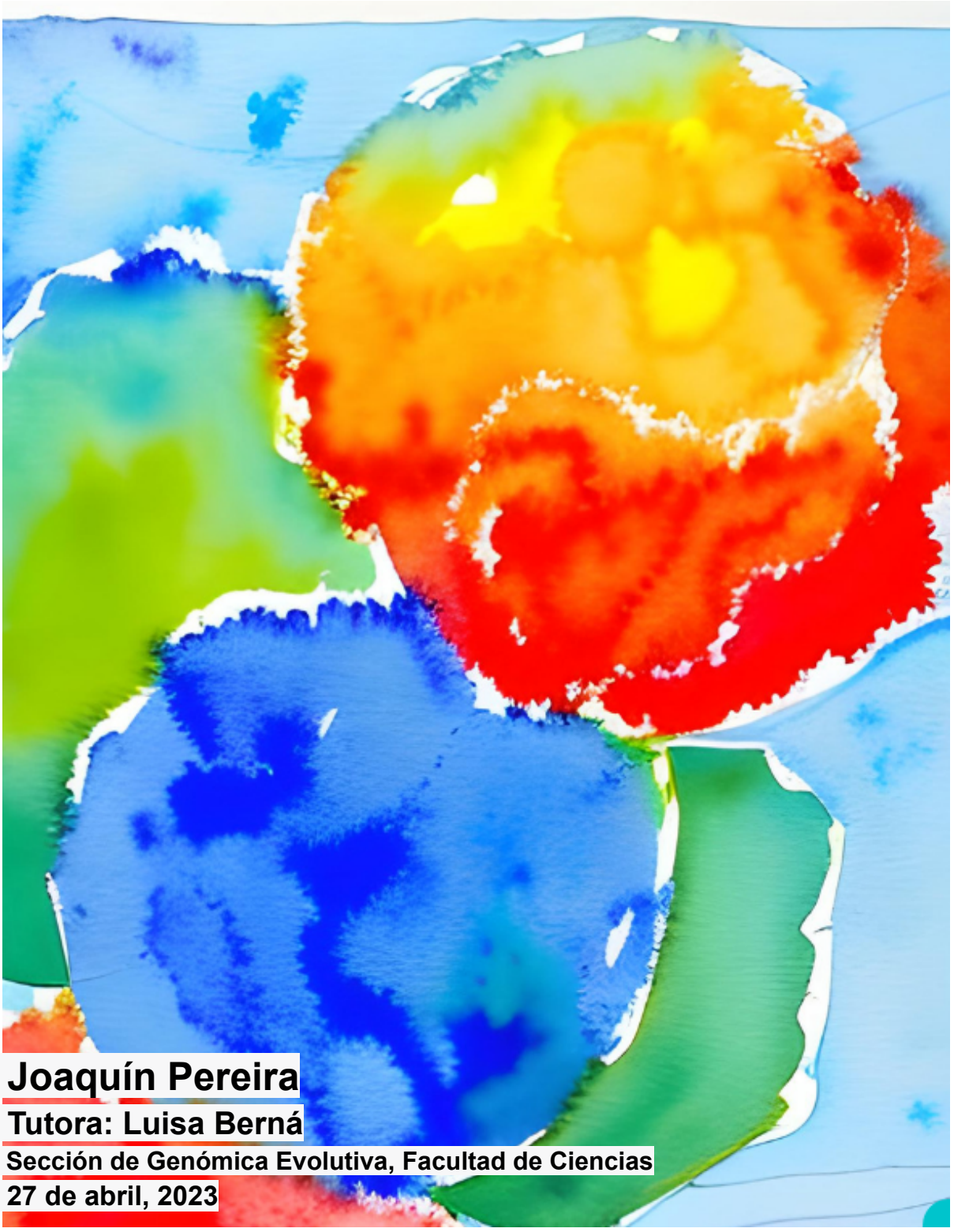
Análisis de familias multigénicas de *Toxoplasma gondii* y *Neospora caninum*



Joaquín Pereira

Tutora: Luisa Berná

Sección de Genómica Evolutiva, Facultad de Ciencias

27 de abril, 2023

Agradecimientos

Primero que nada, quiero agradecer a mi tutora de tesina, la Dra. Luisa Berná, por haber sido mi guía y mentora durante todo el proceso de escritura. Agradezco especialmente por compartir su tiempo, conocimiento y experiencia conmigo.

Asimismo, deseo agradecer a mi familia por darme todo su amor, en especial a mi abuela, mis padres y mis hermanas; así como también a mis cuñados y los diversos animales de la casa.

Por último, quiero agradecer y reconocer a la maestra Mabel Quintana y al profesor Sebastián Rodríguez por haber jugado un rol fundamental en mi educación durante mi niñez y adolescencia.

Resumen

El phylum Apicomplexa es un grupo diverso de parásitos protozoarios intracelulares que incluye más de 6,000 especies descritas. Este phylum se compone de cuatro grupos taxonómicos importantes que comparten la presencia de un complejo apical en una o más etapas de su ciclo de vida. Algunos de estos parásitos tienen un impacto significativo en la salud humana y animal, como *Toxoplasma gondii*, *Neospora caninum* y *Plasmodium falciparum*. La diversificación, la duplicación y la expansión de los loci son comunes en los genomas de Apicomplexa, especialmente en los genes que codifican proteínas presentes en las rutas de secreción del parásito y/o que son expresados en su superficie.

Las diferencias fenotípicas entre diferentes especies de coccidios pueden estar relacionadas a las familias de genes GRA, ROP y SRS. La familia SRS es considerada una de las familias de proteínas más divergentes dentro de los Apicomplexa. La expansión y la duplicación génica en las familias multigénicas, como la familia SRS, se han utilizado para estudiar las diferencias fenotípicas y las relaciones filogenéticas entre especies de Apicomplexa, como *T. gondii* y *N. caninum*. Estudios recientes han encontrado que *N. caninum* tiene más genes SRS que *T. gondii*, lo que sugiere que la familia SRS es una fuente importante de divergencia entre estas dos especies.

En este estudio, se diseñó un pipeline automático para identificar y caracterizar familias multigénicas utilizando anotaciones genómicas. Este agrupa genes según su homología, y luego realiza una búsqueda para identificar clusters de genes relevantes para el análisis. A partir de estos clusters, se generan tablas de datos, árboles filogenéticos y diversas visualizaciones de apoyo. Este pipeline se aplicó al análisis de las familias multigénicas SRS, ROP, MIC y GRA en *T. gondii* y *N. caninum*, con especial énfasis en la familia SRS, utilizando dos estrategias de clusterización distintas. La estrategia de clusterización conjunta fue particularmente útil, permitiendo el uso complementario de las anotaciones de ambos parásitos, lo cual fue crucial para identificar clusters relacionados a las familias ROP y GRA en *N. caninum*, dado este parásito carecía anotaciones para una gran cantidad de genes pertenecientes a estas familias.

El análisis realizado sobre la familia SRS produjo resultados que concuerdan con los hallazgos de otros autores. Destacándose la expansión de la familia SRS en *N. caninum* respecto a *T. gondii,* la presencia de estos genes en todos los cromosomas de ambos parásitos, estando comúnmente agrupados en tandems; y la mayor similitud entre los genes de un mismo tandems que los pertenecientes de otros. Además se produjo evidencia para incorporar tres genes hipotéticos de *N. caninum* a la familia SRS

En conclusión, esta tesina de grado se centró en el estudio de las familias multigénicas en Apicomplexa, específicamente en *T. gondii* y *N. caninum*. Los objetivos propuestos fueron cumplidos mediante el desarrollo de un pipeline para la identificación de familias génicas y la aplicación de diferentes técnicas de análisis para estudiar la evolución y la divergencia de la familia SRS. Proponiendose utilizar este pipeline para contribuir en el entendimiento de la diversidad y la adaptación de estos parásitos. Siendo relevante para el desarrollo de estrategias de control y prevención de enfermedades causadas por ellos.

Introducción

Apicomplejos

Los apicomplejos son un phylum de parásitos intracelulares protozoos, que contiene alrededor de 6000 especies descritas. Este phylum se compone por cuatro grandes grupos taxonómicos incluyendo coccidia, heamosporidinas, piroplasmidos y gregarines. Todos los miembros de este phylum comparten la presencia de un complejo apical en una o más etapas de su ciclo de vida (Morrison, 2009). Varios integrantes de este phylum tienen un destacado impacto en la salud humana y de otros animales; como es el caso de *Toxoplasma gondii*, *Neospora caninum* y *Plasmodium falciparum*, que causan respectivamente toxoplasmosis, neosporosis y malaria (Khan et al., 2020; Nkumama et al., 2017; Tenter et al., 2000)

Epidemiología de *T. gondii y N. caninum*

T. gondii es uno de los parásitos zoonóticos más exitosos, siendo considerado capaz de infectar cualquier animal de sangre caliente. Se estima que un tercio de la población mundial se encuentra crónicamente infectada con este parásito, causando una variedad de enfermedades. Dentro de las manifestaciones clínicas mas dramáticas se encuentran los abortos y el daño fetal. Dependiendo de la etapa del embarazo en que se encuentre la madre al ser infectada, se han observado síntomas severos, como hidrocefalia, toxoplasmosis congénita, deficiencias neurológicas, sordera, convulsiones, retinocoroiditis que llevan a lesiones oculares y ceguera. En personas con SIDA u otras inmunodeficiencias, *T. gondii* puede también ser fatal. Además de su impacto en la salud humana, *T. gondii* ocasiona abortos espontaneos en el ganado, llevando a pérdidas económicas en todo el mundo. Se sabe que los felinos presentan una elevada prevalencia de este parásito. Además se ha observado que la prevalencia de *T. gondii* en otros animales silvestres se correlaciona con la abundancia de felinos en el ambiente (Robert-Gangneux & Dardé, 2012). El cuadro clínico de la toxoplasmosis en la vida silvestre parece ser similar al de humanos, con afectación en muchos casos de ojos, pulmones y cerebro (Robert-Gangneux & Dardé, 2012; Tenter et al., 2000; Wendte et al., 2011).

N. caninum es el agente causante de la neosporosis; una enfermedad con una rango de hospederos más restringido en comparación con *T. gondii,* pero con el cual comparten muchas similaridades. *N. caninum* es incapaz de infectar humanos, sus principales hospederos son los cánidos, los vacunos y las aves (Dubey & Schares, 2011; McCann et al., 2008). Muchos estudios experimentales confirman que hay una fuerte asociación entre los abortos en vacunos y la infección con *N. caninum*, ocasionando por lo tanto un gran impacto económico en las industrias lácteas y cárnicas (Robert-Gangneux & Dardé, 2012). Varios estudios muestran que las pérdidas globales exceden los US$ 1,298 millones por año, concentrándose aproximadamente dos tercios de esas pérdidas en la industria láctea y un tercio en la industria cárnica (Reichel et al., 2013). Estudios clínicos realizados en vacunos muestran que los fetos infectados pueden morir en el útero, ser reabsorbidos,

momificados, nacer con signos clínicos o nacer clínicamente normales pero con una infección persistente (Dubey & Schares, 2011; Ghanem et al., 2009). Manifestaciones clínicas en terneros incluyen manifestaciones neurológicas, peso por debajo del promedio y ocasionalmente defectos de nacimiento como hidrocefalia y afinamiento de la médula espinal (Dubey & Schares, 2011). Las infecciones con *N. caninum* son frecuentes en poblaciones de cánidos (Barber et al., 1997), aunque los casos clínicos son reportados raramente (McInnes, Ryan, et al., 2006). Normalmente la neosporosis es asintomática en perros adultos y de edad avanzada (Silva & Machado, 2016). Las manifestaciones clínicas de la neosporosis en perros pueden ser tratadas, pero con limitado éxito (Reichel et al., 2007). Las infecciones más frecuentes y severas ocurren en perros jóvenes (menores a los 6 meses de edad). Las manifestaciones clínicas son normalmente las mismas que las observadas en la toxoplasmosis, pero con predominancia de anormalidades musculares y neurológicas (Dubey, 2003). En la etapa crónica, la mayoría de los animales permanecen asintomáticos. Sin embargo, las hembras durante su gestación pueden experimentar inmunosupresión y reactivar la infección, infectando al feto (Quinn et al., 2002).

Ciclo de vida de *T. gondii* y *N. caninum*

Los ciclos de vida de *T. gondii* y *N. caninum* son similares y presentan tres formas infectivas: taquizoitos, bradizoitos y ooquistes (Figura 1.1A y 1.1B). Ambos son heteroxenos, con una etapa asexual en el hospedero intermediario y una etapa sexual en el hospedero definitivo; la cual se produce en las células epiteliales intestinales de su hospedero definitivo, produciendo millones de ooquistes (Barber & Trees, 1998; Dubey, 2007). Ambos parásitos pueden propagarse de forma vertical por la infección transplacentaria de taquizoitos a fetos en desarrollo, y, horizontalmente a través de la ingestión de comida o agua contaminada con quistes o bradizoitos. De todas formas, en *T. gondii* suele primar la transferencia horizontal mientras que en *N. caninum* lo hace la transferencia vertical (Barber & Trees, 1998; Dubey, 2007; Tenter et al., 2000). También se diferencian en sus hospederos definitivos, siendo los felinos en *T. gondii* y los cánidos en *N. caninum* (Dubey, 2007).

