- B-Splines
- Thin plate splines [39, 40]

Y entre las tres mencionadas, la base de B-splines es tal vez la más habitual debido a su estabilidad y atractivas propiedades numéricas.

La regresión por splines tiene ciertas ventajas como:

- Es conceptualmente simple y permite interpretar como trabaja el procedimiento de ajuste de la regresión.
- Puede ser vista como una generalización del modelo de regresión lineal.

Sin embargo presentan también ciertos inconvenientes como la necesidad ya mencionada de fijar el número y ubicación de los nodos.

Smoothing Splines

Una manera de corregir parte de los problemas que presenta la regresión spline es a través de los splines de suavizado. Los splines de suavizado o *smoothing splines* se obtienen como la función $s(x)$ suave (dos veces diferenciable) que minimiza la suma de cuadrados residual más una penalización que mide su rugosidad:

$$\sum_{i=1}^{n}(y_i - s(x_i))^2 + \lambda \int s^{''}(x)^2 dx$$

siendo $0 \leq \lambda < \infty$ el parámetro de suavización. Se prueba que la solución a este problema de optimización, para un parámetro λ prefijado, es un spline cúbico natural con nodos en las observaciones muestrales. Procede indicar que un spline cúbico natural es un spline cúbico que es lineal más allá de los nodos frontera t_1 y t_n. Este resultado es particularmente importante porque evita la selección de número y ubicación de los nodos. Además, la regresión por splines suavizados, que inicialmente se plantea a partir de un problema de optimización de dimensión infinita (toda vez que $s(x)$ se busca en el espacio de funciones de cuadrado integrable y con segunda derivada de cuadrado integrable en $[a, b]$) se transforma con este resultado en un problema de optimización en un espacio de dimensión finita: buscar una combinación adecuada de una base de splines cúbicos naturales con tantos nodos como observaciones. Nótese que el hecho de que haya tantos nodos como observaciones podría pensar en que se alcanzará un modelo sobreparametrizado, pero habrá de tenerse en cuenta el efecto de la constante de penalización λ que asegura que los coeficientes estimados se reduzcan a la linealidad, limitando el número de grados de libertad. λ juega el papel de parámetro de suavizado, buscando el equilibrio requerido entre sesgo y varianza de la estimación, y puede determinarse atendiendo a diferentes criterios tales como validación cruzada o validación cruzada generalizada, entre otros. en cualquier caso y como es lógico, este camino de estimación de la regresión puede conllevar problemas computacionales cuando el número de datos es muy elevado.

P-Splines

Se trata de una versión de los ajustes por splines que basándose en la misma formulación que los splines de suavización restringen los posibles valores de la función $s(x)$ a una familia de funciones paramétrica o a una expresión finita dentro de una base [41].

$$\sum_{i=1}^{n}(y_i - s(x_i))^2 + \lambda \int s^{''}(x)^2 dx$$

donde s se formula como combinación de elementos de una base y por tanto definible como:

$$s(x) = \sum_{j=1}^{k} \beta_j b_j(x)$$

Se pueden entender por tanto como una combinación de la *suavización spline* y *B-Splines*. Aunque atendiendo a que $s(x)$ no siempre está restringido a bases $B-Spline$ sino que solo ha de estar restringido a valores fijos en una base o a una determinada familia paramétrica. En definitiva, los splines con penalizaciones combinan lo mejor de los splines de regresisón y los splines por suavizado: utilizan menos parámetros que los splines de suavizado, pero la selección de los nodos no es tan determinante como en los splines de regresión. En los splines penalizados el tamaño de la base utilizada es mucho menor que el número de datos muestrales, lo que supone una considerable reducción de la dimesnión del problema. El número de nodos, en el caso de los P-splines, no supera los 40, ganado así en eficiencia computacinal respecto a los slpliens de suavizado. Por otro lado, la introducción de penalizaciones relaja la importancia de la elección del número y ubicación de los nodos.

2.3. Resultados

Esta sección se exponen los resultados obtenidos en los análisis de RNRS-1 y RNRS-2. Como breve recordatorio:

- RNRS1 y RNRS2 son dos genes del genoma mitocondrial que codifican los ARNrs de subunidades pequena y grande del ribosoma mitocondrial respectivamente.
- El objetivo de este análisis es tratar de determinar si efectivamente existe una tendencia entre el número de variaciones reportadas para una posición y el índice de conservación asignada sobre la misma.

En ambos casos se han aplicado los modelos ya presentados en la Sección 2.2. Adicionalmente, se han realizado también análisis:

- Usando el logaritmo de la variable respuesta, para tratar de acercar valores asociados a haplogrupos altamente secuenciados y evitar diferencias debido al haplogrupo.
- Filtrando las variaciones con mayor nivel de repetición seleccionando los datos ubicados entre $[0, M]$, $[0, 0.99M]$ y $[0, 0.95M]$ siendo M el máximo de variaciones.

2.3.1. Resultados RNRS-2

A continuación se muestran los resultados obtenidos para la subunidad grande del ribosoma mitocondrial RNRS-2. Primero se muestran los resultados con todos los modelos y a continuación se incluye una discusión acerca de las conclusiones que se derivan de los mismos.

Modelos de regresión paramétricos

Se muestran en primer lugar los resultados alcanzados con los modelos de regresión paramétricos. En este caso es importante destacar que todos los modelos ajustados son modelos generalizados inflados con ceros y las funciones de distribución usadas son poisson, geométrica y binomial.

Las Figuras 2.1, 2.2, 2.3 muestran los ajustes sobre todos los datos, los datos filtrados al 99 % y los datos filtrados al 95 %. Nótese que para el filtrado al 95 % la densidad de las respuestas se concentra de tal forma entre 0 y 1 que no es factible detectar tendencia alguna. En el resto de ajustes puede verse claramente que existe una tendencia definida por todos los ajustes. Siguiendo el orden de la leyenda, los ajustes realizados se corresponden a:

- Poisson 1 - regresión de poisson inflada con ceros estructurales ($GB.Seqs\ CVTOT|1$).
- Poisson 2 - regresión de poisson inflada con ceros que dependen de la variable explicativa $CVTOT$ ($GB.Seqs\ CVTOT|CVTOT$).
- Poisson 3 - regresión de Poisson.

		AIC		
Modelo	df	Sin filtro	Filtro 0.01	Filtro 0.05
Poisson 1	3	83160.651	12743.898	3486.948
Poisson 2	4	83131.503	12734.566	3487.552
Poisson 3	4	83131.503	12734.566	3487.552
Default	3	83131.503	12734.566	3487.552
Geométrica 1	3	6268.170	4723.726	3044.212
Geométrica 2	4	6268.424	4715.671	3046.204
Binomial 1	4	4930.922	4267.834	3045.989
Binomial 2	5	4932.922	4268.026	3046.955

Cuadro 2.1: Modelos lineales generalizados: Valores de AIC para los ajustes (LSU)

- Default
- Geometric 1 - regresión de geométrica inflada con ceros estructurales ($GB.Seqs\ CVTOT|1$).
- Geometric 2 - regresión de geométrica inflada con ceros que dependen de la variable explicativa $CVTOT$ ($GB.Seqs\ CVTOT|CVTOT$).
- Binom 1 - regresión de binomial inflada con ceros estructurales ($GB.Seqs\ CVTOT|1$).
- Binom 2 - regresión de binomial inflada con ceros que dependen de la variable explicativa $CVTOT$ ($GB.Seqs\ CVTOT|CVTOT$).

La Tabla 2.1 agrupa los resultados de los AIC (Criterio de Información de Akaike) de los ajustes realizados. Según estos resultados parece claro que la regresión de Poisson, pese a capturar la tendencia, ofrece un ajuste totalmente inadecuado. Por otro lado, tanto el geométrico como la binomial negativa tienen un ajuste mucho mejor, destacando sobre todo el modelo binomial. Mencionar por último que el uso de un modelo con ceros estructurales o dependientes no aporta realmente gran diferencia al nivel del AIC.

Las Figuras 2.4, 2.5, 2.6 muestran exactamente los mismos ajustes que las comentados en el párrafo anterior salvo que en este caso la variable respuesta es $log(\#variaciones)$.

Modelos GAMLSS: ajustes paramétricos y no paramétricos

Siguiendo con el conjunto de ajustes realizados continuamos con una serie de modelos aditivos generalizados para localización, escala y forma, generalmente conocidos como GAMLSS [43, 44, 45, 46]. En nuestro caso al ser X una variable única los modelos que ajustamos únicamente constan de componente paramétrica o no paramétrica pero nunca ambas y por tanto carecen de parte aditiva. Los modelos GAMLSS asumen que la variable respuesta tiene una función de densidad definida por hasta cuatro parámetros $(\mu, \sigma, \ni, \tau)$ que determinan su posición, escala y forma, y que cada uno de ellos puede variar independientemente de los otros en función de los predictores. Estos modelos aprenden

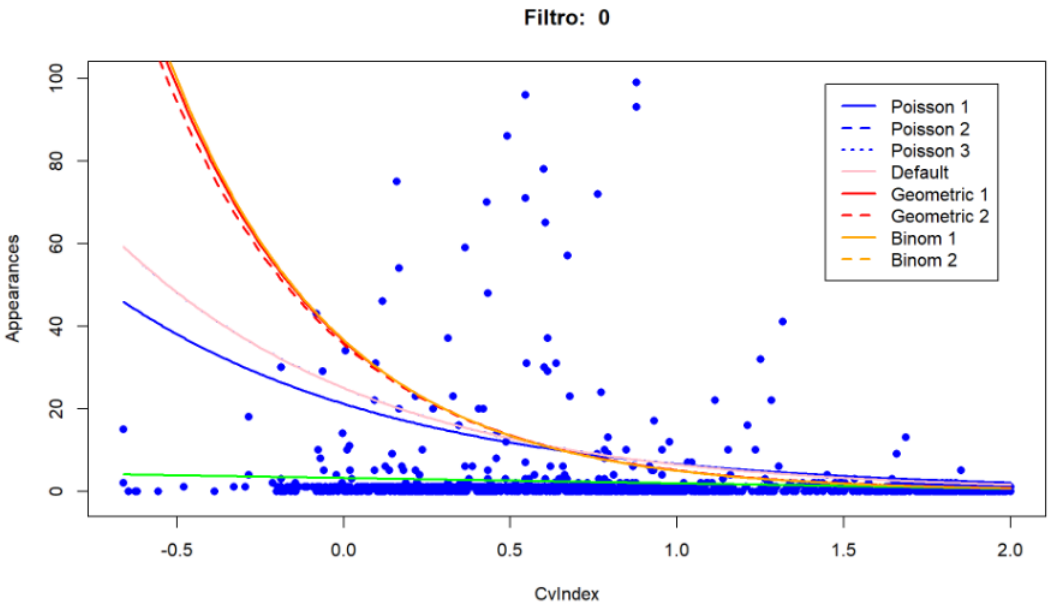

Figura 2.1: RNRS2 - Modelos inflados con ceros sin filtro

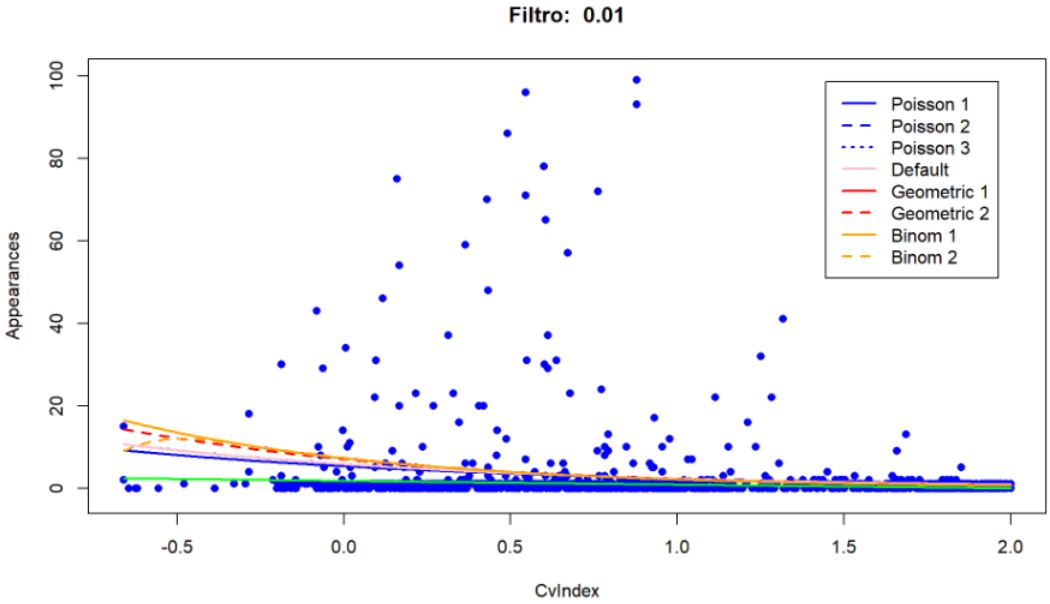

Figura 2.2: RNRS2 - Modelos inflados con ceros filtro al 0.01

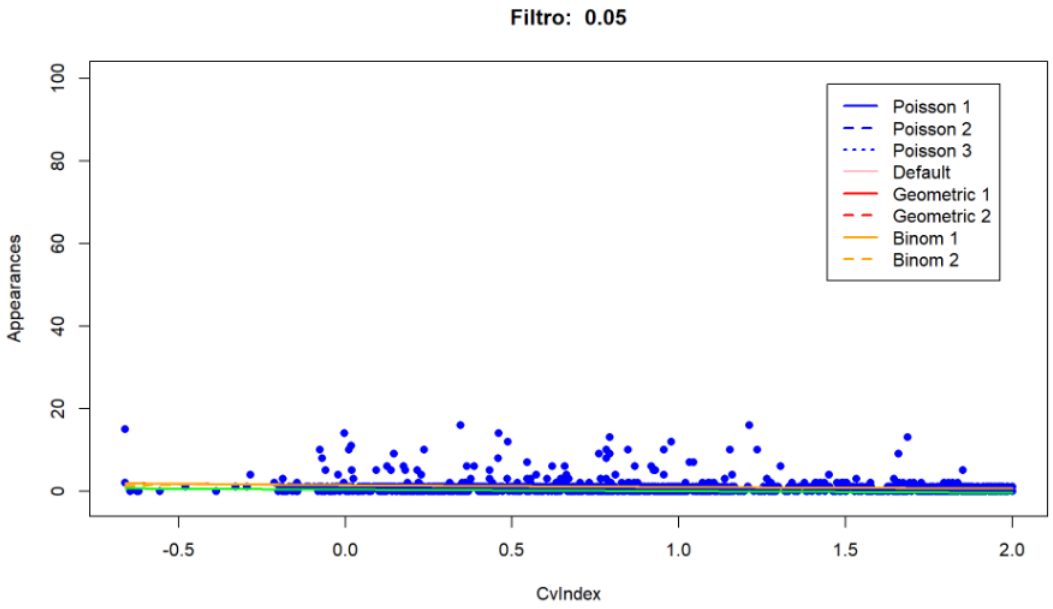

Figura 2.3: RNRS2 - Modelos inflados con ceros filtro al 0.05

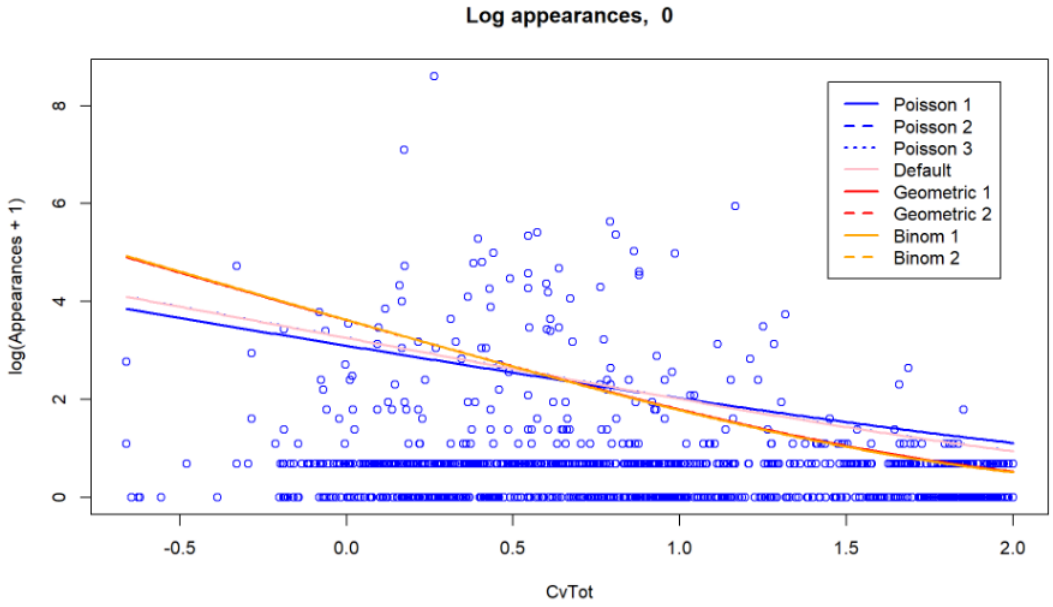

Figura 2.4: RNRS2 - Modelos inflados con ceros sin filtro, versión logarítimica

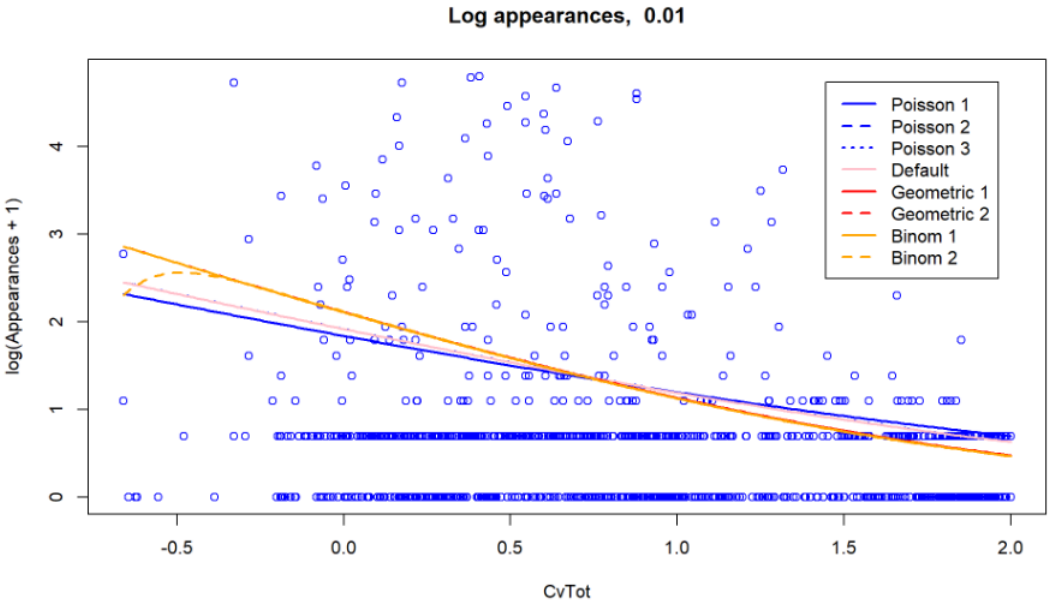

Figura 2.5: RNRS2 - Modelos inflados con ceros filtro al 0.01, versión logarítimica

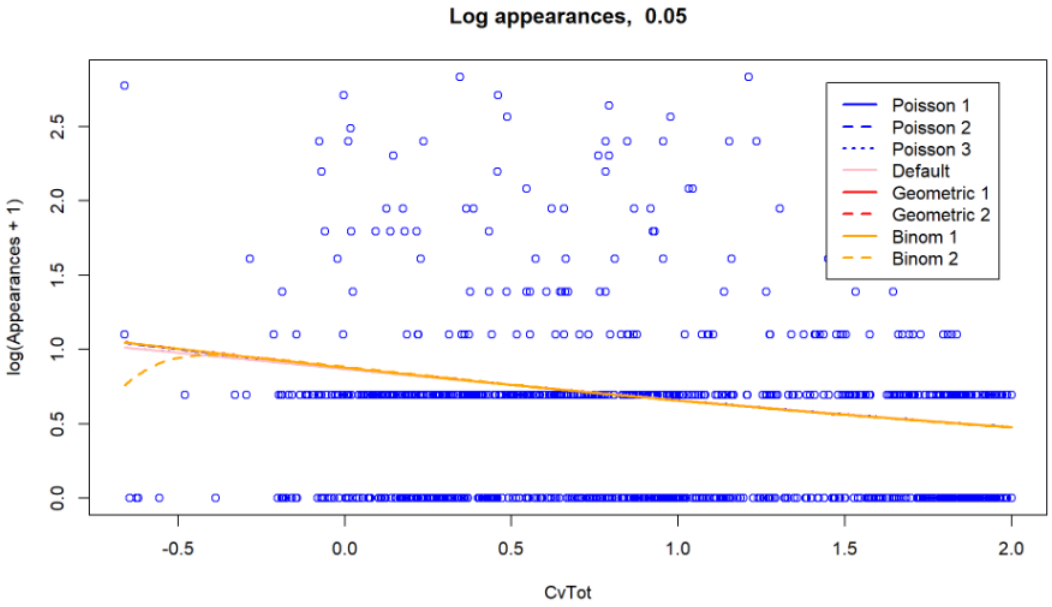

Figura 2.6: RNRS2 - Modelos inflados con ceros filtro al 0.05, versión logarítimica

Modelo	df	AIC
dDEL	3	3934.919
dSI	3	4110.138
dPIG	2	4231.974
dNBI	2	5137.345
dZIP	2	88184.975
dZIP2	2	88184.975
dPO	1	99620.794

Cuadro 2.2: Valores de AIC para los ajustes paramétricos y no paramétricos (RNRS2)

por lo tanto hasta cuatro funciones, donde cada una establece la relación entre las variables predictoras y uno de los parámetros.

Por simplicidad se usó la libreria *gamlss* que ya incorporá métodos de análisis exploratorio que nos permitió seleccionar el modelo GAMLSS que mejor se ajustaba a nuestros datos. Por la naturaleza del problema las funciones de enlace escogidas para evaluar fueron:

- *Poisson*, PO
- *Negative Binomial Type I*, NBI
- *Poisson Inverse Gaussian*, PIG
- *Inverse Gaussian*, SICHEL
- *Zero inflated Poisson*, ZIP
- *Zero inflated Poisson 2*, ZIP2
- *Delaport*, DEL

Los resultados comparativos entre el histograma de la variable respuesta, $\#Apariciones$ se puede ver en la Figura 2.7. Se ve claramente que la disputa se encuentra entre DEL, PIG y $SICHEL$. Su nivel de ajuste con mayor nivel de precisión se puede ver en la Figura 2.9 y en esta ocasión se puede diferenciar que tanto $SICHEL$ como DEL cometen una mayor error para valores más altos de $\#Variaciones$ mientras que PIG parece la mejor opción. La Tabla 2.2 agrupa los valores de los AIC para todos los modelos ajustados. Puede verse claramente que el análisis exploratorio visual sobre los histogramas se confirma destacando los ajustes DEL, PIG y SICHEL. Según los valores de los AIC el orden de selección debería ser *Delaport*, *Inverse Gaussian* y *Poisson Inverse Gaussian*. Tras probar las tres finalmente el ajuste bajo la *Poisson Inverse Gaussian* o PIG es la que captura mejor la tendencia esperada. Además, las diferencias de AIC no parecen significativos especialmente atendiendo a la extrema dispersión que presentan los datos.

