Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Biológicas

Predicciones estructurales de las variantes particulares no-sinónimas en el genoma de la población Comcáac (Seri)

TESIS

PRESENTADA PARA OBTENER EL GRADO DE
LICENCIADA EN BIOTECNOLOGÍA

PRESENTA
Alejandra Paulina Pérez González

OCTUBRE 2022

DIRECTOR DE TESIS: MC. Israel Aguilar Ordóñez
CO-DIRECTORA: Dra. Norma Angélica Caballero Concha
SINODALES
Dra. Wendy Argelia García Suastegui
Dra. Ana Isabel Castillo Orozco

“Estoy entre aquellas que piensan que la ciencia tiene una gran belleza.”

-Marie Curie

I. AGRADECIMIENTOS

A Marcia, mi abuelita Mary, Alejandro y mi Tío Oscar, por la infinita paciencia que me han tenido y por todo el apoyo que he recibido de ustedes. Sin su amor y apoyo, nada de lo que he vivido y hecho habría sido posible.

A Lili, que ha soportado las mil quinientas veces que les he hablado de este trabajo y a pesar de no entenderlo, me ha aplaudido y apoyado con emoción y complicidad.

Al Consejo de Babooshkas y equipo de Sororas en la Ciencia Guadalupe, Irene, Aleida, Lizeth y Alejandra. Ustedes fueron el soporte de mis breakdowns académicos. Gracias por las risas, por el apoyo y por los memes. Han hecho mi vida más feliz.

A Fer, Josué y Edu. Sin ustedes y su increíble capacidad de entendimiento, probablemente este trabajo no hubiera resultado como debiera. Gracias por apoyarme a resolver dudas bioinformáticas y por las trasnoches en Discord escuchando música.

A Isra, por ser un increíble tutor. Gracias por la paciencia, por abrirnos puertas y oportunidades que soñaba imposibles, por su tiempo cada reunión de cada semana, por las comidas y por creer en el potencial de cada miembro del laboratorio virtual. Gracias infinitas.

A la Dra. Norma Caballero. Por haberme mostrado por primera vez la bioinformática y la biología computacional. Gracias también por su apoyo como co-tutora para realizar este trabajo.

A doña Josefina Domínguez O., que me abrió las puertas de su hogar durante parte del desarrollo de este trabajo. Sin su apoyo, este trabajo probablemente se habría visto truncado. Nunca sabré cómo agradecerle.

Al Instituto Nacional de Medicina Genómica, a la Benemérita Universidad Autónoma de Puebla y al Consejo Nacional de Ciencia y Tecnología por brindarme herramientas para el aprendizaje y autorrealización académica.

Por último, infinitas gracias a la comunidad Comcaa'c, especialmente a los 4 individuos cuyos genomas aquí han sido analizados. Ojalá este trabajo, el cual realicé de la manera más respetuosa, les beneficie en un futuro no muy lejano.

Y a ti, que lees esto.

A. Paulina Pérez González

II. ÍNDICE GENERAL

III. ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

La antropología molecular se está convirtiendo en una fuente de información cada vez más importante sobre la historia y desarrollo de las poblaciones humanas. Previamente, a partir de análisis primarios del genoma nativo Comcaa'c, se identificaron características particulares de la población en comparación con otros grupos étnicos previamente estudiados. En el presente trabajo, el genoma de 4 individuos Comcaa'c fue analizado a nivel genómico y proteómico. Mediante el uso de herramientas bioinformáticas, se identificaron variantes presentes en la población pero ausentes en otros grupos nativos. Se identificaron aquellas que generaban cambios no-sinónimos en estructuras proteicas y se exploró la predicción de modelos para las secuencias variantes e invariantes. Se encontró una estructura proteica que podría albergar cambios estructurales: la cadena pesada de la dineína, codificada por el gen DNHD1, perteneciente al complejo de la proteína motora del mismo nombre. Se analizó la función del dominio con el sitio variante y se predice que la variación cae en un dominio AAA+ de unión e hidrólisis de ATP, lo que podría derivar en alguna modificación en el rendimiento del complejo proteico del complejo proteico motor de la dineína. Este proyecto sienta un precedente en el análisis de variantes particulares llevado hasta el nivel estructural, especialmente en grupos nativos.

ABSTRACT

Molecular anthropology is becoming an increasingly important source of information on the history and development of human populations. Previously, from primary analyses of the native Comcaa'c genome, particular characteristics of the population were identified in comparison with other previously studied ethnic groups. In the present work, the genome of 4 Comcaa'c individuals was analyzed at the genomic and proteomic level. Using bioinformatics tools, particular variants present in the population but absent in other native groups were identified. Those generating non-synonymous changes in protein structures were identified and model prediction for variant and invariant sequences was explored. One protein structure was found that could harbor structural changes: the dynein heavy chain, encoded by the DNHD1 gene, belonging to the motor protein complex of the same name. The function of the domain with the variant site was analyzed and it is predicted that the variation falls into an AAA+ domain for ATP binding and hydrolysis, which could lead to some modification in the performance of the protein complex of the dynein motor protein complex. This project sets a precedent in the analysis of particular variants taken to the structural level, especially in native groups.

1. Introducción

1.1. La antropología molecular y la diversidad genética de México

México alberga una gran diversidad cultural, genética y étnica. Según datos del Censo 2020 del INEGI, existe un aproximado de 7,364,645 individuos pertenecientes a alguna comunidad nativa mexicana, repartidos en 68 grupos étnicos, y al menos el 6.71% de la población mexicana aún habla alguna lengua nativa. Estas comunidades, descendientes de los primeros pobladores de América, han existido durante muchos años en el territorio y mantienen su cultura, lengua y herencia genética. En el estado de Sonora, México, existe un aproximado de 62,808 individuos indígenas, entre los cuales se encuentran comunidades Amuzgo, Chatino, Chinanteco, Ch'ol, Coras, y Comcaa'c o Seris, entre otras poblaciones *(INPI, 2015)*.

***Figura** 1. Algunas poblaciones nativas mexicanas previamente estudiadas a nivel molecular. Fotografías recuperadas del Repositorio del IIS-UNAM: 2. Archivo fotográfico «México Indígena» y de Atlas de los Pueblos Indígenas de México*

La antropología molecular es una fuente de información cada vez más importante sobre la historia evolutiva de las poblaciones. Mediante enfoques como la genómica, la proteómica, la inmunología, y otras disciplinas, la antropología molecular analiza evidencia a nivel molecular, es decir, ADN, proteínas y otras moléculas biológicas para obtener información sobre los orígenes humanos, explicar eventos de migración, entender el papel de la selección natural en la evolución humana, el impacto de prácticas culturales particulares en los patrones de variación genética humana, entre otros tópicos *(Stoneking, 2015)*. De esta manera, los eventos antropológicos pueden ser estudiados desde perspectivas biológicas y viceversa. Los estudios genéticos brindan la oportunidad de comprender y complementar la historia evolutiva de la población y las fuerzas que generan la variación en sus genomas *(Korunes & Goldberg, 2021)*. Tales eventos han dado paso a la aparición de particularidades en el genoma de individuos pertenecientes a comunidades nativas, es decir, grupos étnicos con culturas tradicionales pertenecientes al continente americano. A lo largo de los años, diferentes comunidades nativas americanas (NatAm) han sido estudiadas (*fig. 1*) con enfoques moleculares, tales como Mixtecos, Zapotecos, Mayas, Nahuas, Wixarikas, Mixe, Popolucas, etc. *(García-Ortiz et al., 2021; Gómez et al., 2022)*. Entre las comunidades nativas mexicanas, el grupo Comcaa'c había

sido poco estudiado a nivel genético y hasta 2021, no contaba con representación en las bases de datos de secuenciación de genoma completo. A partir de un análisis primario de su genoma *(Aguilar-Ordoñez et al., 2021; Moreno-Estrada et al., 2014)*, se identificaron características que los destacan de otros grupos étnicos previamente estudiados a nivel genómico. Comprender la estructura genómica de una población humana permite el diseño e interpretación de estudios genéticos, que en un futuro cercano, podría derivar en una praxis médica renovada que tenga la capacidad de ajustarse a la información acorde a la diversidad real existente en los diversos grupos humanos *(Moreno-Estrada et al., 2014)*.

En el caso de los grupos nativos americanos, como en otros grupos humanos, los estudios regionales y comunitarios han sido clave para delinear esta misma variación geográfica y cultural. A pesar de que las poblaciones nativas de todo el mundo han estado presentes durante siglos en el territorio, la literatura existente sobre salud, de atención sanitaria, genética y molecular es poca y menos caracterizada que para poblaciones cosmopolitanas *(Hindorff et al., 2018)*.

1.2. Los Comcaa'c: breve contexto histórico y cultural

Los Comcaa'c, Comcaác, Konkaak o Concaác (denominados también como Seris c. 1965 durante la colonización europea) *(Martínez-Tagüeña & Torres Cubillas, 2018)*, son una comunidad indígena que habita la costa central del Desierto del estado de Sonora, México. Comca'ac quiere decir en su lengua **cmiique iitom** "la gente" (*fig. 2*). En cambio, el término "Seri" proviene en cambio de la lengua yaqui y significa "hombres de la arena" *(INPI, 2017)*.

***Figura** 2. Izq. Retrato de mujer Comcaa'c o Seri con pintura facial. Fotografía de 1949. Fotografía de William Neil Smith. Recuperado de Burckhalter, 2013. Derecha. Bandera de la Nación Comcaa'c (Recuperado de Wiki-commons)*

Hoy en día, la población Comcaa'c se encuentra distribuida principalmente en el norte de México en las localidades de Punta Chueca (municipio de Hermosillo) y El Desemboque (municipio de Pitiquito) como principales asentamientos, además de varios campamentos temporales a lo largo de la franja costera, donde las familias habitan durante distintos periodos, según la naturaleza de sus ciclos de pesca *(Rentería-Valencia, 2007)*. La *figura 3* muestra el paisaje geográfico del grupo étnico Comcaa'c.

***Figura** 3 El paisaje geográfico de los Comcáac. Las líneas .--.--.-- demarcan el territorio Comcaa'c actual aproximado. En líneas punteadas más amplias se muestra el territorio Comcaa'c antes de la colonización. Las zonas marcadas con gris muestran la Zona de Territorialidad según la tradición oral Comcaa'c. Mapa recuperado de (Luque, 2006).*

1.3. Las variantes particulares de los grupos nativos

Los grupos nativos americanos presentan variantes genéticas únicas y comunes entre las comunidades indígenas, pero poco frecuentes o incluso inexistentes en otras poblaciones del mundo, debido probablemente a la historia evolutiva y de migración al haber quedado aislados de otras poblaciones humanas en la antigüedad (Moreno-Estrada et al., 2014). La *figura 4* muestra una explicación resumida de lo que se puede entender como Variantes genéticas.

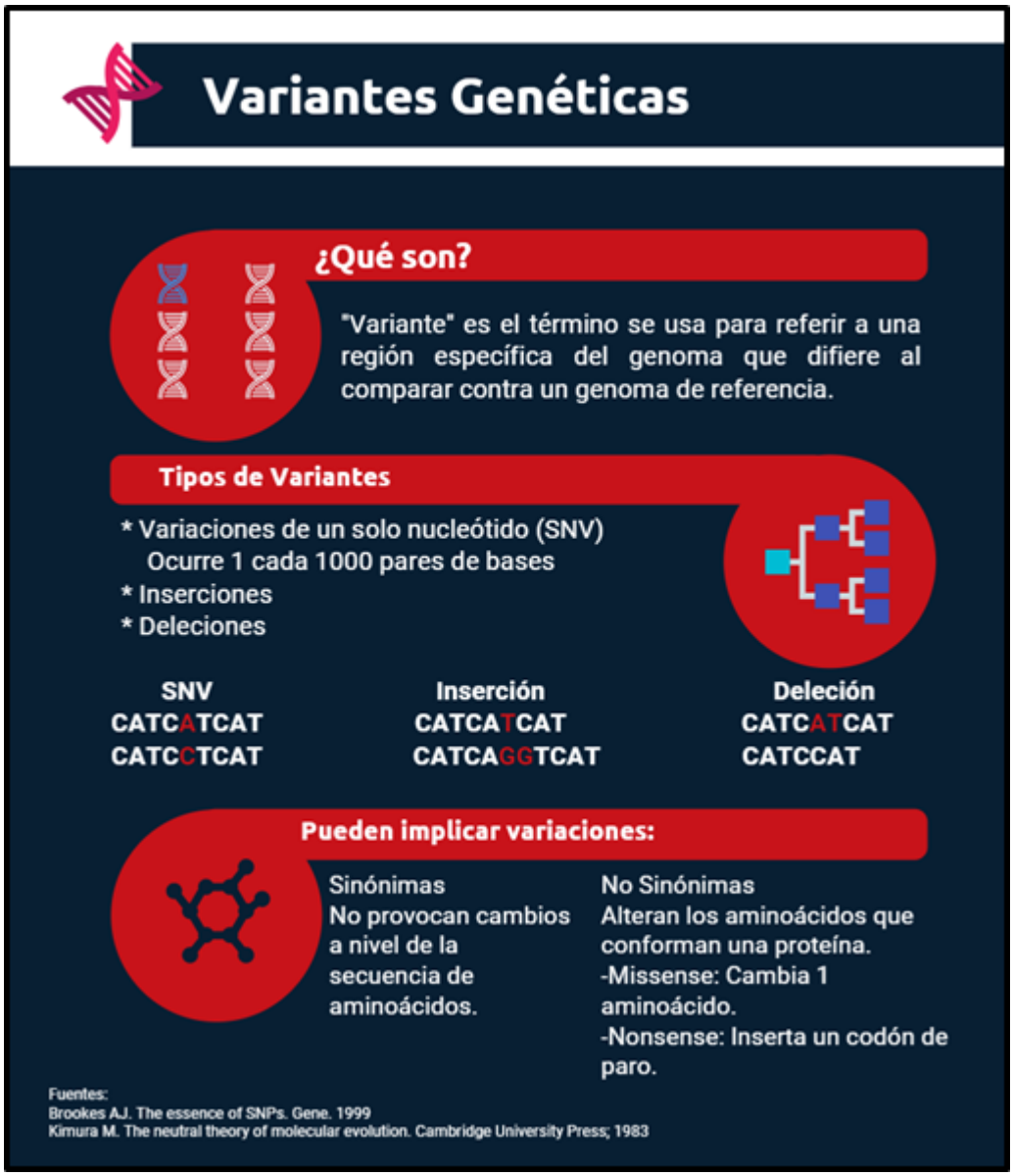


***Figura** 4. Variantes genéticas, clasificación e impacto biológico. Recuperado de Ballesteros-Villascán, 2020.*

Estudiar los genomas de las poblaciones nativas mexicanas proporciona datos para el estudio de fenotipos presentes en las mismas comunidades, que pueden ser variantes funcionales y/o

enfermedades y en grupos derivados, y las variantes particulares de grupos nativos podrían ser relevantes para entender fenotipos, enfermedades presentes en las poblaciones, dinámicas poblacionales y algunas otras características biológicas.

Además, investigaciones han mostrado que los nativos americanos (NatAm) muestran la diversidad genética más baja de cualquier grupo continental, y al mismo tiempo, mantienen gran divergencia entre algunas poblaciones (S. Wang et al., 2007). Como resultado, las poblaciones NatAm actuales (e individuos con ascendencia nativa) pueden albergar alelos locales, ausentes en otras regiones del mundo, incluidas variantes funcionales y médicamente relevantes *(Acuña-Alonzo et al., 2010)*. Así mismo, los nativos americanos tienen estructuras poblacionales genéticas distintivas, que difieren en gran medida de las poblaciones mestizas. De hecho, se estima que la diferenciación de nativos mexicanos es 38 veces mayor que la estimada para los mestizos mexicanos. Mediante una comparación de STRs autosómicos de interés forense para un análisis interpoblacional entre mestizos y otras poblaciones nativas americanas, se identificó que la diferenciación entre los nativos mexicanos también es mayor (aproximadamente el doble) que los grupos étnicos estadounidenses y 38 veces mayor que el estimado entre los mestizos mexicanos *(Martínez-Cortés et al., 2019)*. Lo más probable es que esto se explique por efectos de deriva genética *(Rangel-Villalobos et al., 2013)*.

La *medicina genómica*, otro concepto central a considerar aquí, se ha definido como el uso de la información genotípica de un paciente individual en su atención clínica. Esta definición abarca enfermedades mendelianas, multigénicas y otras enfermedades complejas, siempre con un enfoque de análisis de variabilidad a la que se llega mediante herramientas, técnicas y tecnologías que integran la individualidad y/o características poblacionales del paciente *(Manolio et al., 2013).* La traslación de la genómica a la medicina de precisión depende de la interpretación precisa de la multitud de variantes genéticas observadas en cada individuo *(Glusman et al., 2017).*

Por ejemplo, en los últimos años, los estudios de farmacogenómica, una parte importante de la medicina genómica, han generado información sustancial que es útil en entornos clínicos. Las y los pacientes varían en su respuesta a diferentes fármacos. La idea de que la variación genética se puede utilizar para individualizar la terapia farmacológica, o en su caso, reformular las dosificaciones según grupos genéticos, como pueden ser grupos nativos, abre un panorama al mejoramiento de la calidad de vida de estas comunidades *(Henderson et al., 2018).*

La estructura genética de comunidades nativas americanas (NatAm) se ha analizado previamente utilizando tecnologías de secuenciación de exomas, genotipificaciones por PCR y matrices de microarreglos, estudios de asociación de genoma completo (GWAS), análisis de SNPs, mapeos de mezclas (AM), marcadores genéticos, patrones de desequilibrio de ligamiento, STRs, ADMIXTURE, marcadores farmacogenómicos, inmunoensayos, etc. *(McCarthy et al., 2008; The Wellcome Trust Case Control Consortium, 2007)*, mediante diversidad de técnicas como la PCR, la cromatografía, la secuenciación de exomas, o la secuenciación de genoma completo (*fig.5*). En cuanto a poblaciones infrarrepresentadas en los catálogos de marcadores genéticos y asociación de genotipos existentes, se presentan

limitaciones específicas de poblaciones nativas que originan brechas de conocimiento, que a su vez se traducen en pérdidas de oportunidad para el mejoramiento de la calidad de vida poblacional. Finalmente, debido a que las características genéticas de la población de estudio de este trabajo son extraordinarias (por el contexto e historia cultural), este flujo de trabajo podría servir de ejemplo para el aprendizaje de otras comunidades genéticamente particulares alrededor del mundo.

***Figura** 5. Metodologías de secuenciación y microarreglos de ADN. Tomado de Ballesteros-Villascán, 2020.*

1.4. Bioinformática y Biología computacional

Durante las últimas décadas se han desarrollado nuevas herramientas y enfoques computacionales para el análisis de datos genómicos (*Bolnick et al., 2016*). El número y la diversidad de herramientas, incluidos los recursos de datos, crece día con día, lo cual facilita la comparación e integración de datos de manera que usuarias y usuarios de las herramientas

bioinformáticas son capaces de aprovechar la inmensa cantidad de datos biológicos *(Ison et al., 2013)*. Las herramientas computacionales, a su vez, permiten manejar datos estructurales y de modelado tridimensional en proteínas. El campo de la bioinformática y la biología computacional está revolucionando las preguntas y respuestas de investigación en el ámbito de la genética y genómica humana, haciendo uso de algoritmos y herramientas bioinformáticas *(Lindblom & Robinson, 2011)*. Para este trabajo nos es de interés la información contenida en los genomas humanos.

El análisis de datos biológicos se está consolidando en una actividad fundamental para los laboratorios de todo el mundo. Esta integración permite el hallazgo de nuevo conocimiento biológico a partir de la extracción, normalización, emparejamiento y enriquecimiento de datos iniciales *(Aguilar-Ordoñez et al., 2021; Bernasconi et al., 2021)*. En el presente trabajo se utilizaron herramientas bioinformáticas previamente desarrolladas, así como instrumentos públicos y gratuitos, que pueden encontrarse en repositorios de código abierto.

2. Antecedentes

2.1. Demografía histórica de la comunidad Comcaa'c

Hasta antes de 1944, los Comcáac (*fig. 6)* eran un pueblo nómada cuyo estilo de vida giraba principalmente en torno a recursos acuíferos como el pescado y el marisco. Los Comcaa'c, como otros grupos de nativos mexicanos que se asentaron en el norte, vivían principalmente como cazadores recolectores, un estilo de vida incompatible con grandes tamaños poblacionales (W. W. Taylor, 1972). Es por ello y por otras razones relacionadas al clima y los recursos disponibles en el área, que los Comcaa'c mostraban características como la distribución en clanes y bandas. Al momento del contacto con los primeros exploradores y colonizadores europeos, la sociedad Comcaa'c se hallaba organizada en distintas bandas delimitadas entre sí por sutiles diferencias políticas, económicas y sociales *(Rentería Valencia, 2007)*. Los procesos de colonización, persecución y otros eventos cambiaron profundamente los patrones de asentamiento y organización social durante el período colonial español.

Figura 6. *Campamento de Haj Hax en Tecomate, en el extremo norte de la isla del Tiburón, mirando al oeste con campamento de Xneeelcam más allá de la duna, 1949. Recuperado de Burckhalter, 2013.*

Durante el s. XVIII los Comcaa'c se convirtieron en una amenaza para la seguridad de las colonias provincianas *(Sheridan, 1999*). Debido al *modus vivendi* que sostenían, tendían a robar rancherías y otros establecimientos en las inmediaciones de sus territorios en busca de alimento y bienes. Los enfrentamientos entre nativos y colonizadores desembocaron en un levantamiento. La guerra de resistencia abierta hacia los Comcaa'c, Pimas y otras comunidades nativas provocó que los grupos se refugiaran y finalmente, llegaran a la exclusión definitiva del sistema colonial (Villalpando-Canchola, 1992). A pesar de las campañas de exterminio dirigidas contra los Comcaa'c entre las décadas de 1760 y 1770, nunca fueron sometidos. Entre 1850 y 1860 sucedió lo conocido como las Guerras de Encinas, un intento fallido por incorporar a la comunidad Comcaa'c al trabajo. En esta guerra, se exterminó al menos a la mitad de la población Comcaa'c. La guerra, las enfermedades, y la inadaptación a la población extranjera, diezmó la población hasta llegar a ser insuficiente la cantidad de individuos para sostener las divisiones de clanes y territorios, por lo que la comunidad se refugió en la Isla Tiburón. De esta manera, ante la necesidad comunitaria y el sistema semi-nómada en colapso, los individuos flexibilizaron sus identidades de origen y se fusionaron en un único grupo. El resultado de dicho proceso es el grupo que ha subsistido hasta nuestros días *(Rentería Valencia, 2007)*.

El número de individuos Comcaa'c *(ver sección 4)* ha sido irregular a través de las épocas recientes. Las explicaciones a ello están relacionadas con la historia de la comunidad. Estos cambios demográficos son esenciales para entender algunas características poblacionales que han dado forma a los genomas modernos, como la presencia de alelos que parecen haber sufrido del efecto fundador *(Infante et al., 1999)*. En los últimos tiempos se ha abogado por una conservación de la diversidad genética nativa, cultural y antropológica de grupos nativos debido a su riqueza y particularidades que las hacen únicas a nivel internacional. Es de preocupación que durante años, el crecimiento de la población Comcaa'c se haya mantenido en relativo constante crecimiento y que en registros recientes, del año 2020 a la fecha, se haya visto una reducción poblacional notable. Este hecho podría estar relacionado con el sobrevenimiento de la emergencia sanitaria global, a las nuevas y ahora diferentes metodologías de censo utilizadas por las instancias gubernamentales o por procesos de migración.