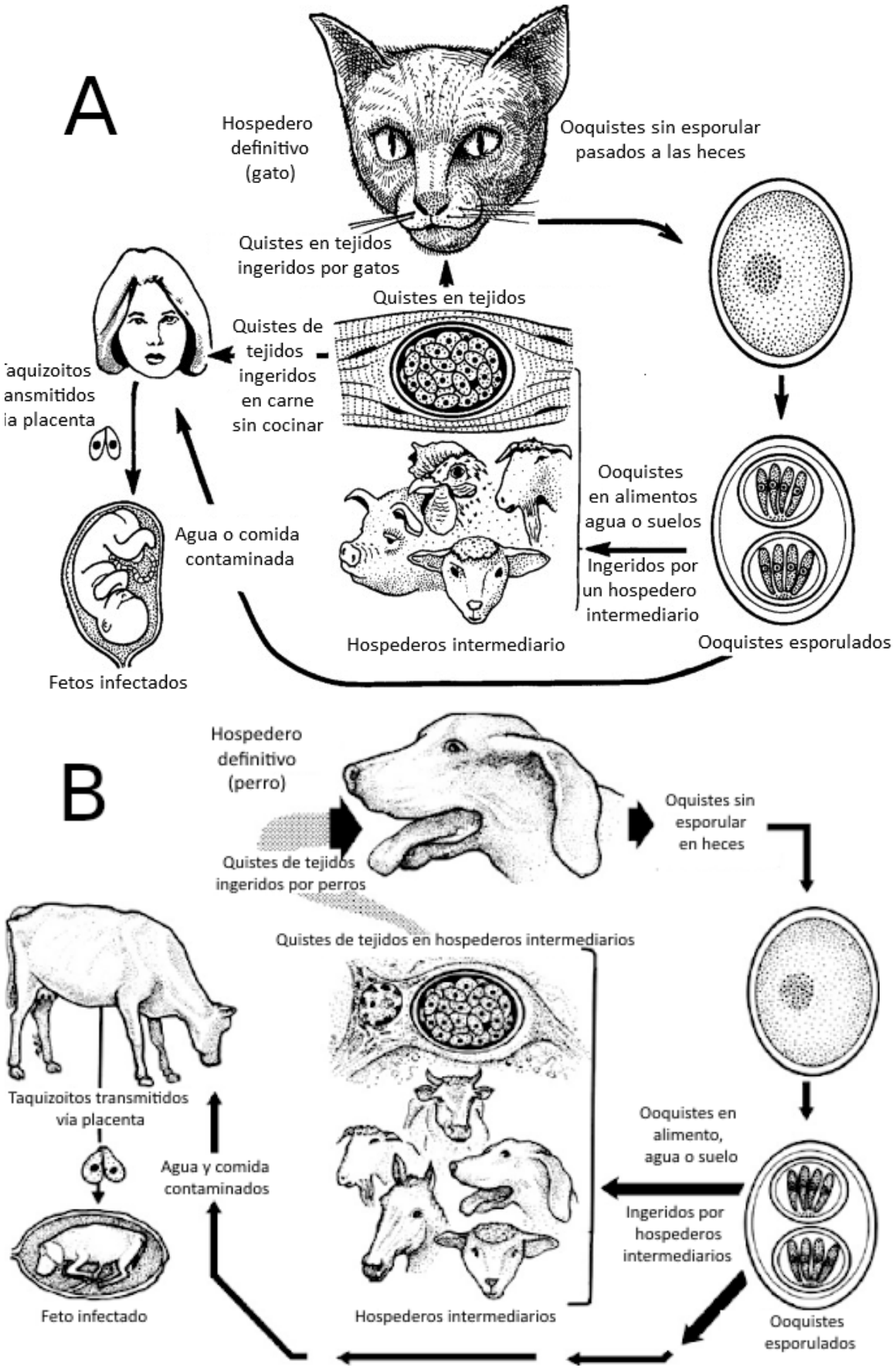


Figura 1.1. (A) Ciclo de vida de *Toxoplasma gondii*, (B) ciclo de vida de *Neospora caninum* (Dubey, 2003, 2004).

Cuando los hospederos intermediarios ingieren sustancias contaminadas con ooquistes esporulados, los esporozoitos son liberados, diferenciándose luego a taquizoitos, los cuales se esparcen a distintos órganos y tejidos a través de la sangre o linfa empezando la fase asexual (Dubey, 2003; Tenter et al., 2000). Los taquizoitos se multiplican asexualmente en vacuolas parasitóforas en sus hospederos intermediarios mediante un proceso llamado endogenia (Gubbels et al., 2008). En la etapa latente de la infección los taquizoitos son convertidos a bradizoitos, los cuales se agrupan en quistes ubicados principalmente en los músculos y el sistema nervioso.

Organelos secretores de *T. gondii* y *N. caninum*

Aunque los apicomplejos infectan una gama muy diversa de huéspedes, su mecanismo de invasión se encuentra relativamente conservado, siendo este esté muy similar en *T. gondii* y *N. caninum*. La mayoría de las etapas invasivas se basan en una forma activa de movimiento llamada motilidad deslizante para impulsarse hacia la célula huésped, mediante un sistema de actomiosina denominado glideosoma (Fr enal et al., 2010). Es esencial para este proceso la liberación oportuna y secuencial del contenido de los organelos secretores ubicados en el extremo apical del parásito (Carruthers & Sibley, 1997). Estos organelos secretores se denominan micronemas, roptrias y gránulos densos (Fig. 1.2) (Paredes-Santoset al., 2012) y presentan diferencias a nivel ultraestructural en *T. gondii* y *N. caninum* (Tabla 1.1).

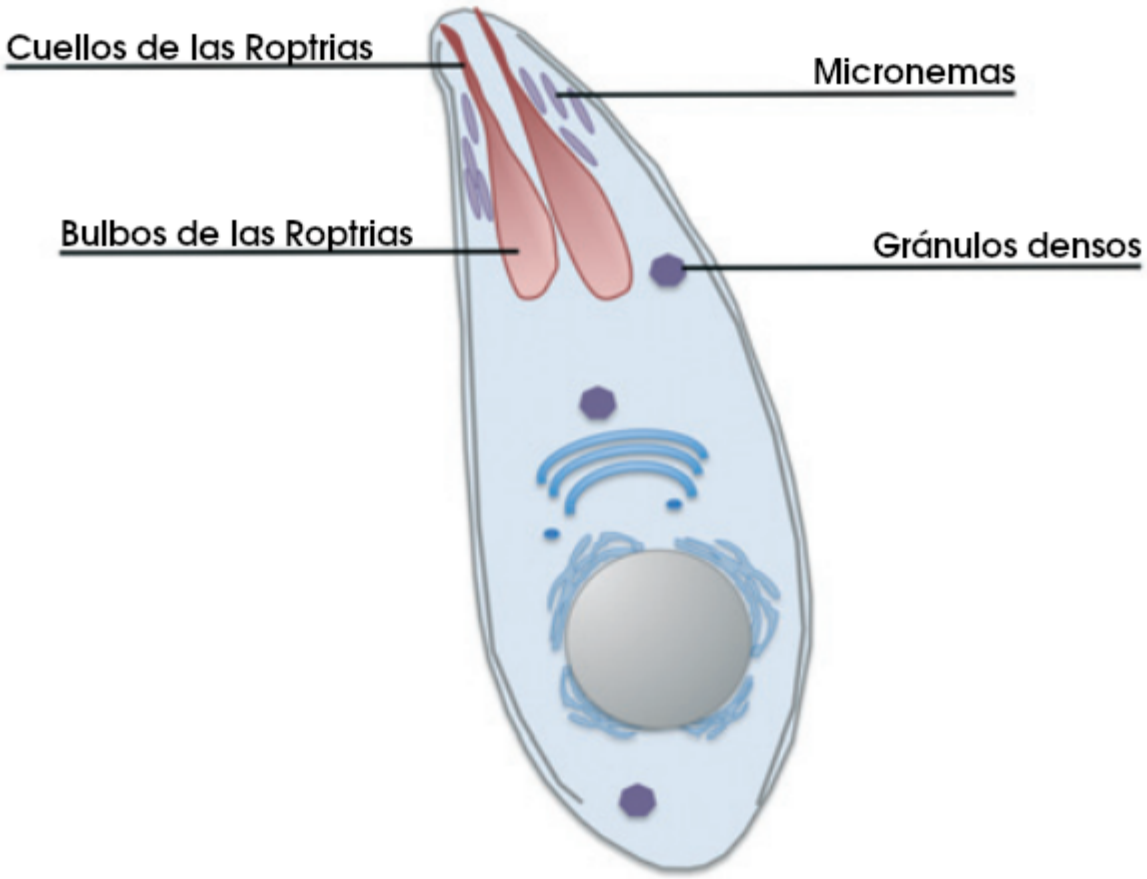


Figura 1.2. Representación esquemática de los organelos secretores de *T. gondii.* Representación esquemática de un taquizoito de *T. gondii* en en donde señalan sus organelos secretores (roptrias, micronemas y gránulos densos) (Kemp et al., 2013).

Tabla 1. Diferencias ultraestructurales entre taquizoitos, quistes de tejido y bradizoitos de *N. caninum y T. gondii* (Speer et al., 1999)

Etapa parasitaria/estructura	*N. caninum*	*T. gondii*
Taquizoito		
Roptrias anteriores[a]	Electrón-denso, 6-16	Laberínticos, 4-10
Roptrias posteriores[b]	Electrón-denso	Raros
Roptrias enrolladas	Electrón-denso, 1-2	Raros
Micronemas anteriores	Muchos	Pocos
Micronemas posteriores	Pocos	Raros
Gránulos densos anteriores	Varios	Varios
Gránulos densos posteriores	Varios	Pocos
Microporos	Raros	Comunes
Pared de quistes de tejido		
Grosor	0.5-0.4μm	< 0.5μm
Contorno	Irregular	Suave
Bradizoitos		
Roptrias anteriores	Electrón densas, 6-12	Laberínticas en quistes jóvenes, electrón-densa en maduros, 6-8
Roptrias posteriores	Electrón densas, raras	Ninguna
Roptrias enrolladas	Ninguna	Electrón densas, 1-3
Micronemas anteriores	Muchas	Muchas
Micronemas posteriores	Pocas	Raras
Gránulos densos anteriores	Varios	Varios
Gránulos densos posteriores	Varios	Raros
Gránulos densos pequeños	4-8	Ninguna
Microporos	Raros	Común

[a] Anterior respecto a la posición del núcleo
[b] Posterior respecto a la posición del núcleo

Micronemas

Los micronemas son los organelos más pequeños encontrados en los parásitos apicomplejos. Las proteínas de los micronemas (MICs) son secretadas en un proceso dependiente de calcio y se encuentran involucradas en la invasión de la célula hospedera, en su unión, y en la motilidad de *T. gondii* y *N. caninum* (Dubremetz & Lebrun, 2012). Casi todas las MICs tienen al menos un dominio adhesivo encontrado en vertebrados, el cual permite la unión a la superficie de la célula hospedera, construyendo un complejo estable de gran importancia en el proceso de invasión (Carruthers, 2002; Dubremetz & Lebrun, 2012)

Roptrias

Las roptrias son organelos secretores que se encuentran mayormente localizadas en el extremo anterior de los taquizoitos de *T. gondii* y *N. caninum* (Tabla 1.1) (Speer et al., 1999). Estos son largos con forma de garrote, teniendo en una región con apariencia de bulbo y otra más fina electrón-densa denominada cuello que se extiende hasta el extremo apical del parásito (Speer et al., 1999). Durante el proceso de invasión, el cuello de la roptria sirve como ducto para secretar el contenido de este organelo (Dubremetz, 2007).

Las proteínas del cuello de las roptrias (RONs) se encuentran conservadas dentro de los apicomplejos. Están involucradas en la formación de la unión móvil y en la formación de la vacuola parasitófora (VP) (Bradley et al., 2005; Lebrun et al., 2005). Las proteínas del bulbo de las roptrias (ROPs) integran la membrana de la VP e influyen sustancialmente en la virulencia del parásito (Beckers et al., 1994; El Hajj et al., 2007). Hasta el momento se conoce una menor variedad de RONs que de ROPs, y además se considera que cada

grupo de proteínas son codificados por genes que pertenecen a familias multigénicas diferentes (Boothroyd & Dubremetz, 2008).