La Figura 2.9 muestra ajuste semiparamétrico a través de la función *PIG* para un modelo lineal, cuadrático y basado en splines.

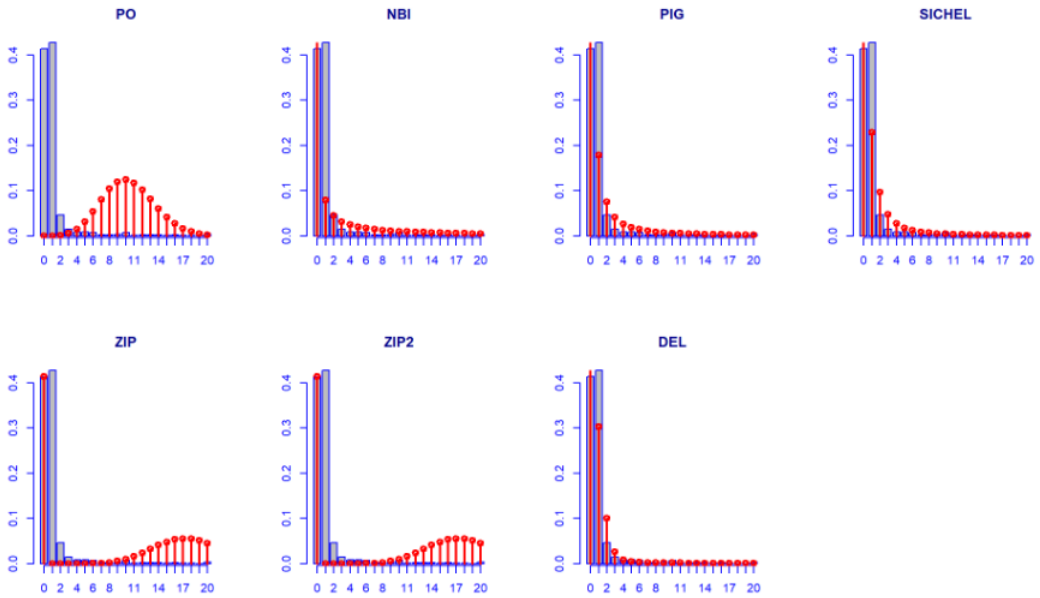

Figura 2.7: RNRS2 - Histogramas y ajuste a las distribuciones GAMLSS

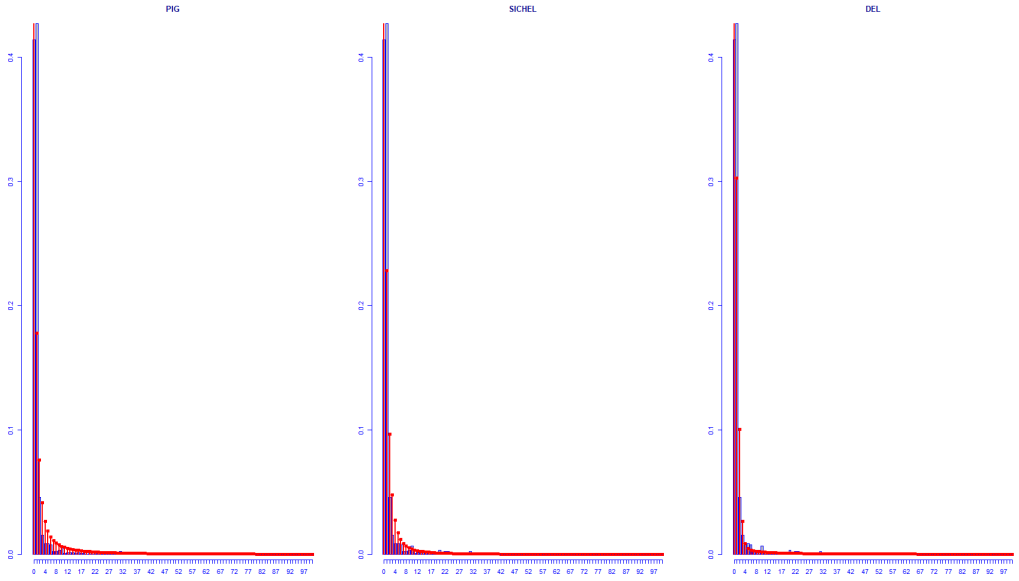

Figura 2.8: RNRS2 - PIG conta SICHEL contra DEL

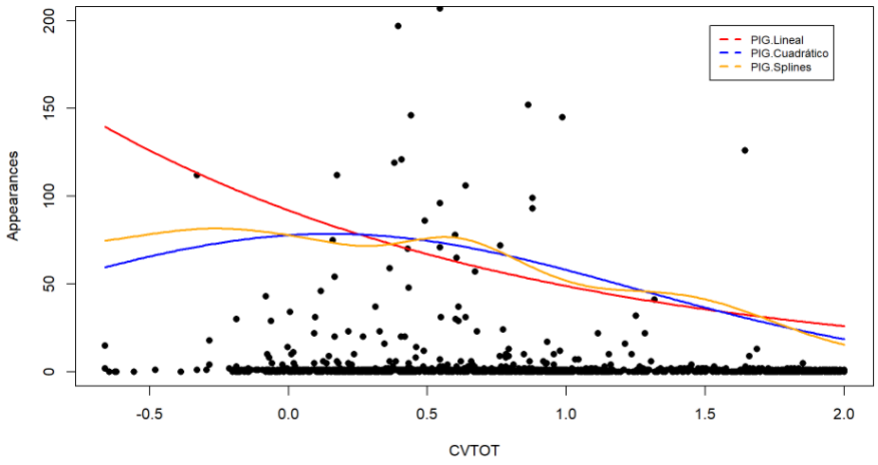

Figura 2.9: RNRS2 - PIG conta SICHEL contra DEL

Estimación via splines

Las estimaciones vía splines las realizamos todas sobre la variable respuesta *variaciones* transformada por su *log*. Esto surgió de la necesidad de homogeneización de los valores de las apariciones para evitar el impacto excesivo que presentaba una variante en concreto que se trata más adelante.

El primero de los casos que se muestra son los *Smoothing Spline Regression* con seis grados de libertad (se hicieron pruebas de selección de los grados de libertad desde dos hasta diez, siendo seis aquel que ofrecía mejores resultados sin un claro sobreajuste). Las Figuras 2.10, 2.11 y 2.12 muestran los ajustes para los filtros 0, 0.01 y 0.05 respectivamente. La segunda aproximación spline que realizamos fue a través de regression spline con base de B-splnes usando tres grados de libertad (en esta ocasión se probó de 3 a 9). Las Figuras 2.13, 2.14 y 2.15 muestran los ajustes para los filtros 0, 0.01 y 0.05 respectivamente.

Finalmente, usamos P-Splines con unos estimadores con un comportamiento prácticamente idéntico al mostrado por los splines por suavización, motivo por el cual omitimos ya los correpondientes gráficos.

Algo a destacar frente a los ajustes de regresión inflados con ceros es que estos, al constar de una mayor flexibilidad, son capaces de capturar cambios de tendencia en la estimación de la curva. Si miramos cuidadosamente a la estimación de la curva se experimenta un cambio en su carácter constante en torno al valor de CV-Index igual a 0 y posteriormente una caída acentuada. Esto viene propiciado por una variación, 2706 G-A, presente en múltiples haplogrupos en una posición con CV-Index igual a uno. Esto es sumamente positivo dado que justamente es lo que se estaba buscando. Pese a que ambos muestran lo mismo, el comportamiento sugerido por los splines se ajusta más a lo establecido por la hipótesis mientras que con los primeros análisis la tendencia era demasiado continua y constante, por tanto, poco realista.

Finalmente, la Tabla 2.3 resume los valores de AIC obtenidos para los tres ajustes no paramétricos con el valor de los parámetros óptimos.

Análisis dicotomizando la variable respuesta

Un análisis sumamente interesante, si queremos eliminar totalmente el efecto derivado del número de secuenciaciones de un haplogrupos en cuestión, es el obtenido a través del ajuste tras haber binarizado la variable respuesta. La premisa en este caso es que el número de variaciones no importa sino que solamente es relevante si ha habido alguna variación registrada o no. En este caso se realizaron dos ajustes para tratar de nuevamente visualizar la tendencia bajo esta nueva idea:

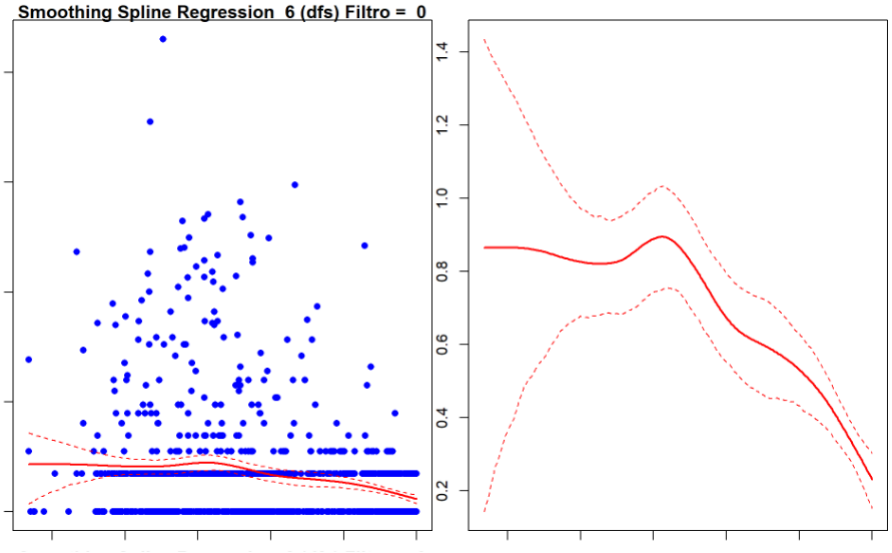

Figura 2.10: RNRS2 - SSR sin filtro

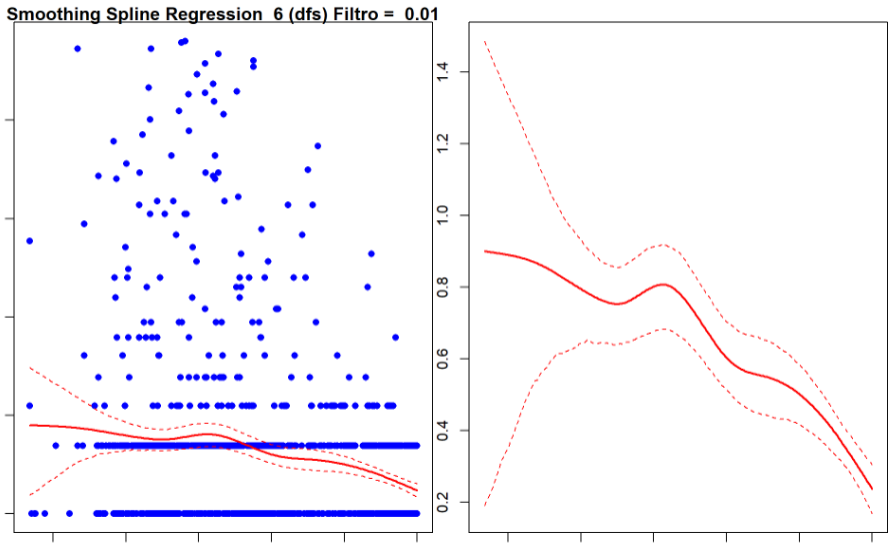

Figura 2.11: RNRS2 - SSR con filtro al 0.01

		AIC		
		Filtro 0	Filtro 0.01	Filtro 0.05
Smoothing splines	df = 6	3255.014	2839.377	1769.851
B-Splines	df = 3	3254.2	2374.048	1768.682
P-Splines	k = 18	3252.413	2370.701	1767.276

Cuadro 2.3: AIC splines RNRS2

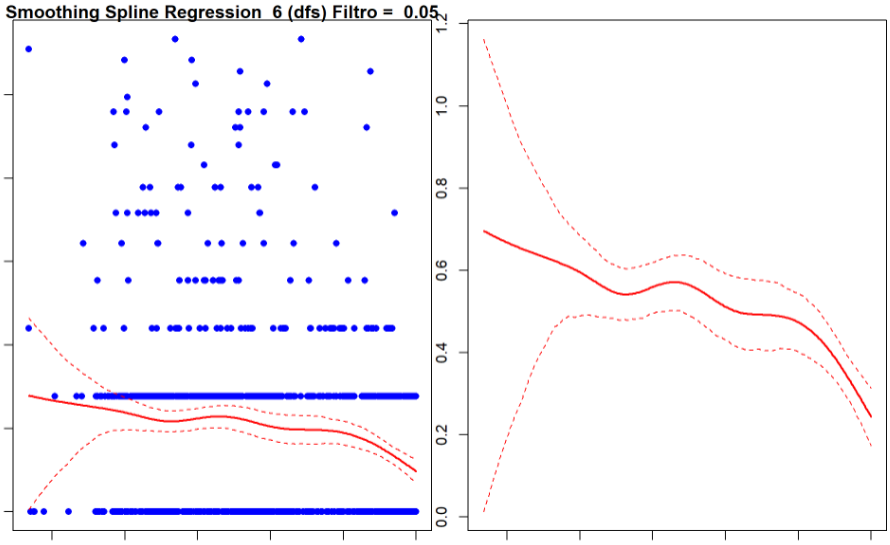

Figura 2.12: RNRS2 - SSR con filtro al 0.05

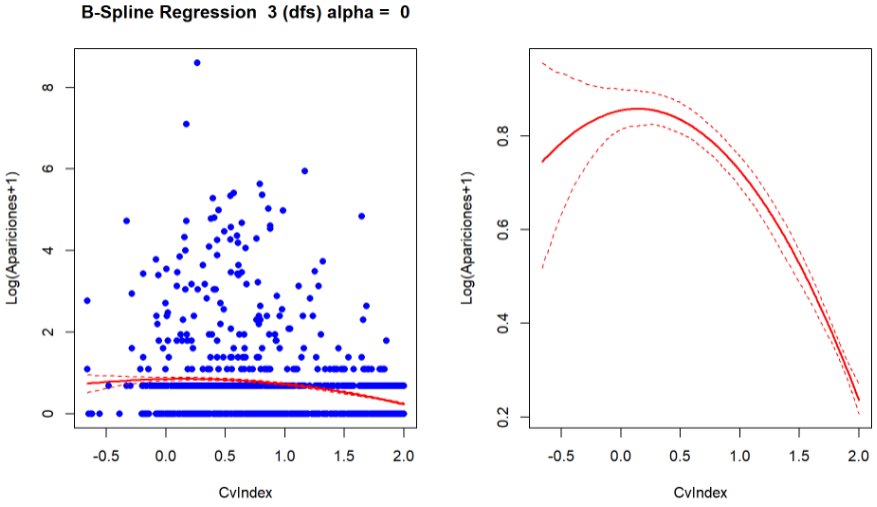

Figura 2.13: RNRS2 - B-Splines sin filtro

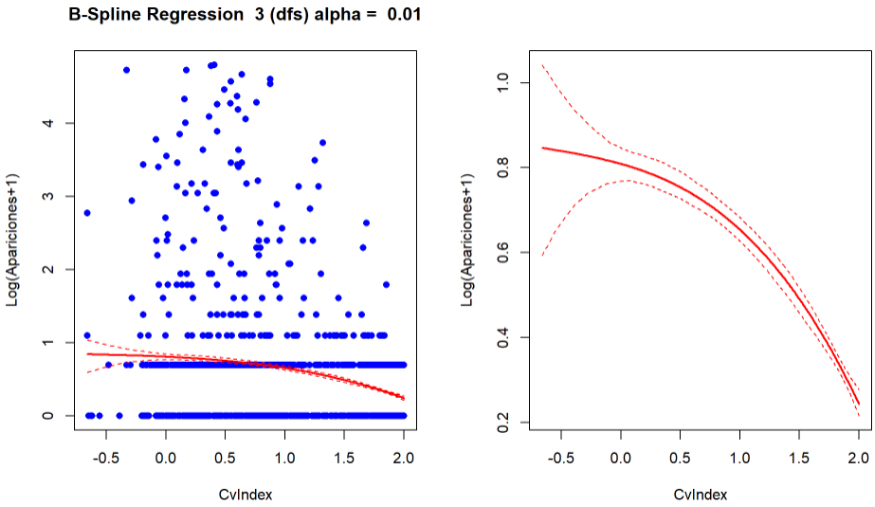

Figura 2.14: RNRS2 - B-Splines con filtro al 0.01

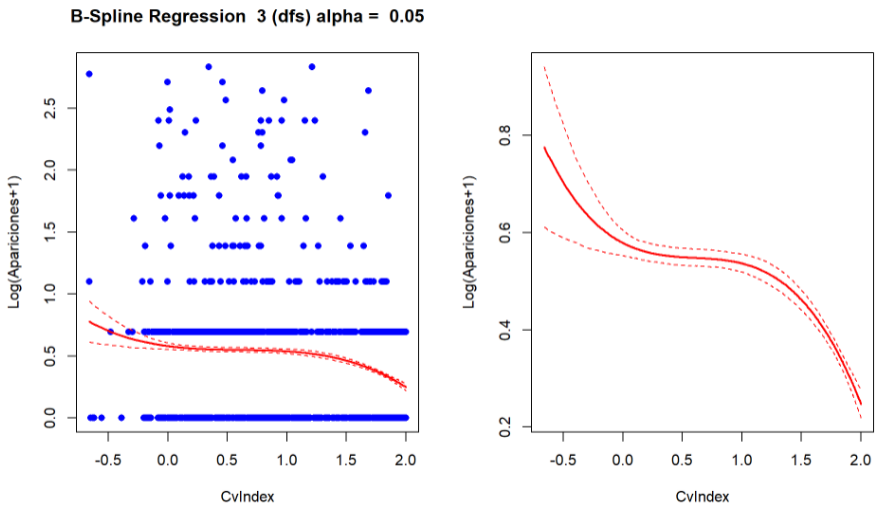

Figura 2.15: RNRS2 - B-Splines con filtro al 0.05

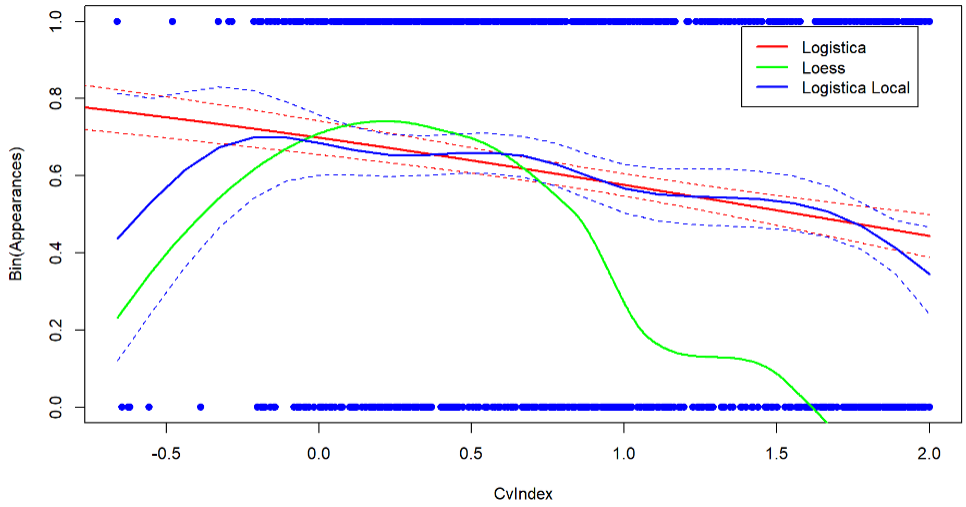

Figura 2.16: RNRS2 - Ajustes sobre la respuesta binarizada

	AIC
Regresión logística	1562.503
Regresión logística local	1560.349
LOESS	1559.515

Cuadro 2.4: AICs para los ajustes con variable respuesta binaria

- Ajuste paramétrico via regresión*logística clásica*
- Ajustes no paramétricos:
  - regresión *logística local*
  - *LOESS*

La Figura 2.16 muestra los ajustes sobre la variable respuesta binarizada. Nótese que todos delatan la tendencia esperada, decreciente a partir de valores de Cv mayores que 0.5-1. La estimación paramétrica, más rígida, de nuevo no dota de la flexibilidad suficiente al estimador para recoger el cambio de tendencia en los valores pequeños de Cv. Los ajustes no parmétricos, sí. Particularmente con el algoritmo LOESS que ejemplifica de manera ideal el comportamiento buscado. Cabe enfatizar que este algoritmo emplea estimación kernel local cuadrática de manera iterativa y con ventana variable, es decir, ventana adaptada a la densidad local de los datos, ganando así en versatilidad. Esta enorme versatilidad posiblemente sea la causa para el descenso tan acentuado acercándose al a frontera, pero más allá de este hecho, refleja claramente el comportamiento esperado en valores menores del Cv. Por otro lado, la Tabla 2.4 muestra los valores de los AIC para los tres ajustes. Los tres son extremadamente similares y decantarse por uno u otro parece algo arbitrario. Aunque nuevamente el ajuste Loess es el que mejor AIC provee.