2.2. Investigaciones previas de la genómica poblacional que incluyen a la población Comcaa'c

Las investigaciones del contexto genómico de los Comcaa'c son relativamente recientes. Los primeros artículos académicos que revisan estos aspectos de la comunidad aparecieron a finales de la década de los 90's *(fig. 7)*. Es hasta 1996 que se hace público el primer artículo que analiza aspectos genéticos en individuos Comcaa'c *(Hector et al., 1996)*. El trabajo analizaba por primera vez a 101 individuos Comcaa'c pertenecientes a 26 familias, enfocándose en la región MHC del brazo corto del cromosoma 6, una zona que se ha utilizado como referente genético debido a que contiene aproximadamente el 0,5% (> 150) de todos los genes codificadores de proteínas conocidos y que es significativa en cuestiones de trasplantes de órganos, estudios poblacionales, estudios de paternidad, e incluso en el análisis de enfermedades autoinmunes (*Charles A Janeway et al., 2001*). Si bien, los estudios genéticos sobre la comunidad Comcaa'c han tenido enfoques que se reducen a unos cuantos loci, permiten tener un panorama que ya habla de un contexto particular en comparación con otros grupos poblacionales.

Infante et al., 1999 expone la discusión en genética del grupo Comcaa'c. En este documento podemos vislumbrar por primera vez características distintivas a nivel genómico: en el análisis del MHCII, se encontró el alelo B27, nunca antes visto en comunidades previamente estudiadas a la fecha, probablemente presente, según el autor, debido al efecto fundador. Además, el alelo A33 estaba ausente en la comunidad Comcaa'c pero presente en todas las poblaciones indígenas antes estudiadas. En *Balladares et al., 2002*, además de la indagación en MHC y de los haplotipos extendidos HLA, se genotipificaron y compararon en individuos Comcaa'c los genes TAP1/TAP2, regiones del MHC II de interés que presentan alta variabilidad entre individuos y alta expresión génica. Todas las combinaciones de haplotipos encontrados en individuos Comcaa'c fueron principalmente de ascendencia nativa, con la excepción de un sujeto que portaba un haplotipo posiblemente caucásico.

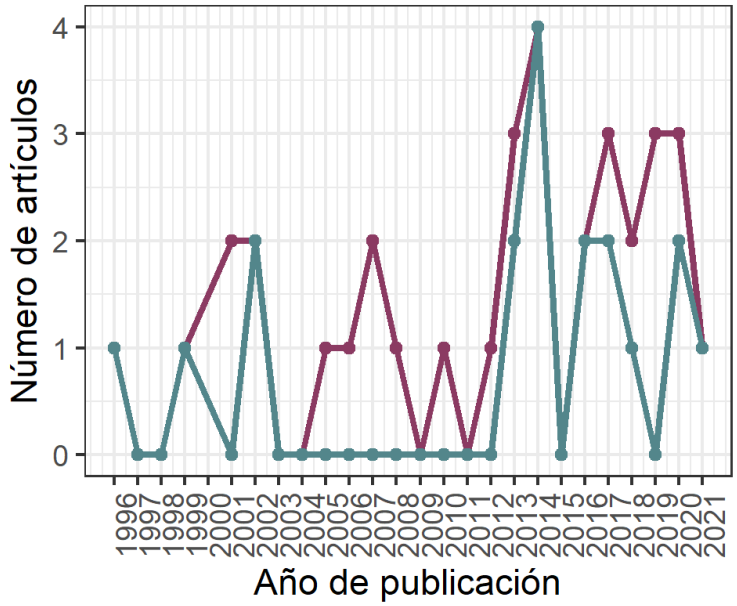


***Figura** 7. Relación de la cantidad de publicaciones académicas que describen el contexto genómico de los Comcaa'c a través de los años. En color verde muestra los artículos que presentan nuevas muestras de genotipificación y la línea color rojo muestra el total de artículos por año.*

Según análisis de genotipos HLA, Los Comcaa'c se encontraron altamente relacionados con las comunidades Guaraní, Wichi, Terena, Toba, Kaikang y Bari *donde además se* concluye que los nativos americanos muestran afinidades genéticas más fuertes con los asiáticos del noreste que las otras poblaciones importantes del mundo, en especial la comunidad Comcaa'c (*Infante et al., 1999)*. El vínculo genético entre los nativos americanos y los asiáticos orientales está respaldado por los patrones de distribución de los alelos y haplotipos HLA característicos (*Tokunaga et al., 2001*). Además, existe una tendencia al estudio de genotipificación de los genes HLA.

Por ejemplo, en un análisis de haplotipos mitocondriales de varias comunidades indígenas, en el que la porción del segmento hipervariable I (HVSI) de la región de control mitocondrial de 8 individuos Comcaa'c fue secuenciado, el grupo mostró una baja diversidad genética *(McCulloh et al., 2016)*, y las comparaciones por pares entre las poblaciones mexicano-amerindias geográficamente emparentadas han mostrado una gran diferenciación genética, especialmente entre las comunidades Tarahumara, Mayo, Comcaa'c y Guarijío *(Rangel-Villalobos et al., 2013)*, geográficamente cercanas. Además se ha identificado una variante de interés forense usando marcadores STR; cabe mencionar que en el mismo trabajo se identificaron grupos como el Cherokee, que mostraron hasta 10 alelos privados *(McCulloh et al., 2016)*.

Martínez-Cortés et al., 2019 también hace una comparación de los mismos STRs para un análisis interpoblacional entre mestizos y otras poblaciones nativas americanas para 21 loci STR autosómicos de interés forense. Curiosamente, la diversidad de este marcador entre los nativos americanos de origen mexicano también es mayor (aproximadamente el doble) que los

grupos étnicos norteamericanos y 38 veces mayor que el estimado entre los mestizos mexicanos.

La comunidad presenta el alelo 9RA, el alelo de 9 repeticiones del locus microsatélite D9S1120 característico de poblaciones nativas americanas *(Schroeder et al., 2007)*. La amplia distribución de un alelo particular privado a las Américas apoya la suposición de que gran parte de la ascendencia genética de los nativos americanos puede derivar de una sola ola de migración, y la presencia en la comunidad Comcáac es muestra de que a pesar de presentar un contexto genómico distintivo, derivan de un grupo genético compartido con otras comunidades nativas americanas, y que no parecen ser un grupo introducido por otro medio más que por la ola de migración proveniente del norte de América.

El cromosoma Y del grupo también ha sido estudiado *(Singh-Malhi et al., 2008)*. Se reportó un 100% de presencia del haplogrupo Q, predominante en América, sin presencia de haplogrupos como el R, que son distintivos de comunidades europeas.

Moreno-Estrada et al., 2014 identificó al grupo Comcaa'c como una de las 3 comunidades más aisladas genéticamente de todo México, mostrando una alta diferenciación poblacional junto con los Lacandones y Tojolabales, siendo también la única población nativa en el estudio que no mostraba ni un solo signo de mezcla con comunidades cosmopolitanas, genomas europeos u otras comunidades nativas. Los Comcaa'c y Lacandones, tomados en algunos trabajos como las comunidades nativas genéticamente más alejadas en México, muestran un nivel alto de diferenciación poblacional medido con el índice de fijación de Wright Fst, siendo éste más alto que el FST entre las poblaciones europeas y chinas en HapMap3, lo que nos habla de la alta diversidad genética existente entre comunidades nativas, debido probablemente a que los grupos han experimentado altos grados de aislamiento entre sí.

En su primer estudio de análisis de genoma completo, el grupo Comcaa'c (Seri) mostró el número más bajo de singletones de los grupos nativos mexicanos estudiados, y el número más alto de variantes novedosas, cifras que reflejan una población aislada. Como la mayoría de las comunidades nativas en el continente, cuando se infiere la autocigosidad usando series de homocigosidad (ROH), los Comcaa'c muestran tractos homocigotos largos: más del 10% del genoma en ROH. Es decir, que al menos el 10% del genoma nativo Comcaa'c es homocigoto. Estas poblaciones son relativamente pequeñas, lo que aumenta los efectos de la deriva genética e impulsa algunos de los valores altos de *Fst* (*Aguilar-Ordoñez et al., 2021*). Asimismo, el grupo se mostró como un clúster aislado en el análisis de ADMIXTURE. Otros trabajos en los que el grupo Comcaa'c se muestra en un clúster aislado de otras poblaciones es en *García-Ortiz et al., 2021* y *Moreno-Estrada et al., 2014.*

2.3. Caracterizaciones genéticas llevadas a la medicina genómica

Se ha invertido mucho esfuerzo en detectar variantes genéticas comunes asociadas con enfermedades complejas, multigénicas, o debido a variantes de nucleótido único y replicar asociaciones fenotipo-genotipo entre genes y poblaciones. Sin embargo, las variaciones

funcionales y médicamente relevantes pueden ser raras o específicas de las poblaciones, lo que requiere estudios de diversos grupos humanos para identificar nuevos factores de riesgo *(The SIGMA Type 2 Diabetes Consortium, 2014)*. En el pasado, incluso, se han intentado utilizar referencias de variantes clínicamente relevantes para poblaciones bien definidas y estudiadas, como la europea, para analizar la presencia de estos mismos rasgos en individuos Comcaa'c *(Costa-Urrutia et al., 2020)*. Los esfuerzos por relacionar las mismas variantes conocidas de otros grupos humanos con los fenotipos patogénicos en individuos Comcaa'c no han sido concluyentes.

2.3.1. Antecedentes de modelado de variantes en secuencias proteicas

Las proteínas son moléculas muy dinámicas, cuya función está intrínsecamente ligada a sus movimientos moleculares *(Rodrigues et al., 2018*). La variación genética entre los individuos puede generar diferencias de cambio de aminoácido en una sola posición de una proteína (*Strausberg et al., 2003*) o en varias posiciones consecutivas o separadas. Las mutaciones puntuales en la secuencia de una proteína podrían provocar un cambio o una pérdida de la estructura nativa, lo que a su vez puede causar cambios como la pérdida de la función o producir diferentes fenotipos. De otra manera, la estructura de la proteína puede adaptarse a una mutación reordenando el entorno espacial del residuo mutado. En el caso de estructuras menos densas, también es posible que una mutación se adapte sin causar ningún desplazamiento o distorsión. Todo depende de las propiedades de las cadenas radicales de los diferentes aminoácidos. Además de las variaciones naturales entre los individuos, los investigadores introducen con frecuencia sustituciones de residuos de aminoácidos por mutagénesis dirigida en el laboratorio para explorar las características estructurales y funcionales de las proteínas (*Feyfant et al., 2007*).

Cuando las mutaciones en una proteína aumentan la susceptibilidad o predisposición a una enfermedad se denominan mutaciones causantes de enfermedades o patogénicas (*Prabantu et al., 202*1). Los enfoques basados en la estructura para predecir el impacto de las mutaciones en la estabilidad utilizan información estructural de la proteína a partir del espacio 3D de una proteína plegada de forma nativa. Aunque estos métodos se basan esencialmente en los mismos datos estructurales, se construyen utilizando enfoques ampliamente diferentes y sofisticados, como los cálculos estadísticos de la energía de la función potencial (*Rodrigues et al., 2018*). A lo largo de los años se han desarrollado herramientas bioinformáticas, muchas de ellas fundamentadas en Machine Learning, que permiten estudiar las estructuras modeladas desde varios puntos interseccionales (*Frazer et al., 2021; K. Li et al., 2020; Pandurangan & Blundell, 2020).*

Existen antecedentes de trabajos que han modelado variaciones en proteínas buscando identificar si los cambios de aminoácidos en secuencias modifican las funciones estructurales, con objetivos variados como analizar uniones proteína-ligando, dinámicas moleculares, efectos de patogenicidad, termoestabilidad, entre otras (*Duan et al., 2020; Kurniawan & Ishida, 2022; Marks et al., 2012; S. Wang et al., 2007*). Por ejemplo, *Soto-Ospina et al., 2021* utilizó un modelo hipotético para analizar los efectos funcionales de las mutaciones Ala246GLu,

Leu248Pro, Leu248Arg, Leu250Val, Tyr256Ser, Ala260Val y Val261Phe, localizadas en el poro catalítico de la proteína Presenilina 1, una unidad catalítica de la γ-secretasa, para la cuál se han identificado más de 200 mutaciones patogénicas como causantes de la enfermedad de Alzheimer. Mediante la generación del modelo, se predijo una zona estructural que no había podido ser elucidada por métodos experimentales clásicos, y concluyó que sí existían cambios topológicos y electrostáticos en la estructura al introducir las variaciones, abriendo una ventana para entender cómo su estructura afecta a su función y a las del complejo γ-secretasa y sus cuatro subunidades

Otro ejemplo relacionado a la medicina humana es la identificación de 175 mutaciones no sinónimas en el ORF3a del SARS-CoV-2, para el que se estudiaron los efectos de estas mutaciones sobre la estabilidad estructural y las funciones de la estructura de la proteína ORF3a (*fig.8*) (*Hassan et al., 2021*).

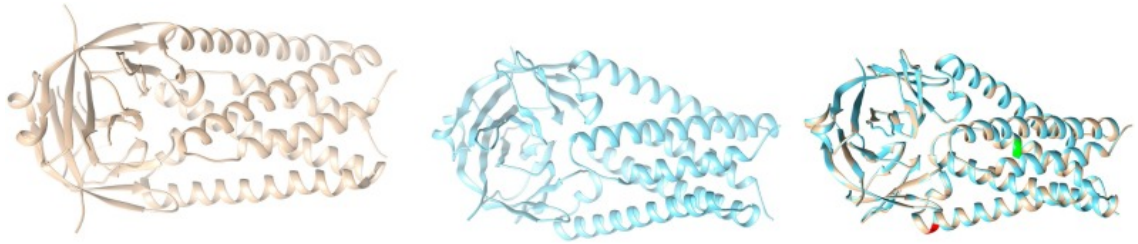


***Figura** 8. Estructuras de ORF3a (referencia coloreada en gris en la izquierda), estructura de ORF3a mutada (coloreada en azul en el centro), y ORF3a superpuesta (imagen de la derecha).*

Un aspecto a considerar, es que a lo largo de los años, los métodos de predicción estructural se han desarrollado y validado utilizando estructuras y mediciones biofísicas experimentales (Pan et al., 2022). Una gran proporción de estructuras proteicas aún no se ha elucidado experimentalmente y muchos estudios han basado sus conclusiones en predicciones realizadas mediante modelos de homología. La información generada por modelos no validados experimentalmente abre oportunidades y propuestas para la exploración experimental.

3. Marco conceptual

3.1. SNVs y Variantes particulares

Un polimorfismo de un solo nucleótido, o SNP (Single Nucleotide Polymorphism) es una variación en una única posición de una secuencia de ADN entre individuos. Los SNP pueden dar lugar a variaciones en la secuencia de aminoácidos, aunque también pueden encontrarse en regiones no codificantes del ADN (*Single Nucleotide Polymorphism / SNP* | *Learn Science at Scitable*, 2015). Para ser considerada un SNP, la variante debe estar presente en al menos el 1% de la población. En caso contrario, la variante puede ser llamada SNV (Single Nucleotide Variant) (*Types of Variants* | *Garvan Institute of Medical Research*, n.d.).

3.1.1. Singletons mundiales, novels y otras variantes

Los singletons mundiales pueden ser definidos como variantes encontradas únicamente en un individuo a nivel mundial al ser comparados con las bases de datos. Por otra parte, las variantes novel pueden ser definidas como aquellas que no se han reportado en las bases de datos públicas como dbSNP (Aguilar-Ordoñez et al., 2021).

3.1.2. Variantes Particulares por grupo/variante discernible

Se pueden definir como SNVs con sesgo de alta frecuencia. En el presente trabajo las SNVs particulares se definen para cada grupo nativo como aquellas que presentan una alta frecuencia (AF > 0.5) en el grupo estudiado, una AF < 0.05 a escala mundial según el *Proyecto 1000 Genomas* y *gnomAD 2.1 whole genome data*, y presente en menos del 5% del resto de las poblaciones NatAm en el estudio 100-GMx. La aparición de estos SNVs podría ser indicativa de una rápida diversificación, aislamiento y/o pequeños tamaños de población (Aguilar-Ordoñez et al., 2021; The 1000 Genomes Project Consortium, 2010).

3.1.3. Tipos de variantes anotadas por la herramienta EnsemblVEP

La herramienta de anotación EnsemblVEP identifica las variantes anotadas, de manera que realiza distinciones entre ciertos tipos de variantes. A continuación se describen brevemente.

- **Variantes intrónicas:** que ocurren dentro de un intrón
- **Variantes intergénicas**: Las regiones intergénicas se encuentran entre los genes que expresan proteínas. Son importantes porque pueden intervenir en la regulación del funcionamiento de los genes.
- **Variantes en regiones reguladoras:** Variantes que se encuentran en regiones reguladoras del genoma, tales como promotores, cajas, operadores, etc.
- **Variantes no-sinónimas:** Variantes que no alteran la secuencia de aminoácidos de una proteína.
- **Variantes start loss:** Estas mutaciones afectan al codón de inicio, es decir, al primer aminoácido de la proteína. Pueden tener efectos en la estructura final de la proteína

- **Variantes de terminación:** Una mutación de terminación (también llamada “sin sentido”, es un cambio en el ADN que hace que una proteína suspenda o termine su traducción antes de lo esperado.
- **Variantes sinónimas:** Las sustituciones sinónimas son aquellas que no alteran las secuencias de aminoácidos y son mutaciones silenciosas.

3.2. Modelamiento de proteínas

Las estructuras tridimensionales de las proteínas proporcionan conocimientos sobre su función a nivel molecular y mantienen varias aplicaciones en la investigación de las ciencias de la vida (Waterhouse et al., 2018). Existen métodos computacionales para el modelado de estructuras proteicas que pueden ser utilizados para complementar la determinación experimental de la estructura. Esto podría integrar un amplio espectro de cuestiones en la investigación biomédica. Los métodos más precisos en la actualidad se basan en el modelamiento por homología, es decir, en la detección de un homólogo de la secuencia objetivo deseada que pueda utilizarse como plantilla para el nuevo modelo. Sin embargo, la determinación experimental de la estructura es un factor limitante y, como consecuencia, el número de entradas en las bases de datos estructurales es mucho menor al número de secuencias de proteínas conocidas, incluídas sus isoformas (Studer et al., 2021).

3.2.1. Estimaciones de calidad del modelo (Model Quality Estimates)

Si bien los métodos de modelamiento han continuado en constante desarrollo hasta convertirse en líneas que pueden generar modelos para casi cualquier proteína de forma automática, la calidad de los modelos generados puede ser muy variable y difícil de predecir en ausencia de observables experimentales, de ahí la importancia de los métodos de estimación de la calidad (). Las estimaciones de calidad pueden ser estimaciones globales de modelos completos, por ejemplo para elegir el mejor modelo en un conjunto de alternativas, o estimaciones locales por residuo. Estas últimas permiten una selección más específica del modelo en los casos en que sólo interesa una parte concreta de la proteína, por ejemplo, un dominio que contiene un sitio activo o una variante de interés (*Studer et al., 2020*). La precisión de un modelo de proteína determina su idoneidad para las aplicaciones biomédicas. Sin embargo, en el momento del modelado, la calidad de un modelo es desconocida y debe predecirse también (*Benkert et al., 2011).*

3.2.2. Matriz BLOSUM de sustitución aminoacídica

BLOSUM (BLOcks of Amino Acid Substitution Matrix, o matriz de sustitución de bloques de aminoácidos) es una matriz de sustitución utilizada para el alineamiento de secuencias de proteínas *(Henikoff & Henikoff, 1992*). La familia de matrices de sustitución BLOSUM, y en particular BLOSUM62 *(fig. 9)*, son los parámetros de sustitución estándar de facto para los alineamientos de proteínas *(Song et al., 2015)*. Estas matrices nos permiten saber si dos secuencias son homólogas (relacionadas evolutivamente) o no, por lo que queremos una

puntuación de alineación que lo refleje. Estas puntuaciones también nos permiten saber si una sustitución aminoacídica es favorable o no evolutivamente (*Sean, 2004*). Números positivos indicarían una sustitución favorable. En cambio, resultados de números negativos señalaría una sustitución no favorable. Estos datos pueden ser calculados mediante una matriz.

	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
A	0	1	0	4																	A
G	-3	0	-2	0	6																G
P	-3	-1	-1	-1	-2	7															P
D	-3	0	-1	-2	-1	-1	6														D
E	-4	0	-1	-1	-2	-1	2	5													E
Q	-3	0	-1	-1	-2	-1	0	2	5												Q
N	-3	1	0	-2	0	-2	1	0	0	6											N
H	-3	-1	-2	-2	-2	-2	-1	0	0	1	8										H
R	-3	-1	-1	-1	-2	-2	-2	0	1	0	0	5									R
K	-3	0	-1	-1	-2	-1	-1	1	1	0	-1	2	5								K
M	-1	-1	-1	-1	-3	-2	-3	-2	0	-2	-2	-1	-1	5							M
I	-1	-2	-1	-1	-4	-3	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-1	-4	-3	-4	-3	-2	-3	-3	-2	-2	2	2	4					L
V	-1	-2	0	0	-3	-2	-3	-2	-2	-3	-3	-3	-2	1	3	1	4				V
W	-2	-3	-2	-3	-2	-4	-4	-3	-2	-4	-2	-3	-3	-1	-3	-2	-3	11			W
Y	-2	-2	-2	-2	-3	-3	-3	-2	-1	-2	2	-2	-2	-1	-1	-1	-1	2	7		Y
F	-2	-2	-2	-2	-3	-4	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	1	3	6	F
	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	

***Figura** 9. Matriz de sustitución BLOSUM-62. Recuperada de Wikicommons.*

3.2.2.1. GMQE, QMEAN y QMEANDisCo

3.2.2.1.1. Global Model Quality Estimate (GMQE)

En el proceso de modelamiento, se realiza una búsqueda primaria de estructuras candidatas para ser plantillas de nuestro nuevo modelo. Una vez finalizada la búsqueda de plantillas, éstas se clasifican según la calidad esperada de los modelos resultantes, estimada por Global Model Quality Estimate (GMQE) y otros estadísticos. El GMQE puede ser utilizado para descartar y escoger las mejores estructuras plantilla para los experimentos de modelado de secuencias (*Waterhouse et al., 2018*).

3.2.2.1.2. QMEAN

El estadístico QMEAN (Qualitative Model Energy ANalysis) es una función de puntuación compuesta que describe los principales aspectos geométricos de las estructuras proteicas (*Benkert et al., 2008*). QMEAN es una puntuación compuesta que califica un modelo con base en las características estructurales predichas a partir de la secuencia *(Studer et al., 2020).*

3.2.2.1.3. QMEANDisCo

Studer et al., 2020 introduce una nueva puntuación de restricción de distancia (DisCo) al estadístico QMEAN. La puntuación DisCo puede considerarse una puntuación de modelo casi

único, ya que deriva su propia información de conjunto con una búsqueda de homología basada en la secuencia, como se ilustra en la *fig. 10* El uso directo de estructuras homólogas permite mantener el tiempo de cálculo bajo, ya que no hay que construir modelos. Si existen muchos homólogos cercanos, se espera que DisCo sea muy preciso. Sin embargo, la precisión disminuye si se pueden identificar pocos o ningún homólogo cercano. El resultado es una puntuación compuesta para la estimación precisa de la calidad local y global del modelo.