Gránulos densos

Los gránulos densos son vesículas esféricas electrón densas que varían en número dependiendo de la etapa del ciclo de vida del parásito (Tabla 1.1). Las proteínas de los gránulos densos (GRAs), son importantes para contribuir y mantener la VP, se ubican también en las paredes de los quistes y se encuentran involucradas en diferentes interacciones que ocurren entre los parásitos y las células hospederas.(Gold et al., 2015; Michelin et al., 2009)

Genomas de *T. gondii* y *N. caninum*

A la fecha existen varios genomas completamente secuenciados para *T. gondii* y *N. caninum* usando lecturas provenientes de ensambladores de segunda y tercera generación, dentro de los genomas completos disponibles de *T. gondii*, la cepa ME49 de *T. gondii* presenta un tamaños de más de 65.6 Mpb, conteniendo ~8330 genes que codifican proteínas y ~600 que transcriben ARNnc. Además su contenido en GC es del 56.4%. A su vez, para la cepa Liverpool de *N. caninum* tiene un genoma nuclear cercano a los 62 Mpb, conteniendo ~7600 genes que codifican proteínas, ~170 que transcriben ARNnc y ~260 pseudogenes. Además su contenido en GC es del 54.8%.Las anotaciones genómicas para ambas especies están disponibles en bases de datos públicas (Benson et al., 2008; Berná et al., 2021; Gajria et al., 2008)

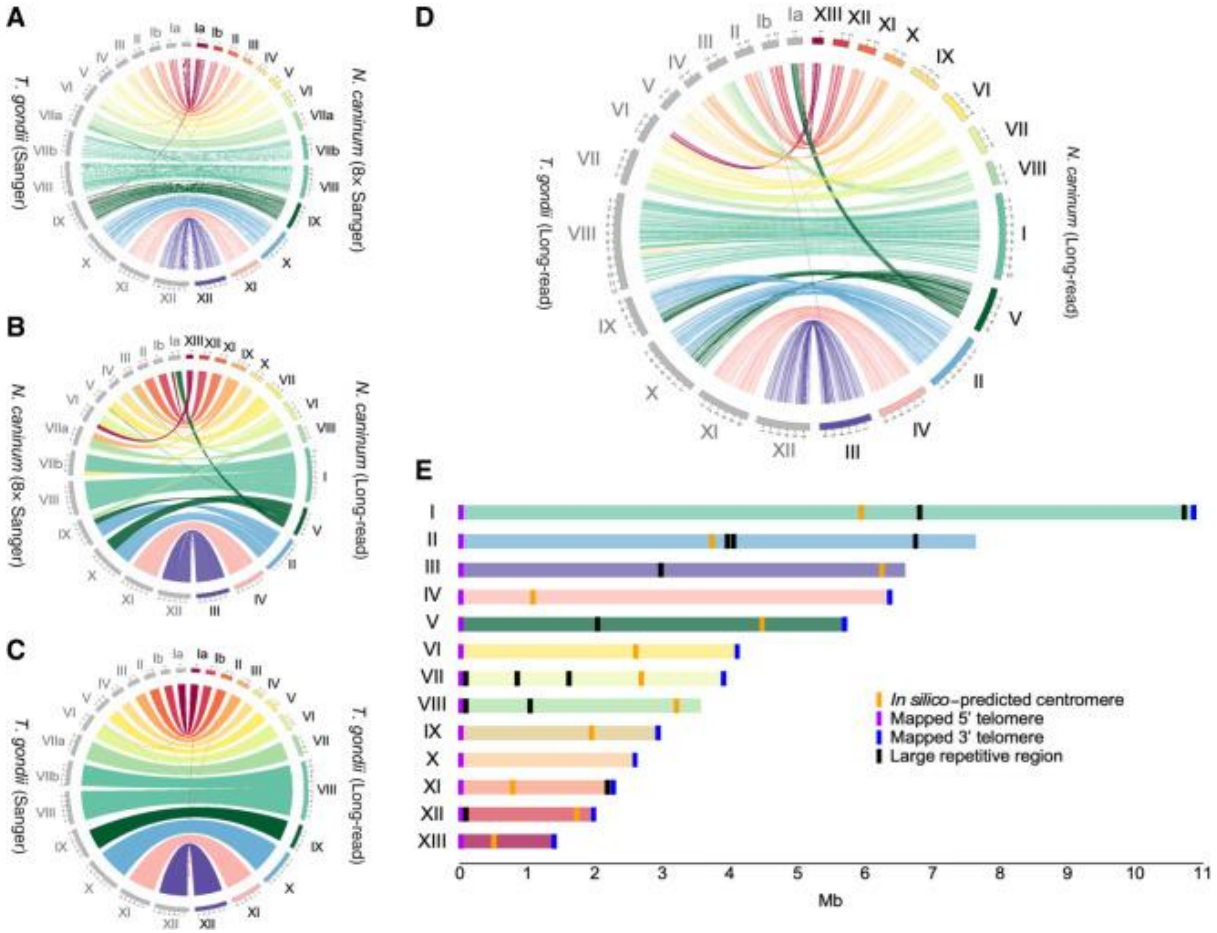


Figura 1.3. Análisis comparativos de los ensamblajes de *Neospora caninum* y *Toxoplasma gondii*, usando información de ensambladores de tercera generación se revelan errores de ensamblaje y diferencias en sus cariotipos. (A) Análisis comparativo entre los ensamblajes de los genomas de las cepas *T. gondii* type II (TgME49) y *N. caninum* Liverpool (NcLiv). (B) Alineamiento comparativo de los ensamblajes del genoma de NcLiv usando Sanger y secuenciadores de tercera generación (con lecturas largas). (C) Alineamiento comparativo del genoma de *T. gondii* tipo II (TgME49) basado en datos de la tecnología Sanger y en datos de secuenciadores de tercera generación (lecturas largas) de *T. gondii* tipo I (TgRH). (D) Alineamiento comparativo de los genomas de *T. gondii* tipo I (TgRH) y NcLiv, basados en secuenciadores de tercera generación (lecturas largas). (E) Cromosomas de *N. caninum*. Se muestran cariotipo, largo cromosómico, telómeros, centrómeros putativos, regiones de repetidos largos.

Parentesco evolutivo entre *T. gondii* y *N. caninum*

Estudios comparativos realizados por Reid et al. 2012 muestran que estos dos parásitos divergieron hace aproximadamente 28 millones de años. Mientras que sus hospederos definitivos divergieron hace más de 55 millones de años (Pontius et al., 2007). Siendo una posibilidad que el ancestro común de ambos parásitos tuviera como hospederos definitivos tanto al gato como el perro, pero al divergir ambos parásitos, *T. gondii* y *N. caninum* hayan quedado restringidos a sus hospederos definitivos actuales (Reid). Los genomas de *T. gondii* y *N. caninum* están disponibles hace varios años. Sin embargo, estudios recientes utilizando las nuevas tecnologías de secuenciación con reads largos, identificaron errores en los ensamblajes, concluyendo que ambas especies presentan 13 cromosomas y no 14,

como se consideraba (Berná et al., 2021). A su vez se identificó que durante la evolución de estos genomas, ocurrieron rearreglos y cambios estructurales significativos, cambiando el paradigma de que los genomas de estos parásitos eran casi iguales (Figura 1.3). A pesar de esto, mantienen un contenido génico y expresión génica similar (Reid et al., 2012). Las diferencias que han sido identificadas entre el contenido génico de estos dos parásitos se concentran principalmente en moléculas que controlan la interacción del parásito con la célula hospedera y regulan su nivel de virulencia, principalmente en genes asociados a ciertas familias multigénicas (ROP, MIC, GRA) y las proteínas de superficie (SRS), estimándose que las diferencias en el número de copias de estas familias génicas y su variabilidad, pueden ser causantes de las diferencias de nichos ecológicos y de rango de hospederos que presentan estos parásitos. Siendo dichas diferencias de particular interés dado el cercano parentesco entre ambos parásitos (Reid et al., 2012)

Estructura Poblacional

Toxoplasma gondii

La mayoría de las muestras obtenidas de *T. gondii* son cepas que se clasifican dentro de tres linajes predominantes denominados como tipo I, II y III, teniendo las cepas de cada linaje niveles distintivos de virulencia (Howe & Sibley, 1995; Sibley & Boothroyd, 1992).

El linaje tipo I tiene una amplia distribución geográfica, pero son encontrados con menor frecuencia en comparación con el linaje de tipo II (Hosseini et al., 2019; Shwab et al., 2014). Ensayos in vivo muestran que todas las cepas del linaje tipo I son altamente virulentas, causando la muerte de todos los ratones infectados con 10 o menos taquizoitos.
Además ensayos in vitro muestran que las cepas del linaje tipo I tienen facilidad para migrar a través de un epitelio polarizado o a través de la matriz extracelular. También muestran una mayor tasa de penetración de las capas del tubo digestivo (Barragan & Sibley, 2002, 2003). En cultivos celulares, las cepas del linaje tipo I crecen más rápido que las del linaje tipo II o III, y se convierten de taquizoito a bradizoíto más lentamente que las cepas del linaje tipo II (Soete et al., 1993). La mayor tasa de crecimiento de los parásitos del linaje tipo I, dada a su mayor tasa de reinvasion, o su mayor distancia de migración, o a su menor tiempo de replicación; también pueden explicar la mayor carga tisular observada en ratones infectados con las cepas de este linaje (Saeij et al., 2005). Las cepas de GTI y RH pertenecen al linaje tipo I.

El linaje tipo II tiene una amplia distribución geográfica, siendo el linaje que se encuentra con mayor frecuencia en las infecciones de humanos y vida silvestre (Hosseini et al., 2019; Shwab et al., 2014). La cepa ME49 pertenece al linaje tipo II y presenta una baja tasa de multiplicación. Se la considera con una virulencia moderada, causando en ratones una infección crónica de quistes de tejido; aunque produce un deterioro progresivo que lleva a la muerte del ratón (Darde et al., 1988). Ensayos in vitro muestran que tiene la capacidad de cambiar entre fases de taquizoito y bradizoito, y produce quistes con relativa facilidad (Soete et al., 1993).

Por último, las cepas del linaje III, en las que se encuentra la cepa VEG, son encontradas frecuentemente en animales salvajes. El linaje III es considerado raro entre las muestras de

USA y Europa. Diferentes ensayos muestran que cepas de este linaje son avirulentas, presentando un rápido crecimiento pero sólo al inicio de la infección, aunque en ratones produce un deterioro progesivo llevándolos a la muerte (Hosseini et al., 2019; Howe & Sibley, 1995; Jerome et al., 1998; Shwab et al., 2014). Nuevos estudios de diversidad indican que la estructura poblacional de *T. gondii* es mayoritariamente clonal en zonas urbanas, pero que adquiere una mayor diversidad de cepas en regiones silvestres. La mayoría de las nuevas cepas se clasifican como atípicas (exóticas) y se encuentran distribuidas mayormente en América del Sur (Lehmann et al., 2004; Shwab et al., 2014; Sibley et al., 2009; Su et al., 2012)

Neospora caninum

En las últimas dos décadas *N. caninum* ha sido extensamente investigado dado su importancia como parásito veterinario. Como se ha mencionado previamente, es sabido que *N. caninum* tiene una distribución global y causa una enfermedad neuromucscular severa en perros, y abortos y mortalidad neonatal en vacas, resultando en devastadoras pérdidas económicas para las industrías cárnica y lactea (Dubey et al., 2007; Dubey & Schares, 2011). *N. caninum* a diferencia de *T. gondii* no presenta una relación tan clara entre su estructura poblacional y su distribución geográfica. Se hipotetiza que esto es una consecuencia de distintos eventos de migración del parásito, producto de la importación de ganado vacuno en pie (Calarco & Ellis, 2020). También se observa que las poblaciones de *N. caninum* tienden a presentar una estructura clonal, lo cual es esperable ya que este parásito se reproduce principalmente de manera asexual, dado a que las infecciones ocurren principalmente por la vía vertical (Barber & Trees, 1998; Calarco & Ellis, 2020; Schock et al., 2001). Sin embargo, existen niveles significativos de variación entre las cepas de *N. caninum*, ya que hay características genotípicas y fenotípicas que no se encuentran rigurosamente conservadas dentro de la especie (Al-Qassab et al., 2010). Mientras que sólo hay diferencias menores en su ultraestructura, los aislados de *N. caninum* parecen variar en sus características biológicas y genéticas, de los cuales varios estudios reportan diferencias (Al-Qassab et al., 2010; Atkinson et al., 1999; McInnes, Irwin, et al., 2006; Rojo-Montejo et al., 2009). Por ejemplo, la cepa altamente virulenta NC-Liverpool causa daño fetal en vacunos (Atkinson et al., 1999), mientras que la cepa NC-Nowra ha sido evaluada para su uso como una vacuna viva atenuada contra la neosporosis bovina, dada su baja virulencia en modelos de ratones (Weber et al., 2013). A su vez, en ratones NC-Liverpool causa una neosporosis severa, caracterizada por encefalitis, parálisis de miembros posteriores y pérdida severa de peso; mientras que otros aislados como NC-SweB1 y NC-nowra presentan una patología similar pero mucho más leve (Atkinson et al., 1999; Miller et al., 2002).

Familias Multigénicas

Principales familias multigénicas de *T. gondii y N. caninum*

La diversificación, la duplicación y la expansión de los loci es ubicua y prevalente en los genomas apicomplejos, lo que es especialmente cierto para genes que codifican proteínas presentes en las rutas de secreción del parásito, y/o que son expresados en su superficie (Blank & Boyle, 2018). Comparando los genomas de 62 aislados geográficamente dispersos entre los que se encuentran *T. gondii* y *N. caninum*, Lorenzi et al. (2016) reportó que estos

parásitos cercanamente emparentados pero fenotípicamente diversos, podrían ser distinguidos en base a la amplificación en tándem y diversificación de determinantes patogénicos secretados. Al comparar los genes ortólogos de las familias GRA, ROP y SRS de los Coccidios, se sugiere que las diferencias entre estos genes pueden ser responsables de las diferencias fenotípicas observadas entre las distintas especies de Coccidios, mientras los genes de la familia MIC se encuentran altamente conservados (Reid, et al 2012, Lorenzi et al., 2016). Estos resultados sugieren que un pequeño set de genes implicados en las interacciones hospedero parásito, han influenciado los nichos ecológicos y las capacidades patogénicas de estas especies.

Particularmente, la familia de antígenos de superficie (SRS) ha sido identificada como una de las familias de proteínas de más rápida evolución y divergencia dentro de este phylum (Adomako-Ankomah et al., 2014; Jung et al., 2004; Reid et al., 2012; Wasmuth et al., 2012). Las SRS junto con otras familias multigénicas han presentado notorios eventos de duplicación y expansión génica, los cuales han sido usados para estudiar las diferencias fenotípicas y las relaciones filogenéticas entre especies de apicomplejos como *T. gondii* y *N. caninum* (Adomako-Ankomah et al., 2014; Lorenzi et al., 2016; Reid et al., 2012). En particular, Reid et. al. (2012) encontró que en *N. caninum* existen más del doble de genes SRS en comparación con *T. gondii,* además de existir una divergencia entre estas especies en sus factores de virulencias excretados, como es el caso de las ROPs. Por lo tanto, se considera a la familia SRS como una fuente de divergencia entre estas dos especies.