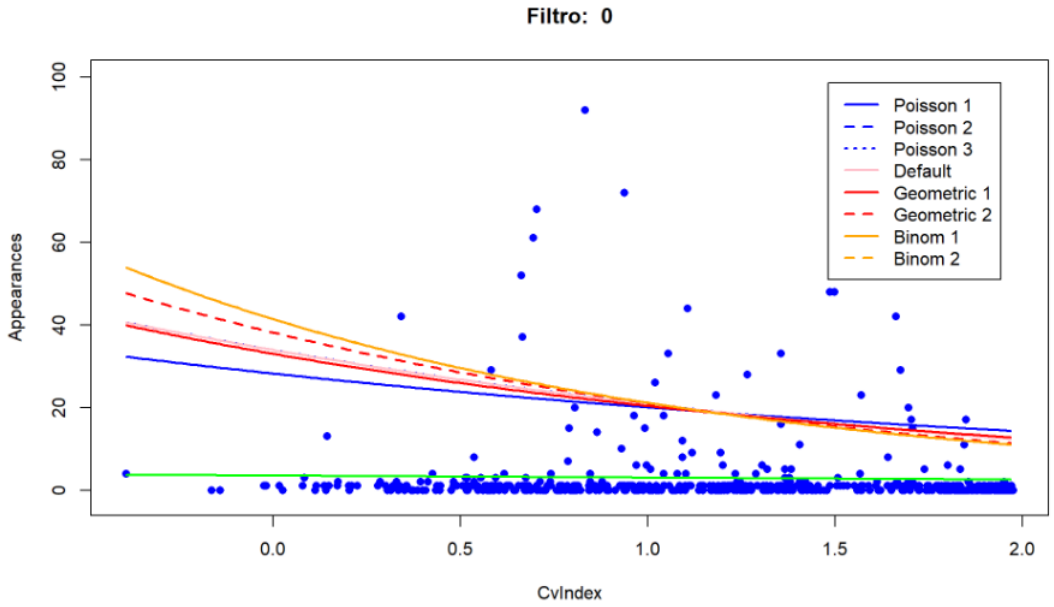

Figura 2.17: RNRS1 - Modelos inflados con ceros sin filtro

2.3.2. Resultados RNRS-1

A continuación se muestran los resultados obtenidos para la subunidad grande del ribosoma mitocondrial RNRS-1. Dado que los modelos usados son los mismos que para la subunidad grande solamente se mostrarán los resultados más relevantes evitando ser innecesariamente redundantes.

Modelos de regresión paramétricos

Nuevamente iniciaremos mostrando los resultados obtenidos con los modelos de regresión paramétricos. A diferencia del apartado anterior únicamente mostraremos los casos sin filtrar los datos pues los resultados siguen la misma tendencia que para la subunidad grande o RNRS2.

Las Figuras 2.17, 2.18 muestran los ajustes de Poisson, Geométrico, binomial negativo y sin regresor en la componente de 0. Como ya venía ocurriendo en el análisis sobre RNRS1 se ve una tendencia más que clara, de hecho en este caso parece incluso más pronunciada que en los ajustes de RNRS2. De hecho si miramos a la Tabla 2.5 podemos observar como los AIC, pese a seguir en unos niveles elevados, son ligeramente inferiores a los que teníamos en el caso de la RNRS2 o LSU.

Modelos generalizados: ajustes paramétricos y no paramétricos

Nuevamente la primera etapa es la selección del modelo GAMLSS que mejor ajusta el número de variaciones. Las funciones evaluadas volvieron a ser las mismas que en el caso RNRS1:

- *Poisson*, PO
- *Negative Binomial Type I*, NBI
- *Poisson Inverse Gaussian*, PIG
- *Inverse Gaussian*, SICHEL
- *Zero inflated Poisson*, ZIP

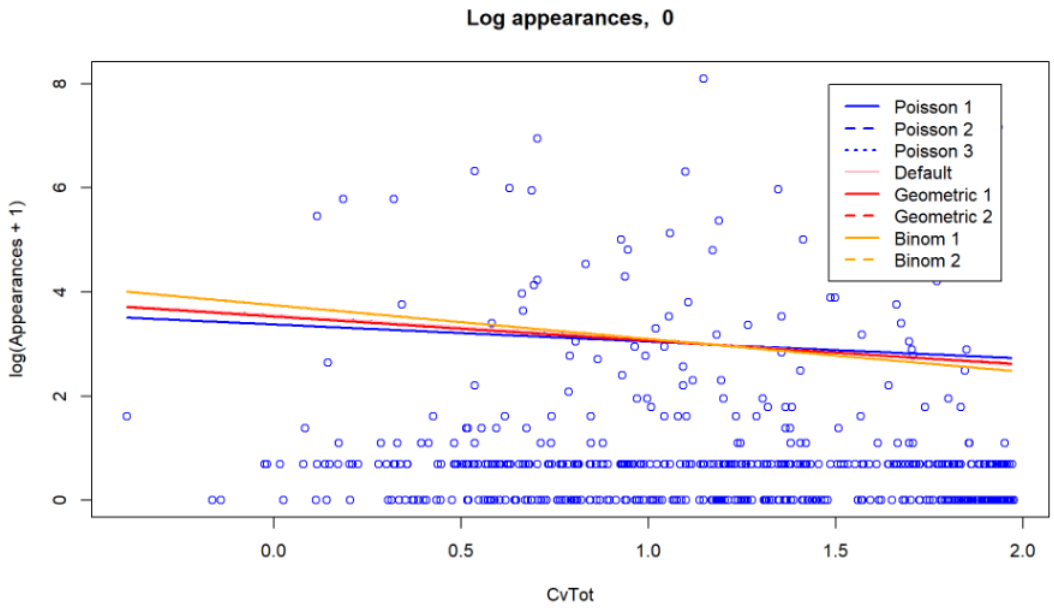

Figura 2.18: RNRS1 - Modelos inflados con ceros versión logarítmica

		AIC		
Modelo	df	Sin filtro	Filtro 0.01	Filtro 0.05
Poisson 1	3	66067.436	17917.799	3039.626
Poisson 2	4	66060.227	17911.189	3035.127
Poisson 3	4	66060.227	17911.189	3035.127
Default	3	66060.227	17911.189	3035.495
Geométrica 1	3	3994.161	3143.599	1985.736
Geométrica 2	4	3988.963	3144.841	1976.719
Binomial 1	4	2902.854	2587.710	1911.416
Binomial 2	5	2904.855	2589.712	1913.416

Cuadro 2.5: Modelos lineales generalizados: Valores de AIC para los ajustes (SSU)

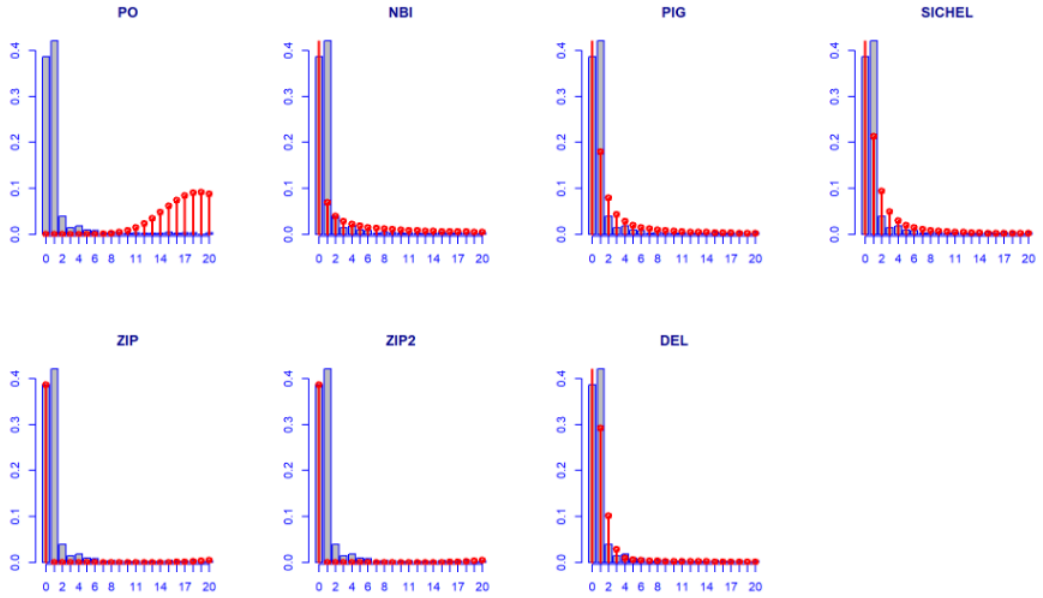

Figura 2.19: RNRS1 - Histogramas y ajuste a las distribuciones GAMLSS

- *Zero inflated Poisson 2*, ZIP2
- *Delaport*, DEL

Los resultados comparativos entre el histograma de la variable respuesta, $\#Apariciones$ se puede ver en la Figura 2.19. Se ve claramente que la disputa se encuentra entre DEL, PIG y $SICHEL$. Su nivel de ajuste con mayor nivel de precisión se puede ver en la Figura 2.20. Nuevamente, mostramos en la Tabla 2.6 y se confirma la intuición planteada donde SICHEL/DEL/PIG presentan valores sumamente similares. En este caso con carácter explicativo se muestran los tres ajustes obtenidos en las Figuras 2.21, 2.22 y 2.23. En estos se observa:

- DEL, el ajuste con el AIC más bajo, no muestra la tendencia esperada sino que se limita a ajustar un modelo que prácticamente no se mueve de la recta $y = 1$ en las tres variantes ajustadas.
- SICHEL, el número dos en términos del AIC, muestra tres curvas algo mejor ajustadas que DEL, especialmente la que presenta una dependencia cuadrática, pero sigue siendo escasa.
- PIG es que de las tres muestra una tendencia más marcada y acorde a lo esperado, a pesar de tener el mayor AIC de las tres.

Estimación via splines

El primero de los casos que se muestra son los *Smoothing Spline Regression* con cuatro grados de libertad. A diferencia del caso de RNRS2 al usar seis grados de libertad la curva construída presentaba un excesivo ajuste a los datos provistos. La Figura 2.24 muestra dicho ajuste.

La segunda aproximación spline que realizamos fue a través de B-splines con tres grados de libertad. La Figura 2.25 muestra dicho ajuste.

De manera análoga a como ocurría en el RNRS1 los ajustes basados en Splines muestran una flexibilidad muy beneficiosa frente a los ajustes inflados con ceros. Básicamente, permiten visualizar

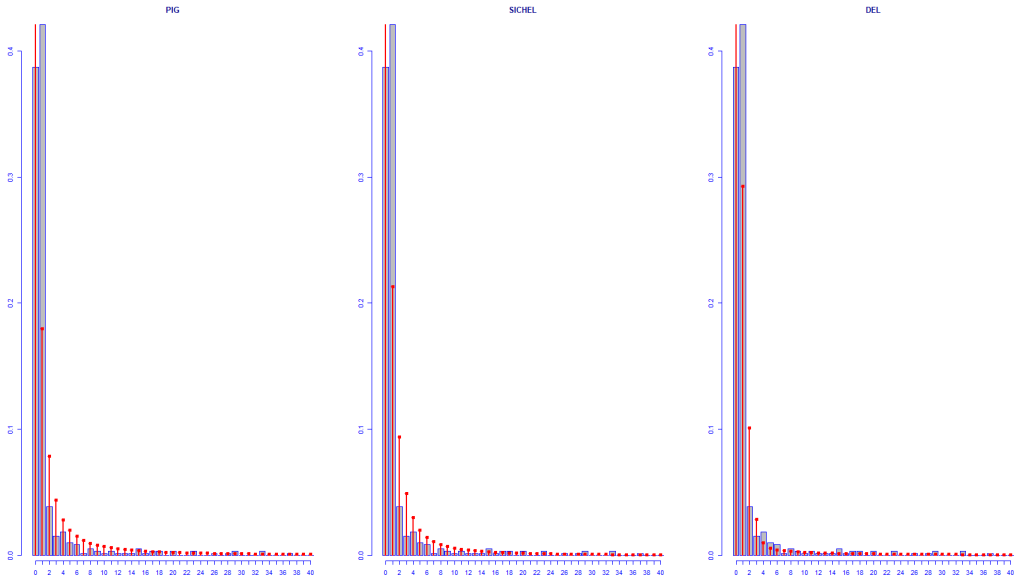

Figura 2.20: RNRS1 - PIG conta SICHEL contra DEL

Modelo	df	AIC
dDEL	3	2307.513
dSI	3	2404.931
dPIG	2	2431.538
dNBI	2	2907.367
dZIP	2	66457.246
dZIP2	2	66457.246
dPO	1	76809.875

Cuadro 2.6: Valores de AIC para los ajustes paramétricos y no paramétricos (RNRS1)

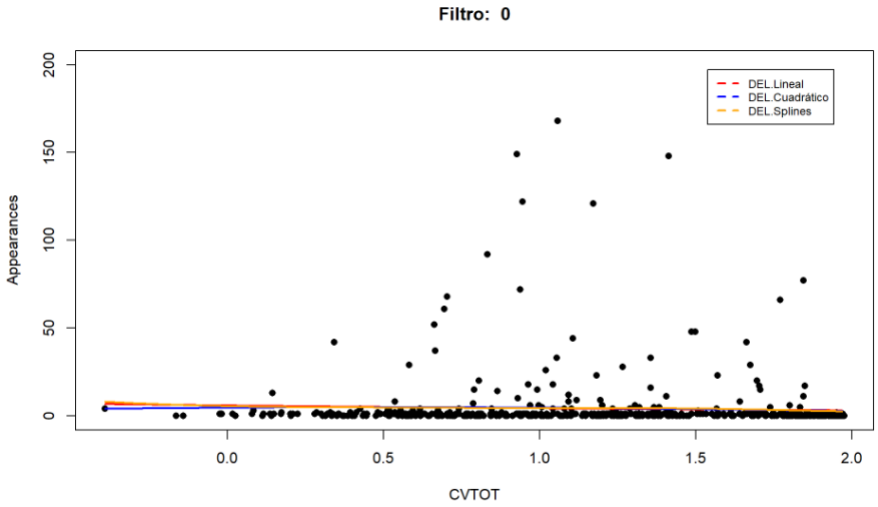

Figura 2.21: RNRS1 - Ajuste DEL

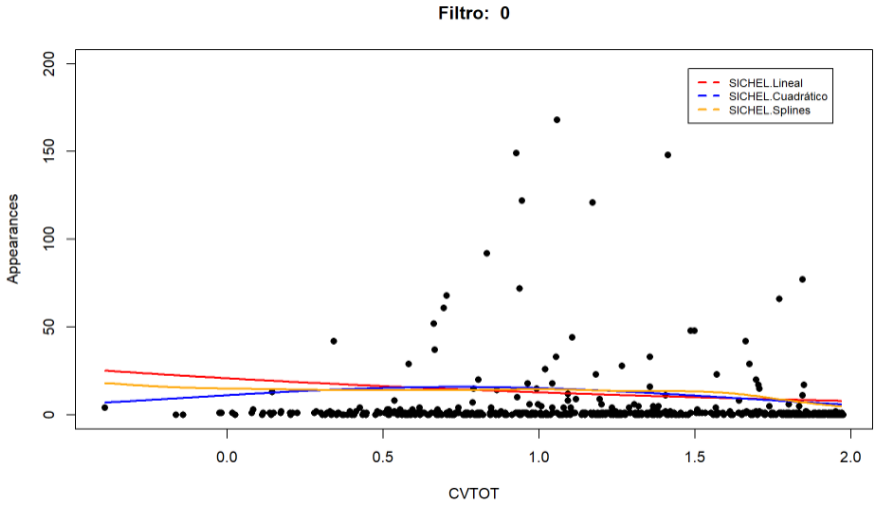

Figura 2.22: RNRS1 - Ajuste SICHEL

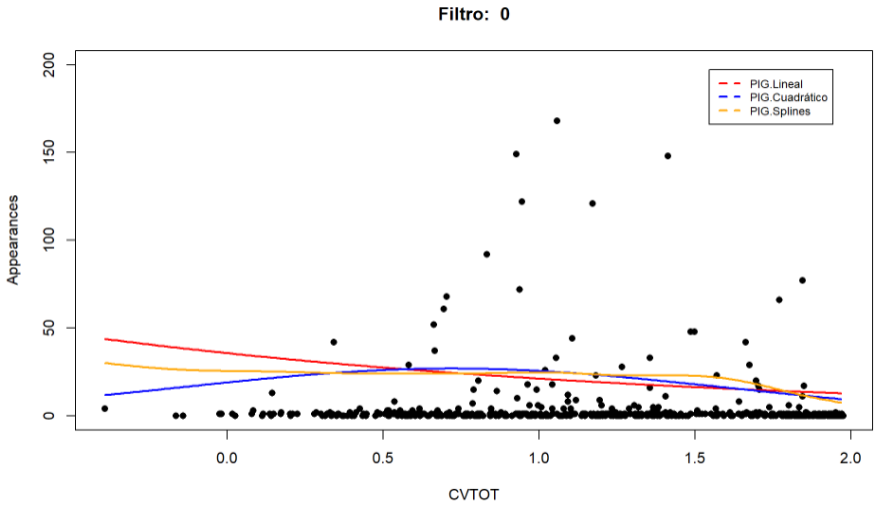

Figura 2.23: RNRS1 - Ajuste PIG

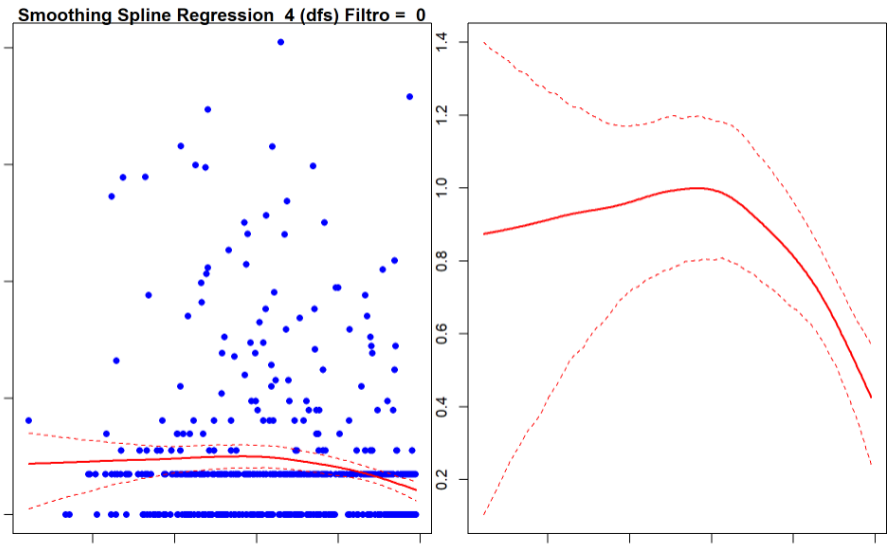

Figura 2.24: RNRS1 - SSR sin filtro

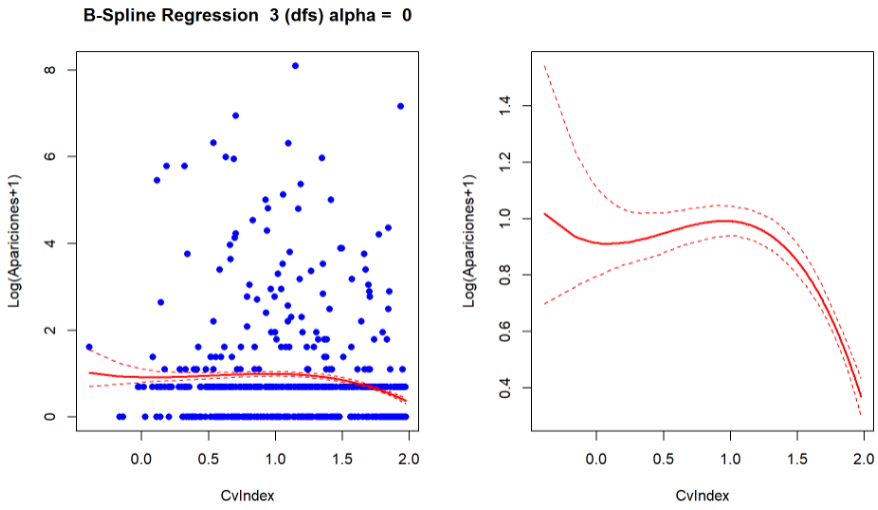

Figura 2.25: RNRS1 - B-Splines sin filtro

claramente la caída en el número de variantes al llegar a cotas más altas de CV-Index mientras que preservan, o incrementan ligeramente, los niveles de variaciones reportadas en las etapas inmediatamente anteriores al comienzo de la caída.

Análisis dicotomizando la variable respuesta

En este caso nuevamente se han realizado los mismos ajsutes que en el caso RNRS2 (regresión logística, regresión logística local y LOESS). Los resultados vuelven a mostrar resultados parejos a los vistos para la sección RNRS2. Sin embargo en este caso parece que LOESS se aparta menos de la regresión logística. Estos resultados se muestran en la Figura 2.26.

2.4. Conclusiones

Se han aplicado diferentes técnicas de regresión paramétricas y no paramétricas, incluyendo modelos sofisticados como los modelos inflados por ceros o los GAMLSS, con la idea básica de corroborar si efectivamente existe una tendencia decreciente entre el el número de variaciones reportadas y el valor del CV-Index. Es decir, apoyar estadísticamente que el CV-Index crece el número de variaciones reportadas se ve reducido. Esta idea surge del propio concepto de CV-Index que implica dicho comportamiento pero cuya veracidad en el caso del genoma mitocondrial no ha sido verificado.

A la vista de los resultados obtenidos podemos asegurar que tanto para la región RNRS1 y como RNRS2 la mencionada relación parece consolidarse de manera significativa. Aún así, no se trata de una relación decreciente estricta pues el comportamiento sugerido por técnicas como la suavizaciónpor splines apunta a un carácter sostenido en el número de variantes seguido de una caída abrupta a partir de valores de CV-Index ≥ 1. Por otro lado, los ajustes logísticos y LOESS, ajustes donde la variable respuesta se encuentra binarizada, parecen apuntar a la existencia de la tendencia aunque con un efecto bastante liviano[1]. De hecho parece que estos apoyan una caída paulatina más que una caída acentuada a diferencia de los splines.

[1]Aunque los ajustes logísticos locales apuntan a una caída acentuada en valores bajos de CV-Index hay que recordar que estos son suscepctibles del efecto frontera que parece jugar un papel importante en estos (tanto al comienzo como al final de la curva estimada)

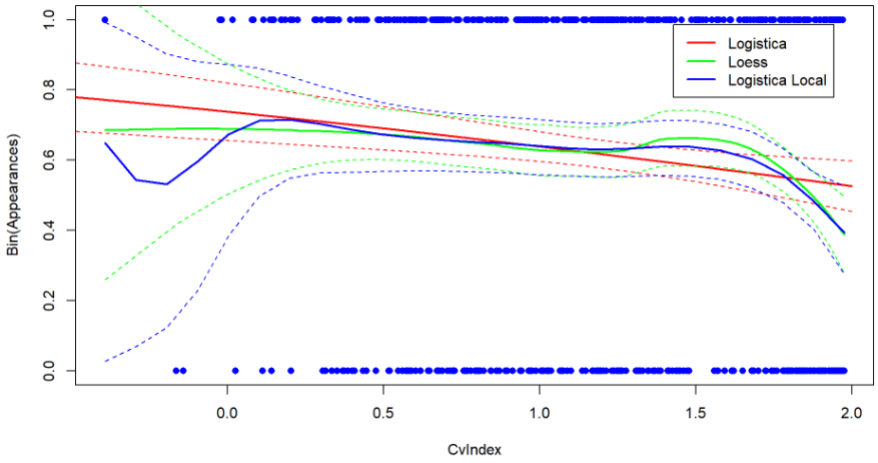

Figura 2.26: RNRS1 - Ajustes sobre la respuesta binarizada

Los resultados son por tanto claros sobre todo a partir del CV > 1 donde caen en picado. Esa es la zona que más nos interesa sobre todo para el cáncer. De ahí para abajo se mantiene mucho más estable, lo cual puede tener sentido. Es posible pues casi todo el RNAr está estructurado, independientemente de su índice de conservación. Para Cvs > 1, la influencia del residuo en la función del ribosoma es la que lo mantiene conservado. De 1 hacia abajo el coste de la variación es el mismo porque los residuos solo tienen función estructural, independientemente de que pertenezcan a estructuras más o menos conservadas en la filogenia. Aún así, los resultados siendo positivos deben ser tenidos en cuenta con cautela y esperar a que existiesen más datos para poder dar un rotundo si, sobre todo atendiendo a que la tendencia derivada de algunas de las técnicas, como la de las regresiones logísticas, es *sutil* o que se experimentan crecimientos injustificados en ciertos valores, curvas splines. Por otro lado, hay que recordar que los CV-Index son asignados por la asociación entre las secciones RNRS1 y RNRS2 con el genoma de *E. coli*. Por lo que no se trata de unos valores específicos calculados para el genoma nuclear humano, de hecho los valores negativos del ajuste provienen de nucleótidos de RNRS1 o RNRS2 que no presentan homólogo.