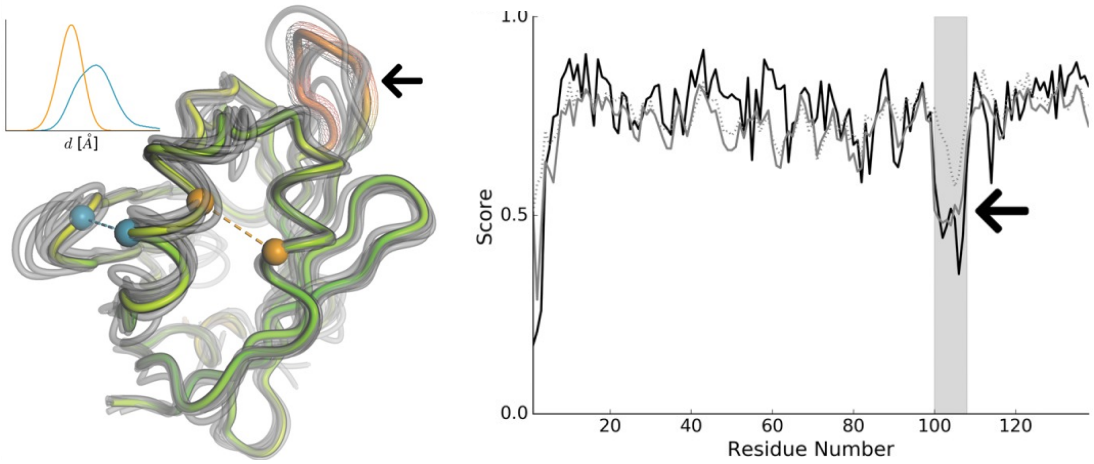


***Figura** 10. Izquierda. Las puntuaciones de QMEANDisCo por residuo se mapean como un gradiente de color rojo a verde en un modelo. Un subconjunto representativo de homólogos se ilustra en gris. Derecha. Los puntos grises representan los valores locales de QMEANDisCo. En ambas imágenes, la flecha muestra el mismo residuo Recuperado de Studer et al., 2020*

4. Planteamiento del Problema

Los Comcaa'c son una comunidad pequeña, con un número aproximado de 733 individuos para 2020, según el Censo de Población y Vivienda 2020 publicado por el INEGI *(fig. 11)*. Como contraste, existen según los artículos revisados, 652 muestras que han sido utilizadas para trabajos de genéticos y genómicos. El grupo muestra particularidades histórico-demográficas que podrían estar relacionadas a las evidencias genéticas encontradas en anteriores investigaciones. Debido a la baja cantidad de individuos, la presencia de variantes que podrían generar cambios fenotípicos podrían impactar más ampliamente en la población, incluyendo variantes perjudiciales o patogénicas.

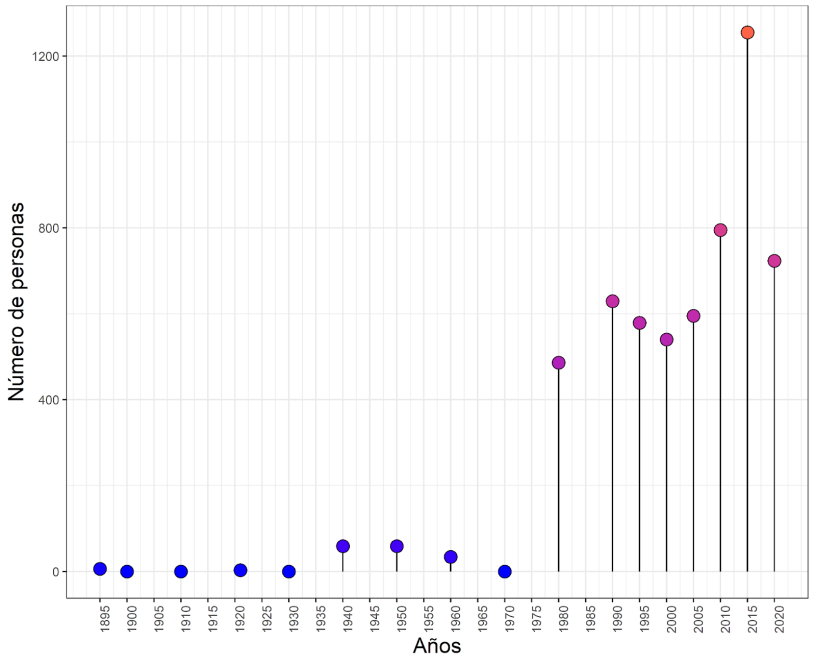


Figura 11. *El gráfico correlaciona los años de censo con el número de hablantes de lengua Comcaa'c y/o individuos pertenecientes a alguna comunidad indígena Comcaa'c. El gráfico ha sido construido con datos de diversos tabulados de censos poblacionales generados por el Instituto Nacional de Estadística y Geografía desde 1895 a 2020. A esta gráfica se le agregó el dato único de conteo poblacional indígena generado por el Instituto Nacional de los Pueblos Indígenas para el reporte de Indicadores Socioeconómicos de los Pueblos Indígenas de México en 2015.*

La *fig. 12* muestra comunidades nativas estudiadas por *Moreno-Estrada et al., 2014*. Cada nodo representa un genoma haploide, y la propagación de cada grupo es indicativa del nivel de parentesco en cada población. La distribución de los genomas haploides del grupo Comcaa'c muestra a una comunidad con alto nivel de parentesco y alta consanguinidad. A pesar de la existencia de trabajos que en conjunto forman un mosaico contextual del genoma Comcaa'c, no existe un análisis integral que permita las pruebas de hipótesis, las conclusiones aplicativas y el discernimiento biomédico.

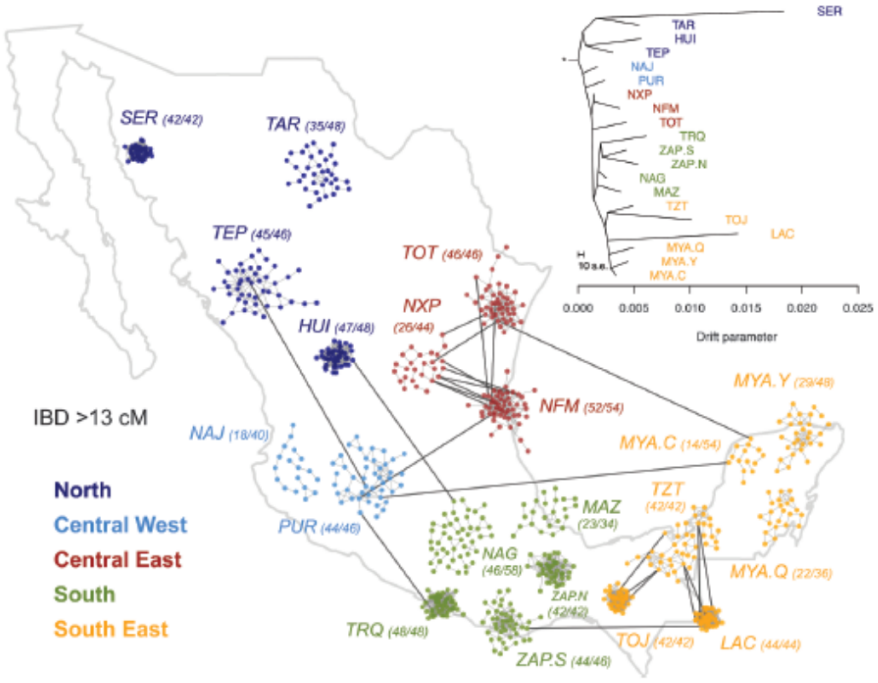


Figura 12. *Izquierda. Localización aproximada de las comunidades indígenas estudiadas. Cada nodo representa un genoma haploide. La comunidad Seri (Comcaa'c) se encuentra representada en la parte superior izquierda en azul con las letras SER. Derecha. El árbol representa los patrones de división de la población de 20 grupos nativos, donde la longitud de la rama es proporcional a la deriva de cada población. Se utilizaron muestras africanas, europeas y asiáticas como grupos externos. Con alta deriva génica, los Comcca'c o Seri se muestran en la rama más superior, en color azul oscuro, y perteneciente al clúster del Norte. Tomado de Moreno-Estrada et al., 2014.*

Como se comentó en la sección 2.2 , los individuos Comcaa'c presentaron altos valores de divergencia genética respecto a otras comunidades nativas, algo previamente observado también por *Moreno-Estrada et al., 2014*. La *fig. 12* muestra dos representaciones gráficas de la divergencia genética entre grupos nativos. A la derecha se muestra un mapa con nodos representando un genoma haploide. La comunidad Seri (Comcaa'c) se encuentra representada en la parte superior izquierda en azul con las letras SER. A la izquierda, el árbol representa los patrones de división de la población de 20 grupos nativos, donde la longitud de la rama es proporcional a la deriva de cada población. Se utilizaron muestras africanas, europeas y asiáticas como grupos externos. Con alta deriva génica, los Comcca'c o Seri se muestran en la rama más superior, en color azul oscuro, y perteneciente al clúster del Norte.

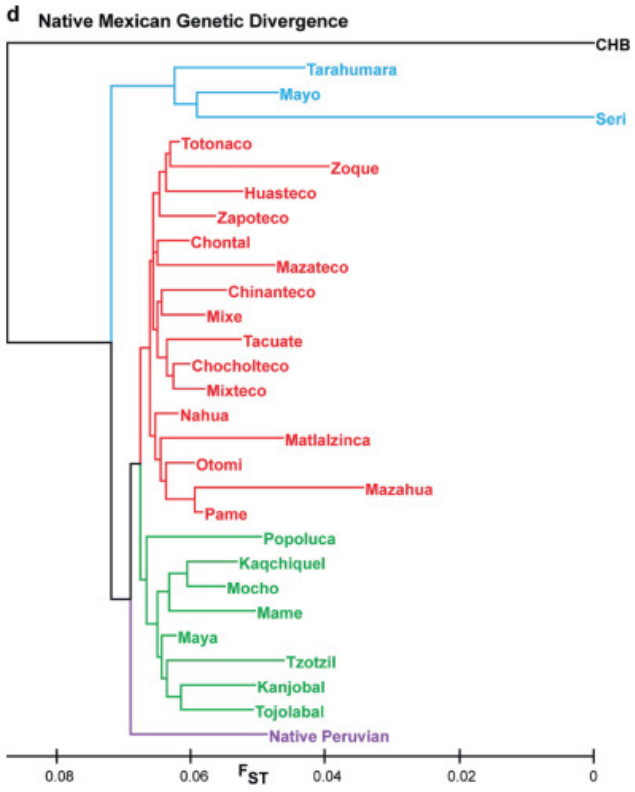


***Figura** 13. Árbol Neighbor-joining basado en el FST entre los 27 grupos Nativos mexicanos y NP de nuestro estudio; los colores indican la región de procedencia. El grupo Seri se encuentra en azul, perteneciente a la Región norte. Es la comunidad más aislada de los grupos analizados en este árbol filogenético. Recuperado de Aguilar-Ordoñez et al., 2021*

La *fig. 13*, tomada de *Aguilar-Ordoñez et al., 2021* es un árbol construido con datos de genoma completo, en el que, de manera similar a la *fig.12*, el grupo Seri (o Comcaa'c) se muestra como la rama más divergente del árbol y perteneciente al cluster del norte. En este mismo estudio, los genotipos específicos de la población Comcaa'c fueron 2.496 SNV particulares, 8 de ellos fijados (presentes en los 4 individuos estudiados), exponiendo un contexto genómico distintivo respecto a las otras poblaciones estudiadas. Con esta información es posible explorar el impacto que tiene el contexto genómico peculiar de los genomas Comcaa'c en la estructura y función de algunas proteínas. Estas variantes encontradas en los individuos Comcaa'c, nunca antes vistas en ningún otro grupo a nivel mundial, podrían caer en regiones expresables que podrían modificar estructuras tridimensionales, y tal vez, funciones en fenotipos expresables. Debido a su naturaleza única en el mundo, probablemente no existan registros en las bases de datos. Además, debido a la baja cantidad de individuos existentes, una variante de interés podría estar en gran parte de la población nativa. Con estos datos de genoma completo, sería posible definir y contextualizar las mismas variaciones atípicas que muestran los análisis previos del patrimonio genético de la comunidad nativa Comcaa'c.

Este trabajo podría fungir como precedente de un estudio de análisis estructural de variantes particulares en grupos particulares, especialmente de otros grupos nativos con características similares a las de la comunidad aquí estudiada.

5. Hipótesis

Existen variantes particulares del genoma Comcaa'c que provocan cambios estructurales en las proteínas afectadas.

6. Objetivos

- Realizar un análisis del contexto genómico de la comunidad nativo-americana Comcaa'c
    - Determinar las variantes particulares del grupo Comcaa'c, mediante análisis de 4 secuencias de genomas completos
- Analizar el impacto estructural de las variantes particulares codificantes.
    - Análisis de la secuencia primaria y estructura tridimensional de las secuencias de los transcritos canónicos de variantes no-sinónimas anotadas.
    - Modelar estructuras proteicas mediante métodos de biología computacional para las variantes antes mencionadas.

7. Métodos

Como parte de un estudio previo, se secuenciaron los genomas de 2 mujeres y 2 hombres Comcaa'c mediante tecnología Illumina HiSeq X Ten, en el que aproximadamente el 95 % del genoma GRCh38 se cubrió con una profundidad media de 24X. Vale la pena señalar que los grupos étnicos se identificaron principalmente por su idioma (*Aguilar-Ordoñez et al., 2021).*

Como parte de este trabajo, un archivo VCF (Variant Calling Format) con los datos de los 4 genomas fue pre-tratado en un pipeline de Nextflow (El nf puede ser leído y descargado en: https://github.com/laguilaror/nf-VEPextended). Posteriormente, el VCF resultante fue utilizado como INPUT para otra herramienta que realiza un recuento y resumen de las variantes del archivo previamente tratado por VEPextended que genera un archivo tsv con el número total de SNV e indels, un archivo tsv con los recuentos por muestra de las variantes de tipo SNV, indel, novel, worldwide singletons, clinvar, gwascat y pharmgkb y un archivo pdf con el número de variantes discernibles en grupos de muestras de interés (El nf puede ser leído y descargado en: https://github.com/laguilaror/nf-100GMX-variant-summarizer). Con lo anterior se obtuvo un archivo VCF con las variantes particulares para el grupo Comcaa'c.

Este VCF contiene variantes tipo SNP, indel, y otras. Se utilizaron comandos de bcftools para dividir el archivo original y de esta manera obtener 3 VCFs distintos, uno conteniendo únicamente SNPs, uno conteniendo únicamente Indels y otro con otras variantes no-SNP y no-Indel. En este proyecto, solo se anotaron los datos de SNPs (SNVs). La anotación fue

realizada con la herramienta EnsemblVEP online (la herramienta puede ser utilizada en https://www.ensembl.org/). Esto generó una tabla con los SNV particularea. La anotación reportó, en forma tabular:

- Localización de la variante
- Alelo referencia y alelo anotado
- Consecuencia de la variante
- Nombre del gen
- Posición cDNA
- Posición en la secuencia proteica
- Codones de cambio
- Aminoácido mutado
- Secuencias canónicas
- Entrada en Swiss-prot y Uniprot
- Isoformas de la variante
- Frecuencias alélicas (si están reportadas)
- Otros datos

Para manejar la gran cantidad de datos generados con la anotación, se desarrolló un script en R (El script, los inputs y los archivos de salida pueden verse y descargarse desde https://github.com/paulinapglz99/proteomic_and_genomic_context) que permite filtrar las variantes, de manera que se obtengan aquellas que generan cambios en residuos de proteínas, que están anotadas como el transcrito canónico y que muestran un acceso del transcrito en UNIPROT, de manera que se pueda obtener la secuencia exacta para los posteriores procesos de modelado.

Se descargaron las secuencias de UNIPROT mediante la API programática en bash y posteriormente, se realizaron las mutaciones correspondientes en las secuencias. Luego se realizaron 14 modelos en SwissModel Expasy (https://swissmodel.expasy.org/) , uno para cada secuencia nativa y una secuencia con la variante anotada previamente.

Para modelar las estructuras se escogió el template con mejor puntaje GMQE y mejor resultado de QMEANDiscO global estructural. Se identificó la posición del aminoácido de interés y se realizó un análisis estructural del Local Quality Estimate. Los modelos que no mostraron relevancia estructural o baja calidad local de modelaje en la posición del aminoácido de interés, fueron descartados para mayores análisis. Por último, se utilizó la herramienta Conserved Domain Search para determinar dominios de interés en estructuras. La *fig.14* muestra un resumen gráfico de la metodología utilizada en este trabajo.

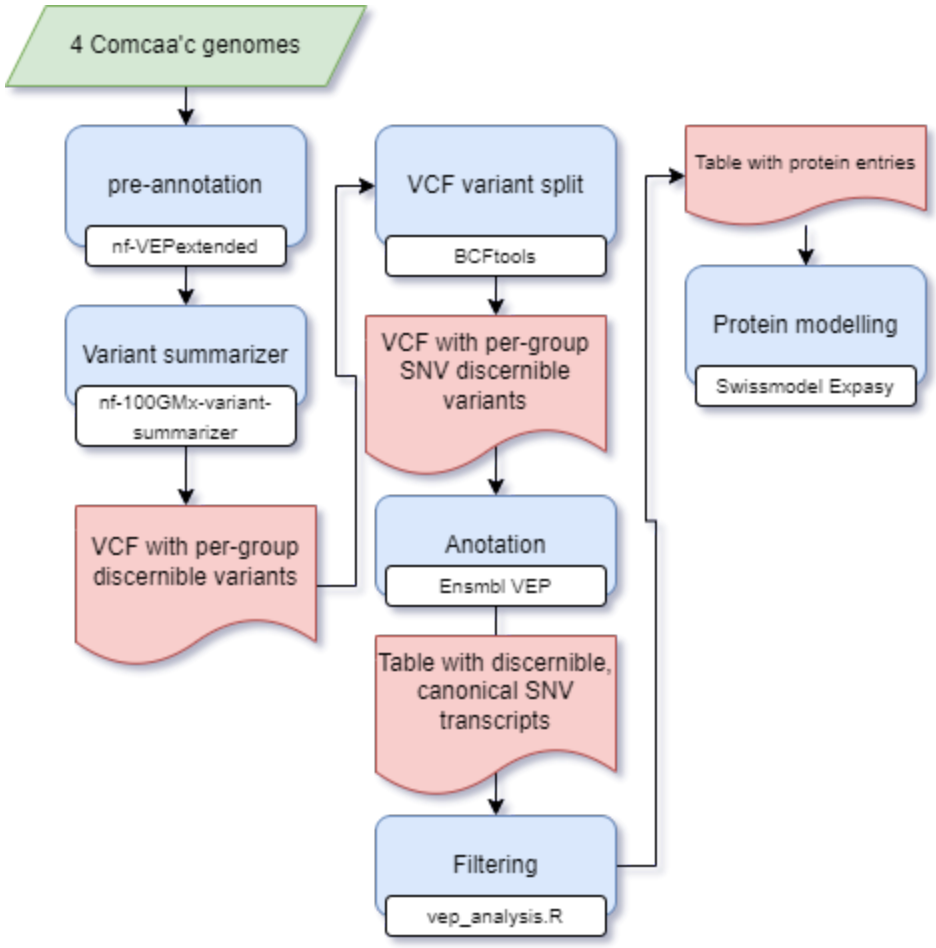


***Figura** 14. Diagrama de flujo de la metodología utilizada para determinar cambios estructurales en proteínas a partir de variantes particulares de la población Seri.*

8. Resultados

8.1. Variantes particulares del grupo Comcaa'c

Se encontraron un total de 1,970 SNVs particulares (https://github.com/paulinapglz99/proteomic_and_genomic_context/blob/main/outputs/SNPs_variantes_Comcaac.csv). Es de notarse que del total de SNVs anotados codificantes, el 65% de ellos corresponde a variantes que generan cambios de aminoácido en alguna secuencia proteica. La *fig. 15* muestra un gráfico con el número de variantes codificantes por cromosoma. Los cromosomas con más variantes fueron el C7 y C3 con 32 variantes, el C1 con 29 variantes y el C12 con 24 variantes SNV particulares. Los cromosomas C5, C6 , C9, C10, C13, C15, C16, C17, C18, C20 y no mostraron variantes SNVs particulares para el grupo de estudio.

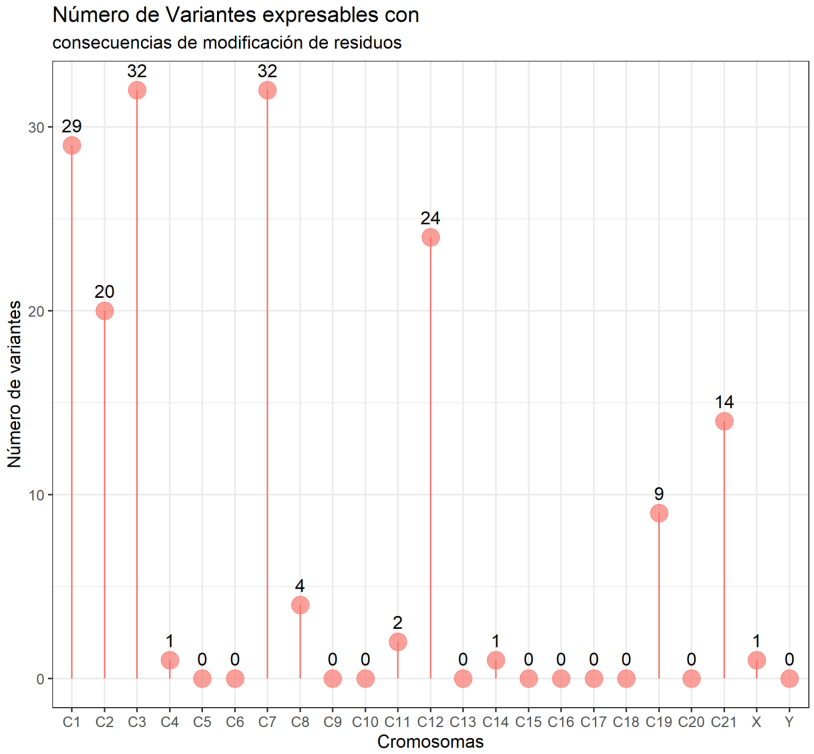


Figura 15 *. Cantidad de SNVs particulares de la población Seri, por cromosoma.*

Por otro lado, la *fig.* 16 muestra la relación de los tipos de consecuencias de los SNV particulares. Las variantes intrónicas conforman el 52% de las variantes anotadas. El 18% corresponde a variantes transcritas no codificantes y el resto corresponde a variantes en genes downstream, variantes intergénicas, variantes en transcrito NMD (variantes en transcritos objetivo de nonsense-mediated mRNA decay), variantes en transcritos exónicos no codificantes, variantes en regiones reguladoras, variantes en sitios de unión a factores de transcripción (TF), y variantes en genes upstream. La *fig. 17* muestra a su vez, las consecuencias de las variantes particulares **codificantes**.