La familia multigénica de SRS

La superficie de los coccidios *T. gondii* y *N. caninum* está cubierta con una gama de antígenos anclados a glicosilfosfatidilinositol (GPI) denominados antígenos de superficie (SAGs), miembros de la familia de secuencias relacionadas a SAG1 (SRSs). El anclaje a GPI también actúa como señalizador, permitiendo que las proteínas SRS se anclen a la superficie del parásito sin necesidad de ser almacenados por un gránulo secretor (Seeber et al., 1998). Todos los genes de esta familia codifican al menos un dominio SRS, de alrededor 20 kDa que forma una única estructura a través de puentes disulfuro entre 4 o 6 residuos de cisteína conservados (He et al., 2002)

La familia SRS se divide en 8 subfamilias (Wasmuth et al., 2012), siendo los miembros de estas muy heterogéneos, con una similitud que varía del 25% al 97% en su secuencia aminoacídica (He et al., 2002). La región más variable de las SRS se encuentra en los dominios de unión a las células del hospedero y la más conservada es la que se encuentra próxima al anclaje GPI en el extremo C-terminal. (Wasmuth et al., 2012). Consistentemente con su variación estructural, los miembros de esta familia cumplen una gran variedad de funciones, como lo son la fijación a las células del hospedero, modulación de la respuesta inmune del hospedero y la regulación de la virulencia parasitaria (Lekutis et al., 2001; Wasmuth et al., 2012)

El estudio de las proteínas SRS es de particular interés debido a su papel en la modulación de la respuesta inmune, tanto del hospedero como del parásito. Esta característica las convierte en posibles objetivos para el desarrollo de nuevas terapias farmacológicas o vacunas (Cruz-Mirón et al., 2021). Siendo de especial importancia facilitar la investigación en este campo, mediante el desarrollo de herramientas bioinformáticas que permitan la

realización de estudios reproducibles y que contribuyan a un mayor entendimiento de la familia de proteínas SRS.

Objetivo general

Desarrollo de un pipeline para identificación de familias génicas en apicomplejos, y su aplicación en la identificación y el estudio evolutivo de familias específicas en *T. gondii* y *N. caninum*.

Objetivos específicos

- Desarrollo de un pipeline automático para la identificación de familias génicas
- Clusterización de genes por homología
- Caracterización de los clusters y generación de estadísticos descriptivos
- Análisis de composición de nucleótidos y aminoácidos de las familias identificadas
- Análisis comparativo de la familia SRS
- Análisis filogenético de la familia SRS de *T. gondii y N. caninum*

Metodología

Datos genómicos

El genoma de *T. gondii* ME49 (TgME49) junto con anotaciones fueron obtenidos de los repositorios públicos GenBank (ABPA02000000) y ToxoDB (www.toxodb.org). El genoma fue secuenciado con una profundidad de cobertura 26.6X usando 454 GS FLX Titanium y tecnología Sanger. Este fue ensamblado con Celera Assembler v. 1.92 (Myers et al., 2000) y las anotaciones estructurales fueron obtenidas con Evidence Modeler (EVM) (Haas et al., 2008). El proceso de secuenciado y anotado fue realizado por Lis Caler en el Instituto J. Craig Venter. Más características de la secuenciación TgME49 se encuentran resumidas en la Tabla 2.1

El genoma de *N. caninum* LIV (NcLIV) junto con sus anotaciones fueron obtenidos de los repositorios públicos se GenBank (GCA_016097395.1) y ToxoDB (www.toxodb.org). El genoma fue secuenciado con una profundidad de cobertura mayor a 100X usando tecnología de PacBio y Oxford Nanopore, además fue corregido con lecturas de Illumina. Las lecturas de PacBio fueron ensambladas usando el programa HGAP Assembly (Chin et al. 2013). Las lecturas de Oxford Nanopore fueron ensambladas usando Canu (Koren et al. 2017). Los ensamblajes fueron unidos usando el programa Quickmerge (Chakraborty et al. 2016). Las anotaciones genómicas fueron realizadas usando la herramienta automática de anotación COMPANION (Steinbiss et al. 2016) usando AUGUSTUS con el apoyo de datos de RNA-seq. El secuenciado y anotado fueron realizados por L. Berná et al. en el Institut Pasteur de Montevideo. Más características de la secuenciación de NcLIV se encuentran disponibles en la Tabla 2.1

Tabla 2.1. Principales características del ensamblaje de *T. gondii* ME49 y *N. caninum* LIV. Los genomas de estos organismos junto con sus pueden ser encontrados en en GenBank (ABPA02000000 y GCA_016097395.1) y en ToxoDB (www.toxodb.com).

	T. gondii ME49	*N. caninum* LIV
Tamaño total	62.4 Mb	61.5 Mb
Cobertura	26.6X	162.0X
Cromosomas	14	13
N° de contigs	2,511	44
N° de scaffolds	2,227	44
Scaffold N50	4.7 Mb	6.4 Mb
Scaffold L50	6	4
Tecnología de secuenciado	454 GS FLX Titanium, Sanger	Illumina HiSeq; MinION;PacBio RS
Método de ensamblaje	Celera Assembler v. 1.92	HGAP Assembly, Canu
Genes	8322	7540
Porcentaje de GC	56.4	54.5

Traducción de secuencias codificantes

A partir de los genomas de *NcLIV* y *TgME49* junto con sus anotaciones correspondientes, usando el programa gffread del paquete EMBOSS (Rice et al., 2000) se obtuvieron los respectivos proteomas de cada organismo.

Clusterización mediante BLASTP-MCL

El proceso de clusterización inicia con la ejecución del BLASTP (Altschul et al., 1997) sobre los proteomas de *N. caninum* y *T. gondii* de forma separada, de donde se obtiene una tabla que relaciona entre sí a todas las secuencias proteicas mediante valores estimadores de identidad media (E-value). A partir de esta tabla, usando el programa mcxload (Dongen, 2000) se ensambló un grafo para cada proteoma, donde cada secuencia representa un nodo, mientras que el logaritmo negativo de los E-values obtenidos (valores de similitud), corresponden a sus vértices. Al incrementarse el valor de similitud entre dos genes, mayor será el largo del vértice que une sus respectivos nodos y más probable que el par de genes presente una relación cercana de homología. Luego a partir de este grafo usando el programa MCL (Dongen, 2000) se obtuvieron clusters de proteínas con valores de similitud elevados, los cuales se esperan que estén enriquecidos en secuencias homólogas. El programa BLASTP fue ejecutado especificando su archivo de salida tenga un formato dado por la opción `-outfmt "6 qseqid sseqid evalue bitscore"`, mientras que para el programa MCL se utilizó una valor de inflación de 4.5, especificado mediante la opción `-I 4.5`.

Extracción, Análisis y Visualización de Datos Genómicos

Para obtener las tablas y valores, así como determinar los atributos básicos de cada gen y las estadísticas relevantes de cada cluster, como la media y el promedio, se analizaron los genomas y sus respectivas anotaciones. Estos datos se extrajeron de archivos en formato '.fasta' y '.gff3' de NcLIV y TgME49. Este proceso de extracción y análisis de datos se realizó a través de la creación de scripts en Bash. Estos scripts incorporaron rutinas de programas especializados en el procesamiento de texto como AWK, Grep, Tr y Sed.

Posteriormente, se elaboraron visualizaciones de los datos a través de la creación de histogramas, gráficos de barras y mapas de calor/dendrogramas. Para esta tarea, se utilizaron los paquetes de Python Numpy, Pyplot y Seaborn. También se incluyó el paquete ChromPlot de R para el mapeo cromosómico de los clusters.

Generación de filogenia

Se usaron un total de 351 secuencias proteicas, pertenecientes a los miembros de los clusters asociados a la familia multigénica SRS. Estas fueron alineadas usando el programa MAFFT v.7.453 (Katoh & Standley, 2013) usando los parámetros por defecto del programa. Luego a partir de este alineamiento se ensambló una filogenia usando el programa IQTREE v.2.0.3 (Minh et al., 2020) con el modelo de sustitución WAG + R6 que fue seleccionado por el programa interno ModelFinder de acuerdo al criterio de información bayesiano (BIC), además se evaluó el sustento de cada nodo usando 5000 réplicas con los métodos de Bootstrap ultra rápido y SH-aLRT mediante las opciones `-bb 5000 -alrt 5000`. La visualización de este árbol filogenético se realiza mediante el paquete Dendextend de R

Desarrollo del pipeline de análisis

Los procesos descritos anteriormente fueron integrados en un pipeline automatizado, facilitando su ejecución de forma reiterada. Este se compone de varios scripts agrupados en varios módulos en los cuales incorporan programas y librerías de Bash, R y Python. Estos módulos fueron diseñados para ser ejecutados en forma secuencial o de forma independiente en caso de que se posean las entradas requeridas. Todos los programas elaborados se encuentran disponibles en el siguiente repositorio de Github (https://github.com/JPereira199/GenomicClusterExplorer).

Resultados

Desarrollo del pipeline

El pipeline desarrollado consta de cuatro módulos principales y uno alternativo usa como entradas a uno o varios genomas en formato '.fasta' junto a sus respectivas anotaciones en

formato ‘.gff3’ (Figura 3.1). La salida del pipeline incluye filogenias, tablas de datos y distintas visualizaciones.

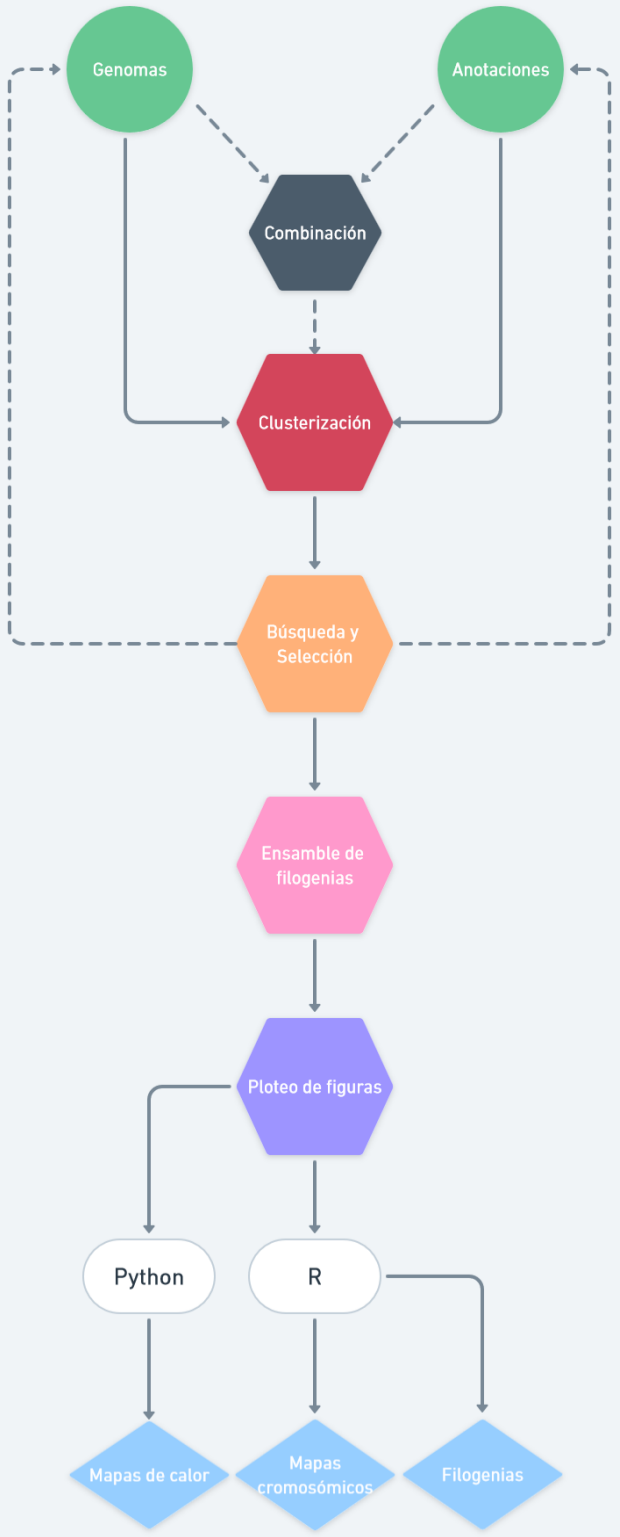


Figura 3.1 Flujo de trabajo de un pipeline automático para la clusterización, búsqueda y análisis de genes homólogos. El pipeline consiste en cuatro módulos principales y uno alternativo que agrupan scripts elaborados en Python, R y Bash. Usa como entradas uno o varios genomas en formato ‘.fasta’ (genome.fasta) con sus correspondientes anotaciones en formato ‘.gff3’ (genome.gff3); este devuelve como salida archivos de texto, tablas de datos y distintas visualizaciones. Las flechas punteadas indican el flujo de trabajo alternativo, mientras que las flechas sólidas indican el flujo de trabajo estándar. (0) Módulo alternativo ‘Combinación’, en caso de usarse permite agrupar a varios genomas y sus anotaciones correspondientes, permitiendo el uso del pipeline sobre conjuntos de genomas y anotaciones. (1) Módulo ‘Clusterización’, según lo descrito en los métodos agrupa a todos los genes de los genomas, en cluster enriquecidos en homólogos. (2) Módulo ‘Búsqueda y Selección’ usando programas nativos de Bash (Sed, tr, grep y AWK) genera tablas en formato tabular (clusters.tab) y otros archivos de texto (clusters.mci, clusters.txt), que contienen información de un subgrupo de clusters que coincide con distintas especificaciones dadas por el usuario. Opcionalmente los genes y anotaciones seleccionadas por este módulo pueden combinarse usando el módulo alternativo con otros genes y anotaciones de interés, según es indicado por las flechas punteadas que marcan el flujo de trabajo alternativo. (3) Módulo ‘Ensamblado de Filogenias’ alinea y ensambla, usando MAFFT e IQTREE a las secuencias codificantes traducidas de los miembros del subgrupo de clusters seleccionado en el módulo anterior. (4) Módulo ‘Visualización de Datos’, empleando las librerías de Numpy, Pyplot y Seaborn de Python, y, Dendextend y ChromPlot de R; junto con archivos creados previamente por el pipeline, se producen distintas de visualizaciones: gráficos de barras, histogramas, mapas de calor, filogenias, mapas cromosómicos.