En definitiva creo que los resultados son alentadores y permiten verificar la hipótesis inicialmente planteada. Esto abre la puerta a estudios posteriores sobre el cáncer, que ya estamos planeando. Al margen creo que el construir un índice de conservación específico para el genoma mitocondrial, independiente de cualquier información nuclear, podría ser sumamente beneficioso. Especialmente en la actualidad donde los estudios mitocondriales están poco a poco ganando más peso.

2.5. Trabajo futuro

Confirmada la hipótesis planteada, queda abierto el comienzo a los estudios de asociación genética[2] sobre el genoma mitocondrial. En la actualidad existen multitud de artículos donde se reportan relaciones entre diversas enfermedades, especialmente en enfermedades como el cáncer, y variaciones en el genoma mitocondrial. De hecho una duda que existe en este contexto es si mutaciones en el genoma mitocondrial ven favorecida la extensión del cáncer o es el cáncer quien favorece la aparición de variaciones mitocondriales. Algo que es sabido es que las células cancerosas suelen tener variaciones

[2]Un estudio de asociación del genoma completo GWAS (Genome-wide association study) es un análisis de las variaciones a lo largo de todo el genoma humano con el objetivo de identificar su asociación a un rasgo observable. Los GWAS suelen centrarse en asociaciones entre los polimorfismos de un solo nucleótido (SNPs) y rasgos como las principales enfermedades.

mitocondriales que modifican los procesos de oxidación de los nutrientes en las células y por tanto de obtención de energía. De hecho parte del trabajo futuro ya bajo estudio es si la tendencia de caída en el intervalo $[1, 2]$ desaparece en células afectadas por el cáncer, es decir que existe una mayor concentración de variaciones en dicho intervalo. Esto significaría que células cancerígenas tienden a tener las regiones conservadas, que recordemos que suelen ser imprescindibles en muchas funciones del orgánulo modificadas, lo que implica cambios[3] en el comportamiento del orgánulo.

En resumen, la siguiente parte a desarrollar será el estudio de asociación en el genoma mitocondrial y el estudio de la tendencia en células con cáncer. Para ello es necesario recopilar la información provista en los artículos de cáncer y mitocondrias realizados hasta el momento y tratar de replicar los experimentos construidos en este estudio.

[3]Matizar que hablamos de cambios no de procesos destructivos. El estudio de células cancerígenas está altamente relacionado con estudios de longevidad.

Capítulo 3

Análisis de coocurrencia de variantes

Este capítulo aborda interrogantes que surgieron a medida que se iba desarrollando el trabajo. Específicamente, comprobar si existe coocurrencia entre diferentes variaciones y si existe alguna manera de obtener variantes coocurrentes de manera automática y fiable a partir únicamente de información de secuenciación. Estudios *similares* se han realizado en el pasado permitiendo asociar de manera manual variantes a haplogrupos, definiendo así el concepto de *variante de haplogrupo* como una variación común a todos los individuos del mismo haplogrupo. Nuestro estudio va más allá dado que nosotros no buscamos un nivel de asociación tan estricto, es decir, no exigimos que la variante esté en todos los individuos de un mismo haplogrupos. De hecho, pese a que sí usamos el concepto de haplogrupo a modo de control y evaluación, nuestro propósito es tratar de obtener perfiles de variaciones y agrupar aquellas variaciones que presentan perfiles altamente similares. Para ello enfocamos el estudio de dos maneras diferentes:

- A través de medidas de información como son:
  - *Mutual Information*, MI
  - *Information value*, IV

  En ambos casos, entendiendo las variaciones como variables aleatorias iid, independientes e idénticamente distribuidas, y los individuos como observaciones de la misma.
- A través de técnicas de recomendación basadas en factorización matricial:
  - *Singular Value Decomposition* - SVD

La idea original es demostrar que el problema de análisis de coocurrencia, cálculo automático o sistema de evaluación automático de haplogrupos puede ser enfocado desde la perspectiva de los sistemas de recomendación. Para ello se van a comparar el uso de medidas de información que actuarán a modo de control o solución intuita, frente a la aproximación basada en *SVD*. Los resultados experimentales obtenidos confirman que *MI* e *IV* son soluciones válidas y viables para el problema y además que *SVD* no solo es válida sino que además mejora a las otras dos opciones.

El capítulo sigue una estructura similar a los anteriores a través de una descripción de cada una de las técnicas, el porqué de su empleo y además incluye una resolución, similar a lo que se hace en el código, pero a pequeña escala de lo que se haría en una matriz de variantes. Finalmente, se ofrecen unas conclusiones y un trabajo futuro, además en este caso se hace un pequeño inciso en la importancia a nivel biológico del correcto funcionamiento de la aproximación.

El conjunto de datos a emplear se estructura en formato de una matriz de tal forma que incluye en sus filas los individuos secuenciados, en las columnas las variantes y el valor de cada una de las celdas

(i, j) determina si el usuario i contiene la variante j. La matriz que empleamos en las resoluciones posteriores es la siguiente:

$$C = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \tag{3.1}$$

de donde obviamente interpretamos:

- Individuo 1 - sin variantes.
- Individuo 2 - variantes 2, 3 y 4.
- Individuo 3 - variantes 1, 2 y 3.

3.1. Mutual Information y Information Value

Ambas medidas de información, *MI* e *IV*, son ampliamente usadas en el campo del *machine learning* y la inteligencia artificial como métricas para la selección de características. En ambos casos, se trata de métricas que permiten capturar la cantidad de información obtenida sobre una variable aleatoria X a través de la observación de otra Y. Por un lado, la *MI* es especialmente interesante en casos donde las variables explicativas, en nuestro ejemplo X e Y, toman valores discretos no necesariamente dicotómicos, mientras que la *IV* se usa cuando la variable dependiente es binaria. En nuestro caso, donde no hay variable respuesta y estamos realizando una estrategia *one-against-all*, todas las variables deben ser binarias pues todas actúan como dependientes e independientes. Es importante recalcar que se espera que valores altos supongan perfiles de variaciones similares. Aunque en otros contextos valores altos también pueden ser debidos a comportamientos radicalmente opuestos, no sería realista tener en cuenta esta situación en este caso.

3.1.1. Mutual Information

La información mutua de dos variables aleatorias es una medida de la dependencia mutua entre ambas variables. Más específicamente, cuantifica la cantidad de información obtenida sobre una variable aleatoria mediante la observación de la otra variable aleatoria. El concepto de información mutua está íntimamente ligado al de entropía de una variable aleatoria, una noción fundamental en la teoría de la información que cuantifica la cantidad de información esperada contenida en una variable aleatoria.

De manera formal se expresa:

$$I(X;Y) = \int_Y \int_X p_{XY}(x,y) log\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right) dxdy$$

En el contexto de variables discretas, que es el de nuestro interés:

$$(X;Y) = \sum_{y\in Y}\sum_{x\in X} p_{XY}(x,y) log\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right)$$

Ejemplo de aplicación

Para el ejemplo recuperamos la matriz C:

$$C = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \tag{3.2}$$

Matriz de información mutua:

$$M = \begin{pmatrix} 0.6365142 & 0.1744160 & 0.1744160 & 0.1744160 \\ 0.1744160 & 0.6365142 & 0.6365142 & 0.1744160 \\ 0.1744160 & 0.6365142 & 0.6365142 & 0.1744160 \\ 0.1744160 & 0.1744160 & 0.1744160 & 0.6365142 \end{pmatrix} \tag{3.3}$$

De la matriz podemos extraer el Top@1 como:

- La variante más similar a la variante 1 son la 2, 3 y 4.
- La variante más similar a la variante 2 es la 3.
- La variante más similar a la variante 3 es la 2.
- La variante más similar a la variante 4 son tanto la 1 como la 2 como la 3.

Los resultados, pese a no ser óptimos (la variante 4 no es tan parecida a la variante 1 como la 2 y la 3), si permiten observar que la técnica es sólida, puesto que los casos claros de similitud 2 y 3 son reportados como tal.

3.1.2. Information Value

El IV es una medida de información que permite cuantificar el poder predictivo de una variable independiente (continua o discreta) X para una variable dependiente Y binaria. En nuestro contexto particular para cada variación X vamos a calcular el IV que presenta para el resto de varianciones. De esta manera resulta obvio que aquellas variaciones con mayor nivel de IV serán las más susceptibles de presentar un comportamiento similar entre ellas.El índice IV se define como:

$$IV = \sum_{i \in \{0,1\}} (g_i - b_i) ln \frac{g_i}{b_i}$$

donde,

- b_i es $\frac{P(X=i|Y=0)}{P(Y=0)}$, proporción de casos donde, siendo Y igual a 0, X toma valor i.
- g_i es $\frac{P(X=i|Y=1)}{P(Y=1)}$, proporción de casos donde, siendo Y igual a 1, X toma valor i.

Ejemplo de aplicación

Desafortunadamente el ejemplo de aplicación no se puede realizar por la naturaleza de la métrica y por la librería usada en R. Por un lado, la librería requiere de mayor tamaño en la matriz para poder ejecutar el algoritmo de la manera apropiada. Por otro, debido a la naturaleza de la métrica con matrices de juguete es probable obtener valores de IV de 0 debido a la falta de combinaciones en la misma.

3.2. ¿Por qué SVD?

La descomposición en valores singulares, SVD, es una técnica de factorización matricial con múltiples aplicaciones en estadística y otras disciplinas, como la recuperación de información. A diferencia de muchos otros sistemas de factorización matricial, SVD no exige que se trate de una matriz cuadrada y tampoco tiene que ser simétrica. De igual manera a como ocurre en sistemas de recomendación, en un análisis de variantes son pocas las ocasiones donde un sistema funcional maneja el mismo número de filas y columnas, individuos y variantes, y la simetría realmente no es algo factible.

El porqué de usar SVD viene de la analogía del problema con los sistemas de recomendación. En un sistema de recomendación se tiene, generalmente, una matriz de unos y ceros donde las filas corresponden a los usuarios, las columnas a los *ítems* (por ejemplo, películas) y cada celda determina si un item ha gustado o no a un usuario. Adicionalmente a las características comentadas en el párrafo anterior es evidente que las matrices de recomendación van a ser sumamente dispersas pues rara vez un usuario va a proporcionar *feedback* para un número sustancial de *ítems*. Por otro lado, en el análisis de variantes podemos construir una matriz también de unos y ceros donde las filas son los individuos secuenciados, las columnas las variantes y cada celda determina si el individuo tiene o no la variación. Además, de igual manera a como ocurre en los sistemas de recomendación es esperable que estas matrices sean sumamente dispersas pues un individuo (vivo) rara vez va a agrupar múltiples variaciones simultáneamente.

Pese a las similitudes que presentan ambos problemas estos tienen objetivos completamente diferentes. Por un lado, los sistemas de recomendación tratan de ofrecer a un nuevo usuario (con unos *ítems* asociados) un abanico de *ítems* diferentes que han gustado a otros usuarios con un perfil parejo al suyo. Por otro lado, en un análisis de coocurrencia de variantes buscamos agrupar aquellas variaciones que presentan unos perfiles similares. Pese a todo, los resultados intermedios que se obtienen en las descomposiciones matriciales, es decir los factores latentes (en un sistema de recomendación los *géneros* de los *ítems* e idealmente en un análisis de variantes el *haplogrupo*), ofrecen la misma interpretación en ambas casuísticas.

3.3. Singular Value Decomposition

Teorema: Dada una matriz de dimensiones M x N y con rango r. Existe una descomposición en valores singulares de la forma:

$$C = U\Sigma V^T$$

donde:

- U es una matriz $M \times M$ cuyas columnas son los autovectores ortogonales de CC^T.
- V es una matriz $N \times N$ cuyas columnas son los autovectores ortogonales de C^TC.
- Σ es una matriz M x N en la que:
  - $\Sigma_{ii} = \sigma_i$ para $1 \leq i \leq r$ y 0 en cualquier otro caso con $1 \leq i \leq r$,
  - $\sigma_i = \sqrt{\lambda_i}$ con $\lambda_i \geq \lambda_i + 1$, donde $\lambda_1, ..., \lambda_r$ son los autovalores de CC^T o C^TC.
  - $\sigma_{ij} = 0$ para $i \neq j$.

3.4. SVD en el análisis de variaciones

Los sistemas basados en modelos, más en concreto los modelos basados en factorización matricial buscan obtener información que relacione a individuos secuenciados y a variantes. A partir de dicha información se quieren encontrar los factores latentes con los que poder capturar perfiles similares. En este tipo de sistemas partiremos de una matriz C, $M \times N$, de $Individuos \times Variantes$ donde:

- Cada fila representa a un individuo secuenciado.
- Cada columna representa a una variante.
- El valor de cada celda C_{ij} determina si el individuo i tiene la variación j.

La matriz C guarda información acerca de individuos y variantes. Pero la información usada es la contenida en $cct = CC^T$ y $ctc = C^TC$:

- cct, matriz MxM muestra en cada celda cct_{ij} el número de variantes comunes entre los individuos i y j.
- ctc, matriz NxN muestra en cada celda ctc_{ij} el número de individuos que comparten la variante i y j.

Los autovectores ortogonales de la matriz C^TC, vectores que forman la matriz V, en SR representan los factores latentes de variantes. Por tanto, la fila i de la matriz V representa la información latente para el item i.

3.4.1. Ejemplo de aplicación

A continuación recuperamos la matriz originalmente planteada y resolvemos un pequeño ejemplo.

$$C = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \tag{3.4}$$

A partir de aquí podemos obtener las versiones cct y ctc:

$$cct = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 3 & 2 \\ 0 & 2 & 3 \end{pmatrix} \qquad ctc = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 2 & 2 & 1 \\ 1 & 2 & 2 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

Que nuevamente podemos interpretar:

- cct:
  - Diagonal indica el número de variantes de cada individuo
  - Por ejemplo, $cct_{2,3} = 2$ esto implica que el individuo 2 y el individuo 3 comparten 2 variantes. Hecho correcto y se trata de las variantes 2 y 3.
- ctc:
  - Diagonal indica el número de individuos que poseen una variante.
  - Por ejemplo, $ctc_{2,3} = 2$ esto indica que la variante 2 y la variante 3 coocurren en 2 individuos. Que nuevamente es correcto y se trata de los individuos 2 y 3.

En nuestro caso solo nos interesa la matriz V de variantes:

$$V = \begin{pmatrix} -0.7071068 & -1.414214 & -1.414214 & -0.7071068 \\ 0.7071068 & 7.021667e-17 & -4.080563e-17 & -0.7071068 \end{pmatrix}$$

A partir del rango de esta matriz, en este caso dos, obtenemos dos factores latentes y con ello cuatro vectores con las características de cada variación.

A partir de aquí podemos calcular la proximidad de las variaciones a partir de la matriz de la raíz de las diferencias cuadráticas medias entre los vectores de factores latentes por variante:

$$RMSE(x, y) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}}$$

Por tanto nos quedaría:

$$V = \begin{pmatrix} 0.0000000 & 0.7071068 & 0.7071068 & 1.0000000 \\ 0.7071068 & 0.0000000 & 7.850462e-17 & 0.7071068 \\ 0.5773503 & 7.850462e-17 & 0.0000000 & 1.0000000 \\ 1.0000000 & 0.7071068 & 0.7071068 & 0.0000000 \end{pmatrix}$$

De donde podemos observar que bajo un criterio *Top@1*:

- La variante más similar a la variante 1 son las variantes 2 y 3.
- La variante más similar a la variante 2 es la 3.
- La variante más similar a la variante 3 es la 2.
- La variante más similar a la variante 4 son la 2 y la 3 por igual, cosa de nuevo lógico por que la variante 4 no la tiene nadie y la 2 y la 1 están menos presentes que la 3.

Pese a que se trata de un ejemplo de juguete ya se ve la primera limitación, heredada de los sistemas de recomendación, podemos dar un *Top* pero eso no indica que exista similitud real. Las variantes 2 y 3 en nada se parecen a la 4 pero por ser las que menos error presentan podrían ser reportadas como similares.

3.5. SVD truncado

Las matrices de variantes suelen ser grandes y por tanto los modelos son costosos. Cuando el sistema maneja miles de individuos y decenas de miles de variaciones, el proceso de descomposición matricial puede llegar a ser excesivamente pesado. Para evitar estos problemas existen las aproximaciones de reducción de rango o *Low-rank approximation*.

Las aproximaciones de reducción de rango calculan un valor k que permita mantener la información con la mínima pérdida y que a su vez permitan mejorar el sistema en términos de eficiencia. Por tanto, se busca construir una matriz Ck que minimice la diferencia de información con C, siendo k menor que $r = rank(C)$.Para obtener el valor de k se minimiza la norma de Frobenius para la diferencia de las matrices C y Ck:

$$argmin_k \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{k} C_{ij} - Ck_{ij}}$$

Para obtener la matriz Ck PureSVD hace uso de la matriz Σ. Dicha matriz es la matriz diagonal de autovalores de C^TC ordenados de mayor a menor. Dado que están ordenados de mayor a menor aquellos de menor valor se consideran menos informativos y se eliminan obteniendo una matriz Σ_k. En este punto se recomputa Ck como la aproximación de $rango_k$ de C.

$$C_k = U_k \Sigma_k V_k^T$$

Las dimensiones de las matrices cambiarían a:

- U_k, $M \times k$
- Σ, $k \times k$
- V_k, $N \times k$

donde tanto Uk y Vk contienen la información de U y V menos la información contenida en las M-k últimas columnas y N-k últimas columnas respectivamente. Una de las cosas más llamativas de la aplicación de técnicas de low-rank es que no solo las recomendaciones pueden no bajar en exceso su calidad sino que con el valor adecuado de k estas pueden hasta mejorar. De hecho se demostró que en ciertos casos los valores adecuados de k permitían obtener mejoras en la precisión de los sistemas.

En nuestro caso particular hemos definido el valor de k de una forma más relajada. Pese a que calculamos igualmente el valor de k óptimo eliminando los autovalores menos informativos, nosotros podemos determinar k a partir del conocimiento biológico del dominio. En nuestro caso nos interesa:

$$k' = min(k, F) \tag{3.5}$$

Donde F es el menor número de variantes más explicativas, entendiendo como más explicativa aquella que mayor prevalencia tenga en el grupo de individuos secuenciados, que permite obtener al menos una variación no asociada a haplogrupos. A través de esta matriz C $M \times k'$ seremos capaces de:

- Confirmar el correcto funcionamiento de la técnica respondiendo a la pregunta ¿es la asociación de variantes coherente con los haplogrupos descritos en la bibliografía?
- Si la primera parte es confirmada podremos suscitar estudios biológicos acerca de si las asociaciones sugeridas son o no correctas y en caso de serlo porqué no fueron previamente, existiendo así la posibilidad de haber definido el primer método automático de asociación de variantes a haplogrupos.

Como trabajo futuro quedaría extender el estudio a una arquitectura cluster, debido a que el propio cálculo del rango de la matriz ya es un auténtico reto a nivel computacional, y analizar el conjunto completo de variantes[1].

3.6. Criterios de Evaluación

En un sistema basado en *recomendaciones* donde no tratamos de predecir un valor esperado, regresión, o una clase en particular, clasificación, las metodologías de evaluación suelen denominarse *Top-N Recommendation*[2]. Las *Top-N recommendation* son tareas de recomendación que funcionan generando rankings de *ítems* en principio relevantes para un usuario. El uso de rankings es útil cuando la tarea de recomendación consiste en mostrar un número N de *ítems* de toda la colección. Este conjunto de N elementos de la colección se encuentran ordenados en base a su relevancia.

[1] En este hilo notar que $k' = F$, al menos en con los datos existentes en la actualidad.

[2] Existe también el concepto de *rating prediction* que no solo busca obtener recomendaciones sino que trata de estimar el valor de rating para un item en particular, en este caso donde trabajamos sobre valores 0/1 esto no interesa.

Este tipo de sistemas usan métricas heredadas de *Information Retrieval*, IR, que determinan la relevancia o no-relevancia de los *ítems* recomendados. Se usan métricas como MAP[3], nDCG[4], precisión o recall, entre otras.

3.6.1. Métricas

Existen múltiples métricas recogidas de la IR. Aquí únicamente se tratarán aquellas que presentan cierta utilidad para el problema abordado.

- *Precision at N* o precisión a un nivel de corte suele ser representado normalmente como

$$P@N = \frac{rel^N}{N}$$

  donde:

  - N representa el número de *ítems* en el top.
  - rel^N representa el número de *ítems* relevantes que se encuentran en el top N.

  . Obviamente, valores bajos de precisión representan que dentro de los N *ítems* recomendados solamente una pequeña fracción son relevantes para el usuario.

- *Recall at N* o recall a un nivel de corte representado normalmente como

$$R@N = \frac{rel^N}{\#rel_u}$$

  donde:

  - N representa el número de *ítems* en el top.
  - rel^N representa el número de *ítems* relevantes que se encuentran en el top N.
  - $\#rel_u$ representa el número de relevantes del usuario u.

  El recall permite determinar que tanto por ciento de *ítems* relevantes se han encontrado en el top-N para el usuario u, es decir la sensibilidad.

- MAP, representa el valor medio de las precisiones de las recomendaciones realizadas a cada uno de los usuarios. Para poder explicar esta métrica de manera adecuada es necesario explicar el AP[5].

  - AP, representa la puntuación media de precisión para cada usuario.

$$AP_u = \frac{\sum_{i \in rel_u} P@irel_u[i]}{\#rel_u}$$

    donde,

    - $P@i$ representa la *precision at i*.
    - $rel_u[i]$ representa si el item i es relevante para el usuario u.
    - $\#rel_u$ representa el número de *ítems* relevantes para el usuario u.