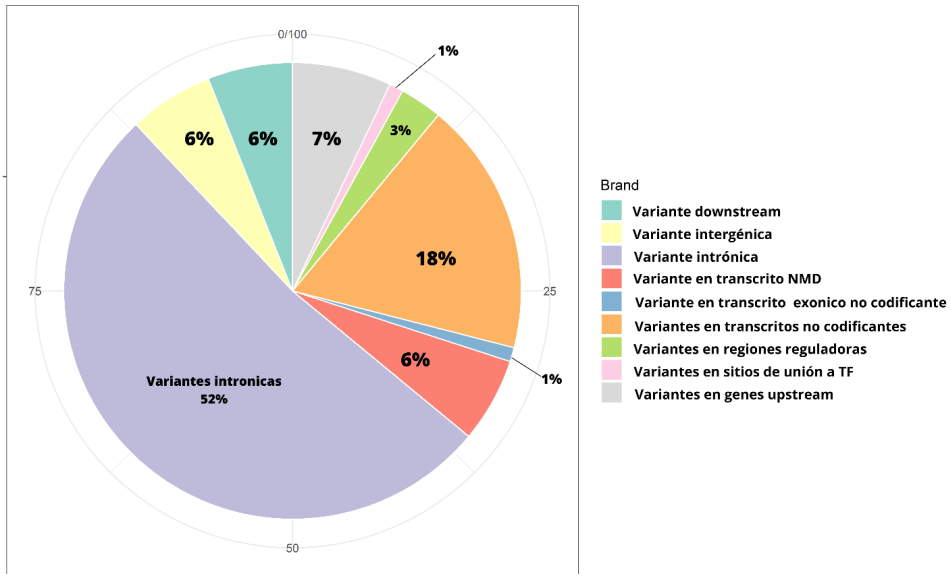


***Figura** 16.Consecuencias de los SNVs, incluyendo las codificantes y no-codificantes. La mayor parte de las variantes anotadas (52%) corresponde a variantes intrónicas.*

Los resultados de la anotación mostraron únicamente la presencia de variantes no sinónimas y variantes sinónimas. Hay que recordar que estas variantes únicamente corresponden a SNVs particulares codificantes del genoma. Los genes afectados por las variantes no-sinónimas son en total 22: MCOLN3, OR2T29, CEP290, TRAV8-7, HCN2, PEAK3, C19orf47, ERCC2, SIGLEC14, ATAD2B, ZSWIM2, SLC37A1, INPP4B, PDLIM2, MROH5, AGRN, PLA2G2F, PUM1, DNHD1, ATP2B2, LMOD3 y MAGI2.

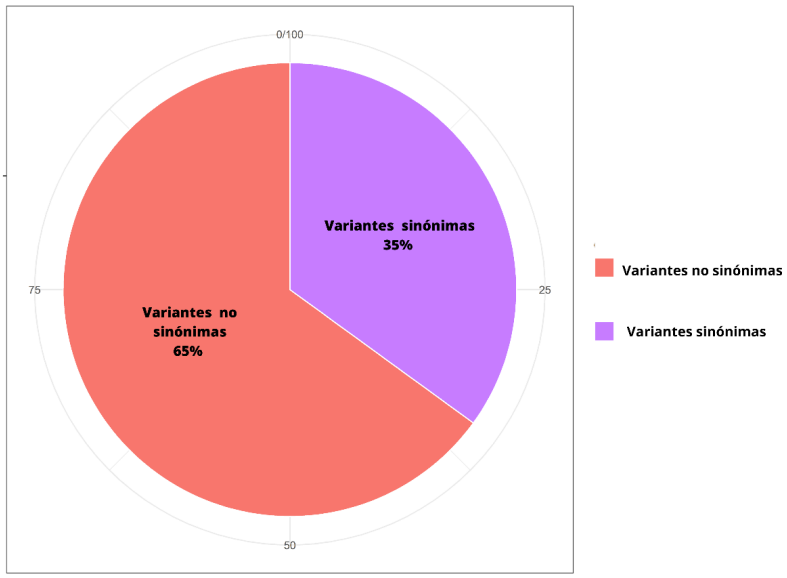


***Figura** 17. Consecuencias codificantes de los SNVs particulares encontrados en los 4 genomas Comcaa'c.*

Una vez identificadas las variantes codificantes, se utilizó un script en R para encontrar un identificador de acceso para UNIPROT. Solo se encontraron 7 genes variantes con identificadores UNIPROT, que además cumplieran con todos los requisitos previos de selección. Las siguientes secciones describen los posibles cambios estructurales.

8.2. Modelos realizados según SNVs particulares de interés

Se validaron los aminoácidos de referencia en las posiciones indicadas para cada una de las secuencias obtenidas mediante la API de Uniprot, y posteriormente se modificaron desde el archivo .fasta. Las mutaciones correspondientes se realizaron según la *tabla 1*. Posteriormente, se realizaron 14 modelos para 14 secuencias, de las cuales, 7 corresponden a las secuencias nativas y 7 corresponden a secuencias con los aminoácidos variantes anotados por la herramienta. Se predijeron las estructuras tridimensionales de **AGRN** (UNIPROT: O00468-6), **PLA2G2F** (Q9BZM2-2), **PUM1** (Q14671-3), **DNHD1** (Q96M86-3), **ATP2B2** (Q01814-1), **LMOD3** (Q0VAK6-1) y **MAGI2** (Q86UL8-1). Para cada proteína se predijo un modelo invariante (referencia) y un modelo variante (con el cambio de aminoácidos).

***Tabla** 1. Mutaciones predichas en los genomas estudiados.*

Nombre del gen	Isoforma en Uniprot	Posición de mutación en proteína	Variación existente (dbSNP)	Aminoácido de referencia	Aminoácido variante
AGRN	O00468-6	1012	rs144781935	A	T
PLA2G2F	Q9BZM2-2	75	rs370472527	V	I
PUM1	Q14671-3	1142	rs772139393	V	I
DNHD1	Q96M86-3	1714	rs201274362	S	L
ATP2B2	Q01814-1	61	rs772521297	K	R
LMOD3	Q0VAK6-1	495	rs780071005	R	H
MAGI2	Q86UL8-1	1013	rs773558417	N	K

8.2.1. AGRIN (Proteína Agrina)

La Agrina *(fig. 18)* es un proteoglicano de sulfato de heparán que se requiere para la formación y el mantenimiento de las uniones neuromusculares y que dirige eventos clave en la diferenciación postsináptica. Durante el desarrollo, las neuronas motoras secretan agrina para desencadenar la agregación local de receptores de acetilcolina (AChR) y otras proteínas en la fibra muscular, que juntas componen el aparato postsináptico *(Denzer et al., 1997).*

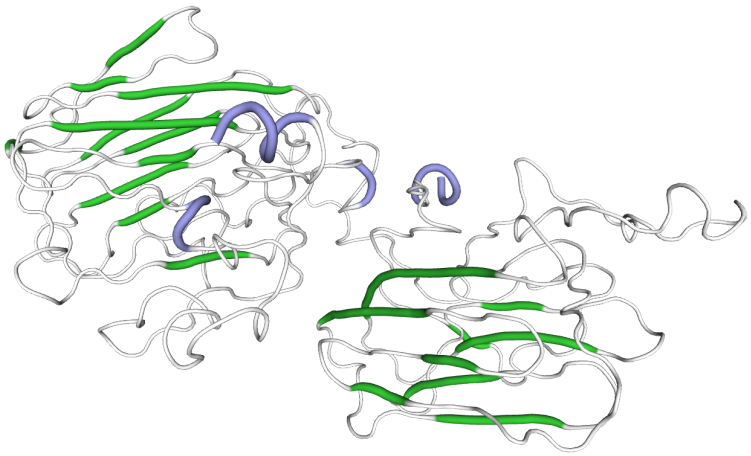


***Figura 18**. Modelo generado por Swiss-Model, basado en el template 6cw1.1.A Los colores violetas muestran zonas plegadas en alfa hélice, y las zonas blancas muestran loops. Recuperado de UNIPROT: O00468.*

8.2.1.1. Modelo de la Agrina | AGRIN

La *fig. 19* muestra el modelo predicho para la secuencia de la Isoforma 6 de la Agrina (O00468-6), también denominada “Agrin y(0)z(0)”, ya que carecen de inserto 'z'. Esta isoforma derivada del splicing alternativo es específica del músculo y puede estar involucrada en la diferenciación de células endoteliales (*AGRN - Agrin - Homo Sapiens (Human)* | *UniProtKB* | *UniProt*, n.d.). Para esta secuencia de proteína en específico no existen modelos experimentales, únicamente predicciones con AlphaFold. El modelo aquí presentado fue modelado con la estructura de la neurexina 1 alfa (3r05.1.A), una proteína elucidada mediante difracción de rayos X a 2.95 Å con el que mantiene una identidad de secuencia del 23.61% y un valor GMQE de 0.14. La secuencia insertada tiene una longitud de 2045 aminoácidos (aa), mientras que el modelo tiene una longitud de 654 aa en el rango 1372-2026, que corresponde a una cobertura de 0.29. La plantilla modelo sugiere que la proteína cuenta con uniones no covalentes con moléculas de NAG(2-acetamido-2-deoxy-beta-D-glucopyranose-(1-4)-2-acetamido-2-deoxy-beta-D-glucopyranose). Información más detallada puede encontrarse en la *tabla 2*.

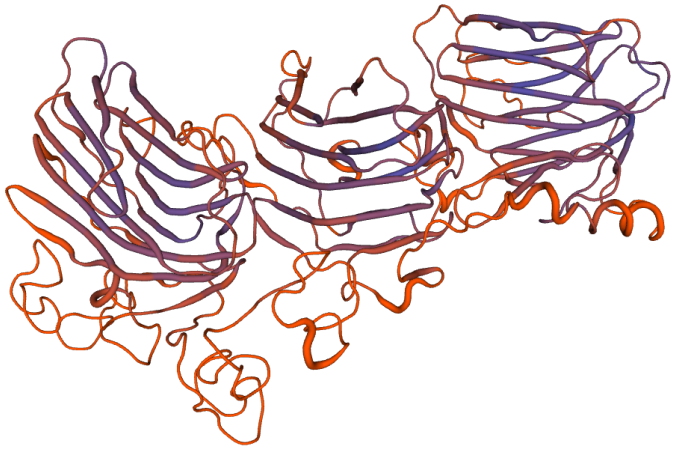


Figura 19..*Estructura modelada para la secuencia O00468-6 correspondiente a la Isoforma 6 el transcrito de la proteína Agrina (AGRN). Las estructuras en color azul muestran una mejor confianza para el modelo. Las estructuras en naranja muestran valores más bajos de confianza para la región modelada.*

La variante registrada para proteína Agrina fue descartada para mayores análisis debido a que ninguno de los modelos realizados en Swiss-prot contuvo la posición del residuo de interés. Es por ello que no se muestra el modelo de la variante insertada.

Tabla 2 *Especificaciones y características para el modelo estructural de la secuencia nativa de O00468-6, correspondiente a la proteína Agrin y(0)z(0)*

Características del modelo para AGRIN	
Template usado	6a69.1.A Plasma membrane calcium-transporting ATPase 1
Método	Microscopía electrónica 4.11 Å
Oligo-State	Monomer
Modelo nativo	
GMQE	0.68 (nat)
QMEANDisCo Global	0.73 ± 0.05 (nat)
Confidence position	0.27 (nat)
Modelo de sustitución BLOSUM	V / V / 4
Modelo variante	
GMQE	0.68 (mut)

QMEANDisCo Global	0.73 ± 0.05 (mut)
Confidence position	0.17
Modelo de sustitución BLOSUM	I / V / 3

8.2.2. PLA2G2F (Fosfolipasa A2 dependiente del calcio (PA2GF)) (UNIPROT: Q9BZM2-2)

La proteína PA2GF (*fig. 20*), derivada del gen PLA2G2F, es una fosfolipasa A2 dependiente del calcio que metaboliza principalmente fosfolípidos extracelulares. Hidroliza el enlace éster del grupo acilo graso unido en la posición sn-2 de los fosfolípidos, incluyendo moléculas como fosfatidilgliceroles, fosfatidiletanolaminas, fosfatidilcolinas, fosfatidilserinas, etc.(Valentin et al., 2000). Además, por similitud con otras proteínas, se predice que pueda desempeñar un papel en la producción de mediadores lipídicos en condiciones inflamatorias, proporcionando ácido araquidónico a ciclooxigenasas y lipooxigenasas.

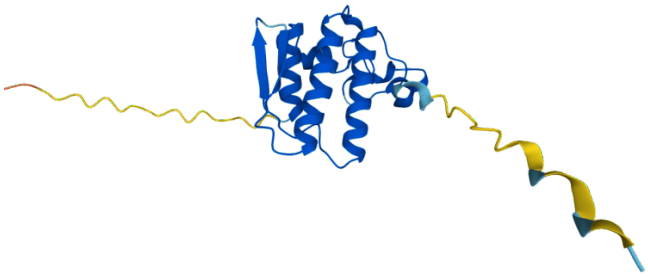


***Figura** 20. Predicción estructural creada por AlphaFold. Número de identificador AlphaFold AF-Q9BZM2-F1. Los colores azules muestran zonas con buena confianza de modelado. Las zonas amarillas muestran una zona con una baja confianza de modelado. Recuperado de https://alphafold.ebi.ac.uk/search/text/%20AF-Q9BZM2-F1*

8.2.2.1. Modelo de PA2GF | PLA2G2F

La *fig. 21* muestra el modelo predicho para la secuencia de Q9BZM2-2, correspondiente a la Isoforma 2 de PA2GF, la proteína codificada por el gen PLA2G2F. Para esta secuencia de proteína en específico no existen modelos experimentales, únicamente predicciones con AlphaFold. La predicción estructural aquí presentada fue modelada con la estructura de la Fosfolipasa IIE secretada humana A2 (5wzs.1.A) como template, una proteína reportada como elucidada mediante difracción de rayos X a 2.30 Å que tiene una identidad de secuencia de 45.45%, y un valor GMQE de 0.42 para la secuencia insertada, la cual tiene una longitud de 211 (aa), mientras que el modelo tiene una longitud de 466 aa en el rango 64-188, que corresponde a una cobertura de 0.57. La plantilla modelo sugiere que la proteína cuenta con uniones no covalentes con iones de calcio, cloro y moléculas como 7W9 (2-[2-methyl-1-(naphthalen-1-ylmethyl)-3-oxamoyl-indol-4-yl]oxyethanoic acid). Información más detallada para ambos modelos puede encontrarse en la *tabla 3.*

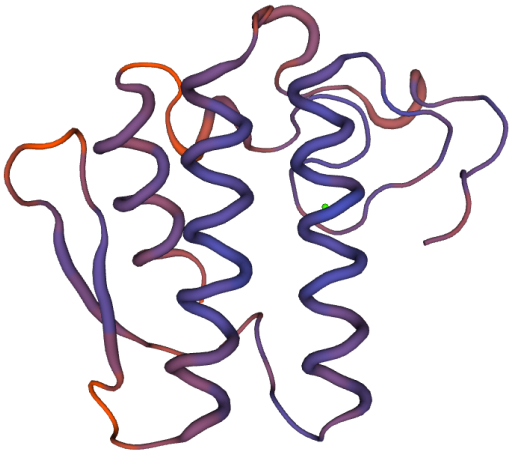


***Figura 21**. Modelo de la proteína PA2GF para la secuencia del transcrito Q9BZM2-2 correspondiente a la Isoforma 2 del transcrito del gen PLA2G2F. La estructura 5wzs.1.A fue utilizada como template. Las estructuras en color azul muestran una mejor confianza para el modelo. Las regiones en color naranja muestran valores más bajos de confianza para la región modelada.*

La *fig. 22* compara el modelo generado para la secuencia nativa o invariante, el cual muestra una valina en la posición 75 y a la derecha, el modelo generado para la proteína con la variación insertada, que presenta una Isoleucina. La proteína fue descartada para mayores análisis debido a que el aminoácido no parece generar cambios estructurales relevantes.

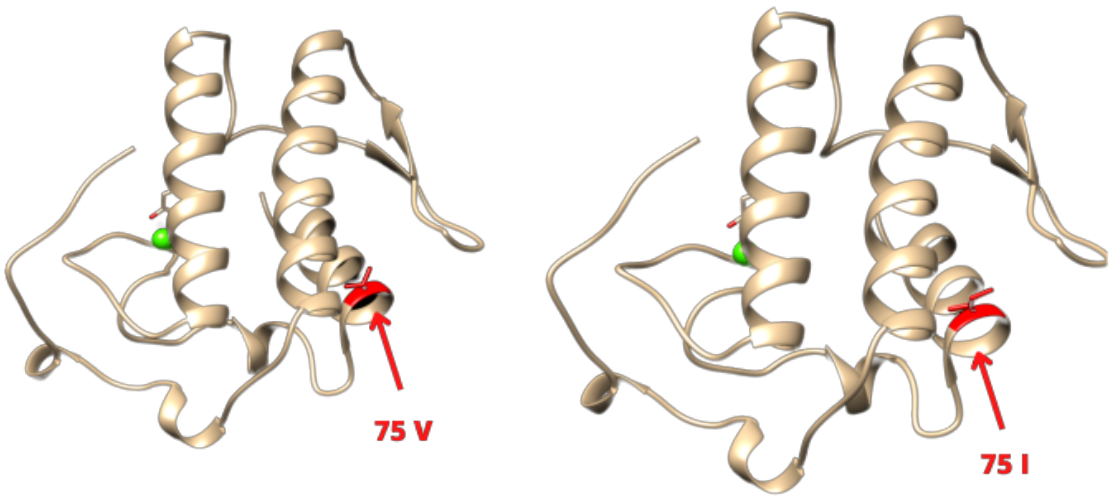


***Figura** 22. Izquierda. Modelo estructural de PA2GF generado para la secuencia nativa, el cual muestra una valina en la posición 75. Derecha. Modelo para la PA2GF con la variación insertada, correspondiente a una Isoleucina. Ambos modelos muestran interacciones no-covalentes con un ión calcio, mostrado en esfera de color verde.*

***Tabla** 3. Especificaciones y características para el modelo estructural de la secuencia nativa y variante de Q9BZM2-2, correspondiente a la proteína PA2GF.*

Características del modelo para PLA2G2F	
Template usado	5wzs.1.A Group IIE secretory phospholipase A2. Crystal structure of human secreted phospholipase A2 group IIE with Compound 8
Method	X-RAY DIFFRACTION 2.30 Å
Oligo-State	Monomer
Modelo nativo	
GMQE	0.42
QMEANDisCo Global	0.70 ± 0.07
Confidence position	0.71
Modelo de sustitución BLOSUM	V / M / 1
Modelo variante	
GMQE	0.42
QMEANDisCo Global	0.69 ± 0.07
Confidence position	0.75
Modelo de sustitución BLOSUM	I / M / 1
Alineamiento estructural	
Confidence	0.71 (nat) / 0.71 (var)
Consistency	1.00 (nat) / 1.00 (var)

8.2.3. PUM1. Proteína Pumilio 1 (UNIPROT: Q9BZM2-2)

PUM1 *(fig. 23)* es una proteína de unión a ARN de secuencia específica que actúa como represor post-transcripcional al unirse al 3'-UTR de las dianas de ARNm (Filipovska et al., 2011). Es un mediador de la represión post-transcripcional de los transcritos a través de diferentes mecanismos. PUM1 se une con alta afinidad a la secuencia consenso UGUANAUA (Van Etten et al., 2012).

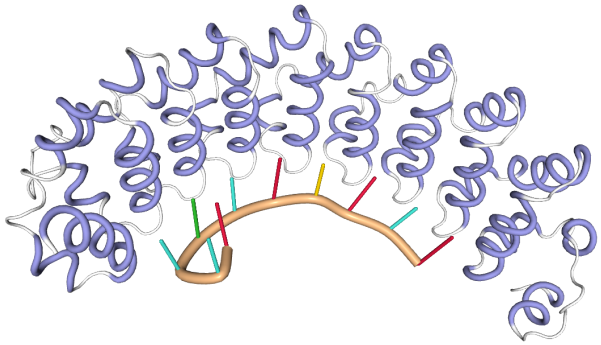


***Figura 23**. Estructura cristalina del dominio Pumilio 1 en complejo con una estructura de RNA. Los colores violetas muestran zonas plegadas en alfa hélice, y las zonas blancas muestran loops. Recuperado de RSCB PDB: 1M8Y.*

8.2.3.1. Modelo de la proteína Pumilio 1 | PUM1

La *fig. 24* muestra el modelo predicho para la secuencia transcrito de la Isoforma 3 derivado del splicing alternativo del gen PUM1 o Pumilio (Q14671-3).

El modelo aquí presentado fue modelado con una estructura elucidada por X-RAY DIFFRACTION a 2.60 Å (1m8y.2), con el que mantiene con el que mantiene una identidad de secuencia del 100.00% y un valor GMQE de 0.17. El modelo tiene una longitud de 654 aa en el rango 828-1171, que corresponde a una cobertura de 0.29. Más información sobre la formación estructural puede ser encontrada en la *tabla 4.*

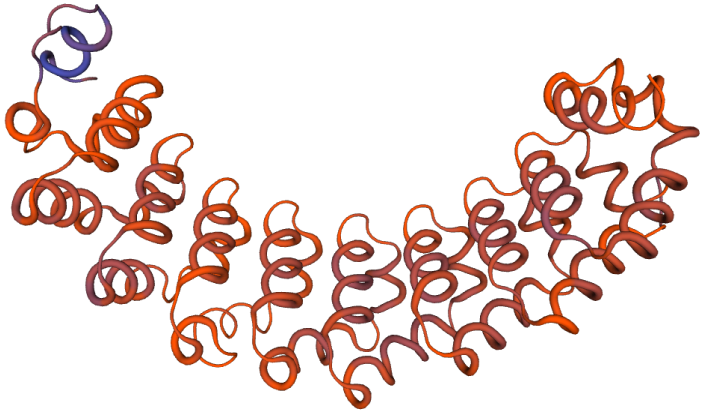


***Figura 24**.Estructura nativa modelada en Swiss-Prot para la secuencia Q14671-3, correspondiente a la proteína Pumilio1 o PUM1*

La *fig. 25* compara el modelo generado para la secuencia nativa para PUM1, el cual muestra una valina en la posición 1142 y a la derecha, el modelo generado para la proteína con la variación insertada, que presenta una Isoleucina en la misma posición. La proteína fue descartada para mayores análisis debido a que el aminoácido no parece generar cambios estructurales relevantes.