Módulos

0) Combinación

Este módulo se usa en caso que se implemente el flujo de trabajo alternativo (Figura 3.1), para realizar análisis conjuntos sobre varios grupos de genomas con sus correspondientes anotaciones. Se pueden usar como entradas los archivos producidos por los módulos 1 y 2 del pipeline. Este módulo también agrega el atributo 'Org_name' a las anotaciones, facilitando distinguir a qué grupo pertenece cada gen. Las salidas de esta función pueden ser usadas como entrada del módulo 1, y seguir el flujo de trabajo estándar (Figura 3.1).

1) Clusterización

El primer módulo agrupa a todos los genes de los genomas usados, en clusters enriquecidos con genes homólogos. Como entradas este módulo usa a los genomas de interés en formato '.fasta' (genome.fasta), con sus respectivas anotaciones en formato '.gff3' (genome.gff3). A partir de estos, siguiendo los lineamientos descritos en los métodos se usan los programas gffread (paquete EMBOSS), BLASTP, mcxload y mcl; creándose distintos archivos de salida:

- genome.faa: Secuencias codificantes traducidas de los genomas usados como entrada
- genome.mci: Listado con los miembros de cada cluster
- E_value.abc: Tabla en formato '.abc' que interrelaciona entre sí mediante valores de E-value, a todos los genes presentes en genome.faa.

2) Búsqueda y selección

Este módulo genera tablas de texto en donde se caracteriza a un grupo de clusters de interés, con distintos atributos indicados por el usuario. Para lograr esto se genera una tabla de texto que asocia a cada uno de los miembros de cada cluster distintos atributos presentes en las anotaciones de formato '.gff3' (ID, product, Name, signature_desc, etc.), según lo especifique el usuario. Este archivo es posteriormente complementado con datos específicos de cada gen que incluyen largos proteicos -en aminoácidos- de sus regiones codificantes traducidas, contenido de GC y ubicación cromosómica. Posteriormente, según patrones de búsqueda introducidos por el usuario y otro tipos de restricciones que éste especifique (tamaño de cluster, porcentaje proteínas hipotéticas, etc), se realiza una selección de clusters de interés y se transforma el archivo texto en distintos formatos de tabla. Todo lo anterior es logrado a partir de la elaboración de varios scripts que incorporan programas nativos de la interfaz de Bash como Sed, tr, grep y AWK. Los archivos resultantes de la ejecución de este módulo son:

- clusters.txt: Archivo de texto resumido con las principales características de cada cluster seleccionado, ideado para ser visto por el usuario desde la terminal. Incluye contenido de GC promedio, largo en aminoácidos promedio, ubicaciones cromosómicas y atributos especificados por el usuario de los miembros de cada clúster.
- clusters.tab: Tabla de columnas separadas por tabulación -formato '.tsv'- que que indica para cada miembro de los clusters seleccionados: contenido de GC, largo en aminoácidos, ubicación cromosómica, cluster de pertenencia y atributos especificados por el usuario.

- clusters.mci: Listado de los miembros que integran los clusters seleccionados.

Alternativamente el conjunto de genes seleccionados por este módulo junto con sus correspondientes anotaciones pueden ser usados como entradas por el módulo de 'Combinación', permitiendo que puedan ser reintroducidos al pipeline con otros grupos de genes y anotaciones de interés.

3) Análisis filogenético
   En este módulo los genes de clusters seleccionados son alineados y ensamblados de forma automática usando los programas MAFFT e IQTREE (Katoh & Standley, 2013; Minh et al., 2020) según lo especificado en la sección de Métodos. Como entradas este módulo usa los archivos: clusters.mci y genome.faa. El resultado principal de este módulo es una filogenia elaborada con las secuencias peptídicas de los genes presentes en clusters.mci, la cual es guardada un archivo en formato Newick denominado clusters.treefile.

4) Visualización de datos
   En el último módulo, la información generada previamente es usada como entrada para producir distintas visualizaciones. Se emplean distintos scripts que incorporan librerías como Numpy, Pyplot y Seaborn de Python, y, Dendextend y ChromPlot de R. Como entradas este módulo emplea los archivos clusters.tab y clusters.trefile. Las posibles salidas resultantes de este módulo constituyen: histogramas, gráficos de barras, mapas de calor, dendrogramas y mapas cromosómicos.

Clusterización de los genomas de *T. gondii* y *N. caninum*

La ejecución del Módulo 1 del pipeline desarrollado, sobre los datos genómicos de *T. gondii* y *N. caninum* produjo una gran cantidad de clusters. A partir de 8322 genes de *T. gondii* se produjeron a un total 2353 clusters, de los cuales 5.9% (140) se compone de más de 10 genes, 10.6% (249) tiene entre 10 y 6 genes, 32.6% (767) tienen entre 5 y 2 genes, y el 50% (1197) tienen un solo gen (Figura 3.3A). Los clusters con màs de 10 genes concentran el 35.4% (2945) del total de los genes, los de entre 10 y 6 genes corresponden al 22% (1833), los de entre 5 y 2 tienen 28.2% (2347), mientras que los clusters solitarios representan sólo 14.4% (1197) de los genes (Figura 3.2B). Por otro lado, a partir de 7364 genes de *N. caninum* se produjeron 1943 clusters, de los cuales el 6.3% (123) presentan más de 10 genes, el 9.8% (190) tiene entre 10 y 6 genes, el 33.9% (659) tiene entre 5 y 2 genes, mientras que el 50% (971) tienen un solo gen (Figura 3.4A). Análogamente para *N. caninum* los clusters mayores 10 genes concentran el 40.3% (2965) de todos los genes, los de entre 10 y 6 genes concentran el 19.3% (1424), los de entre 5 y 2 genes el 27.2% (2004), y los clusters de un solo gen tienen el 13.2% (971) de los genes (Figura 3.4B). Se puede observar que la clusterización inicial es similar en ambos genomas, con sutiles diferencias en los porcentajes pero manteniendo un patrón similar en lo sustancial.

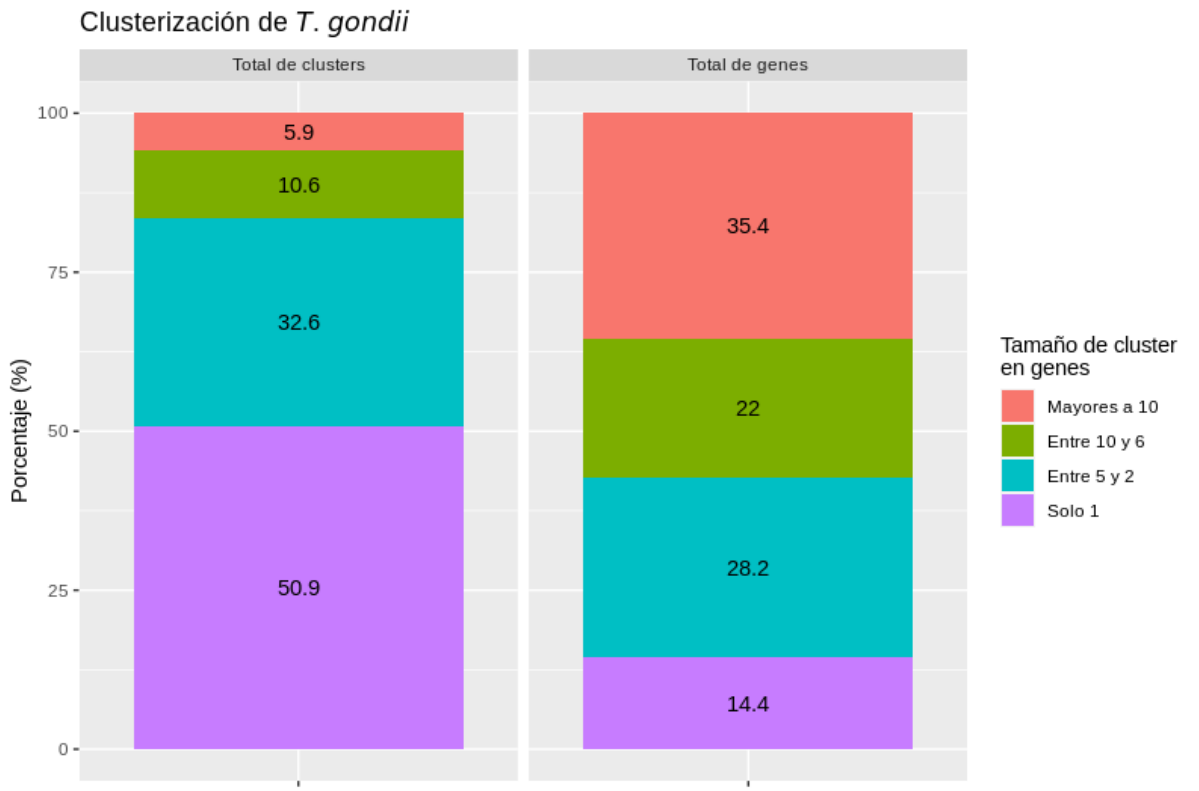


Figura 3.2. Distribución de genes y tamaños de cluster para *T. gondii*. El gráfico de barras A indica sobre el total de cluster obtenidos, qué porcentaje de estos son mayores a 10 genes (5.9%), presentan entre 10 y 6 genes (10.6%), presentan entre 5 y 2 genes (32.6%) y clusters de un solo gen (50.9%). (B) Gráfico de barras indicando sobre el total de genes cuantos pertenecen a clusters mayores a 10 genes (35.4%), a clusters de entre 10 y 5 genes (22%), a clusters de entre 5 y 2 genes (28.2%) y clusters de un solo gen (14.4%).

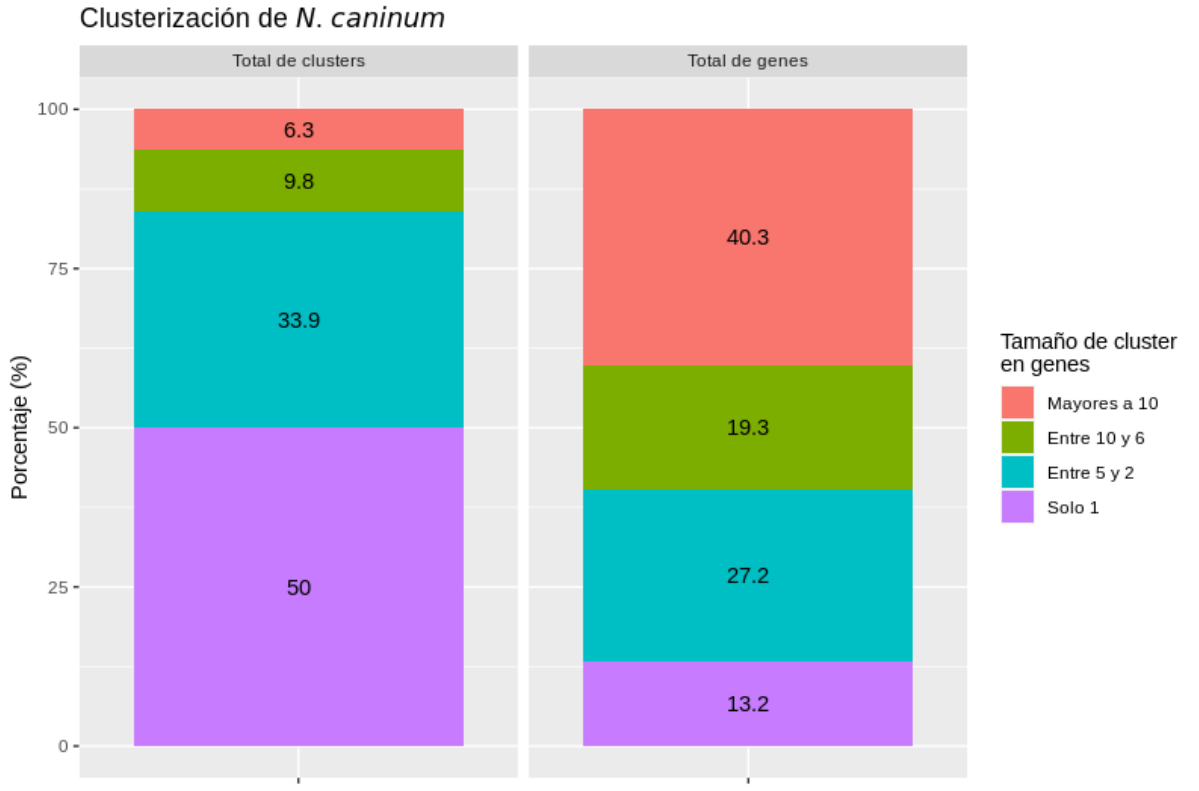


Figura 3.3. Distribución de genes y tamaños de cluster para *N. caninum*. El gráfico de barras A indica sobre el total de cluster obtenidos, qué porcentaje de estos son mayores a 10 genes (6.3%), presentan entre 10 y 6 genes (9.8%), presentan entre 5 y 2 genes (33.9%) y presentan un solo gen (50%). (B) Gráfico de barras indicando sobre el total de genes cuantos pertenecen a clusters mayores a 10 genes (40.3%), a clusters de entre 10 y 5 genes (19.3%), a clusters de entre 5 y 2 genes (27.2%) y clusters de un solo gen (13.2%).