[3]Mean Average Precision
[4]Normalize Discounted Cumulative Gain
[5]Average Precision

El MAP promedia los valores de AP de todos los usuarios que se estén evaluando:

$$MAP = \frac{\sum_{u \in U} AP_u}{\#U}$$

donde,

- U representa el conjunto de usuarios.
- $\#U$ representa el número de usuarios total.

3.6.2. Evaluación

Los sistemas de recomendación generalmente son evaluados con una *N-Fold Cross Validation* donde se parte el dataset en N datasets de igual tamaño al original y posteriormente cada dataset es dividido en entrenamiento y test. Habrá de asegurarse que todos los elementos del dataset original se han evaluado y que los conjuntos de test de los distintos datasets son disjuntos. Por lo tanto, es evidente ver que el conjunto de test ha de ser de un tamaño $\frac{100}{N}$ %. Para obtener si *ítem i* de test es o no relevante para el usuario u se realiza el producto interior de sus vectores latentes $\hat{r}_{ui} = U_u V_i^T$, lógicamente, r_{ui} se encuentra fuera del conjunto de entrenamiento.

En nuestro caso el proceso pese a ser similar tiene ciertos matices diferenciadores. Por un lado, nuestro experimento no trata de asociar nuevas variaciones a individuos sino agrupar por similitud de perfiles latentes las variaciones. Por tanto, es evidente que el *ranking* que nosotros generaremos será un Top de variantes con perfiles latentes parejos y que esto no se va a poder evaluar con la matriz de partida. De hecho nuestra evaluación necesita conocer si las asociaciones sugeridas son correctas en el mundo real, y esta información se encuentra en la información de variantes asociadas a haplogrupos. Esto por un lado simplifica la evaluación eliminando la necesidad de particionar los datos, pero por otro complica el problema pues una variante puede pertenecer a múltiples haplogrupos y por tanto es necesario comparar los perfiles de haplogrupos de las variantes recomendadas.

El último detalle comentado obliga a abstraer ligeramente el concepto de *relevancia*. En un sistema de recomendación clásico la relevancia de los *ítems* sugeridos se contraste a través del valor que cada uno de ellos presenta en la muestra de test para el usuario bajo evaluación. Sin embargo, como ya se ha comentado anteriormente, no es aplicable en nuestro caso:

- Una variación puede pertenecer a distintos haplogrupos.
- Cada variación presenta un perfil de haplogrupos potencialmente diferente.

Por lo tanto, se necesita una métrica que permita dar valores cercanos a 1 cuando los perfiles sean parejos y que otorgue valores cercanos a 0 en casos donde sean altamente dispares. La opción escogida en nuestro caso ha sido el Índice de Jaccard

$$J(A, B) = \frac{\#(A \cap B)}{\#(A \cup B)}$$

donde dadas dos variantes i y j, $J(i, j) = 1$, si $haplogrupos(i) = haplogrupos(j)$, y $J(i, j) < 0$ si $haplogrupos(i) \neq haplogrupos(j)$. Además, el valor será tanto más cercano a 0 a medida que sus vectores de haplogrupos difieran.

Con el *recall* tenemos problemas similares a la precisión. El concepto de *variantes relevantes* es desconocido y por tanto no existe ninguna noción real de que variantes son relevantes entre ellas, impidiendo así poder determinar el top@20, top@15, top@10 y top@5 de variantes relevantes. Además, estamos en un sistema dicotómico y no basado en *ratings* por lo que ordenar estos *top* no es tarea sencilla. Para establecer una medida objetiva de la relevancia de las variantes decidimos proceder como sigue. Primero definimos una función evaluadora que computa para cada variante el índice de Jaccard contra el resto de variantes sobre la base de la información de asociaciones dado el ground truth. Entonces ordenamos los valores de Jaccard y calculamos los umbrales de corte en las posiciones 5, 10, 15 y 20. En otros términos, seleccionamos los índices de Jaccard que determinan los top 5, 10, 15 y 20, respectivamente.

3.6.3. Limitaciones de la evaluación

Los sistemas de recomendación presentan un cierta complejidad en la evaluación debido a la naturaleza de las técnicas. A diferencia de un sistema de aprendizaje automático o un sistema inteligente clásico donde o se clasifica, por lo que o aciertas o fallas la clase, o se predice un valor, donde puedes estar más cerca o más lejos, en este tipo de sistema se ofrece un número de *recomendaciones* que supuestamente son de interés para los usuarios. Desafortunadamente la evaluación de estas sugerencias depende de la información contenida, pero excluída en el entrenamiento, en el conjunto de datos de partida. Por lo tanto, sugerir un *ítem* y que este no aparezca como relevante al usuario puede deberse a: (i) que el usuario no conocía el *ítem* y por tanto no lo había valorado; (ii) que el usuario conoce el *ítem* pero simplemente no hace valoraciones, caso muy típico; o (iii) que al usuario efectivamente no le gusta el *ítem*. Por lo tanto, a efectos de evaluación cuando $rel_u[i] = 0$ implica que el *ítem* no es recomendable para el usuario u. Esto es obviamente impreciso dado que el usuario u no ha tenido porqué valorar todos los *ítems* que le han gustado o simplemente no tiene porqué estar al tanto de todos. Por lo que es sumamente esperable que este tipo de sistemas presenten ratios de precisión sustancialmente bajos, especialmente a medida que se incrementa el corte, frente a modelos de aprendizaje de otras disciplinas. De igual manera en nuestro estudio pueden pasar cosas similares:

- Una variación puede no estar asociadas a ningún haplogrupo. De hecho este es uno de los puntos clave del estudio, tratar de determinar si existen variaciones no asociadas con alto nivel de prevalencia entre los individuos secuenciados.
- Dos variaciones pueden presentar perfiles de haplogrupos diferentes pero a su vez ser recomendadas. Esto implica que ambas variaciones son similares a nivel de individuos secuenciados, pero que curiosamente no se encuentran asociados a haplogrupos comunes. ¿Por qué?
- Si una variación es única de un haplogrupo, por defecto las metodologías de *Top N recommendation* van a otorgarle N recomendaciones independientemente de la magnitud de la estimación de la relevancia. Esto genera pérdidas en las métricas globales.

Estos puntos son elementos de interés que pueden resultar sumamente interesantes en estudios posteriores. De hecho una vez confirmado el correcto funcionamiento de la aproximación habría que estudiar las variaciones sugeridas fuera de los haplogrupos asignados a través de estudios biológicos.

3.7. Resultados

En esta sección mostramos los resultados obtenidos para las tres métricas explicadas anteriormente:

- Precision@N
- Recall@N
- MAP

A mayores, para cada metodología se construye un *mapa de calor* que permite visualizar la distancia entre cada variante y el resto de variantes. El propósito de este mapa es descartar resultados triviales como que todas las variantes son extremadamente parecidas o simplemente no tienen nada en común.

Para la medida de los Top@ nos hemos basado en cuatro valores clásicos 5, 10, 15 y 25. En todo caso conviene precisa que en el problema que nos ocupa resulta complejo determinar los valores adecuados dado que, dependiendo de la variación, 5 pueden ser pocos, una variante perteneciente a un haplogrupo altamente secuenciado o compartida entre muchos haplogrupos, y 20 pueden ser excesivos, en casos donde una variante sea única por ejemplo.

A mayores se han establecido filtros crecientes para la selección del número de variantes bajo estudio. Como se introduce en la Sección 3.5 estos procesos son costosos computacionalmente y por tanto resulta conveniente en problemas reales la selección de un valor $k \leq rank(C)$ que permita

Filtro	Variantes Totales	% Variantes asociadas
20 %	18	100.0 %
10 %	25	100.0 %
5 %	50	100.0 %
2.5 %	134	83.6 %

Cuadro 3.1: Resumen de variantes y asociaciones por nivel de filtrado

simplificar los cálculos. Pese a ser posible calcular k a través de optimización, en nuestro caso podemos simplificarlo pues en primera instancia tratamos de demostrar el correcto funcionamiento del método y por tanto necesitamos de variaciones asociadas para poder contrastar los resultados obtenidos. Además, estas variaciones asociadas son por definición las más frecuentes dentro del conjunto de individuos y por lo tanto su obtención es simple y se basa en eliminar todas aquelas variaciones que no se encuentren en un porcentaje de los individuos secuenciados. En nuestro caso usamos los porcentajes: 20 %, 10 %, 5 % y 2.5 %[6]. El Cuadro 3.1 agrupa los resultados de variantes y asociaciones.

- Con un filtro del 20 % tenemos una matriz de 51433x18.
- Con un filtro del 10 % tenemos una matriz de 51433x25.
- Con un filtro del 5 % tenemos una matriz de 51433x50.
- Con un filtro del 2.5 % tenemos una matriz de 51433x134.

3.7.1. Métricas globales

Se muestran en primer lugar los mapas de calor obtenidos, de modo que a mayor intensidad del color corresponde un más alto grado de similitud. Obviamente, sin un conocimiento profundo del problema en cuestión estos gráficos no son más que gráficos de colores donde díficilmente se puede intuir si la metodología es o no adecuada. De todos modos se añaden por el interés biológico que pueden suscitar. Pues muchas de estas variaciones son populares entre los especialistas y su similitud es conocida. Sin embargo, parte de las asociaciones sugeridas por nuestras técnicas son novedosas por lo que a ojos de experto su cercanía en el mapa puede ser sorprendente e interesante. Precisamente por estos arguemntos, prestaremos especial atención en esta sección a medidas más cuantitativas que permitan confirmar el correcto funcionamiento de los tres sistemas.

Las Figuras 3.1, 3.2 y 3.3 muestran la asociación entre las variantes seleccionadas con un nivel de filtrado del 20 % sugerido por las tres técnicas. Claramente se aprecia como la tendencia es similar para las tres técnicas. Lógicamente, los valores de la diagonal son los más elevados pues una variante siempre es similar a si misma (esto no es así para el IV que para vectores idénticos devuelve un valor 0). Por otro lado, a medida que nos alejamos de la diagonal los valores van siendo más pequeños, destacando en este caso SVD cuya caída es más paulatina que por ejemplo MI donde la caída es mucho más abrupta.

Las Figuras 3.4 a 3.12 muestran los niveles de asociación para el resto de filtros con las tres técnicas. A ojos inexpertos interpretar las variantes y sus asociaciones es complicado pues es necesario un conocimiento profundo de la materia (por ejemplo a que gen pertenece cada una de las variantes, estos genes que codifican, hay interacción entre las proteínas que generan, etc). Sin embargo, de manera

[6]Porcentaje en el que aparece la primera variación no asociada a ningún haplogrupo

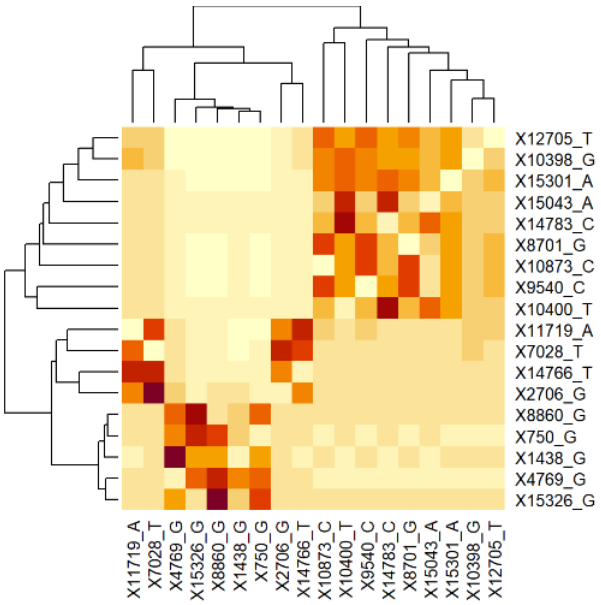

Figura 3.1: Heatmap Information Value filtro 20 %

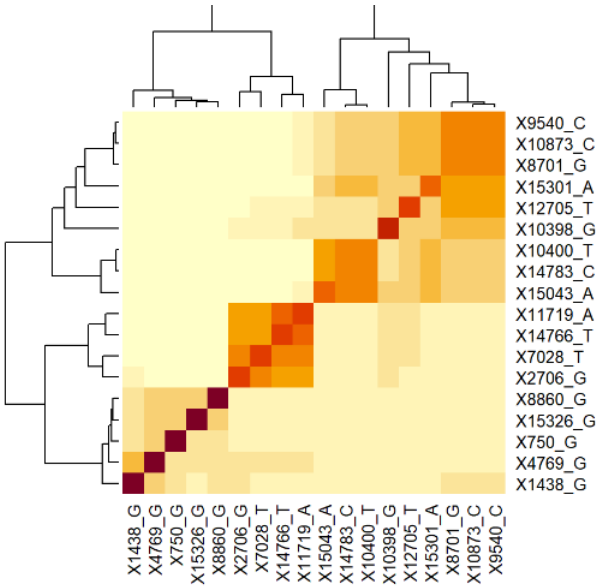

Figura 3.2: Heatmap Mutual information filtro 20 %

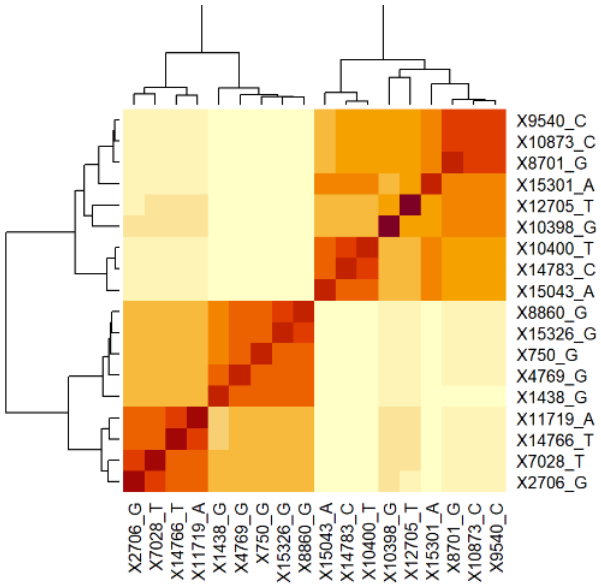

Figura 3.3: Heatmap SVD filtro 20 %

objetiva podemos ver como la tendencia de las tres es similar. Por un lado, el mayor nivel de similitud es con uno mismo de ahí que exista una franja oscura en las diagonales tanto para MI y SVD (matizar que en el caso de IV la IV con uno mismo es siempre 0). Además, la conducta general, con algunas excepciones, es de reducir el nivel de intensidad, es decir la similitud, a medida que nos separamos de la diagonal los niveles de similitud se reducen. Por otro lado, otro detalle interesante es que mientras que IV y MI tienden a sobredimensionar la pérdida de similitud, esto puede verse en el contraste entre la diagonal y el resto de valores, mientras que SVD parece que suaviza bastante más la pérdida de similitud.

En la Tabla 3.2 se muestran los valores de *precisión* y *recall* para las distintas técnicas, con los distintos niveles de filtrado y cortes. De entrada se observa como a medida que el corte se incrementa la precisión baja de manera notable. Esto es consecuencia de las limitaciones ya comentadas en la Sección 3.6.3 para cada tipo de técnica evaluada. Antes de analizar en profundidad los resultados procede destacar que se han empleado las medianas y no las medias pues en este tipo de problemas son frecuentes casos con valores de 0 sin realmente demasiada importancia. Un ejemplo claro de este hecho serían las variantes únicas para las que se hacen recomendaciones. Estas recibirán el mismo número de recomendaciones que el resto, dependiendo del corte empleado, pero dado que son únicas para un haplogrupo su perfil de haplogrupo va a ser totalmente dispar al resto.Por tanto, sus *índices de Jaccard* van a ser bajos pero como consecuencia de la necesidad de recomendación.

De los resultados de la Tabla 3.2 destaca *IV* que se presenta como la técnica con los niveles más bajos de *precisión* y *recall*. Por otro lado, SVD y MI parecen sorprendentemente similares a nivel global siendo capaces ambas de capturar las variantes más similares (a cortes bajos niveles elevados de precisión) Incluso a niveles más altos de corte presentan también niveles de precisión aceptables lo cual es relevante dado que a estos niveles altos de corte es sumamente probable que se estén infraestimando las precisiones debido a un claro exceso de recomendaciones[7]. De igual manera ocurre con la métrica

[7]Esto es un problema clásico en SR y es consecuencia nuevamente de tratar de recomendar *ítems* bajo un top fijo. Perfiles extraños o gente con escasas valoraciones es frecuente que reciban recomendaciones de elementos que no se ajustan a sus criterios o gustos. En el estudio de variantes esto es incluso más factible pues variantes únicas van a recibir el mismo número de recomendaciones que variantes sumamente compartidas entre haplogrupos. Sin embargo, en el caso primer caso estas recomendaciones nunca podrán ser correctas mientras que en el segundo caso la bondad de estas dependerá de la técnica empleada.