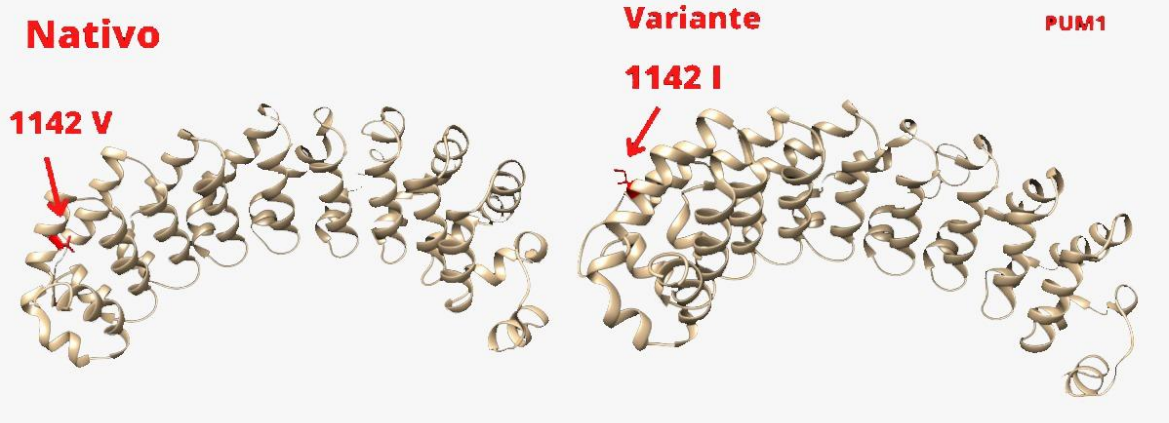


***Figura** 25. Izquierda. El modelo correspondiente a la secuencia nativa de PUM1, el cual muestra una valina en la posición 1142. Derecha. Modelo para la proteína PUM 1 con la variación insertada, correspondiente a una Isoleucina en la misma posición mencionada.*

***Tabla** 4. Especificaciones y características para el modelo estructural de la secuencia nativa y variante de Q14671-3, correspondiente a la proteína PUM1. El modelo de sustitución BLOSUM se calculó para la secuencia modelada respecto a la secuencia modelo del template utilizado.*

Características del modelo para PUM1	
Template usado	1m8y.2.B Pumilio 1 CRYSTAL STRUCTURE OF THE PUMILIO-HOMOLOGY DOMAIN FROM HUMAN PUMILIO1 IN COMPLEX WITH NRE2-10 RNA
Method	Difracción de rayos X a 2.60 Å
Oligo-State	Monomer
Modelo nativo	
GMQE	0.17
QMEANDisCo Global	0.50 ± 0.05
Confidence position	0.60
Modelo de sustitución BLOSUM respecto a secuencia modelo	V / V / 4
Modelo variante	
GMQE	0.17
QMEANDisCo Global	0.51 ± 0.05
Confidence position	0.58
Modelo de sustitución BLOSUM	I / V / 3

Alineamiento estructural	
Confidence	0.58 (nat) / 0.60 (var)
Consistency	1.00(nat) / 1.00 (var)

8.2.4. ATP2B2. ATPasa de bombeo de Ca(2+) de la membrana plasmática (Ca(2+)-ATPasa). UNIPROT: Q01814-1

La ATPasa bomba de Ca(2+) de la membrana plasmática (Ca(2+)-ATPasa) (*fig. 26)*, codificada por el ten ATP2B2 es responsable de mantener la homeostasis del calcio en las células eucariotas. La regulación de los niveles de calcio citosólico libre, esencial para el buen funcionamiento de todas las células eucariotas, se realiza mediante varios mecanismos, entre ellos esta ATPasa. La Ca2+-ATPasa bombea Ca2+ del citosol al espacio extracelular con la hidrólisis concomitante de ATP (Brandt et al., 1992). Las isoformas de la ATPasa de calcio de la membrana plasmática de los mamíferos están codificadas por al menos cuatro genes distintos y la diversidad de estas enzimas se ve incrementada por el splicing alternativo de los transcritos. La expresión de las diferentes isoformas y variantes de splicing está regulada de manera específica para el desarrollo, los tejidos y los tipos de células, lo que sugiere que estas bombas están funcionalmente adaptadas a las necesidades fisiológicas de células y tejidos concretos (Smits et al., 2019).

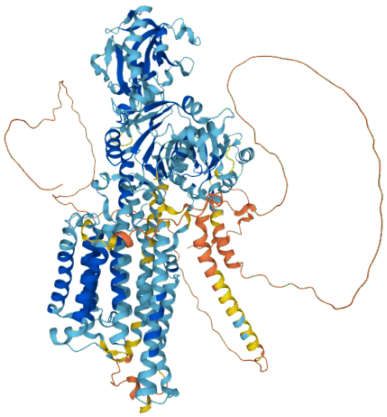


***Figura** 26. Predicción estructural creada por AlphaFold para el acceso Número de identificador AlphaFold AF-Q01814-F1 (https://alphafold.ebi.ac.uk/entry/Q01814). Los colores azules muestran zonas con buena confianza de modelado. Las zonas amarillas y naranjas muestran una zona con una baja confianza de modelado.*

8.2.4.1. Modelo de la Ca(2+)-ATPasa | ATP2B2

La *fig. 27* muestra el modelo generado para la secuencia transcrita de la Isoforma 1 (correspondiente de la isoforma canónica) de Ca(2+)-ATPasa, proteína codificada por el gen ATP2B2. Esta isoforma es derivada del splicing alternativo. La predicción de la estructura presentada fue modelada usando como template una estructura 6a69.1, elucidada por Microscopía electrónica a 4.11 Å, correspondiente a la estructura de una ATPasa de tipo P con el que mantiene una identidad de secuencia del 80.89% y un valor GMQE de 0.68.

El modelo tiene una longitud de 1084 aa en el rango 2-1086 de la secuencia original, que corresponde a una cobertura de 0.98. Más información sobre la formación estructural puede ser encontrada en la *tabla 5.*

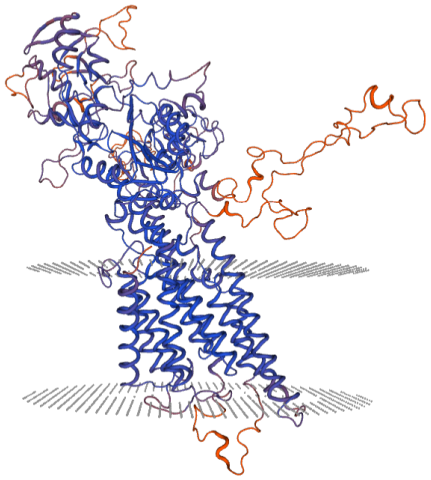


***Figura** 27.Estructura nativa modelada en Swiss-Prot para la secuencia Q01814-1, correspondiente al una ATPasa de Calcio.*

La *fig. 28* compara el modelo generado para la secuencia nativa para la secuencia Q01814-1, el cual muestra una lisina en la posición 61 y a la derecha, el modelo generado para la proteína con la variación insertada, que presenta una arginina, en la misma posición. Las estructuras fueron modeladas con el mismo template. A primera vista, no se encuentran modificaciones estructurales relativas, debido a que la variación se encuentra en una región en loop no modelada, rodeada de aminoácidos con baja confianza estructural, consistente en otros modelos para esta secuencia. Por lo anterior no se realizaron más análisis en la secuencia o estructura.

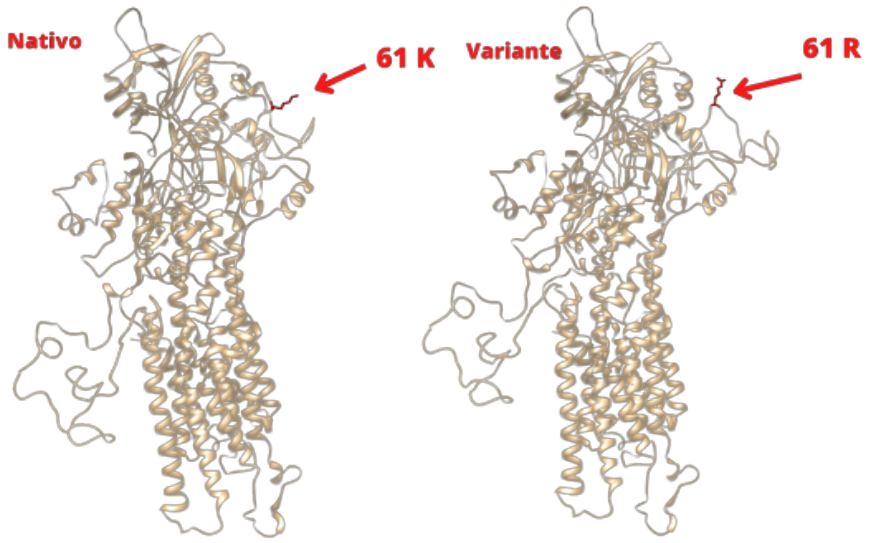


***Figura** 28. Izquierda. El modelo generado para la secuencia nativa de la ATPasa, el cual muestra una Lisina en la posición 61. Derecha. Modelo para la secuencia variante con la variación insertada, correspondiente a una arginina.*

***Tabla** 5. Especificaciones y características para el modelo estructural de la secuencia nativa y variante de la secuencia Q01814-1, correspondiente a una ATPasa de Calcio.*

Características del modelo para ATP2B2	
Template usado	6a69.1.A Plasma membrane calcium-transporting ATPase 1
Method	Microscopía electrónica a 4.11 Å
Oligo-State	Monomer
Modelo nativo	
GMQE	0.68
QMEANDisCo Global	0.73 ± 0.05
Confidence position	0.27
Modelo de sustitución BLOSUM	V / V / 4
Modelo variante	
GMQE	0.68
QMEANDisCo Global	0.73 ± 0.05
Confidence position	0.17

Modelo de sustitución BLOSUM	I / V / 3
Alineamiento estructural	
Confidence	0.27 (nat) / 0.17 (var)
Consistency	0.46 (nat) / 0.30 (var)

8.2.5. LMOD3. Leiomodina-3. UNIPROT: Q0VAK6-1

LMOD3 codifica la leiomodina-3 (*fig. 29*), una proteína de 65 kDa que se expresa en el músculo esquelético y cardíaco. Algunas mutaciones de este gen pueden provocar cardiopatías y miopatías (Yuen et al., 2014). Esta proteína contiene tres dominios de unión a actina, un dominio de tropomiosina, un dominio de repetición rico en leucina y un dominio de homología de la proteína 2 del síndrome de Wiskott-Aldrich (WH2). Se ha observado la localización de esta proteína en los extremos de filamentos celulares, y hay pruebas de que esta proteína actúa como catalizador de la nucleación de la actina, además de que es importante para la organización de los filamentos finos sarcoméricos en los músculos esqueléticos.

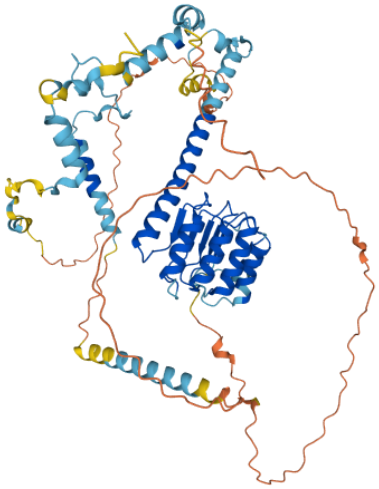


***Figura** 29. Predicción estructural creada por AlphaFold para el número de identificador AlphaFold AF-Q0VAK6-F1 (https://alphafold.ebi.ac.uk/search/text/AF-Q0VAK6-F1). Los colores azules muestran zonas con buena confianza de modelado. Las zonas amarillas y naranjas muestran una zona con una baja confianza de modelado.*

8.2.5.1. Modelo de la Leiomodina-3 | LMOD3

La *fig. 30* muestra el modelo generado para la secuencia transcrita de la Isoforma 1 (correspondiente de la isoforma canónica), de la Lemoidina-3, proteína codificada por el gen LMOD3. Esta isoforma es derivada del splicing alternativo.
La predicción de la estructura presentada fue modelada usando como template una estructura 4rwt.1.B, elucidada por X-RAY DIFFRACTION a 2.98 Å, correspondiente a la estructura de la Leiomodina-2 con el que mantiene una identidad de secuencia del 40.69% y un valor GMQE de 0.45.

El modelo tiene una longitud de 320 aa en el rango 240-560 de la secuencia original, que corresponde a una cobertura de 0.82. Más información sobre la formación estructural puede ser encontrada en la *tabla 6*

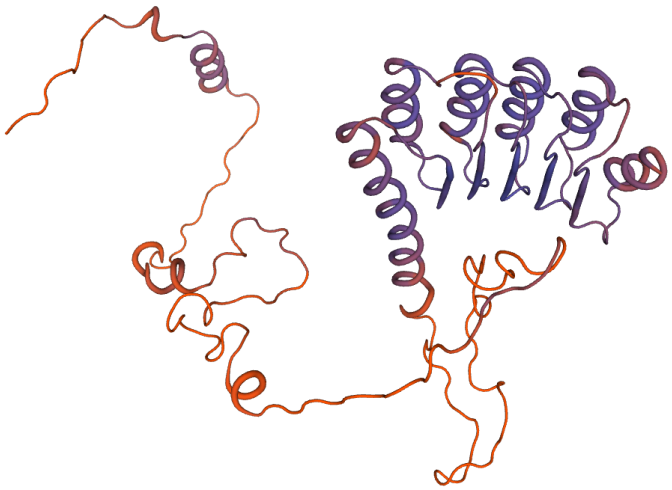


***Figura** 30.Estructura nativa modelada en Swiss-Prot para la secuencia Q0VAK6-1, correspondiente al una ATPasa de Calcio.*

La *fig. 31* compara el modelo generado para la secuencia nativa para la secuencia Q01814-1, el cual muestra una arginina en la posición 495 y a la derecha, el modelo generado para la proteína con la variación insertada, que presenta una histidina, en la

misma posición. Las estructuras fueron modeladas con el mismo template.

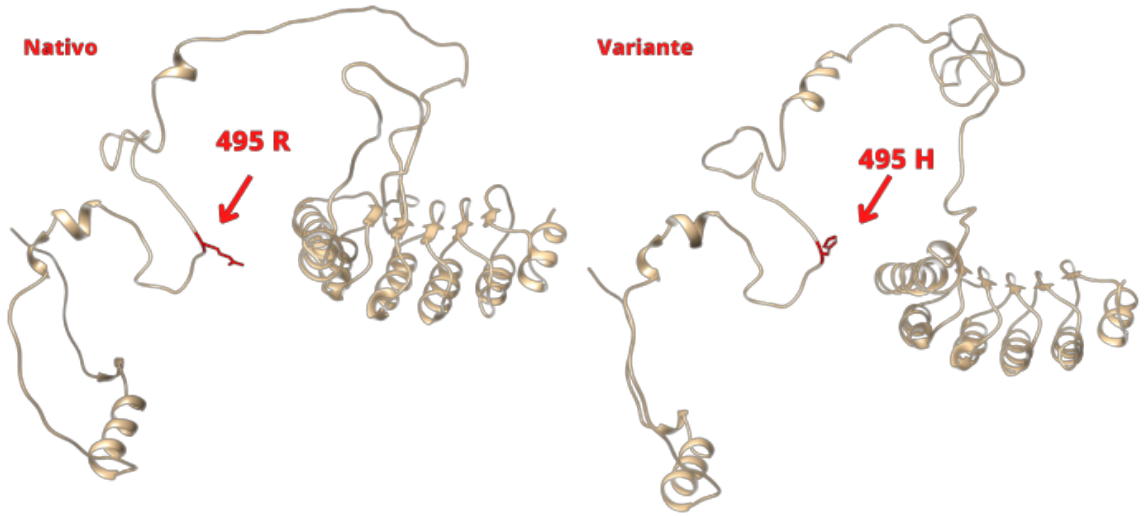


***Figura** 31. Izquierda. El modelo generado para la secuencia nativa de la secuencia de la Lemoidina-3, el cual muestra una arginina en la posición 495. Derecha. Modelo para la secuencia variante con la variación insertada, correspondiente a una histidina.*

Este modelo se descartó para subsecuentes análisis ya que el aminoácido de interés cae en una zona no modelada, en loop, rodeada de aminoácidos con baja confianza estructural.

***Tabla** 6. Especificaciones y características para el modelo estructural de la secuencia nativa y variante de la secuencia Q01814-1, correspondiente a la Lemoidina-3*

Características del modelo para LMOD3	
Template usado	4rwt.1.B Leiomodin-2 Structure of actin-Lmod complex
Method	Difracción de rayos X a 2.98 Å
Oligo-State	Monomer
Modelo nativo	
GMQE	0.45
QMEANDisCo Global	0.58 ± 0.05
Confidence position	0.37
Modelo de sustitución BLOSUM	G / R / -2
Modelo variante	
GMQE	0.45

QMEANDisCo Global	0.57 ± 0.05
Confidence position	0.40
Modelo de sustitución BLOSUM	G / H / 0
Alineamiento estructural	
Confidence	0.37 (nat) / 0.40 (var)
Consistency	1.00 (nat) / 1.00 (var)

8.2.6. MAGI2. Guanilato quinasa asociada a membrana. UNIPROT: Q86UL8-1

La proteína MAGI2 (*fig. 32)* se caracteriza por dos dominios WW, un dominio similar a la guanilato cinasa, y múltiples dominios PDZ. Esta proteína es codificada por el gen MAGI2.

***Figura** 32. Estructura elucidada por NMR de Guanilato quinasa asociada a membrana, codificada por el gen MAGI2. Recuperado de RSCB PDB: 1UEP (https://www.rcsb.org/structure/1UEP)*

8.2.6.1. Modelo de la Guanilato quinasa asociada a la membrana, que contiene el dominio WW y PDZ 2 | MAGI2

La *fig. 33* muestra el modelo generado para la secuencia Q86UL8-1 transcrita de la Isoforma 1 (correspondiente de la isoforma canónica), de la Guanilato cinasa 2 asociada a la membrana. La predicción de la estructura presentada fue modelada usando como template una estructura 3zrt.1, elucidada por X-RAY DIFFRACTION a 3.40 Å, correspondiente a la estructura de la PSD-95 PDZ1-2 humana con el que mantiene una identidad de secuencia del 40.69% y un valor GMQE de 0.45.

El modelo tiene una longitud de 309 aa en el rango 920-1229 de la secuencia original, que corresponde a una cobertura de 0.12. Más información sobre la formación estructural puede ser encontrada en la *tabla 7*

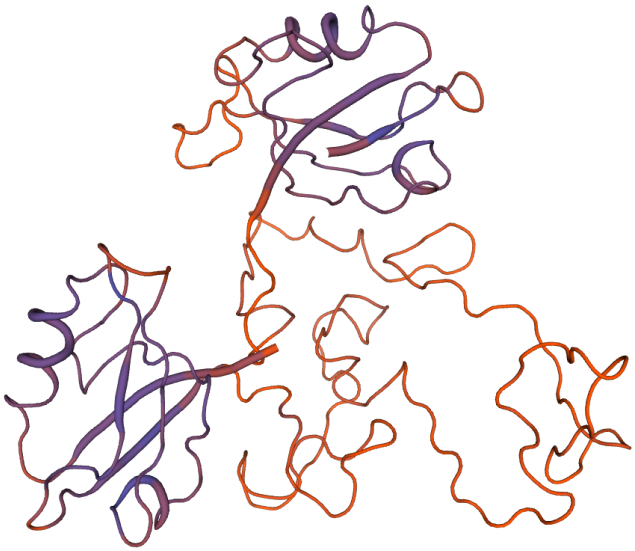


***Figura** 33.Estructura nativa modelada en Swiss-Prot para la secuencia Q0VAK6-1, correspondiente a una ATPasa de Calcio.*

La *fig. 34* compara el modelo generado para la secuencia nativa de Q86UL8-1, el cual muestra una arginina en la posición 1013 y a la derecha, el modelo generado para la proteína con la variación insertada, que presenta una lisina, en la misma posición. Las estructuras fueron modeladas con el mismo template. Este modelo se descartó para subsecuentes análisis ya que el aminoácido de interés cae en una zona no modelada, sin estructura secundaria, rodeada de aminoácidos con baja confianza estructural. Más información relativa a la confianza de construcción del modelo puede ser encontrada en la *tabla 7.*

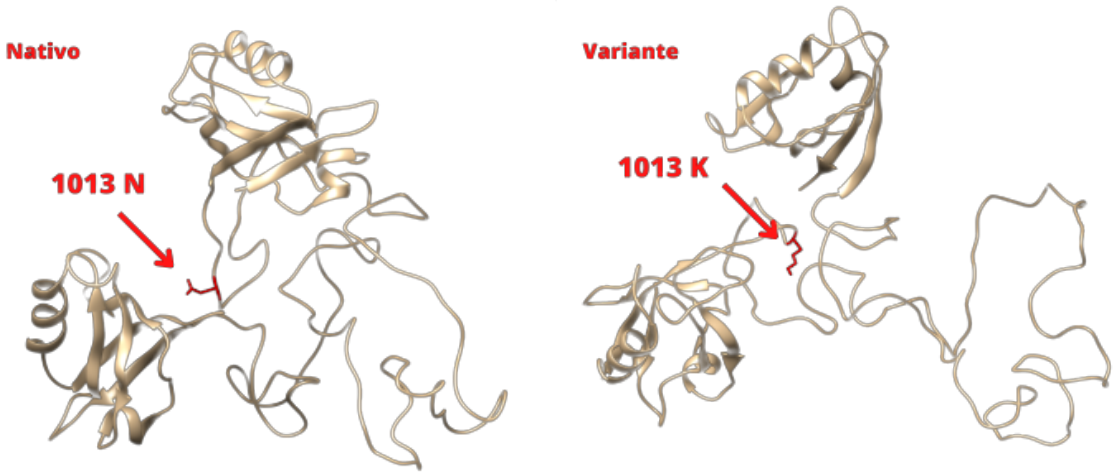


***Figura** 34. Estructuras modeladas de la secuencia Q86UL8-1. A la izquierda, el modelo generado para la secuencia nativa de la Guanilato cinasa, el cual muestra una asparagina en la posición 1013. Derecha, modelo para la secuencia variante con la variación insertada, correspondiente a una lisina.*

***Tabla** 7. Especificaciones y características para el modelo estructural de la secuencia nativa y variante de la secuencia Q86UL8-1, correspondiente a la proteína expresada por MAGI2.*

Características del modelo para MAGI2	
Template usado	3zrt.1.A DISKS LARGE HOMOLOG 4.
Method	Difracción de rayos X a 3.40 Å
Oligo-State	Monomer
Modelo nativo	
GMQE	0.06
QMEANDisCo Global	0.56 ± 0.05
Confidence position	0.34
Modelo de sustitución BLOSUM	N / A / 0
Modelo variante	
GMQE	0.06
QMEANDisCo Global	0.57 ± 0.05
Confidence position	0.28
Modelo de sustitución BLOSUM	K / A / -1

Alineamiento estructural	
Confidence	0.59 (nat) / 0.59 (var)
Consistency	1.00 (nat) / 1.00 (var)

8.2.7. DNHD1. Cadena pesada de la dineína. (UNIPROT: Q96M86-3)

La dineína fue identificada y nombrada por primera vez por Ian Gibbons en la década de 1960 como una ATPasa que podía extraerse de los cilios y los flagelos (Hirose, 2019).

Las dineínas son máquinas macromoleculares alimentadas por ATP que impulsan todos los procesos de transporte de carga molecular en los microtúbulos en células eucariotas y desempeñan papeles esenciales en una amplia variedad de funciones celulares (Grotjahn & Lander, 2019). Por ejemplo, la dineína citoplasmática promueve la migración nuclear, la organización del huso mitótico, la separación de los cromosomas durante la mitosis, y el posicionamiento y la función de muchos orgánulos intracelulares (Oiwa & Sakakibara, 2005). La *fig. 35* muestra una estructura monomérica de un modelo para la cadena pesada de la dineína.

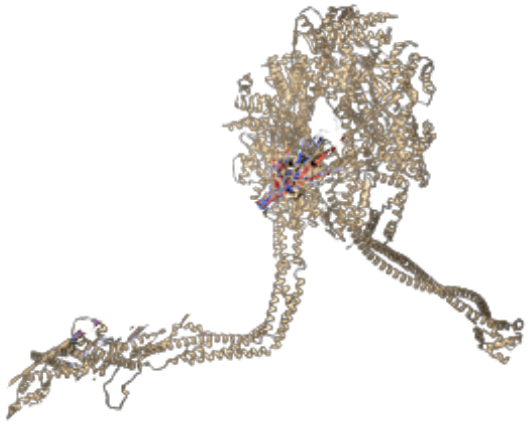


Figura *35.Estructura de la Cadena Pesada de la dineína, correspondiente al gen DNHD1. Recuperada de RSCB PDB: 7ZF8.1 (https://swissmodel.expasy.org/templates/7z8f.1)*

8.2.7.1. Modelo del dominio de la Cadena Pesada de la Dineína | DNHD1

La *fig. 36* muestra el modelo estructural predicho para la secuencia transcrita de la Isoforma 3 (correspondiente de la isoforma canónica), derivado del splicing alternativo del gen DNHD1, propio del dominio de la cadena pesada de la dineína (Q96M86-3).