Búsqueda de familias génicas en *T. gondii* y en *N. caninum*

Con el fin de identificar clusters de genes relacionados a las familias multigénicas asociadas a proteínas de superficie y de secreción de *T. gondii* y *N. canium*, a partir de los resultados previos, se ejecutaron en varias instancias el módulo 2 del pipeline descrito. Específicamente a partir de las anotaciones genómicas de *T. gondii* y *N. caninum* se buscaron clusters asociados a las familias SRS, ROP, MIC y GRA. Para esto se utilizaron de forma respectiva como argumentos de búsqueda las siguientes expresiones regulares: 'SRS|srs|SAG', 'ROP|rhoptria', 'MIC|microneme' y 'GRA|granul|Granul'.

Búsqueda de familias génicas en *T. gondii*

Para *T. gondii* se encontraron varios clusters asociados a las familias multigénicas SRS, ROP, MIC y GRA (Tabla 3.1). Específicamente la búsqueda de clusters asociados a la familia multigénica SRS arrojó un total de 18 clusters, conteniendo 130 genes en total. De estos 130 genes, 111 corresponden a SRS, 15 a proteínas hipotéticas y 4 están anotados con otras funciones. Además, el contenido de GC promedio para los genes de estos clusters es de 52.8% y el largo promedio proteico es de 377 aminoácidos.

La búsqueda de miembros de la familia multigénica ROP devolvió 26 clusters compuestos por 270 genes en total. De estos 270 genes, 138 están anotados como genes hipotéticos, 52 como ROPs, 16 como quinasas varias, 9 como RONs y 55 con otras funciones. De los 52 genes anotados como ROPs, 10 son putativos o tienen secuencias incompletas. El contenido de GC promedio para los genes de estos clusters es de 55.9% y el largo proteico promedio es de 982 aminoácidos.

La búsqueda de miembros de la familia multigénica MIC arrojó un total de 12 clusters, con un total de 138 genes. De estos, 26 están anotados como proteínas MIC, 6 de los cuales están anotados como putativos. También se encuentran 61 genes hipotéticos, 10 con dominio EGF, 8 con dominio PAN y 33 anotados con otras funciones. El contenido de GC promedio para los genes de estos clusters es de 56.8% y el largo proteico promedio es de 769 aminoácidos.

La búsqueda de miembros de la familia multigénica GRA arrojó un total de 16 clusters, con un total de 185 genes. De estos, solamente 19 están anotados como proteínas GRA. En el resto se encuentran 90 genes hipotéticos, 11 fosfatasas, 10 con dominios dedos de Zinc, 9 con dominio Rap y 46 anotados con otras funciones. El contenido de GC promedio para los genes de estos clusters es de 56.8% y el largo proteico promedio es de 918 aminoácidos.

Tabla 3.1. Búsquedas de clusters asociados a las familias multigénicas SRS, ROP, MIC y GRA en de *T. gondii*. Se describen para cada familia el número de clusters, total de genes, número de genes anotados previamente a la familia buscada, largo proteico promedio (AA) y porcentaje promedio de GC

	SRS	*ROP*	*MIC*	*GRA*
N° de clusters	18	26	12	16
Total de genes	130	270	138	185
N° de genes anotados	111	52	26	19
Largo proteico promedio	377	982	1017	918
%GC promedio	52.8	55.9	56.2	56.8

Búsqueda de familias génicas en *N. caninum*

Para *N. caninum* se encontraron varios clusters asociados a las familias multigénicas SRS, MIC y GRA; y ninguno asociado a la familia ROP. La búsqueda de clusters asociados a la familia multigénica ROP no arrojó resultados (Tabla 3.2). Específicamente, se detectaron 13 clusters SRS, conteniendo un total de 241 genes. De estos 241 genes, 217 corresponden a SRS, 11 a proteínas hipotéticas y 13 están anotados con otras funciones. De las 217 SRS, 23 están anotadas como miembros putativos. Además, el contenido de GC promedio para genes de estos clusters es de 53.39% y el largo proteico promedio es de 463 aminoácidos.

La búsqueda de miembros de la familia multigénica MIC arrojó un total de 4 clusters, con un total de 28 genes. De estos, solo 5 están anotados como proteínas MIC putativas, mientras que los 23 genes restantes, 12 están anotados con otras funciones putativas y 11 son hipotéticos. El contenido de GC promedio para los genes de estos clusters es de 54.90% y el largo proteico promedio es de 769 aminoácidos.

Por último se encontró sólo 1 cluster asociado la familia multigénica GRA, el cual presenta 2 miembros, uno corresponde a una proteína hipotética y otro un miembro putativo de esta familia. Su contenido promedio de GC es de 56.00% y su largo proteico promedio es de 234 aminoácidos.

Tabla 3.2. Resultados de las búsquedas de clusters asociados a las familias multigénicas SRS, ROP, MIC y GRA en *N. caninum*. Se describen para cada familia el número de clusters, total de genes, número de genes anotados previamente a la familia buscada, largo proteico promedio (AA) y porcentaje promedio de GC

	SRS	*ROP*	*MIC*	*GRA*
N° de clusters	13	0	4	1
Total de genes	241	0	28	2
N° de genes anotados	217	0	5	1
Largo proteico promedio	463	-	769	234
%GC promedio	53.4	-	54.9	56.0

Clusterización conjunta de los genomas de *T. gondii* y *N. caninum*

La aplicación del Módulo alternativo 0 y del Módulo 1, usando como entrada los datos genómicos en conjunto de *T. gondii* y *N. caninum* produjo una gran cantidad de clusters de genes. Específicamente entre ambos organismos se produjeron un total de 6314 clusters a partir de 15686 genes, de los cuales el 3.5% son clusters mayores a 4 genes, el 20.9% tienen 3 o 4 genes, el 73.9% tienen 2 genes, mientras que solo el 1.7% tiene un gen (Figura 3.4A). Además para estos parásitos se observa que el 12.7% del total de los genes se encuentran en los clusters mayores a 4 genes, el 27.1% de los genes están en clusters de entre 3 y 4 genes, el 59.5% en clusters de 2 genes y solo el 0.7% están en clusters de un solo gen (Figura 3.4B). Del total de clusters se observa que el 97.7% (6168) tienen genes pertenecientes a *T. gondii* y *N. caninum*, quedando solamente 2.3% (147) de los clusters con genes exclusivamente de *T. gondii* (1.8%) o *N. caninum (0.5%)*. Estos clusters especie-especìficos contienen un total de 229 genes de los cuales 112 son de *T. gondii* y 51 de *N. caninum*

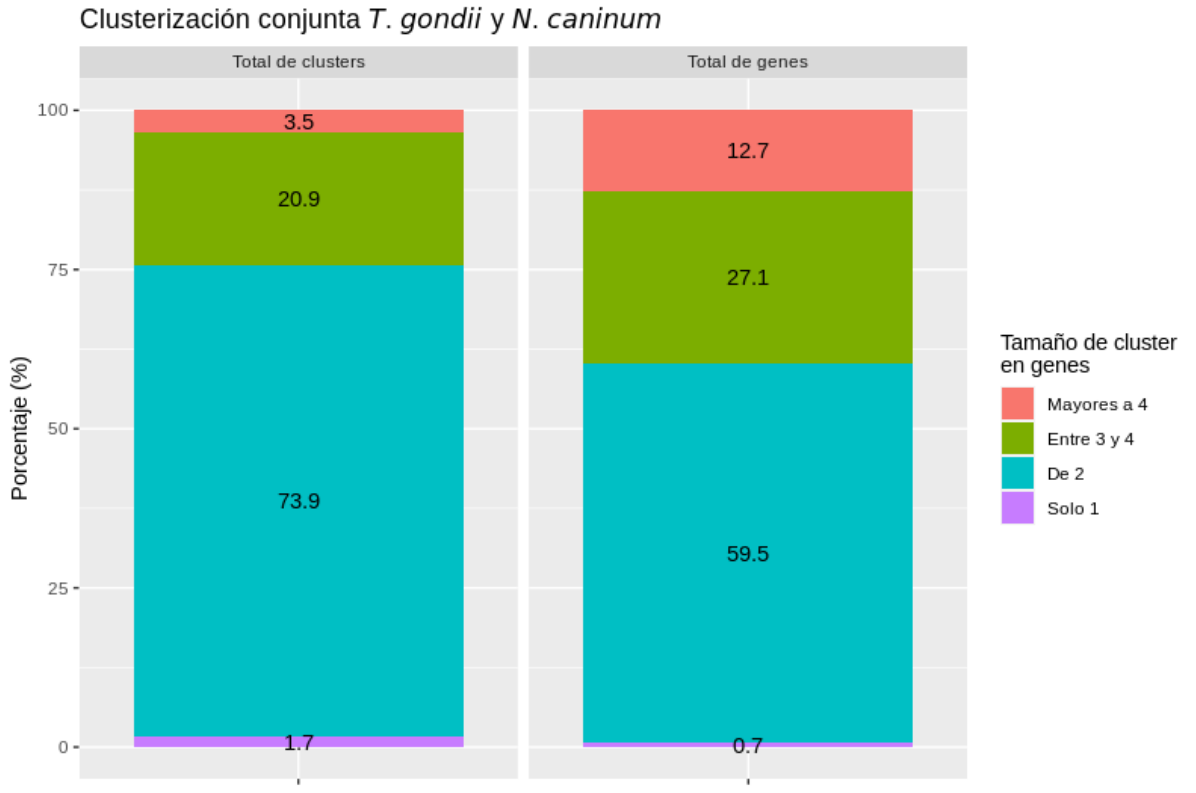


Figura 3.4. Distribución de genes y tamaños de cluster en conjunto para *T. gondii* y *N. caninum*. En el gráfico de barras A indica sobre el total de cluster obtenidos, qué porcentaje de estos son mayores a 4 genes (3.5%), presentan entre 3 y 4 genes (20.9%), presentan 2 genes (73.9%) y clusters de un solo gen (1.7%). (B) Gráfico de barras indicando sobre el total de genes cuantos pertenecen a clusters mayores a 4 genes (12.7%), a clusters de entre 3 y 4 genes (27.1%), a clusters de entre de 2 genes (59.5%) y clusters de un solo gen (0.7%).

Búsqueda de familias génicas conjuntas en *T. gondii* y *N. caninum*

Las búsquedas realizadas con las expresiones regulares mencionadas, encontraron varios clusters asociados a las multigénicas SRS, MIC, GRA y ROP. La mayoría de estos clusters presentan genes de ambos parásitos (Tabla 3.3). Cabe resaltar que para las familias SRS, ROP y GRA se encontraron un total de 3 clusters conformados únicamente por genes de *T. gondii*; no se encontraron clusters conformados únicamente por genes de *N. caninum*.

La búsqueda de clusters asociados a la familia multigénica SRS arrojó un total de 38 clusters, 37 conformados por genes de *T. gondii* y *N. caninum* y 1 conformado únicamente por genes de *T. gondii*. En total los clusters de SRS encontrados tienen 345 genes (122 pertenecientes a *T. gondii* y 223 a *N. caninum*), de estos, 329 se encuentran como miembros de esta familia (111 en *T. gondii* y 218 en *N. caninum*). Además se encontraron en conjunto 13 proteínas hipotéticas y 3 genes anotados con otras funciones. El largo promedio proteico es de 406 aminoácidos y el contenido de GC 52.9%, siendo ambas medidas muy similares en ambos parásitos (Tabla 3.5).

La búsqueda de clusters asociados a la familia multigénica ROP arrojó un total de 46 clusters, 45 conformados por genes de *T. gondii* y *N. caninum, y* uno conformado únicamente por genes de *T. gondii*. Los cluster encontrados presentan un total de 148 genes, 83 pertenecientes a *T. gondii* y 65 a *N. caninum*. De los 148 genes encontrados, 52 presentan anotaciones asociadas a esta familia estando 18 anotadas de forma putativa o son secuencias incompletas. Se destaca que 50 de los 51 genes anotados como integrantes de la familia ROP pertenecen a *T. gondii*, 5 clasificadas como putativas y 12 incompletas. En *N. caninum* se destaca la presencia de 54 genes sin clasificar distribuidos entre todos los clusters. También se encontraron 22 proteínas hipotéticas, 11 RONs y 9 proteínas con función variada. Los miembros de estos clusters presentan en promedio un largo proteico de 679 aminoácidos -620 en *T. gondii* y 752 en *N. caninum* -, y un contenido promedio de GC de 54.5%.

La búsqueda de miembros de la familia multigénica MIC arrojó un total de 20 clusters, con miembros de *T. gondii* y *N. caninum*. Los clusters encontrados presentan 82 genes - 43 pertenecientes a *T. gondii* y 39 a *N. caninum* -. De estos, solo 32 están anotados como proteínas MIC - 26 en *T. gondii* y 6 en *N. caninum* -. Entre las anotaciones restantes se encuentran 10 proteínas hipotéticas, 16 genes con dominios PAN y 7 genes con otras funciones. También se encuentran 17 genes sin clasificar pertenecientes a *N. caninum*. El contenido de GC promedio para los genes de estos clusters es de 54.9% y el largo promedio es de 769 aminoácidos.

Por último en la búsqueda de clusters asociados a la familia GRA, se encontró un total de 19 clusters; 18 conformados por genes de *T. gondii* y *N. caninum*, y uno conformado únicamente por genes de *T. gondii*. Los clusters GRA se componen por 45 genes - 26 en *T. gondii* y 19 en *N. caninum* -. De los genes encontrados 19 tienen anotaciones asociadas a la familia GRA, 7 son proteínas hipotéticas, 11 sin clasificar y 8 presentan funciones varias. Se destaca que la gran mayoría de los genes anotados (17) en la familia GRA son pertenecientes a *T. gondii*, que todos los genes sin clasificar forman parte de *N. caninum*. El

contenido de GC promedio para los genes de estos clusters es de 56.7% y el largo promedio es de 667 aminoácidos.