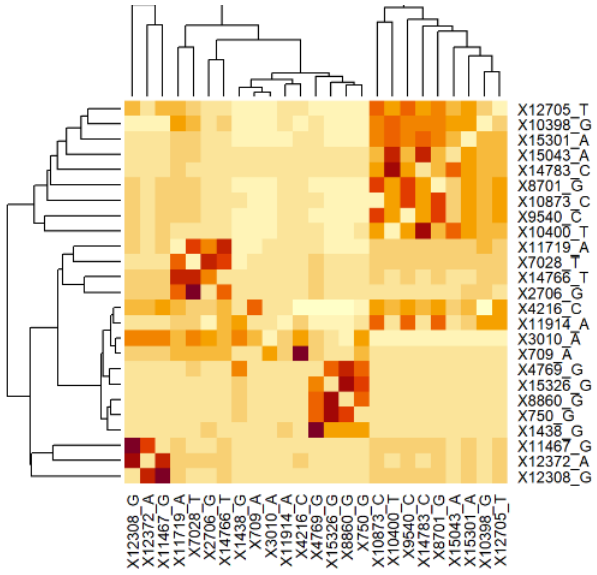

Figura 3.4: Heatmap Information Value filtro 10 %

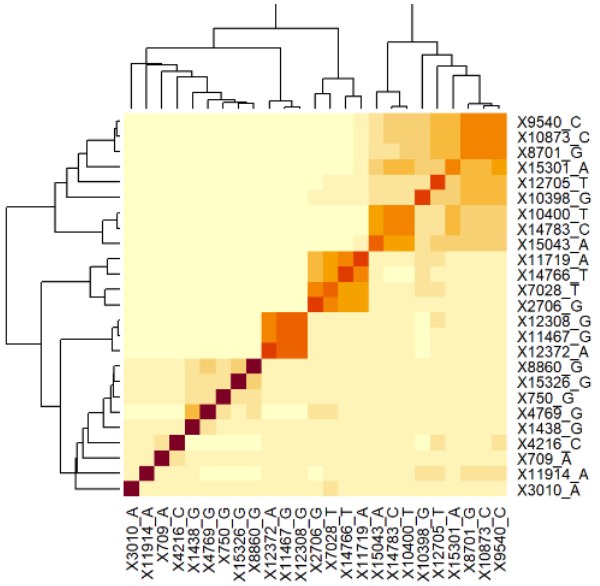

Figura 3.5: Heatmap Mutual information filtro 10 %

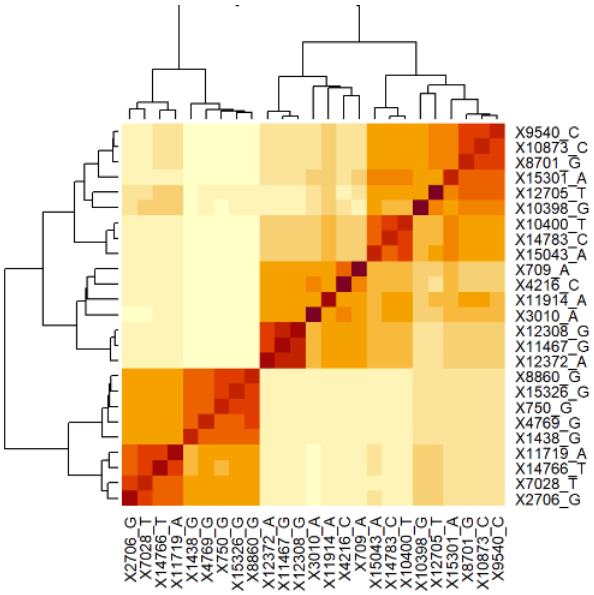

Figura 3.6: Heatmap SVD filtro 10 %

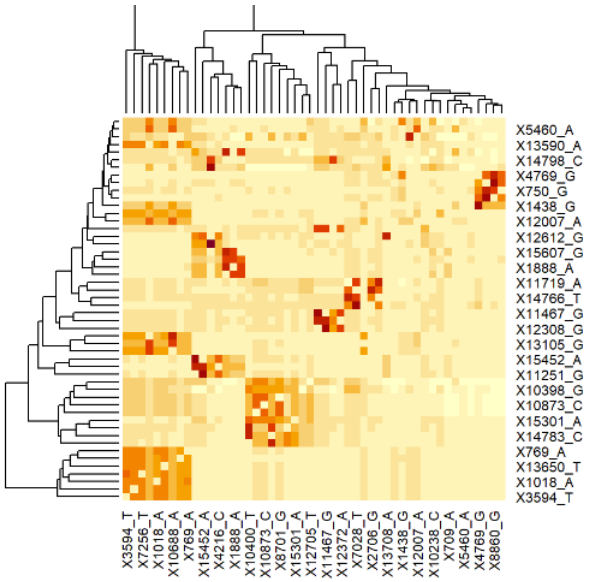

Figura 3.7: Heatmap Information Value filtro 5 %

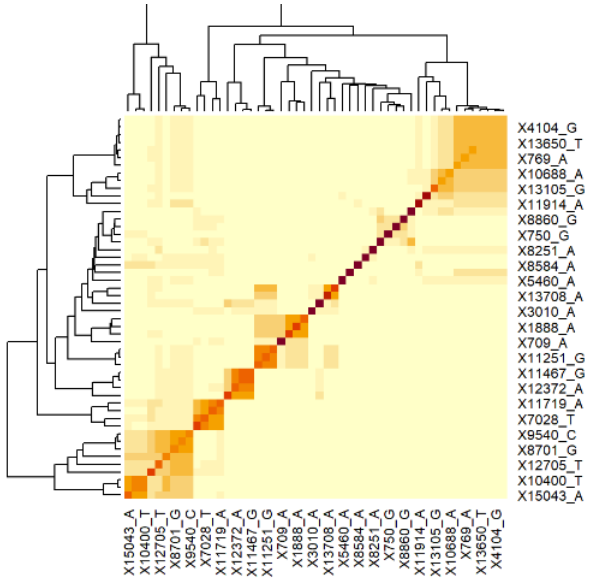

Figura 3.8: Heatmap Mutual information filtro 5 %

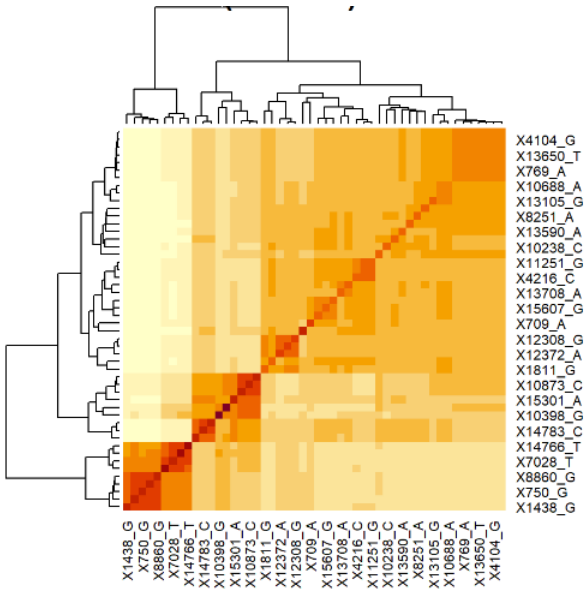

Figura 3.9: Heatmap SVD filtro 5 %

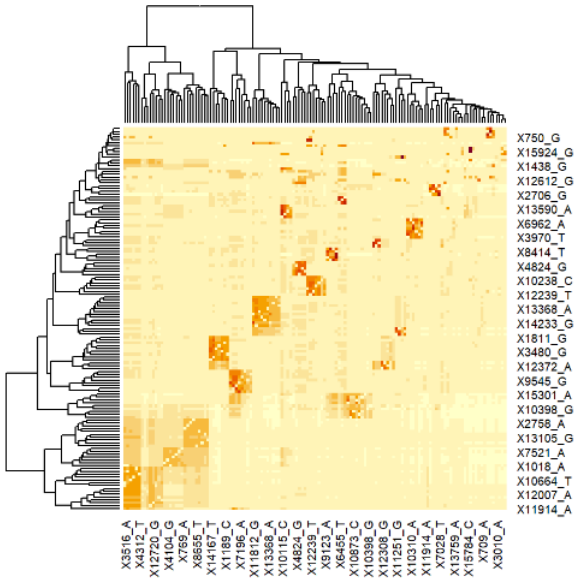

Figura 3.10: Heatmap Information Value filtro 2.5 %

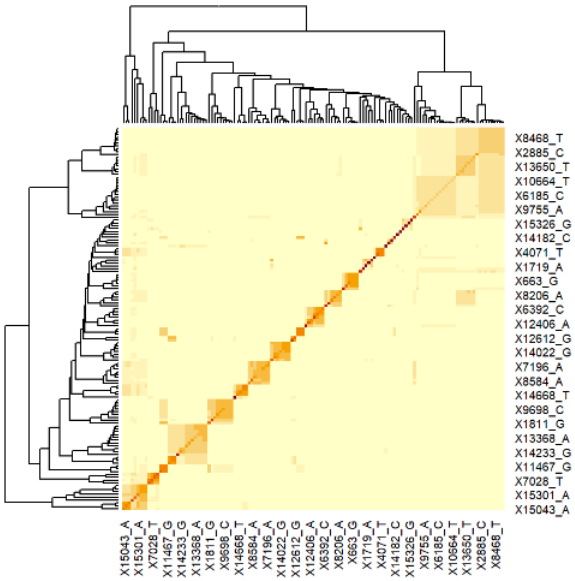

Figura 3.11: Heatmap Mutual information filtro 2.5 %

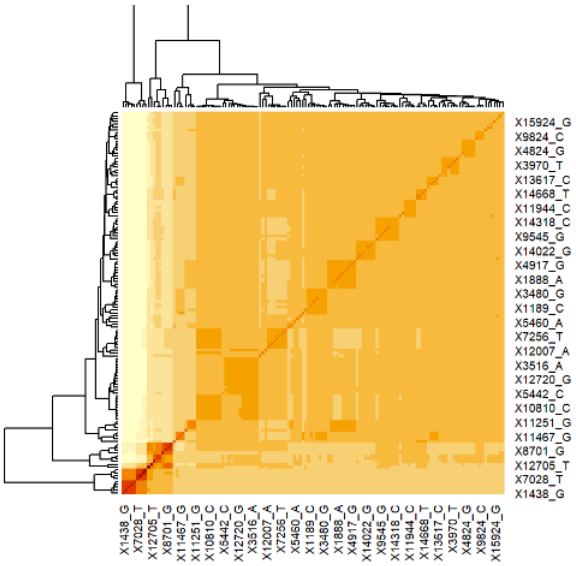

Figura 3.12: Heatmap SVD filtro 2.5 %

de *recall* aunque en este caso somos más escépticos dado que el criterio de relevancia es autoasignado y sería necesario tener una base real provista tras pruebas en laboratorio para poder asegurar estas conclusiones.

A la vista de los resultados parece bastante atinado afirmar que las tres técnicas tienen un buen comportamiento. Destacar especialmente a *SVD* y *MI* dado que ambas superan con creces a *IV* que sin dejar de ser adecuado no alcanza las cotas de bondad de sus dos contendientes. Aún así, cabe destacar que las medidas globales no son del todo fiables en este tipo de estudios dado que cada *ítem* tiene su propio comportamiento particular. Un *ítem* particular, como puede ser a título de ejemplo una película poco conocida o de género de autor, es probable que le guste a gente de perfiles variopintos y por tanto su perfil latente posiblemenmte sea dispar al del resto de *ítems* más conocidos e incluso a los desconocidos (dado que el perfil de gente que lo conoce y gusta puede variar mucho). Esto produce lógicamente casos extremos de valores de similitud bajos. Por otro lado, tenemos las superproducciones que en si ya seguramente sean un factor latente y todos los *ítems* de este estilo presentarán perfiles muy similares y de ahí niveles de similitud elevados. En definitiva, si tenemos un conjunto de datos con muchas superproducciones y pocas películas de autor es probable que nuestras métricas globales sean buenas pero el comportamiento a nivel individual sea malo. Por este último detalle es necesario hacer un análisis adicional y evaluar el comportamiento de las recomendaciones a nivel local o individual.

3.7.2. Métricas locales

En esta sección se desgranan los resultados globales vistos en el apartado anterior. Para cada métrica global obtenida se muestra un gráfico análogo que separa la métrica en sus valores individuales[8]. En esta sección la importancia radica en ver que los resultados no son extremos, es decir, que no tenemos un conjunto de valores cercanos a 1 y otros cercanos a 0 sistemáticamente sino que las medidas individuales se distribuyen de manera uniforme a lo largo del intervalo $[0, 1]$.

Las Figuras 3.13 a 3.16 muestran los resultados de precision y recall para todas las variantes a nivel individual y con un nivel de filtrado del 20 % (es decir para variantes que al menos están presentes en

[8] A mayores en los apéndices se incluyen las tablas con los valores exactos y las asociaciones entre variaciones sugeridas. Por problemas lógicos de espacio solo se incluyen las asociaciones para niveles de filtrado más altos pero el código está preparado para reportar en formato *csv* todas las asociaciones y niveles de similitud para todos los niveles deseados.

	Median Precision				Median Recall			
	20 %				20 %			
Técnica	@5	@10	@15	@20	@5	@10	@15	@20
SVD	0.90	0.78	0.66	0.63	0.7	0.75	0.73	0.75
IV	0.79	0.65	0.59	0.63	0.5	0.6	0.56	0.75
MI	0.86	0.68	0.61	0.63	0.6	0.7	0.63	0.75
	10 %				10 %			
Técnica	@5	@10	@15	@20	@5	@10	@15	@20
SVD	0.80	0.64	0.50	0.43	0.6	0.7	0.6	0.6
IV	0.72	0.53	0.44	0.41	0.4	0.5	0.46	0.55
MI	0.80	0.64	0.49	0.42	0.6	0.5	0.46	0.55
	5 %				5 %			
Técnica	@5	@10	@15	@20	@5	@10	@15	@20
SVD	0.80	0.59	0.42	0.33	0.4	0.5	0.46	0.45
IV	0.64	0.46	0.33	0.27	0.2	0.3	0.36	0.4
MI	0.79	0.58	0.41	0.33	0.4	0.35	0.4	0.45
	2.5 %				2.5 %			
Técnica	@5	@10	@15	@20	@5	@10	@15	@20
SVD	0.88	0.70	0.53	0.43	0.15	0.4	0.33	0.45
IV	0.78	0.57	0.41	0.34	0.06	0.3	0.26	0.4
MI	0.87	0.65	0.48	0.39	0.15	0.4	0.4	0.5

Cuadro 3.2: Resultados globales obtenidos con SVD, IV y MI para los distintos niveles de filtrado y corte

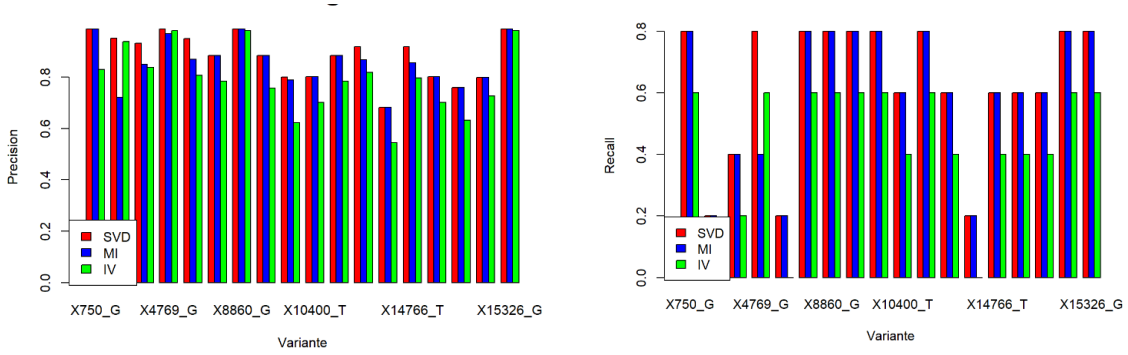

Figura 3.13: a) Precision@5 b) Recall@5 con nivel de filtrado del 20 %

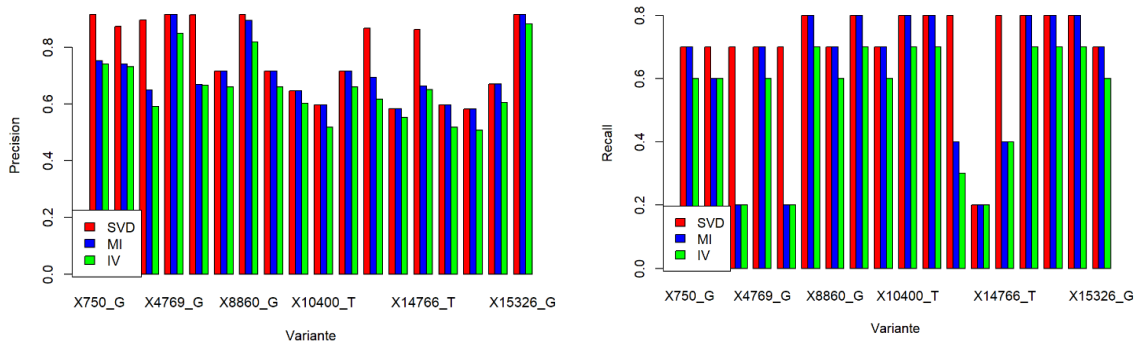

Figura 3.14: a) Precision@10 b) Recall@10 con nivel de filtrado del 20 %

el 20 % de los individuos). En estos gráficos se puede ver como todos los resultados son adecuados y no existen los casos extremos que se mencionaban anteriormente. El único hecho relevante a enfatizar es que a pesar de que los resultados con las medidas globales SVD y MI no diferían demasiado, en este caso puede verse una ligera superioridad sistemática por parte de *SVD*, con alguna diferencia sustancial en ocasiones. En Figura 3.16 que las tres métricas ofrecen exactamente los mismos valores, algo que también se muestra en el Cuadro 3.2. La explicación es sencilla: la matriz consta de 18 variantes y estamos recomendando 20 por tanto dado que las métricas empleadas no dependen del orden de sugerencia todos recomiendan todo y por tanto todos tienen los mismos resultados. Además, viendo estos resultados podemos romper una lanza a favor de las técnicas. La bondad de las técnicas se muestra al ver que en cortes inferiores o iguales 5 están obteniendo niveles de precisión cercanos al 90 % esto supone que 9 de cada 10 variantes sugeridas son correctas. Sin embargo, en la recomendación naive (recomendar todo lo disponible) en promedio únicamente 6 *ítems* son relevantes por variante. Aún así, las tres técnicas están por un lado limitadas a nivel de funcionamiento por la escasa información y además están sesgadas al alza porqué las variantes más comunes tiene sentido que coocurran.

Las Figuras 3.17 a 3.20 muestran los resultados de precision y recall para todas las variantes a nivel individual y con un nivel de filtrado del 10 %. Los resultados obtenidos son similares a los anteriores. De nuevo *SVD* se posiciona como el ligeramente superior en el cuadro general, aunque en este caso existen algunas situaciones donde *MI* lo supera ligeramente, con alguna diferencia notable en algún caso. En la parte derecha de los paneles a) estas figuras (Figuras 3.17, 3.18, 3.19 y 3.20) se ve el problema del exceso de recomendaciones. Existen variaciones que probablemente sean únicas y que por tanto no

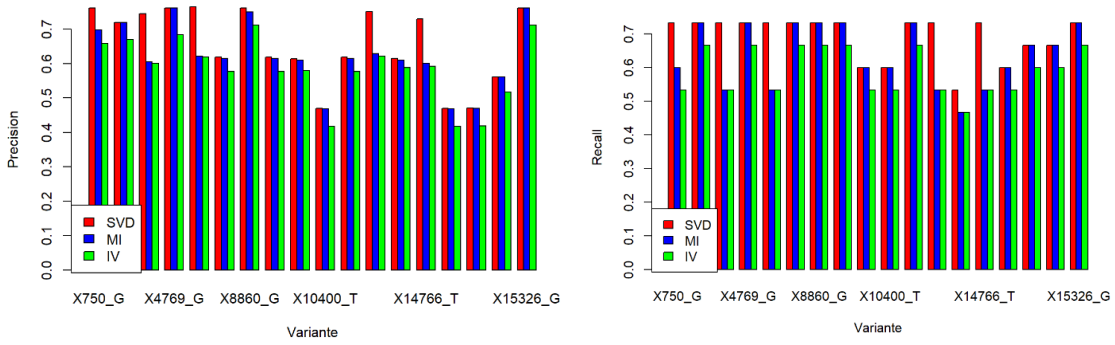

Figura 3.15: a) Precision@15 b) Recall@15 con nivel de filtrado del 20 %

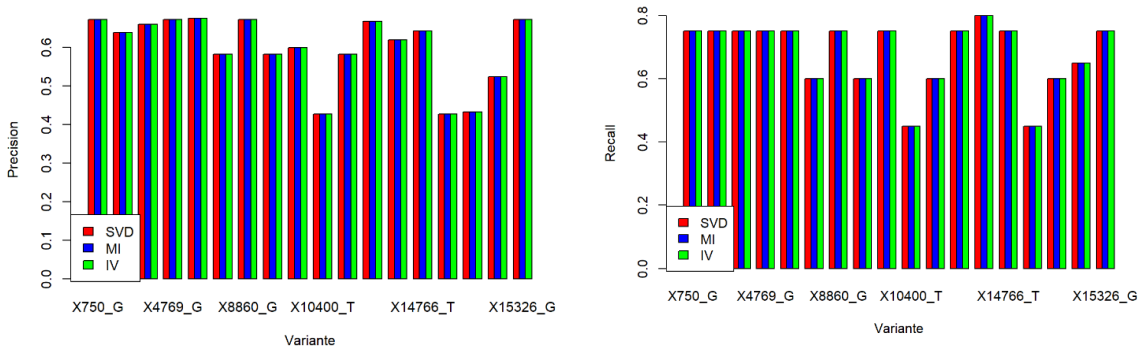

Figura 3.16: a) Precision@20 b) Recall@20 con nivel de filtrado del 20 %

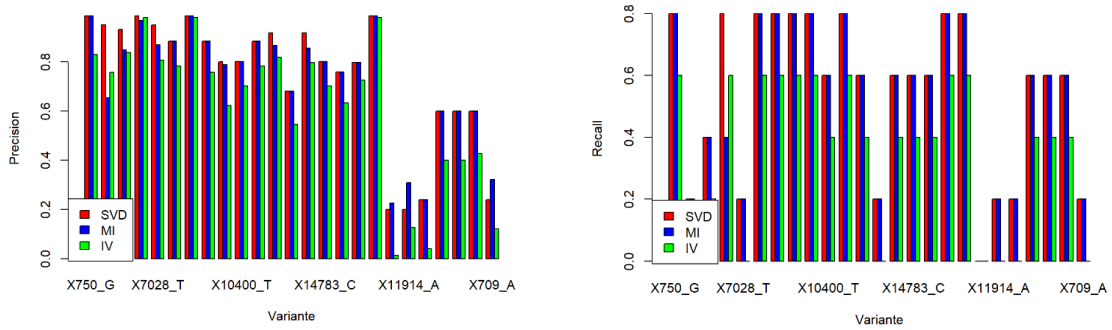

Figura 3.17: a) Precision@5 b) Recall@5 con nivel de filtrado del 10 %

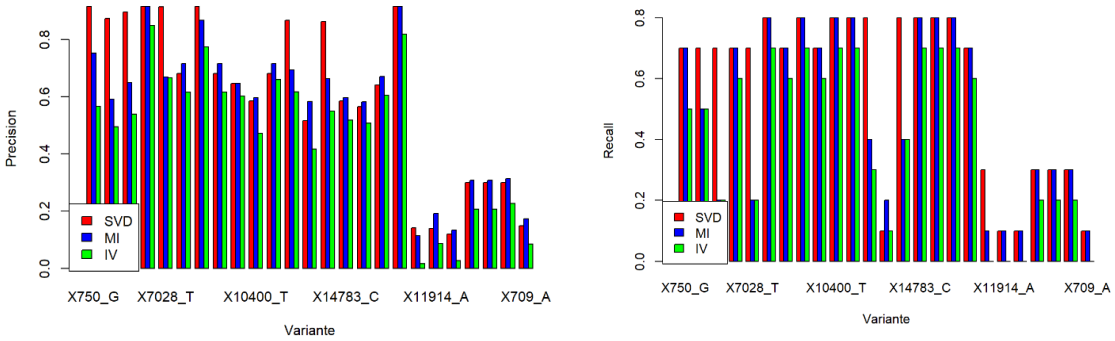

Figura 3.18: a) Precision@10 b) Recall@10 con nivel de filtrado del 10 %

presenten variaciones asociadas o presenten pocas, en el *ground truth* vigente actualmente. Esto se comprueba claramente al ver que dichas variantes con niveles bajos de precisión presentan sin embargo niveles altos de recall. Esto puede ser debido a dos situaciones:

- Carácter excesivamente relajado del recall adaptado.
- Escaso y acertado número de variantes recomendado, es decir, presenta pocas variantes sin embargo nosotros hemos recomendado 5, 10, 15 y 20 de ahí los bajos niveles de precisión. Pero resulta que en esas recomendaciones las variantes reales se encuentran y por tanto se tienen niveles altos de recall. Este hecho se ve acentuado en el caso del top@5, donde la precisión ya es baja pero sin embargo el recall es elevado. Dicho de manera coloquial *acertamos a la primera.*

Por último, los resultados de precisión y recall con un nivel de filtrado al 5 % se muestran en las Figuras 3.21 a 3.24. Nuevamente se puede ver que no se tiene una sistemática de casos extremadamente buenos y casos extremadamente malos. Por lo tanto, podemos concluir que los resultados globales de la Tabla 3.2 son adecuados y pueden interpretarse de la manera que se ha realizado.