El modelo aquí presentado fue modelado con una estructura elucidada por Microscopía electrónica, 7z8f.1.0 del dominio de la cadena pesada de la Dineína, con el que mantiene una identidad de secuencia del 18.27% y un valor GMQE de 0.20. El modelo

tiene una longitud de 4236 aa en el rango 514-4750 de la secuencia original, que corresponde a una cobertura de 0.75. Más información sobre la formación estructural puede ser encontrada en la *tabla 8*

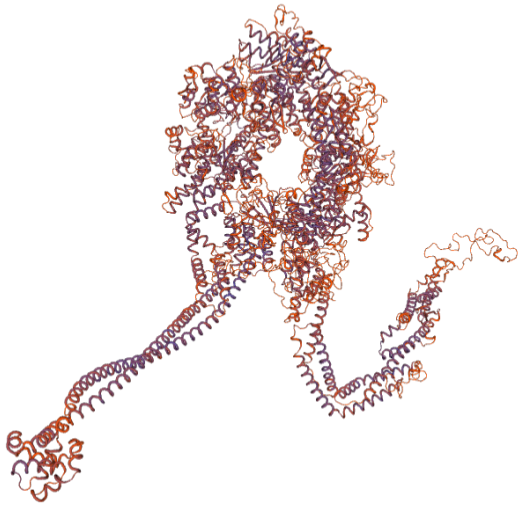


***Figura** 36.Estructura nativa modelada en Swiss-Prot para la secuencia Q96M86-3, correspondiente al dominio de la cadena pesada de la Dineína*

La *fig. 37* compara el modelo generado para la secuencia nativa para DNHD1, el cual muestra una Serina en la posición 1714 y a la derecha, el modelo generado para la proteína con la variación insertada, que presenta una Leucina, en la misma posición. Las estructuras fueron modeladas con el mismo template. A primera vista, se encuentran modificaciones estructurales relativas.

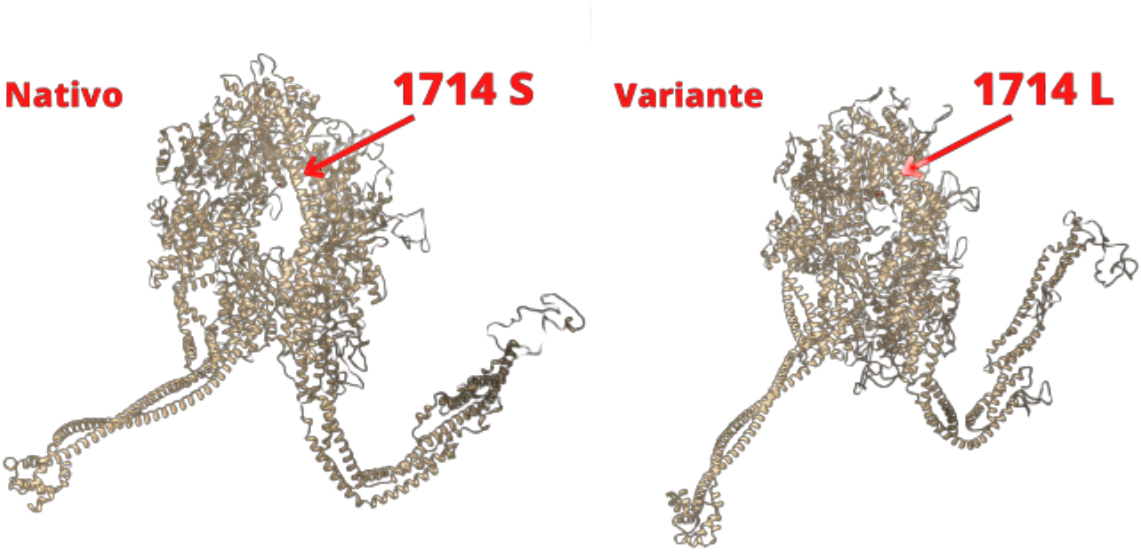


***Figura** 37. Izquierda. El modelo generado para la secuencia nativa de DNHD1, el cual muestra una Serina en la posición 1714. Derecha. Modelo para la secuencia variante de DNHD1 con la variación insertada, correspondiente a una Leucina.*

Respecto al posible cambio estructural provocado por la variante de interés, la *fig. 38* muestra a la izquierda y en color rojo una Serina. Este corresponde al residuo nativo en el cuál se encontró la variante, mostrada a la derecha en color naranja (1714 L).

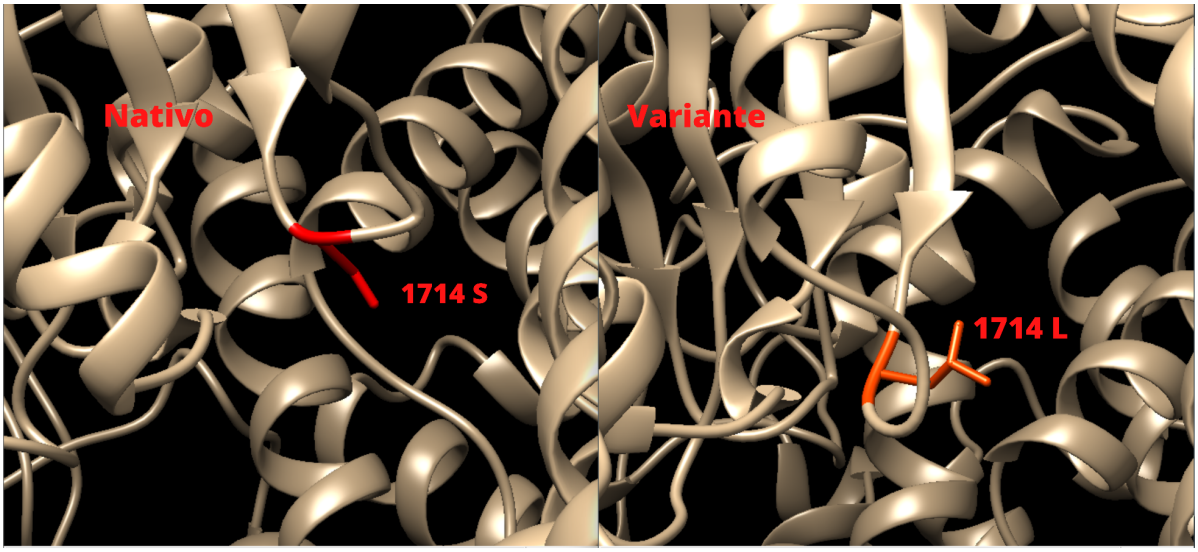


***Figura** 38.Izquierda. Close-up al al aminoácido nativo (1714 S), mostrado en color rojo. Derecha. Modelo variante para el gen DNHD1, en el que se modificó el aminoácido 1714 a Leucina., representada en color naranja.*

***Tabla** 8. Especificaciones y características para el modelo estructural de la secuencia nativa y variante de Q96M86-3, correspondiente al dominio de la cadena pesada de la dineína.*

Características del modelo para DNHD1	
Template usado	7z8f.1.0
Method	ELECTRON MICROSCOPY
Oligo-State	Monomer
Modelo nativo	
GMQE	0.20
QMEANDisCo Global	0.53 ± 0.05
Confidence position	0.47
Modelo de sustitución BLOSUM	S / D / 0
Modelo variante	
GMQE	0.15

QMEANDisCo Global	0.53 ± 0.05
Confidence position	0.42
Modelo de sustitución BLOSUM	L / D / -4
Alineamiento estructural	
Confidence	0.46 (nat) / 0.49 (var)
Consistency	0.67

9. Discusión

9.1. Sobre la relación de la demografía histórica con los registros genéticos

Algunas de las características genómicas de la comunidad podrían ser explicadas desde un punto de vista antropológico y demográfico. Los Comcaa'c se organizaron durante una gran temporada como grupos nómadas (Rentería Valencia, 2007), hasta el contacto con los grupos colonizadores, evento que los llevó a modificar los patrones de asentamiento y organización social, y que junto a los eventos bélicos, provocaron que la población mantuviera durante muchos años una población reducida (Sheridan, 1999). La aparición de variantes no presentes en otras poblaciones relacionadas, la reducción de la variabilidad comunitaria, la alta diferenciación genética con otras comunidades, y la alta homocigocidad del genoma probablemente podría ser explicado por el efecto fundador o "Founding Effect" (*fig. 39*), precisamente derivado de la demografía histórica que ha sufrido la población (Barton & Charlesworth, 1984).

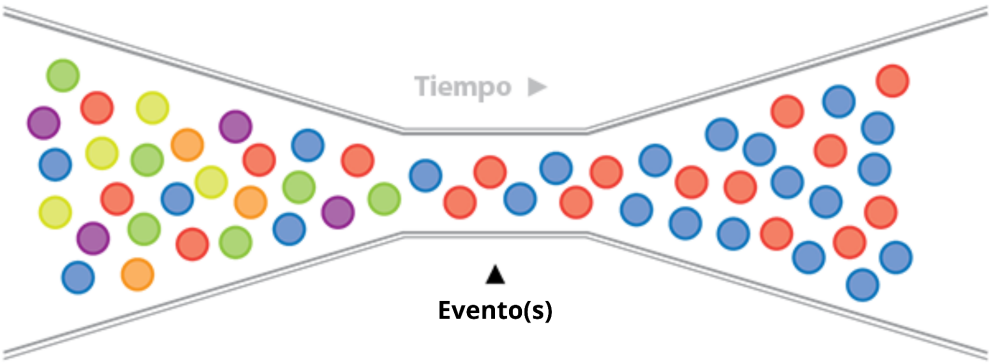


***Figura** 39. Representación gráfica del evento de cuello de botella, el cual puede dar paso al llamado Efecto Fundador. Debido a uno o varios eventos en el tiempo, puede suceder una reducción de la variabilidad (representada por colores diferentes en la imagen), generando poblaciones provenientes de únicamente unos cuantos individuos.*

El efecto fundador puede aumentar la frecuencia de ciertos trastornos raros, mientras que otros alelos de enfermedad característicos de la población parental pueden desaparecer (Kivisild, 2013). Un efecto fundador puede ser el resultado del establecimiento de una nueva población a partir de individuos derivados de una población mucho mayor o de una reducción extrema del tamaño de la población. Los alelos presentes en alguna población pueden encontrarse en una frecuencia mucho mayor después del acontecimiento fundador o del cuello de botella, y pueden alcanzar frecuencias aún más altas debido a la fuerte deriva genética que se produce mientras la población es aún pequeña. Esto podría ser probado mediante estudios estadísticos como el que ha sido utilizado anteriormente en otras poblaciones (Slatkin, 2004)*.* Esta no es la primera vez que se propone que la población haya sufrido del Efecto fundador. En *1999, Infante et al.* propuso la aparición de un alelo perteneciente al HLA bajo el efecto fundador.

Hay que se cuenta con una muestra de únicamente 4 individuos, que representa a aproximadamente ~0.5% de la población al tomar en cuenta que los últimos resultados de los censos del INEGI para 2020 marcan un aproximado de 723 individuos. A pesar de eso, hay que tener en cuenta que el grupo se conoce por ser genéticamente poco diverso y por mostrar un alto porcentaje del genoma en estado homocigoto, la cual presenta una ventaja en la representatividad de las muestras, ya que es más probable encontrar variantes de interés que impacten en la comunidad.

9.2. Sobre las variantes encontradas

En *Aguilar-Ordoñez et al., 2021*, se muestran un total de 2,496 SNV particulares para los 4 genomas secuenciados, también utilizados en este trabajo. Los resultados de este estudio mostraron 3,451 SNVs particulares, de los cuales se anotaron 3,428. La diferencia en los números se debe a que en este trabajo se utilizaron versiones más recientes de las bases de datos de referencia. En la anotación, se identificaron un alto número relativo de variantes en algunos cromosomas y al mismo tiempo, cromosomas que no presentan ninguna variante, tal como se observa en la *fig.15*.

De las variantes no sinónimas estudiadas, algunas han sido relacionadas con funciones neuro-musculares, tales como AGRN, LMOD3, y MAGI2 *(Bierzynska et al., 2017; Denzer et al., 1997; Yuen et al., 2014*), dos con regulación de expresión génica (PUM1, ERCC2) (Galgano et al., 2008; E. M. Taylor et al., 1997), dos con canales iónicos (MCOLN3, HCN2) (Curcio-Morelli et al., 2010; M. Li et al., 2018), dos ATPasas con dominios AAA (ATAD2B, ATP2B2) (Brandt et al., 1992; Ota et al., 2004), dos con procesos metabólicos (SLC37A1, INPP4B) (Bartoloni et al., 2000; Norris et al., 1997) una con actividad motora de microtúbulos (DNHD1) (Asai & Koonce, 2001), uno relacionado a procesos cancerígenos (PDLIM2) (Zeng et al., 2022), una con receptores olfativos (OR2T29) (B. Wang et al., 2019), una con adhesión celular (SIGLEC14) (Angata et al., 2006), una cinasa (PEAK3) (Lopez et al., 2019), una con acción fosfolipasa (PLA2G2F) (Murakami et al., 2002), una relacionada al Complejo Principal de Histocompatibilidad (TRAV8-7) (Koop et al., 1994), una a procesos apoptóticos (ZSWIM2) (Nishito et al., 2006), un pseudogen (MROH5) (Wojczynski et al., 2013) y uno con funciones no identificadas (C19orf47).

9.3. Sobre las variantes modeladas

Se encontraron 7 genes con variantes particulares de la población Comcaa'c, que además fueran variantes no-sinónimas, anotadas sobre el transcrito canónico y con identificador del transcrito en UNIPROT. A pesar de que únicamente estas 7 variantes fueron modeladas, otras presentes en los registros de anotación podrían ser candidatas a modelado estructural. La mayoría de las estructuras modeladas muestran una alta confianza global de modelado pero no muestran relevancia estructural debido a varios factores.

En el caso del modelo de la proteína Agrina del gen **AGRN**, el modelo y la variante fueron descartadas para mayores análisis en este estudio debido a que ninguno de los modelos

realizados contuvo la posición del residuo de interés. En el caso de los modelos de la proteína **PA2GF** del gen PLA2G2F y la proteína Pumilio 1 del gen **PUM1**, los modelos y la variantes fueron descartados para mayores análisis debido a que el aminoácido no parece generar cambios estructurales relevantes. En el caso de modelos de la proteína Ca(2+)-ATPasa, correspondiente al gen **ATP2B2**, la Lemoidina-3, correspondiente al gen **LMOD3**, y la Guanilato cinasa asociada a membrana WW/PDZ 2, correspondiente al gen **MAGI2,** fueron descartados para mayores análisis debido a que el aminoácido no parece generar cambios estructurales relevantes o cae en una región con baja confianza estructural, sin estructura secundaria o en un loop. Para los modelos de **PA2GF, PUM1 y ATP2B2,** se observa una relativa buena favorabilidad. **AGRN, LMOD3** y **MAGI2** muestran una favorabilidad de sustitución neutral, y **DNHD1** una baja mala favorabilidad evolutiva. La *tabla 9* muestra brevemente los resultados de la matriz de sustitución.

Varias de las estructuras aquí modeladas proteína parecen no contar con las estructuras en loop en las cuales cae el aminoácido de interés en un contexto biológico. Más investigación estructural de estas proteínas debe de ser llevada a cabo para concluir si el aminoácido de interés cae en una región real de la proteína. Tal es el caso de **LMOD3** y **MAGI2.**

***Tabla** 9. Favorabilidad de sustitución para cada variante de interés, obtenida mediante la matriz BLOSUM62.*

Nombre del gen	Aminoácido de referencia	Aminoácido variante	Resultados de la matriz de sustitución	Favorabilidad
AGRN	A	T	0	Neutral
PA2GF	V	I	3	Buena
PUM1	V	I	3	Buena
DNHD1	S	L	-2	Baja
ATP2B2	K	R	2	Buena
LMOD3	R	H	0	Neutral
MAGI2	N	K	0	Neutral

Finalmente, el dominio de la cadena pesada de la Dineína, correspondiente al gen **DNHD1,** fue el único modelo presentado que presentó la mutación en un espacio tridimensional en donde se esperaría que pudiera generar algún cambio de interés estructural. La predicción presenta una variación S/L en la posición 1714 de la cadena peptídica. La mutación se encuentra probablemente cerca de un dominio de unión al ATP, aunque la posición 1714 no parece interactuar con el ATP en otros modelos conocidos

9.4. Características de la proteína expresada por DNHD1

La sección 8.3.4 menciona algunas características del gen DNHD1, que expresa a la Cadena pesada de la dineína, con coordenadas GRCh38: 11:6,497,280-6,572,020 y una localización citogenética 11p15.4 (*fig.40*). Este gen ha sido relacionado anteriormente con astenoteratozoospermia, definida como motilidad espermática reducida y morfología espermática anormal que a su vez, puede causar infertilidad en varones (Tan et al., 2022).

***Figura** 40. Localización del gen DNHD1, que expresa a la Cadena pesada de la dineína, con coordenadas GRCh38: 11:6,497,280-6,572,020 y una localización citogenética 11p15.4. Recuperado de UCSC Genome Browser on Human (GRCh38/hg38)*

El gen también puede encontrarse en los registros como DHCD1; CCDC35; DNHD1L; SPGF65; C11orf47. Según *Fagerberg et al., 2014*, este gen se ve más expresado en testículos, pero también se ve expresado en otros órganos del cuerpo, debido probablemente a la necesidad motriz de ciertos tejidos en el cuerpo humano (fig. 41).

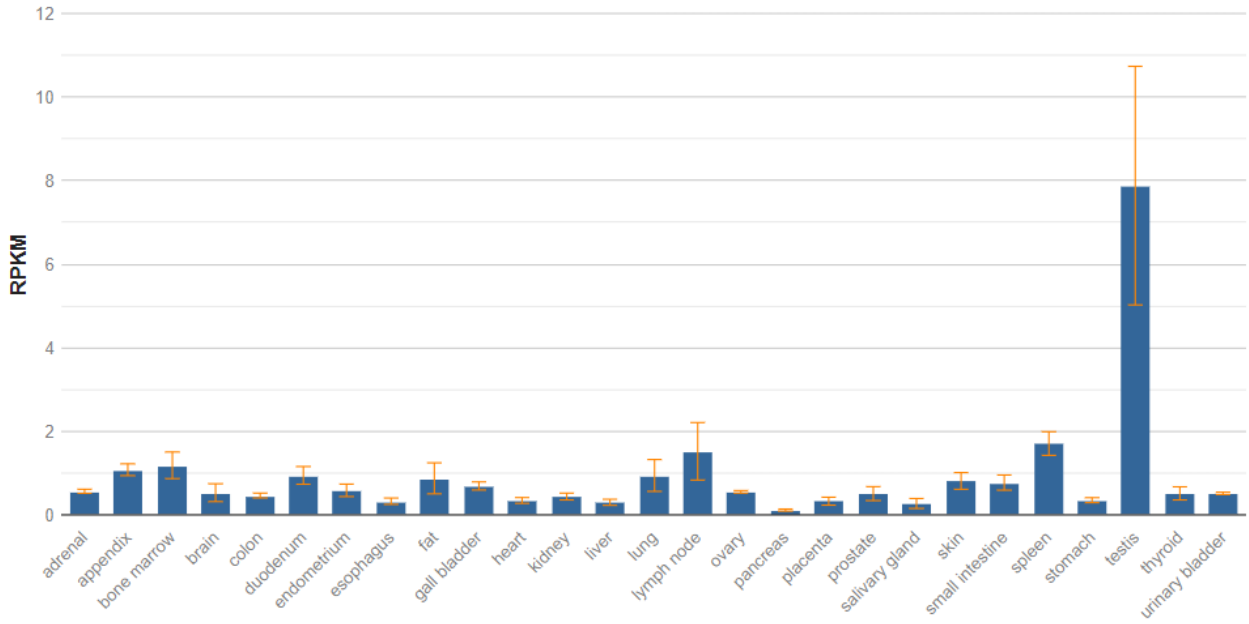


***Figura** 41. Niveles de expresión del gen DNHD1, medidos en RPKM (reads per kilobase of transcript per million reads mapped), una unidad de expresión génica que mide la abundancia de mRNA. Recuperado de (Fagerberg et al., 2014)*

9.5. Sobre la estructura de la Dineína

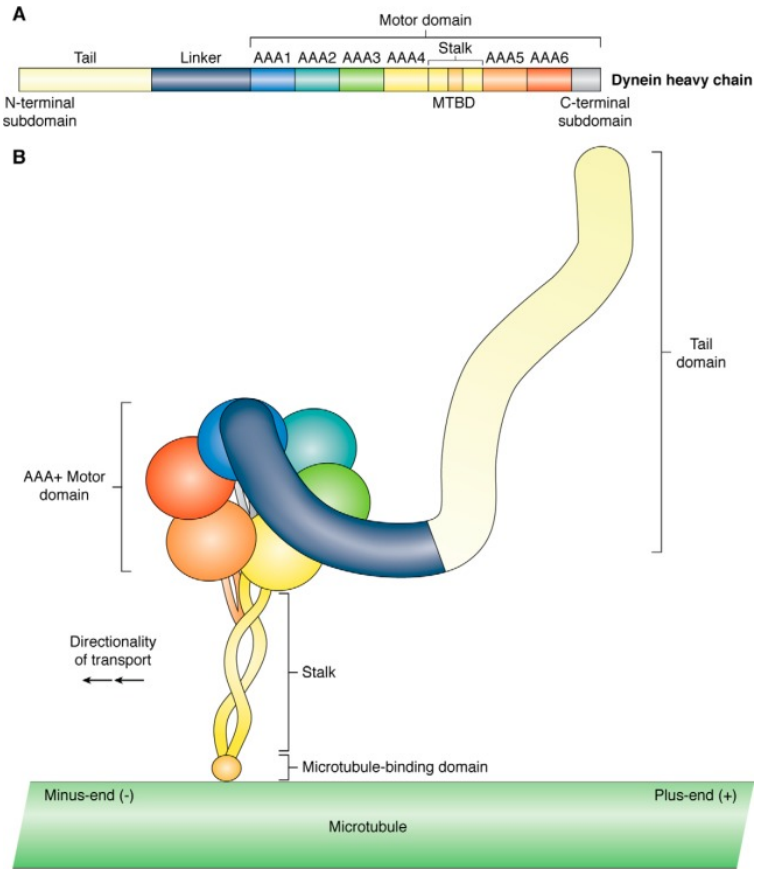
La cadena pesada de la dineína (*fig. 42)* es una máquina molecular de ~0,5 MDa que vincula la hidrólisis del ATP a los ciclos de unión, movimiento y liberación de los microtúbulos. La cadena pesada de la Dineína pertenece a la gran y funcionalmente diversa superfamilia de las AAA+

ATPasas, que a su vez forman una subdivisión de NTPasas de bucle P en forma de anillo (*Hirose, 2019*). Las isoformas citoplasmáticas transportan diversas cargas en las células (*Vallee et al., 200*4)y se ensamblan en homodímeros que pueden coordinar cientos de pasos sucesivos a lo largo de a lo largo de los microtúbulos sin desprenderse.

La EM estableció que cada cadena pesada de dineína se pliega para formar una "cabeza" globular y anular con dos estructuras alargadas, la "cola" y el "tallo", que emergen de ella (*fig. 51 B)* (Clark & Rose, 2006). Los dominios de la cadena pesada incluyen la cola, el dominio del tallo, el dominio de la cabeza (que incluye los anillos AAA+), el dominio linker, y la secuencia C-terminal (*Hirose, 2019*).