Tabla 3.5. Clusters asociados a las familias multigénicas SRS, ROP, MIC y GRA para los genomas de *T. gondii* y *N. caninum*. Se describen para cada familia el número de clusters, total de genes, número de genes anotados previamente a la familia buscada, largo proteico promedio (AA) y porcentaje promedio de GC.

	SRS			***ROP***		
	Conjunto	*T. gondii*	*N. caninum*	Conjunto	*T. gondii*	*N. caninum*
Nº de clusters	38	38	37	46	46	45
Total de genes	345	122	223	148	83	65
Nº de genes anotados	329	111	218	52	51	1
Largo proteico promedio	406	408	405	679	620	752
%GC promedio	52.9	52.6	53.0	52.9	52.6	53.0
	MIC			***GRA***		
	Conjunto	*T. gondii*	*N. caninum*	Conjunto	*T. gondii*	*N. caninum*
Nº de clusters	38	38	37	46	46	45
Total de genes	345	122	223	148	83	65
Nº de genes anotados	329	111	218	51	51	1
Largo proteico promedio	406	408	405	679	620	752
%GC promedio	52.9	52.6	53.0	52.9	52.6	53.0

Análisis de la familia multigénica SRS en *T. gondii* y *N. caninum*

Para profundizar en el estudio de la familia multigénica SRS se ejecutaron los módulos 3 y 4 del pipeline descrito. Como entrada se usaron los resultados de búsqueda de clusters asociados a la familia multigénica SRS, obtenidos por el módulo 2 a partir de la clusterización conjunta de *T. gondii* y *N. caninum*. A partir de esto se obtuvieron distintas visualizaciones entre las que se incluyen un árbol árbol filogenético, mapas cromosómicos y mapas de calor para los distintos clusters asociados a la familia multigénica SRS.

Visualización de clusters de la familia multigénica SRS en *T. gondii* y *N. caninum*

Para cada cluster se obtuvo un mapa de calor, el cual indica las relaciones de similitud existentes entre los distintos genes integrantes del cluster (Figura 3.5, véase https://github.com/JPereira199/GenomicClusterExplorer/blob/main/Supplementary_Data/HeatMaps). Se evidencian relaciones de similitud de variada complejidad entre los integrantes de los distintos clusters. Para facilitar su comprensión los genes de cada cluster son ordenados de forma jerárquica, usando el método del promedio en una métrica euclídea. Los valores de cada celda de los mapas de calor corresponden a los valores de similitud existentes entre las secuencias integrantes del clúster, obtenidos al realizar un BLASTP

entre los productos proteicos de los integrantes de cada clúster. Al incrementarse el valor de cada celda se incrementa la probabilidad de que los genes que esta relaciona presenten una relación cercana de homología.

Como consecuencia de este ordenamiento emergen distintos agrupamientos de genes que pueden ser clasificados de diferente manera. En la mayoría de los clusters se observa un grupo central de genes, en el cual todos sus miembros presentan elevados valores de similitud (>50) entre sí (Figuras 3.5A). Este grupo suele estar integrado por todos o la mayoría de los miembros del clúster. Con menos frecuencia se puede distinguir uno o varios grupos secundarios, donde sus integrantes presentan elevados valores de similitud entre sí, pero se componen por una cantidad minoritaria de los miembros del clúster. Además en ocasiones se presentan genes periféricos al cluster, caracterizados por tener índices de similitud moderado o bajo (<30) con todos los demás miembros del clúster (Figuras 3.5B). De todos modos, hay varios clusters en donde la relaciones de similitud entre sus integrantes son más complejas no pudiendo ser descritas con las clasificaciones empleadas Figuras 3.5C).

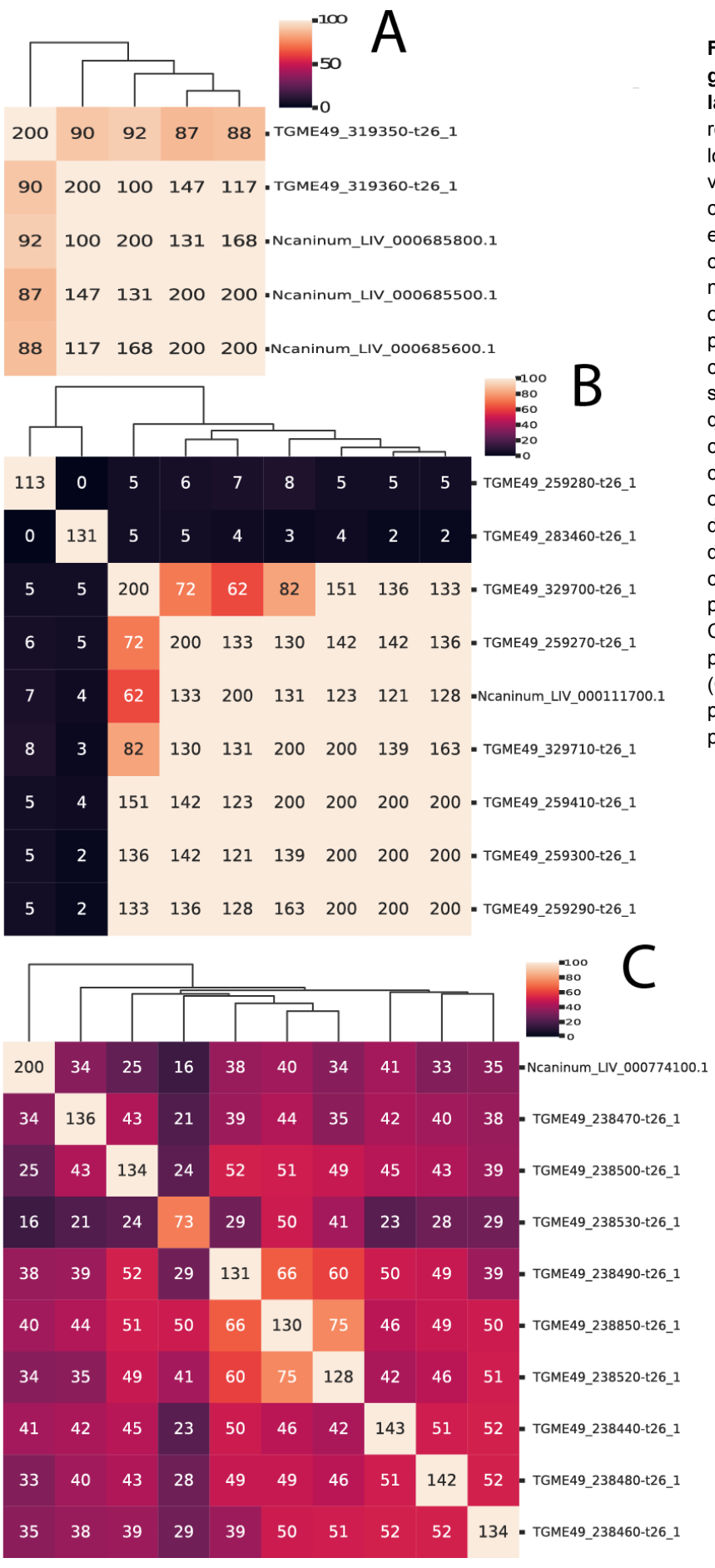


Figura 3.5. Mapas de calor de clusters de genes de *T. gondii* y *N. caninum* asociados a la familia multigénica SRS. Se observan las relaciones valores de similitud que existen entre los diferentes miembros de cada cluster. Los valores de cada celda de los mapas de calor corresponden a los valores de similitud existentes entre las secuencias integrantes del cluster, estos valores corresponden logaritmo negativo de los 'Expect values' (E-values) obtenidos al realizar un BLASTP entre los productos proteicos de los integrantes de cada cluster. Al incrementarse el valor de cada celda se incrementa la probabilidad de que los genes que esta relaciona presenten una relación cercana de homología. Basado en los valores de cada celda los integrantes de cada cluster fueron ordenados de forma jerárquica en el dendograma que se observa en la parte superior de cada mapa de calor. (A) Cluster 208, conformado por 5 secuencias proteicas, 2 pertenecientes a *T. gondii* y 3 a *N. caninum*. (B) Cluster 43 conformado por 9 genes, 8 pertenecientes a *T. gondii* y uno a *N. caninum*. (C) Cluster38 conformado por 10 secuencias proteínas, 9 pertenecientes a *T. gondii* y una perteneciente a *N. caninum*.

Distribución cromosómica de la familia Multigénica SRS en *T. gondii* y *N. caninum*

Al analizar la distribución de los genes SRS en los respectivos genomas, se observa que están presentes en todos los cromosomas de ambos parásitos (Figuras 3.6 y 3.7). La mayoría de estos genes se encuentran agrupados en tandems (~82%), siendo esto más prevalente en *N. caninum* (~86%) respecto a *T. gondi* (~75%). Ambos parásitos presentan cada uno un total de 31 genes no agrupados en ningún tándem. También presentan un número similar de tándems, teniendo *T. gondii* 26 y *N. caninum* 28. La gran mayoría de los tandems se componen por genes de un mismo cluster, a excepción de 6 tandems, de los cuales 2 pertenecen a *T. gondii* y 4 a *N. caninum*. El número de genes por tándem se distribuye de forma diferente en ambos parásitos, presentando *T. gondii* tandems de entre 2 y 6 genes, mientras que en *N. caninum* estos tienen entre 2 y 19 genes (Tabla 3.6 y 3.7).

A pesar de que todos los cromosomas de estos parásitos presentan genes asociados a la familia multigénica SRS, existen varios cromosomas que no presentan tandems (Tabla 3.6 y 3.7). Destacando el cromosoma V de *N. caninum* con 65 genes y 9 tandems (Tabla 3.6). El número de genes SRS en proporción al largo del cromosoma que pertenecen es variable, siendo el cromosoma IV de *T. gondii* y el V de *N. caninum* son los que presentan una mayor proporción de genes (Tabla 3.6 y 3.7).

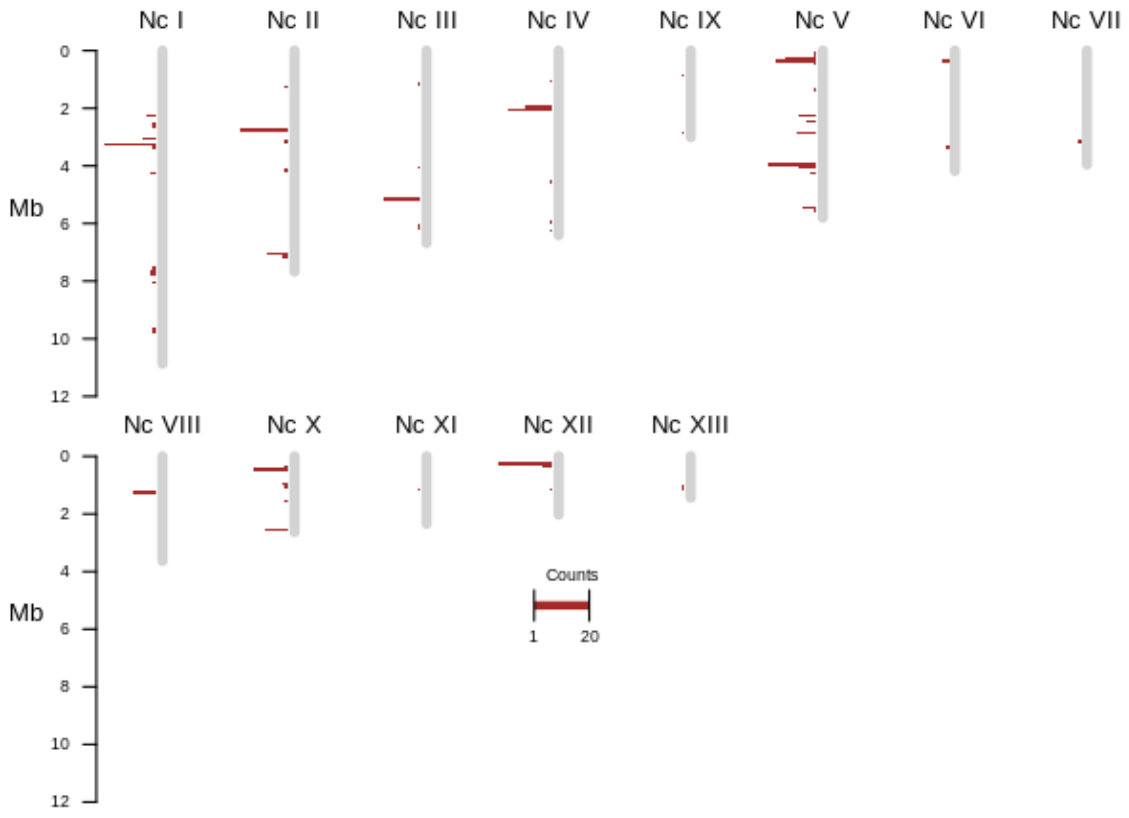


Figura 3.6. Mapa cromosómico de la familia multigénica SRS en *N. caninum*.
Para todos los cromosomas de *N. caninum*, el número de genes dentro de una región de 100Kb es observado en forma de histograma.