3.8. Conclusiones

A la luz de los resultados alcanzados, tanto de carácter global como local, y como ya se ha venido adelantando en las dos secciones previas, cabe concluir que nuestra evaluación de niveles de asociación de variantes con los haplogrupos es alentadora. Queda tan solo pendiente el trabajo de laboratorio

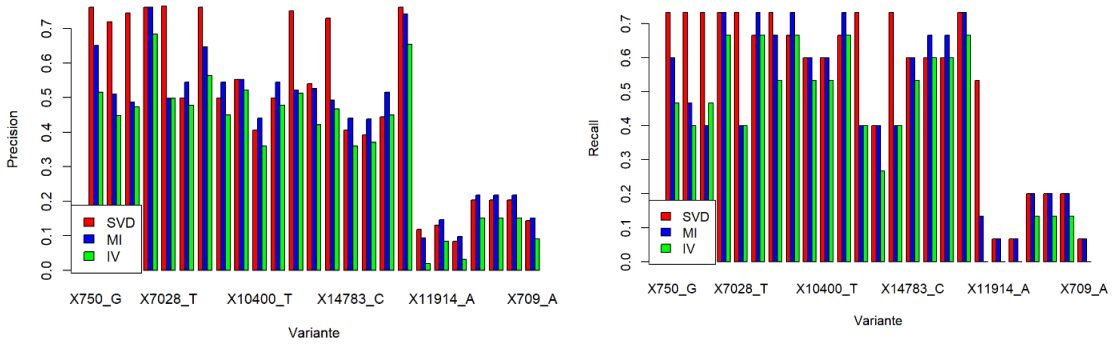

Figura 3.19: a) Precision@15 b) Recall@15 con nivel de filtrado del 10 %

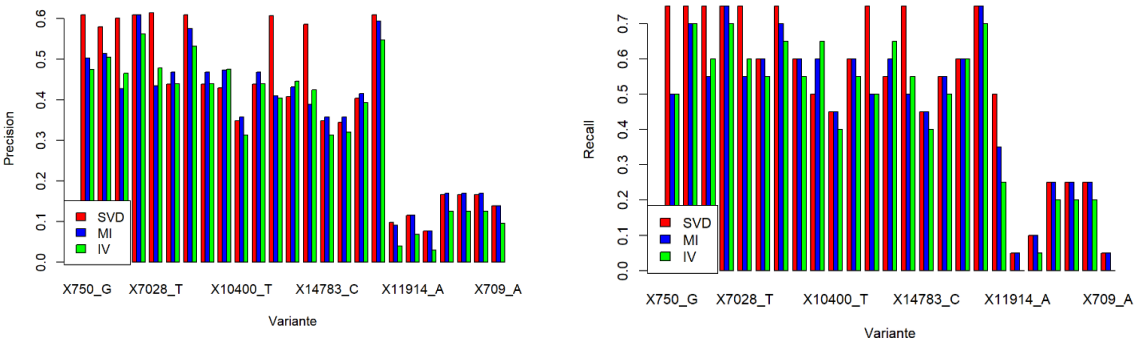

Figura 3.20: a) Precision@20 b) Recall@20 con nivel de filtrado del 10 %

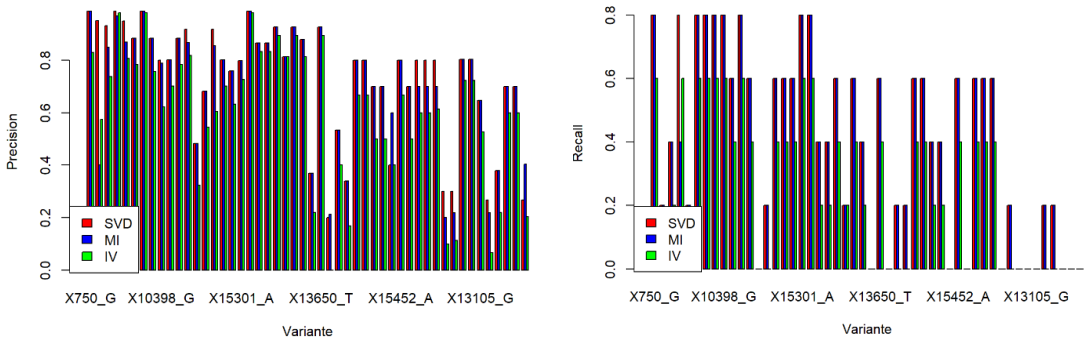

Figura 3.21: a) Precision@5 b) Recall@5 con nivel de filtrado del 5 %

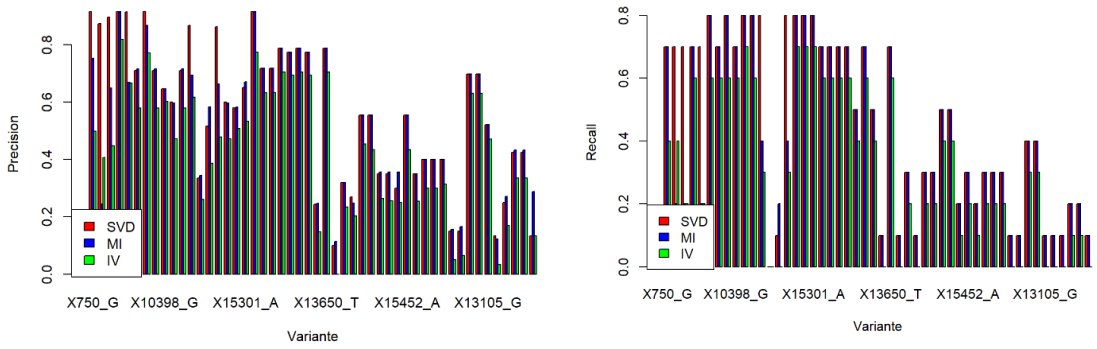

Figura 3.22: a) Precision@10 b) Recall@10 con nivel de filtrado del 5 %

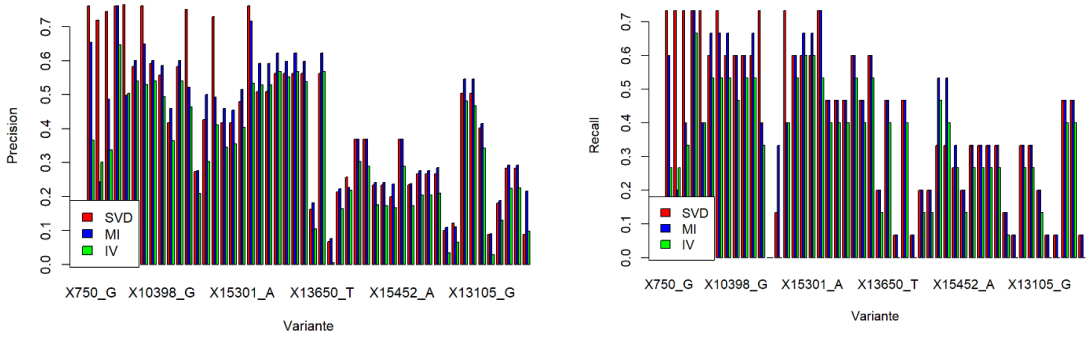

Figura 3.23: a) Precision@15 b) Recall@15 con nivel de filtrado del 5 %

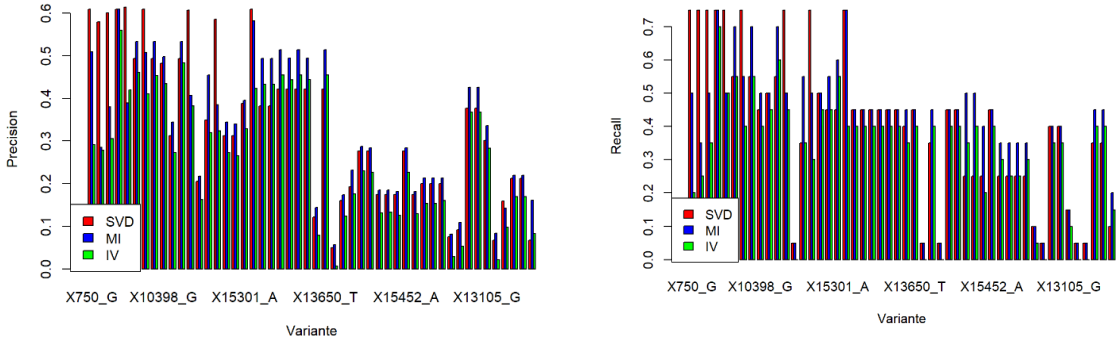

Figura 3.24: a) Precision@20 b) Recall@20 con nivel de filtrado del 5 %

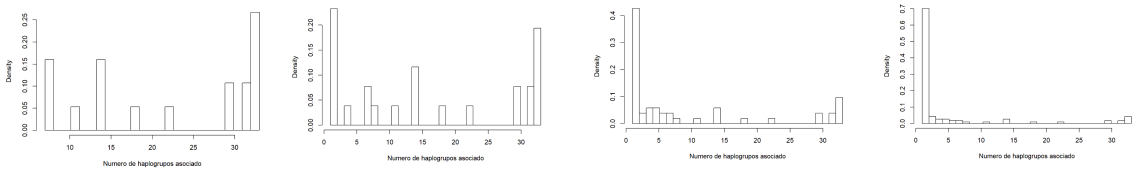

Figura 3.25: Distribuciones de las variantes asociadas a haplogrupos con filtros (a) 20 % (b) 10 % (c) 5 % y (d) 2.5 %

que permita confirmar si los casos con bajos niveles de precisión son debidos al exceso de exigencia de la metodología de asociación actual o si realmente se trata de casos de asociación incorrectos. Independientemente de este resultado, ha quedado demostrado que la selección de *MI* y *IV* fue correcta y adecuada. Además de que *SVD* no solo es competitiva en sus precisiones con las otras opciones sino que incluso las mejora en la mayoría de contextos.

Finalmente, quedan por comentar los valores de recall que hasta el momento se les ha dado poca importancia. Desafortunadamente del recall poco podemos decir salvo que la métrica ofrece valores relativamente malos sobre todo a medida que el número de variantes se incrementa. Esto es fácilmente explicable a través de los histogramas de los perfiles de haplogrupos mostrados en la Figura 3.25. Estos histogramas muestan que a medida que el filtro decrece, lógicamente más variantes entran en el problema. Lo que no es tan esperable es que el número de variantes con un solo haplogrupo asociado crezcan de semejante manera. Esto provoca que los valores del *Índice Jaccard* para cada par de variantes con un solo haplogrupo asociado vaya ser 1 o 0, dependiendo de si casualmente están asociadas al mismo o no, respectivamente. Por tanto los valores van a ser poco fiables provocando que los valores en los cortes 5, 10, 15 y 20 no sean significativos. Por último recalcar que esto no es defecto de la técnica sino defecto de como las variantes se han asociado a los haplogrupos[9]. Esta asociación se ha hecho de manera extremadamente restrictiva obligando a que todos los individuos de un haplogrupo deban contener la variante para poder ser asociada. De hecho, nuestras técnicas son una alternativa realista, automática y fiable, estando esta última propiedad pendiente de ser validada por estudios de expertos en el dominio, para la asociación y coocurrencia de variantes.

3.9. Alcance y trabajo futuro

Tras los resultados del análisis numérico y gráfico procede naturalmente valorar el alcance e interés que estas técnicas introducidas pueden tener en el ámbito biológico. Los estudios de asociación de variantes genómicas en el genoma humano es indudablemente tema de sumo interés especialmente en estudios filogenéticos o médicos. Desafortunadamente el tamaño del genoma humano, 3000 millones de pares de bases, hace inviable la adaptación de técnicas como la selección de características, descomposición de matrices u otras semejantes para hacer estudios de agrupación. De hecho, populares estudios de *Polimorfismos de nucleótido único* o *SNPs* por lo general asumen que las diferentes variaciones en una muestra de N individuos son independientes e idénticamente distribuidas y por tanto son estudiadas de manera independiente unas de otras.

Por otro lado, el genoma mitocondrial es el *hermano pequeño desatendido* y por tanto tiende a recibir lo ya aplicado y verificado en el genoma nuclear. Cabe destacar que en el estado del arte no hay referencia alguna que proporcione una aproximación automática de asociación de variantes en este tipo de genomas. Más aún, la persona responsable de la plataforma MitoMap realiza las asociaciones a haplogrupos de manera totalmente manual, siguiendo una técnica naive que supone asociar una variante a un haplogrupo si y solo si todos los individuos contienen dicha variación. En este contexto

[9]Nosotros nos comparamos contra la asociación provista por MitoMap y por el momento no existen estudios alternativos.

resulta evidente que el desarrollo de una herramienta fiable y precisa que permita automatizar este proceso es de enorme interés. Es evidente que el desarrollo de una herramienta fiable que permita automatizar este proceso es de interés. Especialmente ahora que el número de genomas mitocondriales ensamblados y reportados es cada vez más elevado. La duda que puede surgir en este punto es ¿por qué en el genoma mitocondrial sí y en el genoma nuclear no? La respuesta es clara y es el tamaño. Donde el genoma nuclear tiene 3000 millones de pares de bases el genoma mitocondrial presenta únicamente 30000. Lo que habilita la posibilidad de usar técnicas más complejas y sofisticadas que las existentes para el genoma nuclear.

En cuanto a las técnicas seleccionadas, el objetivo fue siempre usar era usar técnicas previamente consideradas en el ámbito biológico como son la información mútua [27, 28] o el information value. Ambas técnicas han sido usadas en el contexto del análisis de la expresión génica en muestras de micromatrices. En ambos casos el fundamento es construir una matriz con filas como un conjunto de individuos y columnas los niveles de expresión (discretos) de los distintos genes bajo evaluación. En la literatura técnicas basadas en estos criterios han sustituido a los tradicionales contrastes múltiples. Dado que nuestro problema no es más que una abstracción[10] de problema, ambas técnicas parecen sumamente adecuadas. Adicionalmente se propone una técnica innovadora en este contexto como es el empleo de SVD. En efecto, aplicamos MI y VI a este problema por vez primera pero es bien cierto que ambas técnicas ya han sido previamente usadas en contextos biológicos, al contrario que SVD que, hasta donde nosotros sabemos, no se ha usado nunca en problemas biológicos d eesta índole y es por tanto una aproximación totalmente original. Recordar en este sentido que en la Sección 3.2 se revisan las analogías entre los tres sistemas de recomendación y el problema tratado, permitiendo argumentar y justificar el interés en el uso de SVD.

Los resultados obtenidos evidencian que SVD se postula como una opción más que viable para la asociación de variantes, obviamente seguido de un proceso de exploración biológica. Además, se ha abierto la puerta a su uso en estudios de asociación simultánea de variantes en enfermedades como el cáncer, donde el propósito no sería asociar una variación a un cáncer como tradicionalmente se plantea, sino automatizar el proceso de asociar múltiples variaciones a los distintos tipos de cáncer. Por otro lado y de manera totalmente secundaria se abre la posibilidad de usar técnicas de recomendación para estudios de expresión genética donde *MI* e *IM* han probado ser de mucha utilidad en el pasado.

[10] cambiar los genes por variantes y los niveles de expresión por valores 1/0 que pueden representar gen expresado o no expresado

Apéndice A

Software

El trabajo desarrollado se encuentra depositado en un código en github `https://github.com/borjaf696/mtDNAProject` siendo posible replicar todos los análisis y resultados obtenidos en el proyecto.

A.1. Estructura

El repositorio se estructura en:

- Control - esta carpeta contiene la información necesaria básica para evaluar los resultados obtenidos en el análisis de variantes.
- Data - esta carpeta contiene la información necesaria para realizar los análisis de coocurrencia y de variantes
- output_r - salida de los resultados de r, necesarios para posteriores análisis biológicos
- output - salida del código de python donde se encuentra toda la información correctamente procesada
- r
  - analisis_rmd.Rmd - código del análisis de la subunidad grande.
  - analisis_rmd_SSU.Rmd - código del análisis de la subunidad pequeña.
  - coocurrencia_rmd.Rmd - código para el análisis de coocurrencia.
- src - carpeta con el código python necesario para la descarga y parseo de tanto las variantes y los haplogrupos de *MitoMap*. En este punto incidir en que no proveen de una plataforma de acceso y es necesaria la descarga página a página de las variantes y su consecuente parseo para la obtención de los resultados.

Apéndice A

Resultados extendidos

A continuación se muestran los resultados para un nivel de filtrado de 20 % y 10 % para los top5 y top10. El propósito de esto es observar las magnitudes de las similitudes para cada uno de los top.

	X1	X2	X3	X4	X5	Similitud X1	Similitud X2	Similitud X3	Similitud X4	Similitud X5
X750_G	X750_G	X1438_G	X4769_G	X8860_G	X15326_G	1	0.747734368886101	0.831346372915075	0.856305581597305	0.862506071269093
X1438_G	X750_G	X1438_G	X4769_G	X8860_G	X15326_G	0.747734368886101	1	0.776864314853309	0.750065390005824	0.749567610962614
X2706_G	X2706_G	X4769_G	X7028_T	X11719_A	X14766_T	1	0.495839126611222	0.827332854711486	0.723431609773309	0.719363880536587
X4769_G	X750_G	X1438_G	X4769_G	X8860_G	X15326_G	0.831346372915075	0.776864314853309	1	0.841943230878749	0.836064369245444
X7028_T	X2706_G	X4769_G	X7028_T	X11719_A	X14766_T	0.827332854711486	0.520527742024664	1	0.768267465338602	0.763427548739303
X8701_G	X8701_G	X9540_C	X10873_C	X12705_T	X15301_A	1	0.915451952635614	0.909339188184368	0.665026823734163	0.729486757935362
X8860_G	X750_G	X1438_G	X4769_G	X8860_G	X15326_G	0.856305581597305	0.750065390005824	0.841943230878749	1	0.876334995045535
X9540_C	X8701_G	X9540_C	X10873_C	X12705_T	X15301_A	0.915451952635614	1	0.932856817701363	0.667826871537777	0.735303926064407
X10398_G	X8701_G	X9540_C	X10398_G	X10873_C	X12705_T	0.597666771567988	0.602556176213775	1	0.60138614033957	0.541017133367401
X10400_T	X9540_C	X10400_T	X14783_C	X15043_A	X15301_A	0.609670051840734	1	0.945557743876789	0.818678376368708	0.692919005466179
X10873_C	X8701_G	X9540_C	X10873_C	X12705_T	X15301_A	0.909339188184368	0.932856817701363	1	0.668089414558443	0.733849411218247
X11719_A	X2706_G	X4769_G	X7028_T	X11719_A	X14766_T	0.723431609773309	0.476664736757296	0.768267465338602	1	0.841785725339774
X12705_T	X8701_G	X9540_C	X10873_C	X12705_T	X15301_A	0.665026823734163	0.667826871537777	0.668089414558443	1	0.58039406290625
X14766_T	X2706_G	X4769_G	X7028_T	X11719_A	X14766_T	0.719363880536587	0.470302123757717	0.763427548739303	0.841785725339774	1
X14783_C	X9540_C	X10400_T	X14783_C	X15043_A	X15301_A	0.60798271578668	0.945557743876789	1	0.816019259704134	0.690857629150182
X15043_A	X9540_C	X10400_T	X14783_C	X15043_A	X15301_A	0.578617067644538	0.818678376368708	0.816019259704134	1	0.654737841912674
X15301_A	X8701_G	X9540_C	X10400_T	X10873_C	X15301_A	0.729486757935362	0.735303926064407	0.692919005466179	0.73384941218247	1
X15326_G	X750_G	X1438_G	X4769_G	X8860_G	X15326_G	0.862506071269093	0.749567610962614	0.836064369245444	0.876334995045535	1

Cuadro A.1: Resultados de asociación para variantes corte 5, filtro 20 %

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Similitud X1	Similitud X2	Similitud X3	Similitud X4	Similitud X5	Similitud X6	Similitud X7	Similitud X8	Similitud X9	Similitud X10
X750_G	X1438_G	X2706_G	X4769_G	X7028_T	X8860_G	X10400_T	X11719_A	X14783_C	X15043_A	X15326_G	0.023026641280082	0.001534507040696	0.079743423791095	0.001745566988448	0.112615326884333	0.003028957034458	0.001156831521168	0.003388147766989	0.002761197548107	0.132435295396495
X1438_G	X750_G	X2706_G	X4769_G	X7028_T	X8701_G	X8860_G	X9540_C	X10873_C	X12705_T	X15326_G	0.082746338806388	0.012980646111069	0.247608767354477	0.002323521686978	0.006799076749742	0.006936667254838	0.006673826396707	0.006705731807034	0.003597342466415	0.079994091921962
X2706_G	X4769_G	X7028_T	X10398_G	X10400_T	X11719_A	X12705_T	X14766_T	X14783_C	X15043_A	X15301_A	0.10437968635844	0.651513944983246	0.066154552578953	0.048411977193022	0.352434390921595	0.057406892019172	0.348145471312947	0.048360679018867	0.053275723016225	0.063381957175069
X4769_G	X750_G	X1438_G	X2706_G	X7028_T	X8860_G	X10398_G	X11719_A	X14766_T	X15301_A	X15326_G	0.121771135961867	0.105219886875498	0.012140287839243	0.015551926545756	0.175295371334584	0.004025954816789	0.010215436554834	0.010117456223427	0.003892564551004481	0.128478803691695
X7028_T	X2706_G	X4769_G	X8701_G	X9540_C	X10398_G	X10873_C	X11719_A	X12705_T	X14766_T	X15301_A	0.68826682544876	0.124829815458107	0.127242885540133	0.132275102973521	0.154181306259249	0.13852818894403	0.47746377251683	0.149412118481682	0.503901084077474	0.108832517897094
X8701_G	X7028_T	X9540_C	X10398_G	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15043_A	X15301_A	0.184018976381374	0.807741354101559	0.322969428094172	0.469660860130068	0.842666107167679	0.192186398739658	0.45052184727177	0.448964888551049	0.25206732223637	0.504276338238842
X8860_G	X750_G	X1438_G	X2706_G	X4769_G	X7028_T	X11719_A	X14766_T	X15043_A	X15301_A	X15326_G	0.094648585372625	0.021268924636086	0.003005389207413	0.096478308221187	0.003075061007724	0.001556956675707	0.001558142089779	0.000185067296002503	0.000276013417837198	0.166709676789808
X9540_C	X7028_T	X8701_G	X10398_G	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15043_A	X15301_A	0.191245569661843	0.807509581310047	0.341112743380848	0.534909096277312	0.959921966339645	0.20662284263697	0.47514473057772	0.4989874190834315	0.257435990779589	0.527729206730402
X10398_G	X7028_T	X8701_G	X9540_C	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15043_A	X15301_A	0.248220061110093	0.359531933673376	0.379830492558941	0.440145991268409	0.377905045316752	0.268636583259027	0.18718724181660 25	0.362014794976306	0.281639144118491	0.322159573657861
X10400_T	X7028_T	X8701_G	X9540_C	X10398_G	X10873_C	X11719_A	X12705_T	X14783_C	X15043_A	X15301_A	0.11452374743679	0.356185925122629	0.405818838031786	0.299891210752602	0.384059766775751	0.12097500223314	0.268704463240989	1	0.662282899938707	0.487254409286187
X10873_C	X7028_T	X8701_G	X9540_C	X10398_G	X10400_T	X11719_A	X12705_T	X14783_C	X15043_A	X15301_A	0.209221015916682	0.842166340463028	0.959611599878476	0.339272795201868	0.506055449481624	0.20258014749179	0.48108867093802	0.477345559621427	0.253080521266345	0.519411532194492
X11719_A	X2706_G	X7028_T	X8701_G	X9540_C	X10398_G	X10400_T	X10873_C	X12705_T	X14766_T	X15301_A	0.370735906159689	0.537967908318087	0.149730229971347	0.161020438380913	0.188007832169278	0.124262312486534	0.15792152219009	0.175745992874148	0.604837389497182	0.130083321423514
X12705_T	X7028_T	X8701_G	X9540_C	X10398_G	X10400_T	X10873_C	X11719_A	X14783_C	X15043_A	X15301_A	0.236461825330926	0.493017084355115	0.520068362350326	0.184012066175957	0.387684064066545	0.526770708267714	0.24685461900526 6	0.362110160721891	0.306073819125 63	0.378571422065958
X14766_T	X2706_G	X4769_G	X7028_T	X8701_G	X9540_C	X10398_G	X10873_C	X11719_A	X12705_T	X15301_A	0.372701443320698	0.093123365426 2039	0.577969892653576	0.075794129393528	0.076181362686548	0.10938126578126 9	0.07245691694992 05	0.61553476831594	0.100695286668928	0.062950157757302
X14783_C	X7028_T	X8701_G	X9540_C	X10398_G	X10400_T	X10873_C	X11719_A	X12705_T	X15043_A	X15301_A	0.115220660096701	0.340372058695894	0.378396151742378	0.246542495392219	0.099536092423826	0.362102904027252	0.120876693017106	0.250802721358404	0.641212817802825	0.462428782158656
X15043_A	X7028_T	X8701_G	X9540_C	X10398_G	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15301_A	0.11758329658 3973	0.205745855838393	0.210183995704736	0.206505448191903	0.712762468635381	0.207371671196366	0.12471489031 5644	0.228294066480274	0.690593573427169	0.280912820565115
X15301_A	X7028_T	X8701_G	X9540_C	X10398_G	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15043_A	0.144862310809189	0.464126345218 2	0.485814745825931	0.266356890334171	0.591262540890431	0.478340127938044	0.153674710949168	0.318397276374401	0.561398133108464	0.316755844287969
X15326_G	X750_G	X1438_G	X2706_G	X4769_G	X7028_T	X8860_G	X10398_G	X11719_A	X12705_T	X14766_T	0.091907479443615	0.015448536898597	0.001478171047048 51	0.058388741688913 9	0.002073869909601401	0.137657291688003	0.000502389034933861	0.001270317469966 7	0.000482589410053863	0.001353791529585 16