Los ~1.300 aminoácidos N-terminales de la cadena pesada forman el dominio de la cola, según se demostró con digestión enzimática, biología molecular y EM *(Koonce & Samsó, 1996; Mocz & Gibbons, 1993*). Las colas de las diferentes isoformas de dineína contienen numerosos motivos a-helicoidales cortos, pero la estructura y topología del polipéptido dentro de la cola siguen siendo inciertos. En la dineína citoplásmica, la cola media la dimerización de la cadena pesada y la unión de proteínas asociadas reguladoras de la dineína y que unen a las moléculas de carga.

El tallo es un dominio alargado con estructura intramolecular de espiral antiparalela, con un dominio globular en su extremo que se une y libera a los microtúbulos de forma sensible a los nucleótidos (*Amos, 1989*).



.

***Figura** 42. A, esquema de la secuencia de dominios de la cadena pesada de dineína. Los dominios comprenden la cola N-terminal, el enlazador, los seis dominios AAA+ (AAA1-AAA6), el tallo, el dominio de unión a microtúbulos y el dominio C-terminal. B. Representación cartográfica de la cadena pesada de dineína unida a un microtúbulo (verde) y orientada para su transporte hacia el extremo negativo del microtúbulo. Recuperado de Grotjahn & Lander, 2019*

El grueso de la secuencia del dominio motor forma la cabeza, que tiene un aspecto anular con un diámetro de unos 13 nm. La cabeza puede subdividirse en el anillo AAA+, la secuencia C-terminal y el dominio linker, que juntos dan lugar a la aparición de siete lóbulos de densidad que se observan en los promedios de EM (Roberts et al., 2009; Samsó et al., 1998). El análisis de la secuencia muestra que la cabeza contiene seis módulos AAA+ (denominados AAA1-AAA6). El módulo AAA+ es una región de ~200-250 aminoácidos que suele contener varios motivos característicos implicados en la hidrólisis del ATP. Sólo las cuatro primeras (AAA1-AAA4) contienen motivos de unión e hidrólisis de nucleótidos y, entre ellas, la AAA1 es el principal lugar de hidrólisis de ATP relacionado con la actividad motora. Los AAA2-AAA4 parecen tener una función reguladora. En particular, AAA3 desempeña un papel importante en el mecanismo, ya que las mutaciones que impiden la unión e hidrólisis de nucleótidos aquí paralizan la capacidad de la dineína para liberarse de los microtúbulos (Cho et al., 2008; Silvanovich et al., 2003). El dominio del tallo se encuentra entre AAA4 y AAA5.

Finalmente, en la mayoría de las cadenas pesadas de dineína, la región C-terminal de AAA6 es una región de ~400 aminoácidos que contiene una mezcla de a-hélices y hojas en beta. El papel exacto de la secuencia C en el mecanismo de la dineína es actualmente un misterio, pero se sabe que regula la actividad de unión con los microtúbulos (Hirose et al., 2006; Telzer & Haimo, 1981).

Después del análisis de dominios de la secuencia **Q96M86-3**, se identificó que la región en la que se encuentra la variación estudiada S1714L y rs201274362 corresponde a un dominio AAA_6 (*fig. 43*), con un intervalo de 1649-1999 y un e-value de 8.40e-15. En los modelos generados, la posición 1714 no parece interactuar directamente con moléculas de ATP.

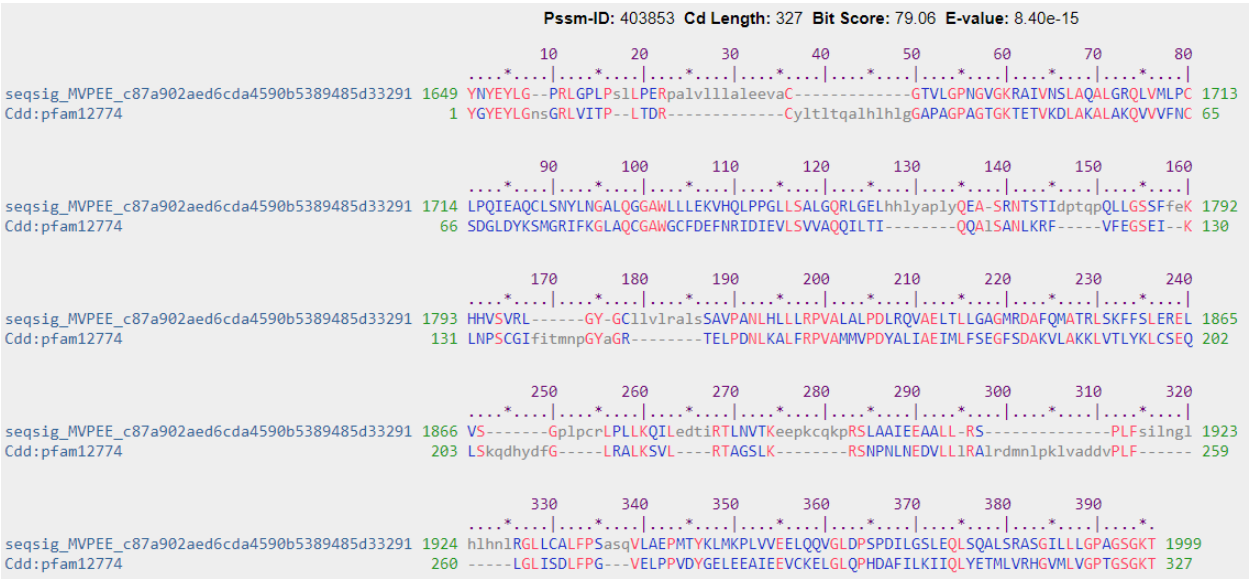


Figura 43. *Análisis de dominios conservados para la secuencia Q96M86-3, correspondiente al dominio de la cadena pesada de la dineína. La flecha roja muestra la posición del aminoácido variante leucina 1714. La secuencia superior corresponde a la estructura primaria de la proteína problema. La secuencia inferior corresponde a la referencia del dominio AAA6.*

El hecho de que la variación se encuentre en un dominio AAA podría implicar que la capacidad de unión de moléculas de ATP a los aminoácidos del dominio podría modificarse. Esta hipótesis debe ser probada a nivel bioquímico y experimental. Por otro lado, se debe de tener en cuenta la capacidad de la proteína y de esta familia de dominios a tolerar variaciones sin modificar significativamente sus funciones. También es posible que la mutación se adapte sin causar ningún desplazamiento o distorsión relevante.

Figura 44. *Izquierda. Proteína predicha por los algoritmos de modelado estructural para la secuencia Q96M86-3 del dominio de la cadena pesada de la dineína. Izq. Imágenes de microscopía electrónica que muestra hacia la izquierda, la cara del dominio linker y hacia la derecha, la cara del dominio C-terminal. La imagen pretende demostrar que el modelo tiene congruencia con evidencia experimental como lo es la microscopía electrónica.*

Según los registros del UCSC Genome Browser para la variante **rs201274362** correspondiente a la variación aquí estudiada, S1714 es un aminoácido no tan conservado, esto después de realizar una comparación con otros 100 vertebrados (más puede ser consultado en https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A6545980%2D6546180&hgsid=1483832663_C3q28wVv6tAubG7iz0hkhQVvY5uO).

9.6. Homólogos del gen DNHD1 en humanos

El gen DNHD1 no es el único en la especie humana que codifica para la Cadena pesada de la dineína. La *tabla 10* muestra los genes que expresan cadenas pesadas de la dineína, así como algunas características y consideraciones respecto a su expresión. La mayoría de los genes homólogos de la dineína corresponden a variantes axonemales, es decir, que se encuentran asociados con axonemas, estructuras citoesquelética basada en microtúbulos que forman el núcleo de un cilio o flagelo. El hecho de que existan tantos homólogos podría ser indicio de que la función asociada a la proteína no se ve afectada tan fácilmente por mutaciones. En el caso específico de la cadena pesada de la dineína, hay que tener en cuenta varios factores, entre los que se encuentran la densidad aminoacídica de la zona donde cae la

mutación, la cantidad de genes homólogos que se expresan en las mismas regiones anatómicas humanas, el tamaño de la proteína, y su asociación con el complejo motor proteico. Un análisis de coeficientes de variación genética de este grupo de genes sería útil para predecir qué tanta resistencia tiene la estructura a las variaciones.

Tabla 10. *Genes en la especie humana que codifican cadenas pesadas del complejo proteico de la dineína. La mayoría corresponden a dineínas axonemales.*

ID	Nombre	Localización
DYNC1H1	dynein cytoplasmic 1 heavy chain 1	14q32.31 NC_000014.9 (101964573..102056443)
DYNC2H1	dynein cytoplasmic 2 heavy chain 1	11q22.3 NC_000011.10 (103109426..103479863)
DNAH5	dynein axonemal heavy chain 5	5p15.2 NC_000005.10 (13690328..14011818)
DNAH1	dynein axonemal heavy chain 1	3p21.1 NC_000003.12 (52310920..52400492)
DNAH11	dynein axonemal heavy chain 11	7p15.3 NC_000007.14 (21543039..21901839)
DNAH8	dynein axonemal heavy chain 8	6p21.2 NC_000006.12 (38715311..39030792)
DNAH9	dynein axonemal heavy chain 9	17p12 NC_000017.11 (11598470..11969748)
DNAH17	dynein axonemal heavy chain 17	17q25.3 NC_000017.11 (78423697..78577396)
DNAH10	dynein axonemal heavy chain 10	12q24.31 NC_000012.12 (123762301..123935720)
DNAH6	dynein axonemal heavy chain 6	2p11.2 NC_000002.12 (84459572..84819589)
DNAH14	dynein axonemal heavy chain 14	NC_000001.11 (224929654..225399286)
DNAH2	dynein axonemal heavy chain 2	NC_000017.11 (7717744..7833742)
DNAH7	dynein axonemal heavy chain 7	NC_000002.12 (195737703..196068837)
DNAH3	dynein axonemal heavy chain 3	NC_000011.10 (6497280..6572020)
DNAH12	dynein axonemal	NC_000003.12 (57293700..57556034, complement)

	heavy chain 12	

10. Conclusiones

En el presente trabajo describimos las variantes particulares presentes en 4 genomas pertenecientes a individuos de la etnia Comcaa'c previamente secuenciados de genoma completo. Se realizaron filtros y caracterizaciones a nivel estructural en búsqueda de propiedades de interés que permitieran definir desde una perspectiva genómica y proteómica el contexto particular de la comunidad Comcaa'c o Seri. Aún así, el genoma completo de los nativos de Comcaa'c permanece en gran medida sin describir, ya que no se ha indagado en variantes indel, no-codificantes, variantes intergénicas, transcritos exónicos no-codificantes, sitios de unión a TF, entre otras características de interés biomédico, genético y poblacional, que pudieran estar modificando algunos fenotipos de interés.

Respecto al modelado de proteínas variantes, se realizaron un total de 14 modelos, 7 con las secuencias nativas y 7 con las secuencias variantes. Después del análisis de las anotaciones y de las estructuras modeladas, únicamente el dominio de la cadena pesada de la dineína, codificado por el gen DNHD1 fue candidato a ser analizado a nivel estructural debido a las características de posicionamiento del aminoácido de interés. El gen DNHD1 codifica para el dominio de la cadena pesada de la dineína, una estructura importante dentro de un complejo proteico motor presente en cilios y flagelos. La variante anotada por las herramientas bioinformáticas fue encontrada en una secuencia muy probablemente perteneciente a un dominio AAA+ de unión a ATP. Esto podría sugerir alguna modificación en la eficacia de la unión de nucleótidos a este dominio proteico, y por lo tanto, algún cambio en la funcionalidad estructural final de la cadena. A pesar de eso, es notable la cantidad de homólogos de este gen en el genoma humano, y el considerable tamaño del péptido, por lo que es probable que la variación no genere cambios relevantes en sus funciones cuando tomamos en cuenta el contexto celular completo.

Este estudio propone un panorama contextual genómico y proteómico al definir las variantes particulares presentes en la población, así como predicciones del comportamiento estructural de estas. No obstante, más análisis son necesarios para concluir si las variantes aquí presentadas podrían tener impactos funcionales a nivel biológico, médico o antropológico. Finalmente, este estudio tiene potencial de ser utilizado como referente para el análisis de otras poblaciones con características semejantes.

11. Perspectivas

El genoma completo de los nativos de Comcaa'c permanece en gran medida sin describir. De las 10,771 mutaciones que inicialmente fueron procesadas y anotadas después del primer filtro de identificación de variantes particulares, únicamente 7 fueron estudiadas a profundidad estructural. Entre las variantes aún no estudiadas, se encuentran de diversas índoles y regiones, tales como variantes en regiones regulatorias, intrónicas, exónicas, variantes indel,

no-codificantes, variantes intergénicas, transcritos exónicos no-codificantes, sitios de unión a TF, entre otras; implicando relaciones y cambios que aún no se han descrito para el genoma Comcaa'c. El análisis de INDELS y otras variantes como duplicaciones, inversiones, modificaciones en STRs y otros tipos de variaciones, podría complementar el panorama contextual del genoma estudiado.

Este proyecto sienta un precedente en el análisis de variantes particulares llevado hasta el nivel estructural, especialmente en grupos reducidos o particulares, y que podría ser aplicado a otros grupos nativos con características genéticas parecidas a las del grupo Comcaa'c, como Lacandones, Tojolabales, Wixarikas, Tarahumaras, Triquis, que podrían beneficiarse de un estudio como este.

Una de las grandes limitantes que tuvo este trabajo fue la falta de estructuras cristalinas para el modelado de estructuras, por lo que existe un nicho de oportunidad para estructurólogas y estructurólogos que planeen realizar modelos variantes e invariantes de péptidos de interés antropológico, médico y biotecnológico.

Finalmente, es de remarcar que estos estudios exploratorios en comunidades nunca antes estudiadas, sientan bases para el futuro desarrollo de medicina personalizada, ya que nos permite explorar las diferencias con las comunidades ya estudiadas bajo esta perspectiva, pudiendo derivar a su vez, a mediano o largo plazo, en asociaciones certeras de polimorfismos genéticos a fenotipos médicos, tratamientos farmacogenéticos personalizados, estimaciones de riesgo poligénico, epigenética médica, desarrollo de terapias génicas, terapias preventivas, estudios en cáncer, estudios de enfermedades raras, y un largo etcétera, que solo se verá resuelto a lo largo del desarrollo de la investigación en genomas nativos. La era genómica apenas empieza para los grupos nativos americanos. Se espera que en unos años, estas comunidades tengan las mismas oportunidades de ser atendidas en el marco de la medicina genómica gracias a los estudios previos en los genomas, proteomas y otros aspectos moleculares de interés biológico, médico y antropológico de los individuos originarios de todo el mundo.

12. Referencias

Acuña-Alonzo, V., Flores-Dorantes, T., Kruit, J. K., Villarreal-Molina, T., Arellano-Campos, O., Hünemeier, T., Moreno-Estrada, A., Ortiz-López, M. G., Villamil-Ramírez, H., León-Mimila, P., Villalobos-Comparan, M., Jacobo-Albavera, L., Ramírez-Jiménez, S., Sikora, M., Zhang, L.-H., Pape, T. D., Granados-Silvestre, M. de Á., Montufar-Robles, I., Tito-Alvarez, A. M., … Canizales-Quinteros, S. (2010). A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. *Human Molecular Genetics*, *19*(14), 2877–2885. https://doi.org/10.1093/hmg/ddq173

AGRN - Agrin—Homo sapiens (Human) | *UniProtKB* | *UniProt*. (n.d.). Retrieved September 20, 2022, from https://www.uniprot.org/uniprotkb/O00468/entry

Aguilar-Ordoñez, I., Pérez-Villatoro, F., García-Ortiz, H., Barajas-Olmos, F., Ballesteros-Villascán, J., González-Buenfil, R., Fresno, C., Garcíarrubio, A., Fernández-López, J. C., Tovar, H., Hernández-Lemus, E., Orozco, L., Soberón, X., & Morett, E. (2021). Whole genome variation in 27 Mexican indigenous populations, demographic and biomedical insights. *PLOS ONE*, *16*(4), e0249773. https://doi.org/10.1371/journal.pone.0249773

Amos, L. A. (1989). Brain dynein crossbridges microtubules into bundles. *Journal of Cell Science*, *93 (Pt 1)*, 19–28. https://doi.org/10.1242/jcs.93.1.19

Angata, T., Hayakawa, T., Yamanaka, M., Varki, A., & Nakamura, M. (2006). Discovery of Siglec-14, a novel sialic acid receptor undergoing concerted evolution with Siglec-5 in primates. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, *20*(12), 1964–1973. https://doi.org/10.1096/fj.06-5800com

Asai, D. J., & Koonce, M. P. (2001). The dynein heavy chain: Structure, mechanics and evolution. *Trends in Cell Biology*, *11*(5), 196–202. https://doi.org/10.1016/S0962-8924(01)01970-5

Balladares, S., Alaez, C., Pujol, J., Duran, C., Navarro, J. L., & Gorodezky, C. (2002). Distribution of TAP gene polymorphisms and extended MHC haplotypes in Mexican Mestizos and in Seri Indians from northwest Mexico. *Genes & Immunity*, *3*(2), 78–85. https://doi.org/10.1038/sj.gene.6363835

Ballesteros-Villascán, J. (2020). *Desarrollo de una herramienta bioinformática para el análisis poblacional de la variación genómica* [Tesis de Licenciatura]. Benemérita Universidad Autónoma de Puebla.

Bartoloni, W., J, K., A, B., K, S., K, K., J, W., S, A., I, T., B, B.-T., C, R., J, M., Er, M., S, M., N, S., Hs, S., & Se, A. (2000). Cloning and characterization of a putative human glycerol 3-phosphate permease gene (SLC37A1 or G3PP) on 21q22.3: Mutation analysis in two candidate phenotypes, DFNB10 and a glycerol kinase deficiency. *Genomics*, *70*(2). https://doi.org/10.1006/geno.2000.6395

Barton, N. H., & Charlesworth, B. (1984). Genetic Revolutions, Founder Effects, and Speciation. *Annual Review of Ecology and Systematics*, *15*(1), 133–164.

https://doi.org/10.1146/annurev.es.15.110184.001025

Benkert, P., Biasini, M., & Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, *27*(3), 343–350. https://doi.org/10.1093/bioinformatics/btq662

Benkert, P., Tosatto, S. C. E., & Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins*, *71*(1), 261–277. https://doi.org/10.1002/prot.21715

Bernasconi, A., Canakoglu, A., Masseroli, M., & Ceri, S. (2021). The road towards data integration in human genomics: Players, steps and interactions. *Briefings in Bioinformatics*, *22*(1), 30–44. https://doi.org/10.1093/bib/bbaa080

Bierzynska, A., Soderquest, K., Dean, P., Colby, E., Rollason, R., Jones, C., Inward, C. D., McCarthy, H. J., Simpson, M. A., Lord, G. M., Williams, M., Welsh, G. I., Koziell, A. B., Saleem, M. A., NephroS, & UK study of Nephrotic Syndrome. (2017). MAGI2 Mutations Cause Congenital Nephrotic Syndrome. *Journal of the American Society of Nephrology: JASN*, *28*(5), 1614–1621. https://doi.org/10.1681/ASN.2016040387

Bolnick, D. A., Raff, J. A., Springs, L. C., Reynolds, A. W., & Miró-Herrans, A. T. (2016). Native American Genomics and Population Histories. *Annual Review of Anthropology*, *45*(1), 319–340. https://doi.org/10.1146/annurev-anthro-102215-100036

Brandt, P., Ibrahim, E., Bruns, G. A., & Neve, R. L. (1992). Determination of the nucleotide sequence and chromosomal localization of the ATP2B2 gene encoding human Ca(2+)-pumping ATPase isoform PMCA2. *Genomics*, *14*(2), 484–487. https://doi.org/10.1016/s0888-7543(05)80246-0

Burckhalter, D. (2013). William Neil Smith and the Seri Indians: Photographs, Letters and Field Notes. *Journal of the Southwest*, *55*(1), 1–118.

Charles A Janeway, J., Travers, P., Walport, M., & Shlomchik, M. J. (2001). The major histocompatibility complex and its functions. *Immunobiology: The Immune System in Health and Disease. 5th Edition*. https://www.ncbi.nlm.nih.gov/books/NBK27156/

Cho, C., Reck-Peterson, S. L., & Vale, R. D. (2008). Regulatory ATPase Sites of Cytoplasmic Dynein Affect Processivity and Force Generation. *Journal of Biological Chemistry*, *283*(38), 25839–25845. https://doi.org/10.1074/jbc.M802951200

Clark, S. W., & Rose, M. D. (2006). Arp10p Is a Pointed-End-associated Component of Yeast Dynactin. *Molecular Biology of the Cell*, *17*(2), 738–748. https://doi.org/10.1091/mbc.E05-05-0449

Costa-Urrutia, P., Abud, C., Franco-Trecu, V., Colistro, V., Rodríguez-Arellano, M. E., Alvarez-Fariña, R., Acuña Alonso, V., Bertoni, B., & Granados, J. (2020). Effect of 15 BMI-Associated Polymorphisms, Reported for Europeans, across Ethnicities and Degrees of Amerindian Ancestry in Mexican Children. *International Journal of Molecular Sciences*, *21*(2), 374. https://doi.org/10.3390/ijms21020374

Curcio-Morelli, C., Zhang, P., Venugopal, B., Charles, F. A., Browning, M. F., Cantiello, H. F., &

Slaugenhaupt, S. A. (2010). Functional multimerization of mucolipin channel proteins. *Journal of Cellular Physiology*, *222*(2), 328–335. https://doi.org/10.1002/jcp.21956

Denzer, A. J., Brandenberger, R., Gesemann, M., Chiquet, M., & Ruegg, M. A. (1997). Agrin binds to the nerve-muscle basal lamina via laminin. *The Journal of Cell Biology*, *137*(3), 671–683. https://doi.org/10.1083/jcb.137.3.671

Duan, J., Lupyan, D., & Wang, L. (2020). Improving the Accuracy of Protein Thermostability Predictions for Single Point Mutations. *Biophysical Journal*, *119*(1), 115–127. https://doi.org/10.1016/j.bpj.2020.05.020

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjöstedt, E., Lundberg, E., Szigyarto, C. A.-K., Skogs, M., Takanen, J. O., Berling, H., Tegel, H., Mulder, J., … Uhlén, M. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics: MCP*, *13*(2), 397–406. https://doi.org/10.1074/mcp.M113.035600

Feyfant, E., Sali, A., & Fiser, A. (2007). Modeling mutations in protein structures. *Protein Science : A Publication of the Protein Society*, *16*(9), 2030–2041. https://doi.org/10.1110/ps.072855507

Filipovska, A., Razif, M. F. M., Nygård, K. K. A., & Rackham, O. (2011). A universal code for RNA recognition by PUF proteins. *Nature Chemical Biology*, *7*(7), 425–427. https://doi.org/10.1038/nchembio.577

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., & Marks, D. S. (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature*, *599*(7883), Article 7883. https://doi.org/10.1038/s41586-021-04043-8

Galgano, A., Forrer, M., Jaskiewicz, L., Kanitz, A., Zavolan, M., & Gerber, A. P. (2008). Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system. *PloS One*, *3*(9), e3164. https://doi.org/10.1371/journal.pone.0003164

García-Ortiz, H., Barajas-Olmos, F., Contreras-Cubas, C., Cid-Soto, M. Á., Córdova, E. J., Centeno-Cruz, F., Mendoza-Caamal, E., Cicerón-Arellano, I., Flores-Huacuja, M., Baca, P., Bolnick, D. A., Snow, M., Flores-Martínez, S. E., Ortiz-Lopez, R., Reynolds, A. W., Blanchet, A., Morales-Marín, M., Velázquez-Cruz, R., Kostic, A. D., … Orozco, L. (2021). The genomic landscape of Mexican Indigenous populations brings insights into the peopling of the Americas. *Nature Communications*, *12*, 5942. https://doi.org/10.1038/s41467-021-26188-w

Glusman, G., Rose, P. W., Prlić, A., Dougherty, J., Duarte, J. M., Hoffman, A. S., Barton, G. J., Bendixen, E., Bergquist, T., Bock, C., Brunk, E., Buljan, M., Burley, S. K., Cai, B., Carter, H., Gao, J., Godzik, A., Heuer, M., Hicks, M., … Deutsch, E. W. (2017). Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: A proposed framework. *Genome Medicine*, *9*(1), 113. https://doi.org/10.1186/s13073-017-0509-y

Gómez, R., Tapia-Guerrero, Y. S., Cisneros, B., Orozco, L., Cerecedo-Zapata, C., Mendoza-Caamal, E., Leyva-Gómez, G., Leyva-García, N., Velázquez-Pérez, L., &

Magaña, J. J. (2022). Genetic Distribution of Five Spinocerebellar Ataxia Microsatellite Loci in Mexican Native American Populations and Its Impact on Contemporary Mestizo Populations. *Genes*, *13*(1), 157. https://doi.org/10.3390/genes13010157

Grotjahn, D. A., & Lander, G. C. (2019). Setting the dynein motor in motion: New insights from electron tomography. *Journal of Biological Chemistry*, *294*(36), 13202–13217. https://doi.org/10.1074/jbc.REV119.003095

Hassan, Sk. S., Attrish, D., Ghosh, S., Choudhury, P. P., & Roy, B. (2021). Pathogenic perspective of missense mutations of ORF3a protein of SARS-CoV-2. *Virus Research*, *300*, 198441. https://doi.org/10.1016/j.virusres.2021.198441

Hector, D., Carmen, A., Angelica, O., Janette, P. M., Constanza, D., Navarro, J. L., Victor, J., & Clara, G. (1996). DNA profile of class II loci in Seris: A Mexican Indian tribe. *Human Immunology*, *47*(1–2), 62. https://doi.org/10.1016/0198-8859(96)85025-7

Henderson, L. M., Claw, K. G., Woodahl, E. L., Robinson, R. F., Boyer, B. B., Burke, W., & Thummel, K. E. (2018). P450 Pharmacogenetics in Indigenous North American Populations. *Journal of Personalized Medicine*, *8*(1), 9. https://doi.org/10.3390/jpm8010009

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915

Hindorff, L. A., Bonham, V. L., & Ohno-Machado, L. (2018). Enhancing diversity to reduce health information disparities and build an evidence base for genomic medicine. *Personalized Medicine*, *15*(5), 403–412. https://doi.org/10.2217/pme-2018-0037

Hirose, K. (Ed.). (2019). *Handbook of Dynein* (Second edition). Jenny Stanford Publishing.