Cuadro A.2: Resultados de asociación para variantes corte 10, filtro 20 %

	X1	X2	X3	X4	X5	Similitud X1	Similitud X2	Similitud X3	Similitud X4	Similitud X5
X750_G	X1438_G	X4769_G	X8860_G	X14783_C	X15326_G	0.0222783703330188	0.0771520911471386	0.108955792857265	0.00327804403364893	0.128131694072441
X1438_G	X750_G	X4769_G	X8860_G	X15326_G	X11914_A	0.0800573270486405	0.23956250283078	0.0879810101481106	0.0773946135958369	0.0158982306592892
X2706_G	X4769_G	X7028_T	X10398_G	X11719_A	X14766_T	0.100987776709129	0.630342426711812	0.0640048022482882	0.340918173346344	0.336832178230313
X4769_G	X750_G	X1438_G	X7028_T	X8860_G	X15326_G	0.117814075871593	0.101800674171524	0.0150457462305853	0.169598994172307	0.124303772043032
X7028_T	X2706_G	X10398_G	X11719_A	X12705_T	X14766_T	0.588500660334605	0.149171049199187	0.461948167576005	0.144556840370191	0.487526375461198
X8701_G	X9540_C	X10400_T	X10873_C	X12705_T	X15301_A	0.839543336124044	0.454345817809657	0.815282838061806	0.43588174387162	0.487889435408774
X8860_G	X750_G	X1438_G	X4769_G	X7028_T	X15326_G	0.0915712255227081	0.0205769014829378	0.0933431589226845	0.00297513419754325	0.161292299309202
X9540_C	X8701_G	X10400_T	X10873_C	X14783_C	X15301_A	0.839319459568972	0.517518831192261	0.920020921258277	0.482772123103731	0.510580182325484
X10398_G	X8701_G	X9540_C	X10400_T	X10873_C	X14783_C	0.34784862748854	0.367487566862292	0.425843060428645	0.365624688720554	0.350250806031193
X10400_T	X9540_C	X10873_C	X14783_C	X15043_A	X15301_A	0.392631398207298	0.371579349045982	0.967504121079131	0.640805359698034	0.471420648998364
X10873_C	X8701_G	X9540_C	X10400_T	X12705_T	X15301_A	0.81479940503211	0.919720639369877	0.489610732868023	0.465447531154294	0.502532797934196
X11719_A	X2706_G	X7028_T	X10398_G	X12705_T	X14766_T	0.358688517041505	0.520486168306069	0.18189835241893	0.170033624067716	0.58518265821373
X12705_T	X8701_G	X9540_C	X10400_T	X10873_C	X15301_A	0.476996060875991	0.503168283816815	0.375085930240531	0.509652831112786	0.366269410971601
X14766_T	X2706_G	X7028_T	X10398_G	X11719_A	X12705_T	0.360590182344915	0.559020874789051	0.10582682541223	0.59553242501316	0.0974231048254324
X14783_C	X9540_C	X10400_T	X10873_C	X15043_A	X15301_A	0.366099836211235	0.967055288587383	0.350336051901087	0.620376043712995	0.447401752444104
X15043_A	X9540_C	X10400_T	X12705_T	X14783_C	X15301_A	0.203353882029211	0.689600625755266	0.220875450137578	0.667925538816393	0.271784311556875
X15301_A	X8701_G	X9540_C	X10400_T	X10873_C	X14783_C	0.449044151700004	0.470053630052755	0.57204894495121	0.462796045101114	0.54315500734857
X15326_G	X750_G	X1438_G	X4769_G	X7028_T	X8860_G	0.088920865119693	0.0149465232023327	0.0564913482086491	0.00200647768355305	0.133183996598384
X3010_A	X1438_G	X7028_T	X11467_G	X12308_G	X12372_A	0.0294301728234937	0.0311058853882237	0.0321218333131376	0.0322122769415657	0.0293337741896369
X11914_A	X8701_G	X9540_C	X10398_G	X10873_C	X12705_T	0.054204607570631	0.0538316695016555	0.0381532135377342	0.0540871843151749	0.037165475843661
X4216_C	X8701_G	X9540_C	X12705_T	X11467_G	X709_A	0.0377883794746241	0.0392735487685455	0.0426844237419075	0.0372083647743232	0.0545033925734894
X11467_G	X8701_G	X9540_C	X12705_T	X12308_G	X12372_A	0.0881877341908113	0.094023908212025	0.110380943334373	1	0.723064192867484
X12308_G	X8701_G	X9540_C	X12705_T	X11467_G	X12372_A	0.0904340286861617	0.0975577649283954	0.124022572013268	0.996673342864714	0.759975069917243
X12372_A	X9540_C	X11719_A	X14766_T	X11467_G	X12308_G	0.049348240533655	0.0594124011842818	0.0540622297569993	0.771359861952005	0.813442172587126
X709_A	X2706_G	X11719_A	X14766_T	X3010_A	X4216_C	0.0170946823932646	0.0166926905824877	0.017537464943012	0.0207858297099853	0.0687357460155803

Cuadro A.3: Resultados de asociación para variantes corte 5, filtro 10 %

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Similitud X1	Similitud X2	Similitud X3	Similitud X4	Similitud X5	Similitud X6	Similitud X7	Similitud X8	Similitud X9	Similitud X10
X750_G	X1438_G	X4769_G	X7028_T	X8860_G	X10400_T	X14783_C	X15043_A	X15326_G	X3010_A	X4216_C	0.0222783703330188	0.0771520911471386	0.0016884462980573	0.108955792857365	0.0029304317112631	0.0032780440364893	0.0026714700690769	0.128131694072441	0.0026214509116798	0.0019529738592726
X1438_G	X750_G	X2706_G	X4769_G	X8860_G	X15326_G	X3010_A	X11914_A	X11467_G	X12308_G	X12372_A	0.0800573270486405	0.0125588288969892	0.2395625283078	0.0879810101481106	0.0773946135958369	0.0107213554631836	0.0158982306592892	0.0095206820529332	0.0104656000248045	0.0094706846602035
X2706_G	X4769_G	X7028_T	X10398_G	X11719_A	X12705_T	X14766_T	X15301_A	X11467_G	X12308_G	X12372_A	0.100987776709129	0.630342426711812	0.0640048022482882	0.340918173346344	0.0555414046068986	0.336832178230313	0.0613223047669126	0.0603858788202993	0.0618006497030696	0.0589086354061245
X4769_G	X750_G	X1438_G	X2706_G	X7028_T	X8860_G	X10398_G	X11719_A	X14766_T	X15301_A	X15326_G	0.117814075871593	0.101800674171524	0.0117457784998628	0.0150457462305853	0.169598994172307	0.0038942571227497	0.0098834769649833	0.0097886805918893	0.0053660722050472	0.124303772043032
X7028_T	X2706_G	X4769_G	X8701_G	X9540_C	X10398_G	X10873_C	X11719_A	X12705_T	X14766_T	X15301_A	0.588500660334605	0.149171049199187	0.123180916138679	0.127976707243048	0.149171049199187	0.134026593688978	0.461948167576005	0.144556840370191	0.487526375461198	0.105295909572805
X8701_G	X9540_C	X10398_G	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15043_A	X15301_A	X12308_G	0.839543336124044	0.312474525063682	0.454345817809657	0.815282838061896	0.185941132795976	0.435881743871462	0.434375379853402	0.243876173053069	0.487889435408774	0.181443235192037
X8860_G	X750_G	X1438_G	X2706_G	X4769_G	X7028_T	X11719_A	X14766_T	X15326_G	X4216_C	X12372_A	0.0915712255227081	0.0205769014829378	0.0029077264361916	0.0933431589226845	0.0029751341975432	0.0015063029744706	0.0014975508930804	0.161292299309202	0.0008436934887728	0.0004180228866534
X9540_C	X8701_G	X10398_G	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15043_A	X15301_A	X12308_G	0.839319459568972	0.330027984973578	0.517518831192261	0.920020921258277	0.199084533334218	0.459675459818842	0.482772123103731	0.249070381993342	0.510580182325484	0.197840792066483
X10398_G	X7028_T	X8701_G	X9540_C	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15043_A	X15301_A	0.240153932655529	0.34784862748854	0.367487566862292	0.425843060428645	0.365624688726554	0.259607001375725	0.181104427876044	0.350250806031193	0.372487032591839	0.311660715159076
X10400_T	X8701_G	X9540_C	X10398_G	X10873_C	X14783_C	X15043_A	X15301_A	X11467_G	X12308_G	0.344611350426526	0.392631398207298	0.290145982278552	0.371579349045982	0.259972675538013	0.967504121079131	0.640805359698034	0.471420648998364	0.132339719565677	0.126257426063351
X10873_C	X7028_T	X8701_G	X9540_C	X10398_G	X10400_T	X11719_A	X12705_T	X14783_C	X15043_A	X15301_A	0.193714658027975	0.814799059211	0.919720639369877	0.328247825727843	0.489610732868023	0.195597127547126	0.465447531154294	0.401833796112554	0.245657540702297	0.502532797934196
X11719_A	X2706_G	X7028_T	X8701_G	X9540_C	X10398_G	X10400_T	X10873_C	X12705_T	X14766_T	X15301_A	0.358688517041505	0.520486168306069	0.114864614547404	0.155787937720209	0.181898324189	0.120224299425544	0.152789724037733	0.170033624067716	0.585182654213373	0.125856149558976
X12705_T	X8701_G	X9540_C	X10400_T	X10873_C	X11719_A	X14783_C	X15043_A	X15301_A	X11467_G	X12308_G	0.476996060877091	0.503168283816815	0.375085930240531	0.509652831112786	0.238832861195014	0.35934397858808	0.296127681358476	0.366269410971601	0.244206862759239	0.275306629920138
X14766_T	X2706_G	X4769_G	X7028_T	X10398_G	X10873_C	X11719_A	X12705_T	X11467_G	X12308_G	X12372_A	0.360590182344915	0.0900972208186101	0.559020874789051	0.105826824512223	0.0749398865135192	0.595518242501316	0.0974231048254324	0.0856289259955	0.0887764736607	0.0812338891299835
X14783_C	X7028_T	X8701_G	X9540_C	X10398_G	X10400_T	X10873_C	X11719_A	X12705_T	X15043_A	X15301_A	0.111476414134305	0.329311369742963	0.366099836211235	0.238530880313105	0.967055288587383	0.356336051901087	0.116948698863467	0.242710716739381	0.620376043712995	0.447401752444104
X15043_A	X7028_T	X8701_G	X9540_C	X10398_G	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15301_A	0.113762314340922	0.199059966030859	0.203353882029211	0.199784872159096	0.689600625755266	0.200632946477551	0.120662170340317	0.220875450137578	0.667925538816393	0.271784311556875
X15301_A	X7028_T	X8701_G	X9540_C	X10398_G	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15043_A	0.140154882609936	0.449044151700004	0.470053630052755	0.257706605978297	0.572048944955121	0.462796045101114	0.146869161483964	0.308350677032604	0.543155007348857	0.369462584724511
X15326_G	X750_G	X1438_G	X2706_G	X4769_G	X7028_T	X8860_G	X10398_G	X11719_A	X14766_T	X3010_A	0.0889208651196693	0.0149465232023327	0.0014301365796729	0.0564913482086491	0.0020064776835305	0.133183996598384	0.0004860724692205	0.0012290371714666	0.0013097988804247	0.0005578097943805
X3010_A	X750_G	X1438_G	X2706_G	X7028_T	X11719_A	X4216_C	X11467_G	X12308_G	X12372_A	X709_A	0.0258584514783656	0.0294307238234937	0.0250863400517154	0.0311058853882237	0.0194906923781673	0.0270986339672707	0.0321218333131376	0.0322122769415657	0.0293337741896369	0.0225530897085063
X11914_A	X1438_G	X7028_T	X8701_G	X9540_C	X10398_G	X10873_C	X11719_A	X12705_T	X14766_T	X4216_C	0.0347997401936772	0.0188731087711064	0.0542046075770631	0.0538316695016555	0.0381532135377342	0.0540871843151749	0.0210271973628986	0.0371654758843661	0.0218977734847408	0.0167563289532864
X4216_C	X8701_G	X9540_C	X10873_C	X11719_A	X12705_T	X15301_A	X11467_G	X12308_G	X12372_A	X709_A	0.0377883794746241	0.0392735487685455	0.0371012204972561	0.0335385634591558	0.0426844237419075	0.0349024503671665	0.0372083647743232	0.0364823823452833	0.0345545587738945	0.0545033925734894
X11467_G	X8701_G	X9540_C	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15301_A	X12308_G	X12372_A	0.0881877341908113	0.0940239082120325	0.0863027236241399	0.0869173961496253	0.0678043404946947	0.110380943334373	0.0714567479898716	0.0643902865529958	1	0.723064192867484
X12308_G	X8701_G	X9540_C	X10400_T	X10873_C	X11719_A	X12705_T	X14783_C	X15301_A	X11467_G	X12372_A	0.0904340286861617	0.0975577649283954	0.0820623723461385	0.0768905570623388	0.0710986234664328	0.124022572013268	0.0711111476142656	0.0645263270169692	0.9966733428647 14	0.7599750699172 43
X12372_A	X2706_G	X7028_T	X8701_G	X9540_C	X10873_C	X11719_A	X14766_T	X4216_C	X11467_G	X12308_G	0.0419697125721914	0.0467417105039712	0.0487982200120378	0.0493482405336 55	0.0478552531595664	0.0594124011842818	0.0540622297569993	0.0448364122294302	0.7713598619520 05	0.8134421725871 26
X709_A	X750_G	X2706_G	X7028_T	X11719_A	X14766_T	X3010_A	X4216_C	X11467_G	X12308_G	X12372_A	0.0131756013214288	0.0170946823932646	0.0164089955482405	0.0166926905824877	0.017537464943012	0.0207858297099853	0.0687357460155803	0.0073774638606588	0.0073560024593726 7	0.0076348368634798

Cuadro A.4: Resultados de asociación para variantes corte 5, filtro 10 %

Bibliografía

[1] Smith PM, Elson JL, Greaves LC. (2014) The role of the mitochondrial ribosome in human disease: searching for mutations in 12S mitochondrial rRNA with high disruptive potential. Hum Mol Genet. ;23(4):949-967

[2] Jamie J Cannone, Sankar Subramanian, Murray N Schnare, James R Collett, Lisa M D'Souza, Yushi Du, Brian Feng, Nan Lin, Lakshmi V Madabusi, Kirsten M Müller, Nupur Pande, Zhidi Shang, Nan Yu, Robin R Gutell (2002) The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3.

[3] Gutell, R.R., Larsen, N., and Woese, C.R. (1994). Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiology . Rev. 58, 10-26.

[4] Gutell, R.R., Lee, J.C., and Cannone, J.J. (2002). The accuracy of ribosomal RNA comparative structure models. Curr. Opin. Struct. Biol. 12, 301-310.

[5] Vila, A., Viril-Farley, J., and Tapprich, W.E. (1994). Pseudoknot in the central domain of small subunit ribosomal RNA is essential for translation. Proc. Natl. Acad. Sci. U. S. A. 91, 11148-11152

[6] Poot, R.A., van den Worm, S.H., Pleij, C.W., and van Duin, J. (1998). Base complementarity in helix 2 of the central pseudoknot in 16S rRNA is essential for ribosome functioning. Nucleic Acids Res. 26, 549-553.

[7] Cunningham, P.R., Nurse, K., Weitzmann, C.J., and Ofengand, J. (1993). Functional effects of base changes which further define the decoding center of Escherichia coli 16S ribosomal RNA: mutation of C1404, G1405, C1496, G1497, and U1498. Biochemistry 32, 7172-7180.

[8] Aagaard, C., and Douthwaite, S. (1994). Requirement for a conserved, tertiary interaction in the core of 23S ribosomal RNA. Proceedings of the National Academy of Sciences 91, 2989-2993.

[9] Powers, T., and Noller, H.F. (1990). Dominant lethal mutations in a conserved loop in 16S rRNA. Proc. Natl. Acad. Sci. U. S. A. 87, 1042-1046.

[10] Dong, J., Nanda, J.S., Rahman, H., Pruitt, M.R., Shin, B.S., Wong, C.M., Lorsch, J.R., and Hinnebusch, A.G. (2008). Genetic identification of yeast 18S rRNA residues required for efficient recruitment of initiator tRNA(Met) and AUG selection. Genes Dev. 22, 2242-2255.

[11] Yassin, A., Fredrick, K., and Mankin, A.S. (2005). Deleterious mutations in small subunit ribosomal RNA identify functional sites and potential targets for antibiotics. Proc. Natl. Acad. Sci. U. S. A. 102, 16620-16625.

[12] Cochella, L., Brunelle, J.L., and Green, R. (2007). Mutational analysis reveals two independent molecular requirements during transfer RNA selection on the ribosome. Nat. Struct. Mol. Biol. 14, 30-36.

[13] Kim, D.F., and Green, R. (1999). Base-pairing between 23S rRNA and tRNA in the ribosomal A site. Mol. Cell 4, 859-864.

[14] Leffers, H., and Andersen, A.H. (1993). The sequence of 28S ribosomal RNA varies within and between human cell lines. Nucleic Acids Res. 21, 1449-1455.

[15] Greber, B.J., Boehringer, D., Leibundgut, M., Bieri, P., Leitner, A., Schmitz, N., Aebersold, R., and Ban, N. (2014a). The complete structure of the large subunit of the mammalian mitochondrial ribosome. Nature

[16] Kaushal, P.S., Sharma, M.R., Booth, T.M., Haque, E.M., Tung, C.S., Sanbonmatsu, K.Y., Spremulli, L.L., and Agrawal, R.K. (2014). Cryo-EM structure of the small subunit of the mammalian mitochondrial ribosome. Proc. Natl. Acad. Sci. U. S. A. 111, 7284-7289.

[17] Amunts, A., Brown, A., Bai, X.C., Llacer, J.L., Hussain, T., Emsley, P., Long, F., Murshudov, G., Scheres, S.H., and Ramakrishnan, V. (2014). Structure of the yeast mitochondrial large ribosomal subunit. Science 343, 1485-1489.

[18] Schuwirth, B.S., Borovinskaya, M.A., Hau, C.W., Zhang, W., Vila-Sanjurjo, A., Holton, J.M., and Cate, J.H. (2005). Structures of the bacterial ribosome at 3.5 A resolution. Science 310, 827-834. Sharma, M.R., Koc, E.C., Datta, P.P., Booth, T.M., Spremulli, L.L., and Agrawal, R.K. (2003). Structure of the mammalian mitochondrial ribosome reveals an expanded functional role for its component proteins. Cell 115, 97-108.

[19] Ben-Shem, A., Garreau de Loubresse, N., Melnikov, S., Jenner, L., Yusupova, G., and Yusupov, M. (2011). The structure of the eukaryotic ribosome at 3.0 A resolution. Science 334, 1524-1529.

[20] Ban, N., Nissen, P., Hansen, J., Moore, P.B., and Steitz, T.A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 289, 905-920.

[21] Burger, G., Gray, M.W., and Lang, B.F. (2003). Mitochondrial genomes: anything goes. Trends Genet. 19, 709-716.

[22] Blackstone, N.W. (2013). Why did eukaryotes evolve only once? Genetic and energetic aspects of conflict and conflict mediation. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 368, 20120266.

[23] Wolff, J.N., Ladoukakis, E.D., Enriquez, J.A., and Dowling, D.K. (2014). Mitonuclear interactions: evolutionary consequences over multiple biological scales. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 369, 20130443.

[24] Lynch, M., Koskella, B., and Schaack, S. (2006). Mutation pressure and the evolution of organelle genomic architecture. Science 311, 1727-1730.

[25] Ruiz-Pesini, E., Lott, M.T., Procaccio, V., Poole, J.C., Brandon, M.C., Mishmar, D., Yi, C., Kreuziger, J., Baldi, P., and Wallace, D.C. (2007). An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res. 35, D823-D828.

[26] Smith, P.M., Elson, J.L., Greaves, L.C., Wortmann, S.B., Rodenburg, R.J., Lightowlers, R.N., Chrzanowska-Lightowlers, Z.M., Taylor, R.W., and Vila-Sanjurjo, A. (2014). The role of the mitochondrial ribosome in human disease: Searching for mutations in 12S mitochondrial rRNA with high disruptive potential. Hum. Mol. Genet. 23, 949-956.

[27] Sergio Ramírez-GallegoIago LastraIago LastraDavid MartinezDavid MartinezShow all 7 authorsAmparo Alonso-BetanzosAmparo Alonso-Betanzos, (2016) Fast-mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data: FAST-mRMR ALGORITHM FOR BIG DATA. International Journal of Intelligent Systems. 32

[28] Jorge González-Domínguez,Verónica Bolón-Canedo, Borja Freire, Juan Touriñoa, (2018), Parallel Feature Selection for Distributed-Memory Clusters, Information Science.

[29] Arthur M. Lesk, (2017) Introduction to genomics, Book.

[30] Faraway, J.J. (2006). Extending the linear model with R. Generalized linear, mixed effects and nonparametric regression models. Chapman and Hall.

[31] Chao-Ying Joanne Peng,Kuk Lida Lee, Gary M. Ingersoll, (2010) An Introduction to Logistic Regression Analysis and Reporting, The Journal of Educational Research, 3-14

[32] Mark Lunt, (2015), Introduction to statistical modelling: linear regression, Rheumathology,54, 1137–1140.

[33] Cleveland, W. (1979). Robust locally-weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74:829-836.

[34] Green, P.J. y B.W. Silverman (1994). Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. London: Chapman & Hall.

[35] Fan, J. y I. Gijbels (1996). Local polynomial modelling and its applications. London: Chapman & Hall;

[36] Hall. Hastie, T. and Tibshirani, R. (1987). Generalized additive models: Some applications. Journal of the American Statistical Association, 82:371(386).

[37] Hastie, T., R. Tibshirani y J. Friedman (2001). The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer.

[38] Wasserman, L. (2006). All of Nonparametric Statistics. New York: Springer. Ruppert, D., Wand, M., and Carroll, R. (2003). Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[39] Wood, N. (2003). Thin plate splines regression. Journal of the Royal Statistical Society,65(1):95114.

[40] Wood, S. (2006a). mgcv 1.3. r package. cran.r-project.org.

[41] Wood, S. (2006b). On confidence intervals for gams based on penalized regression splines. Australian and New Zealand Journal of Statistics, 48:445(464).

[42] McCullagh, P. and Nelder, J. (1989). Generalized Linear Models. Chapman & Hall, New York.

[43] Rigby, R. A. and Stasinopoulos D. M. (2005). Generalized additive models for location, scale and shape,(with discussion), Appl. Statist., 54, part 3, pp 507-554.

[44] Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z., and De Bastiani, F. (2019) Distributions for modeling location, scale, and shape: Using GAMLSS in R, Chapman and Hall/CRC.

[45] Stasinopoulos D. M. Rigby R.A. (2007) Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software, Vol. 23, Issue 7, Dec 2007, https://www. jstatsoft.org/v23/i07/.

[46] Stasinopoulos D. M., Rigby R.A., Heller G., Voudouris V., and De Bastiani F., (2017) Flexible Regression and Smoothing: Using GAMLSS in R, Chapman and Hall/CRC.