Hirose, K., Akimaru, E., Akiba, T., Endow, S. A., & Amos, L. A. (2006). Large Conformational Changes in a Kinesin Motor Catalyzed by Interaction with Microtubules. *Molecular Cell*, *23*(6), 913–923. https://doi.org/10.1016/j.molcel.2006.07.020

Infante, E., Olivo, A., Alaez, C., Williams, F., Middleton, D., De la Rosa, G., Pujol, M. j., Durán, C., Navarro, J. l., & Gorodezky, C. (1999). Molecular analysis of HLA class I alleles in the Mexican Seri Indians: Implications for their origin: Class I DNA typing in Mexican Seri Indians. *Tissue Antigens*, *54*(1), 35–42. https://doi.org/10.1034/j.1399-0039.1999.540104.x

INPI. (2017). *Indicadores Socioeconómicos de los Pueblos Indígenas de México, 2015.* gob.mx. http://www.gob.mx/inpi/articulos/indicadores-socioeconomicos-de-los-pueblos-indigenas-de-mexico-2015-116128

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., & Rice, P. (2013). EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, *29*(10), 1325–1332. https://doi.org/10.1093/bioinformatics/btt113

Kivisild, T. (2013). Founder Effect. In *Brenner's Encyclopedia of Genetics* (pp. 100–101).

Elsevier. https://doi.org/10.1016/B978-0-12-374984-0.00552-0

Koonce, M. P., & Samsó, M. (1996). Overexpression of cytoplasmic dynein's globular head causes a collapse of the interphase microtubule network in Dictyostelium. *Molecular Biology of the Cell*, *7*(6), 935–948. https://doi.org/10.1091/mbc.7.6.935

Koop, B. F., Rowen, L., Wang, K., Kuo, C. L., Seto, D., Lenstra, J. A., Howard, S., Shan, W., Deshpande, P., & Hood, L. (1994). The human T-cell receptor TCRAC/TCRDC (C alpha/C delta) region: Organization, sequence, and evolution of 97.6 kb of DNA. *Genomics*, *19*(3), 478–493. https://doi.org/10.1006/geno.1994.1097

Korunes, K. L., & Goldberg, A. (2021). Human genetic admixture. *PLoS Genetics*, *17*(3). https://doi.org/10.1371/journal.pgen.1009374

Kurniawan, J., & Ishida, T. (2022). Protein Model Quality Estimation Using Molecular Dynamics Simulation. *ACS Omega*, *7*(28), 24274–24281. https://doi.org/10.1021/acsomega.2c01475

Li, K., Zhong, Y., Lin, X., & Quan, Z. (2020). Predicting the Disease Risk of Protein Mutation Sequences With Pre-training Model. *Frontiers in Genetics*, *11*, 605620. https://doi.org/10.3389/fgene.2020.605620

Li, M., Maljevic, S., Phillips, A. M., Petrovski, S., Hildebrand, M. S., Burgess, R., Mount, T., Zara, F., Striano, P., Schubert, J., Thiele, H., Nürnberg, P., Wong, M., Weisenberg, J. L., Thio, L. L., Lerche, H., Scheffer, I. E., Berkovic, S. F., Petrou, S., & Reid, C. A. (2018). Gain-of-function HCN2 variants in genetic epilepsy. *Human Mutation*, *39*(2), 202–209. https://doi.org/10.1002/humu.23357

Lindblom, A., & Robinson, P. N. (2011). Bioinformatics for human genetics: Promises and challenges. *Human Mutation*, *32*(5), 495–500. https://doi.org/10.1002/humu.21468

Lopez, M. L., Lo, M., Kung, J. E., Dudkiewicz, M., Jang, G. M., Von Dollen, J., Johnson, J. R., Krogan, N. J., Pawłowski, K., & Jura, N. (2019). PEAK3/C19orf35 pseudokinase, a new NFK3 kinase family member, inhibits CrkII through dimerization. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(31), 15495–15504. https://doi.org/10.1073/pnas.1906360116

Luque, D. (2006). *Naturalezas, saberes y territorios comcáac (seri). Diversidad cultural y sustentabilidad ambiental,*.

Manolio, T. A., Chisholm, R. L., Ozenberger, B., Roden, D. M., Williams, M. S., Wilson, R., Bick, D., Bottinger, E. P., Brilliant, M. H., Eng, C., Frazer, K. A., Korf, B., Ledbetter, D. H., Lupski, J. R., Marsh, C., Mrazek, D., Murray, M. F., O'Donnell, P. H., Rader, D. J., … Ginsburg, G. S. (2013). Implementing genomic medicine in the clinic: The future is here. *Genetics in Medicine*, *15*(4), 258–267. https://doi.org/10.1038/gim.2012.157

Marks, D. S., Hopf, T. A., & Sander, C. (2012). Protein structure prediction from sequence variation. *Nature Biotechnology*, *30*(11), 1072–1080. https://doi.org/10.1038/nbt.2419

Martínez-Cortés, G., Zuñiga-Chiquette, F., Celorio-Sánchez, A. S., Ruiz García, E., Antelo-Figueroa, A. B., Dalpozzo-Valenzuela, V., Valenzuela-Coronado, A., &

Rangel-Villalobos, H. (2019). Population data for 21 autosomal STR loci (GlobalFiler kit) in two Mexican-Mestizo population from the northwest, Mexico. *International Journal of Legal Medicine*, *133*(3), 781–783. https://doi.org/10.1007/s00414-018-1950-1

Martínez-Tagüeña, N., & Torres Cubillas, L. A. (2018). Walking the desert, paddling the sea: Comcaac mobility in time. *Journal of Anthropological Archaeology*, *49*, 146–160. https://doi.org/10.1016/j.jaa.2017.12.004

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369. https://doi.org/10.1038/nrg2344

McCulloh, K. L., Ng, J., Oldt, R. F., Weise, J. A., Viray, J., Budowle, B., Glenn Smith, D., & Kanthaswamy, S. (2016). The genetic structure of native Americans in North America based on the Globalfiler® STRs. *Legal Medicine*, *23*, 49–54. https://doi.org/10.1016/j.legalmed.2016.09.007

Mocz, G., & Gibbons, I. R. (1993). ATP-insensitive interaction of the amino-terminal region of the beta heavy chain of dynein with microtubules. *Biochemistry*, *32*(13), 3456–3460. https://doi.org/10.1021/bi00064a032

Moreno-Estrada, A., Gignoux, C. R., Fernandez-Lopez, J. C., Zakharia, F., Sikora, M., Contreras, A. V., Acuna-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S., Ortiz-Tello, P., Robles, V., Kenny, E. E., Nuno-Arana, I., Barquera-Lozano, R., Macin-Perez, G., Granados-Arriola, J., Huntsman, S., Galanter, J. M., … Bustamante, C. D. (2014). The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science*, *344*(6189), 1280–1285. https://doi.org/10.1126/science.1251688

Murakami, M., Yoshihara, K., Shimbara, S., Lambeau, G., Gelb, M. H., Singer, A. G., Sawada, M., Inagaki, N., Nagai, H., Ishihara, M., Ishikawa, Y., Ishii, T., & Kudo, I. (2002). Cellular arachidonate-releasing function and inflammation-associated expression of group IIF secretory phospholipase A2. *The Journal of Biological Chemistry*, *277*(21), 19145–19155. https://doi.org/10.1074/jbc.M112385200

Nishito, Y., Hasegawa, M., Inohara, N., & Núñez, G. (2006). MEX is a testis-specific E3 ubiquitin ligase that promotes death receptor-induced apoptosis. *The Biochemical Journal*, *396*(3), 411–417. https://doi.org/10.1042/BJ20051814

Norris, F. A., Atkins, R. C., & Majerus, P. W. (1997). The cDNA cloning and characterization of inositol polyphosphate 4-phosphatase type II. Evidence for conserved alternative splicing in the 4-phosphatase family. *The Journal of Biological Chemistry*, *272*(38), 23859–23864. https://doi.org/10.1074/jbc.272.38.23859

Oiwa, K., & Sakakibara, H. (2005). Recent progress in dynein structure and mechanism. *Current Opinion in Cell Biology*, *17*(1), 98–103. https://doi.org/10.1016/j.ceb.2004.12.006

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., Kimura, K., Makita, H., Sekine, M., Obayashi, M., Nishi, T., Shibahara, T., Tanaka, T., Ishii, S., Yamamoto, J., … Sugano, S. (2004). Complete

sequencing and characterization of 21,243 full-length human cDNAs. *Nature Genetics*, *36*(1), 40–45. https://doi.org/10.1038/ng1285

Pan, Q., Nguyen, T. B., Ascher, D. B., & Pires, D. E. V. (2022). Systematic evaluation of computational tools to predict the effects of mutations on protein stability in the absence of experimental structures. *Briefings in Bioinformatics*, *23*(2), bbac025. https://doi.org/10.1093/bib/bbac025

Pandurangan, A. P., & Blundell, T. L. (2020). Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning. *Protein Science : A Publication of the Protein Society*, *29*(1), 247–257. https://doi.org/10.1002/pro.3774

Prabantu, V. M., Naveenkumar, N., & Srinivasan, N. (2021). Influence of Disease-Causing Mutations on Protein Structural Networks. *Frontiers in Molecular Biosciences*, *7*. https://www.frontiersin.org/articles/10.3389/fmolb.2020.620554

Rangel-Villalobos, H., Martínez-Sevilla, V. M., Salazar-Flores, J., Martínez-Cortez, G., Muñoz-Valle, J. F., Galaviz-Hernández, C., Lazalde-Ramos, B. P., & Sosa-Macías, M. (2013). Forensic parameters for 15 STRs in eight Amerindian populations from the north and west of Mexico. *Forensic Science International: Genetics*, *7*(3), e62–e65. https://doi.org/10.1016/j.fsigen.2013.02.003

Rentería Valencia, R. F. (2007). *Seris*. CDI, Comisión Nacional para el Desarrollo de los Pueblos Ind??genas. http://www.cdi.gob.mx/index.php?option=com_docman&task=doc_download&gid=46&Itemid=65

Roberts, A. J., Numata, N., Walker, M. L., Kato, Y. S., Malkova, B., Kon, T., Ohkura, R., Arisaka, F., Knight, P. J., Sutoh, K., & Burgess, S. A. (2009). AAA+ Ring and linker swing mechanism in the dynein motor. *Cell*, *136*(3), 485–495. https://doi.org/10.1016/j.cell.2008.11.049

Rodrigues, C. H., Pires, D. E., & Ascher, D. B. (2018). DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Research*, *46*(W1), W350–W355. https://doi.org/10.1093/nar/gky300

Samsó, M., Radermacher, M., Frank, J., & Koonce, M. P. (1998). Structural characterization of a dynein motor domain. *Journal of Molecular Biology*, *276*(5), 927–937. https://doi.org/10.1006/jmbi.1997.1584

Schroeder, K. B., Schurr, T. G., Long, J. C., Rosenberg, N. A., Crawford, M. H., Tarskaia, L. A., Osipova, L. P., Zhadanov, S. I., & Smith, D. G. (2007). A private allele ubiquitous in the Americas. *Biology Letters*, *3*(2), 218–223. https://doi.org/10.1098/rsbl.2006.0609

Sean, E. (2004). Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology*, *22*(8), 1035–1036. https://doi.org/10.1038/nbt0804-1035

Sheridan, T. E. (Ed.). (1999). *Empire of sand: The Seri Indians and the struggle for Spanish Sonora, 1645-1803*. University of Arizona Press.

Silvanovich, A., Li, M., Serr, M., Mische, S., & Hays, T. S. (2003). The Third P-loop Domain in Cytoplasmic Dynein Heavy Chain Is Essential for Dynein Motor Function and ATP-sensitive Microtubule Binding. *Molecular Biology of the Cell*, *14*(4), 1355–1365. https://doi.org/10.1091/mbc.E02-10-0675

Singh-Malhi, R., Gonzalez-Oliver, A., Schroeder, K. B., Kemp, B. M., Greenberg, J. A., Dobrowski, S. Z., Smith, D. G., Resendez, A., Karafet, T., Hammer, M., Zegura, S., & Brovko, T. (2008). Distribution of Y chromosomes among native North Americans: A study of Athapaskan population history. *American Journal of Physical Anthropology*, *137*(4), 412–424. https://doi.org/10.1002/ajpa.20883

Single nucleotide polymorphism / SNP | Learn Science at Scitable. (2015, November 10). https://web.archive.org/web/20151110112814/http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295

Slatkin, M. (2004). A Population-Genetic Test of Founder Effects and Implications for Ashkenazi Jewish Diseases. *American Journal of Human Genetics*, *75*(2), 282–293.

Smits, J. J., Oostrik, J., Beynon, A. J., Kant, S. G., de Koning Gans, P. A. M., Rotteveel, L. J. C., Klein Wassink-Ruiter, J. S., Free, R. H., Maas, S. M., van de Kamp, J., Merkus, P., DOOFNL Consortium, Koole, W., Feenstra, I., Admiraal, R. J. C., Lanting, C. P., Schraders, M., Yntema, H. G., Pennings, R. J. E., & Kremer, H. (2019). De novo and inherited loss-of-function variants of ATP2B2 are associated with rapidly progressive hearing impairment. *Human Genetics*, *138*(1), 61–72. https://doi.org/10.1007/s00439-018-1965-1

Song, D., Chen, J., Chen, G., Li, N., Li, J., Fan, J., Bu, D., & Li, S. C. (2015). Parameterized BLOSUM Matrices for Protein Alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *12*(3), 686–694. https://doi.org/10.1109/TCBB.2014.2366126

Soto-Ospina, A., Araque Marín, P., Bedoya, G., Sepulveda-Falla, D., & Villegas Lanau, A. (2021). Protein Predictive Modeling and Simulation of Mutations of Presenilin-1 Familial Alzheimer's Disease on the Orthosteric Site. *Frontiers in Molecular Biosciences*, *8*. https://www.frontiersin.org/articles/10.3389/fmolb.2021.649990

Stoneking, M. (2015). *An introduction to molecular anthropology*. Wiley, Blackwell.

Strausberg, R. L., Simpson, A. J. G., & Wooster, R. (2003). Sequence-based cancer genomics: Progress, lessons and opportunities. *Nature Reviews. Genetics*, *4*(6), 409–418. https://doi.org/10.1038/nrg1085

Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., & Schwede, T. (2020). QMEANDisCo—Distance constraints applied on model quality estimation. *Bioinformatics*, *36*(6), 1765–1771. https://doi.org/10.1093/bioinformatics/btz828

Studer, G., Tauriello, G., Bienert, S., Biasini, M., Johner, N., & Schwede, T. (2021). ProMod3—A versatile homology modelling toolbox. *PLoS Computational Biology*, *17*(1), e1008667. https://doi.org/10.1371/journal.pcbi.1008667

Tan, C., Meng, L., Lv, M., He, X., Sha, Y., Tang, D., Tan, Y., Hu, T., He, W., Tu, C., Nie, H., Zhang, H., Du, J., Lu, G., Fan, L., Cao, Y., Lin, G., & Tan, Y.-Q. (2022). Bi-allelic variants

in DNHD1 cause flagellar axoneme defects and asthenoteratozoospermia in humans and mice. *The American Journal of Human Genetics*, *109*(1), 157–171. https://doi.org/10.1016/j.ajhg.2021.11.022

Taylor, E. M., Broughton, B. C., Botta, E., Stefanini, M., Sarasin, A., Jaspers, N. G., Fawcett, H., Harcourt, S. A., Arlett, C. F., & Lehmann, A. R. (1997). Xeroderma pigmentosum and trichothiodystrophy are associated with different mutations in the XPD (ERCC2) repair/transcription gene. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(16), 8658–8663. https://doi.org/10.1073/pnas.94.16.8658

Taylor, W. W. (1972). The hunter‐gatherer nomads of northern Mexico: A comparison of the archival and archaeological records. *World Archaeology*, *4*(2), 167–178. https://doi.org/10.1080/00438243.1972.9979530

Telzer, B. R., & Haimo, L. T. (1981). Decoration of spindle microtubules with Dynein: Evidence for uniform polarity. *The Journal of Cell Biology*, *89*(2), 373–378. https://doi.org/10.1083/jcb.89.2.373

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

The SIGMA Type 2 Diabetes Consortium. (2014). Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*, *506*(7486), 97–101. https://doi.org/10.1038/nature12828

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678. https://doi.org/10.1038/nature05911

Tokunaga, K., Ohashi, J., Bannai, M., & Juji, T. (2001). Genetic link between Asians and native Americans: Evidence from HLA genes and haplotypes. *Human Immunology*, *62*(9), 1001–1008. https://doi.org/10.1016/S0198-8859(01)00301-9

Types of variants | Garvan Institute of Medical Research. (n.d.). Retrieved September 19, 2022, from https://www.garvan.org.au/research/kinghorn-centre-for-clinical-genomics/learn-about-genomics/for-gp/genetics-refresher-1/types-of-variants

Valentin, E., Singer, A. G., Ghomashchi, F., Lazdunski, M., Gelb, M. H., & Lambeau, G. (2000). Cloning and Recombinant Expression of Human Group IIF-Secreted Phospholipase A2. *Biochemical and Biophysical Research Communications*, *279*(1), 223–228. https://doi.org/10.1006/bbrc.2000.3908

Vallee, R. B., Williams, J. C., Varma, D., & Barnhart, L. E. (2004). Dynein: An ancient motor protein involved in multiple modes of transport. *Journal of Neurobiology*, *58*(2), 189–200. https://doi.org/10.1002/neu.10314

Van Etten, J., Schagat, T. L., Hrit, J., Weidmann, C. A., Brumbaugh, J., Coon, J. J., & Goldstrohm, A. C. (2012). Human Pumilio proteins recruit multiple deadenylases to efficiently repress messenger RNAs. *The Journal of Biological Chemistry*, *287*(43),

36370–36383. https://doi.org/10.1074/jbc.M112.373522

Villalpando-Canchola, E. (1992). ¿Encuentro o exterminio? Una historia entre los comcaac. *Memorias Del XVII Simposio de Historia y Antropología de Sonora*, *1*, 1–12.

Wang, B., Xu, X., Yang, Z., Zhang, L., Liu, Y., Ma, A., Xu, G., Tang, M., Jing, T., Wu, L., & Liu, Y. (2019). POH1 contributes to hyperactivation of TGF-β signaling and facilitates hepatocellular carcinoma metastasis through deubiquitinating TGF-β receptors and caveolin-1. *EBioMedicine*, *41*, 320–332. https://doi.org/10.1016/j.ebiom.2019.01.058

Wang, S., Lewis, C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M. V., Molina, J. A., Gallo, C., Mazzotti, G., Poletti, G., Hill, K., Hurtado, A. M., Labuda, D., Klitz, W., Barrantes, R., Bortolini, M. C., Salzano, F. M., … Ruiz-Linares, A. (2007). Genetic Variation and Population Structure in Native Americans. *PLoS Genetics*, *3*(11), e185. https://doi.org/10.1371/journal.pgen.0030185

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, *46*(W1), W296–W303. https://doi.org/10.1093/nar/gky427

Wojczynski, M. K., Li, M., Bielak, L. F., Kerr, K. F., Reiner, A. P., Wong, N. D., Yanek, L. R., Qu, L., White, C. C., Lange, L. A., Ferguson, J. F., He, J., Young, T., Mosley, T. H., Smith, J. A., Kral, B. G., Guo, X., Wong, Q., Ganesh, S. K., … Reilly, M. P. (2013). Genetics of coronary artery calcification among African Americans, a meta-analysis. *BMC Medical Genetics*, *14*, 75. https://doi.org/10.1186/1471-2350-14-75

Yuen, M., Sandaradura, S. A., Dowling, J. J., Kostyukova, A. S., Moroz, N., Quinlan, K. G., Lehtokari, V.-L., Ravenscroft, G., Todd, E. J., Ceyhan-Birsoy, O., Gokhin, D. S., Maluenda, J., Lek, M., Nolent, F., Pappas, C. T., Novak, S. M., D'Amico, A., Malfatti, E., Thomas, B. P., … Clarke, N. F. (2014). Leiomodin-3 dysfunction results in thin filament disorganization and nemaline myopathy. *The Journal of Clinical Investigation*, *124*(11), 4693–4708. https://doi.org/10.1172/JCI75199

Zeng, Y., Lin, D., Gao, M., Du, G., & Cai, Y. (2022). Systematic evaluation of the prognostic and immunological role of PDLIM2 across 33 cancer types. *Scientific Reports*, *12*(1), 1933. https://doi.org/10.1038/s41598-022-05987-1