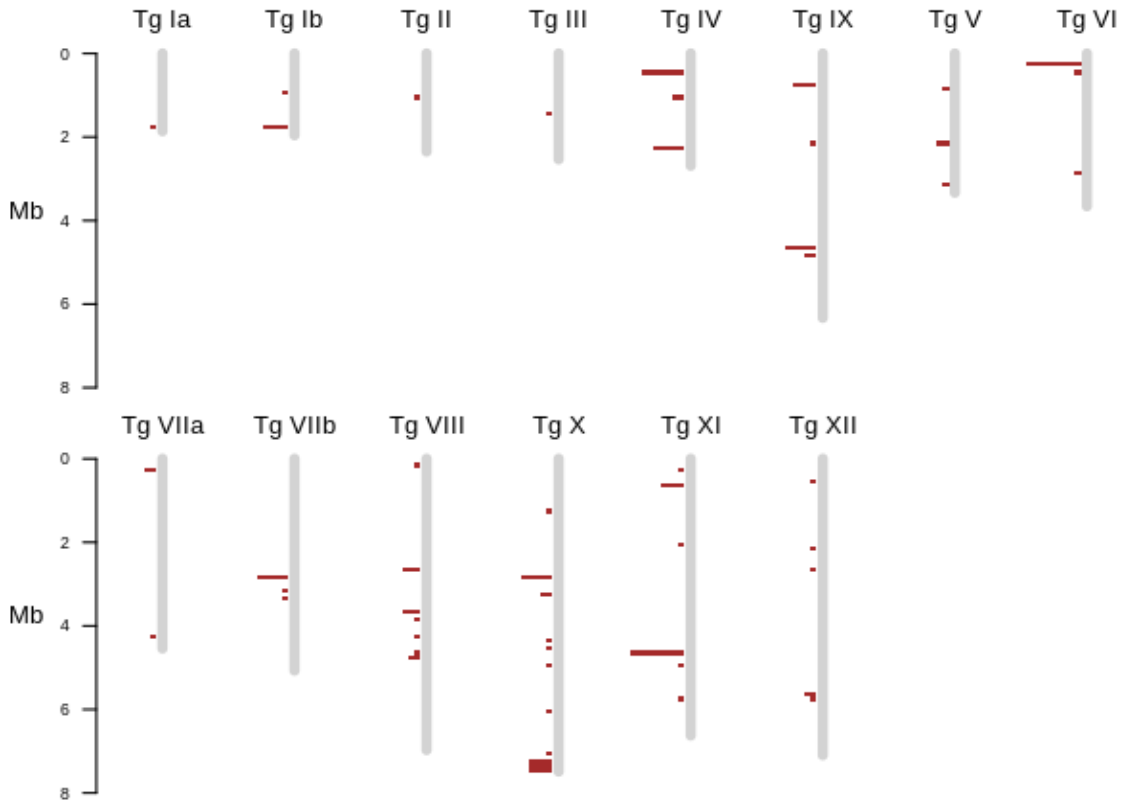


Figura 3.7. Mapa cromosómico de la familia multigénica SRS en *T. gondii*.
Para todos los cromosomas de *T. gondii*, el número de genes dentro de una región de 100Kb es observado en forma de histograma.

Tabla 3.7. Distribución de la familia multigénica SRS en los cromosomas de *T. gondii* y *N. caninum.*
Se describe para cada cromosoma su largo en megabases (10^6 bases), la cantidad de tandems y genes de la familia SRS, así como la cantidad de genes en proporción al largo cromosómico. En *T. gondii* se encontraron 3 genes y un tándem ubicados en los contigs KE139435 y KE139801 que no pertenecen a ningún cromosoma

T. gondii					*N. caninum*				
Cromosoma	**Largo (Mb)**	**Tándems**	**Genes**	**Genes/Mb**	**Cromosoma**	**Largo (Mb)**	**Tándems**	**Genes**	**Genes/Mb**
Tg Ia	1.9	0	1	0.53	Nc I	10.9	6	35	3.21
Tg Ib	2.0	1	5	2.50	Nc II	7.7	3	25	3.25
Tg II	2.3	0	1	0.43	Nc III	6.7	1	10	1.49
Tg III	2.5	0	1	0.40	Nc IV	6.4	2	25	3.91
Tg IV	2.7	4	14	5.19	Nc V	5.8	9	65	11.21
Tg V	3.3	1	4	1.21	Nc VI	4.2	1	3	0.71
Tg VI	3.7	1	11	2.97	Nc VII	4.0	0	1	0.25
Tg VIIa	4.5	2	3	0.67	Nc VIII	3.6	1	7	1.94
Tg VIIb	5.1	1	7	1.37	Nc IX	3.0	0	2	0.67
Tg VIII	7.0	3	12	1.71	Nc X	2.6	4	27	10.38
Tg IX	6.3	3	12	1.90	Nc XI	2.3	0	1	0.43
Tg X	7.5	5	25	3.33	Nc XII	2.0	1	20	10.00
Tg XI	6.6	3	17	2.58	Nc XIII	1.4	0	2	1.43
Tg XII	7.1	1	6	0.85					

Análisis filogenético de proteínas SRS en *T. gondii* y *N. caninum*

Se realizó un análisis filogenético utilizando los miembros de los clusters identificados, el cual generó un árbol filogenético sin raíz compuesto por 345 secuencias peptídicas (véase Figura Suplementaria S1). Este análisis reveló una complejidad significativa en la familia estudiada. Las proteínas que pertenecen al mismo cluster se agrupan comúnmente en la misma rama de la filogenia, y esto ocurre con mayor frecuencia para aquellas proteínas que provienen de genes en tándem. Sin embargo, no se observan agrupamientos basados en la longitud de la secuencia de aminoácidos o el contenido de GC de los genes que codifican estas proteínas.

La rama conformada desde la secuencia peptídica 269 hasta la 284 (Figura 3.8) es una excepción a lo anterior. Esta se compone por varios clusters y la mayoría de sus genes no forman parte de ningún tándem. Contiene proteínas de gran tamaño y los genes que las codifican presentan un elevado contenido de GC. Esta rama también presenta a 6 de los 13 proteínas hipotéticas presentes en esta filogenia y 2 de las 3 secuencias anotadas con una función distinta a la de la familia SRS.

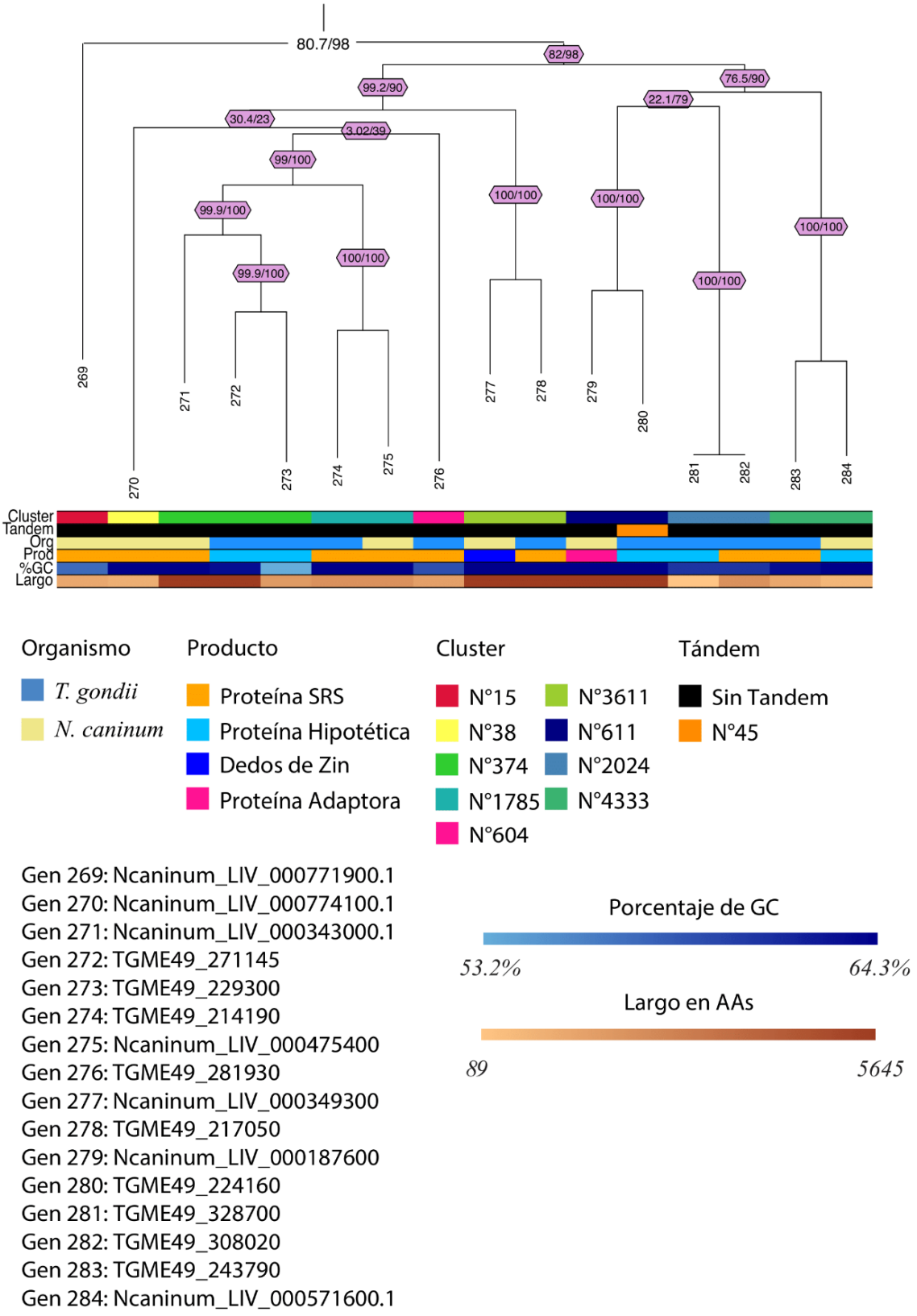


Figura 3.8. Rama de alta complejidad de un árbol filogenético de proteínas SRS en *T. gondii* y *N. caninum*. Filogenia parcial de la filogenia obtenida a partir de todos los integrantes de la familia SRS. Se utilizó el método de máxima verosimilitud. En la parte superior de cada nodo se encuentra su apoyo estadístico obtenido por 5000 iteraciones de UF-bootstrap y sh-aLRT. En la parte inferior de cada rama se encuentra una barra de colores que indica para cada integrante cluster, organismo (Org) y tándem al

que pertenecen; así como su, producto proteico (Prod), porcentaje de GC (%GC) y largo proteico en aminoácidos (Largo). Cada integrante de la filogenia está numerado del 269 al 281, siguiendo el orden de las filas de la tabla usada para generar la figura (véase https://github.com/JPereira199/GenomicClusterExplorer/blob/main/Supplementary_Data/TableS1.tab).

.

Discusión

Desarrollo del pipeline

Se desarrolló un pipeline automático para identificar clusters de familias multigénicas basándose en la anotación de genomas. El pipeline se compone de una serie de programas divididos en 4 módulos, 1) Clusterización, 2) Búsqueda y selección, 3) Ensamblado de filogenias, 4) Visualización de datos. El pipeline es muy flexible, ya que se puede ejecutar de manera total o parcial, empleando múltiples genomas con sus anotaciones respectivas según lo especifique el usuario. Cada módulo cuenta con diferentes opciones de ejecución que brindan mayor versatilidad. La flexibilidad del pipeline es potenciada por su estructura modular (Baldwin & Clark, 1999), facilitando la incorporación de futuras funciones mediante la sustitución o adición de nuevos módulos. Esto permite la ejecución parcial del pipeline y facilita la futura incorporación de funciones mediante sustitución o adición de nuevos módulos.

Durante la elaboración del Módulo 2 se observó que las diferencias internas que pueden llegar a tener los archivos de anotaciones de formato '.gff3', son una de las principales fuentes de errores. Otra desventaja del pipeline respecto a su fiabilidad responde al hecho de que incorpora múltiples programas elaborados por diferentes autores, lo que dificulta la corrección de errores y el mantenimiento del código. Para evitar futuros errores y mejorar la fiabilidad de todo el pipeline va a ser necesario implementar estrategias de garantía de calidad de software (software QA), las cuales son comunes en los procesos de desarrollo informático (Despa, 2014). Además, para continuar avanzando en la validación del pipeline, también va a ser de utilidad emplear como entrada otros conjuntos de datos usados en investigaciones previas, para así poder comparar sus resultados con los que obtenga el pipeline.

En cuanto a la instalación y ejecución del pipeline, los usuarios podrían encontrar dificultades debido a la variedad de programas y librerías empleadas, y a los múltiples prerrequisitos que se necesitan. Esto podría limitar el número de usuarios que puedan utilizar el software. Una solución a este problema podría ser reemplazar el software desarrollado por terceros con software desarrollado internamente. Sin embargo, esto podría ralentizar el desarrollo y sólo aplicarse a ciertos componentes del pipeline, ya que muchos programas provienen de equipos especializados y multidisciplinarios. Una opción más práctica sería emplear contenedores, los cuales son similares a las máquinas virtuales, pero solo incluyen las librerías y programas necesarios para ejecutar la aplicación (Paraiso et al., 2016). De ser necesario estos pueden ejecutarse mediante servicios en la nube como es el caso Amazon Web Services, Google Cloud Platform y Microsoft Azure (Paraiso et al., 2016).

De igual forma, el pipeline propuesto tiene el potencial de ser una herramienta valiosa para futuros estudios de genómica comparativa. Ya que este sistema automatiza el procesamiento de datos genómicos y visualizaciones, simplificando el proceso y reduciendo la posibilidad de errores humanos. Además, al estar disponible en repositorios de acceso público, no solo permite a otros investigadores utilizar y adaptar el pipeline desarrollado para sus propios estudios, sino que además facilita la reproducibilidad de los estudios que lo utilizan.

Clusterización y análisis de los genomas de *T. gondii* y *N. caninum*

Al ejecutar el módulo 1 en los datos genómicos de *T. gondii* y *N. caninum*, se llevó a cabo un análisis comparativo preliminar entre estos dos parásitos, el cual sugiere similitudes a nivel genómico. La clusterización de los genomas de ambos parásitos en instancias separadas generó una cantidad similar de clusters, con un incremento aproximado del 20% en *T. gondii* en comparación con *N. caninum*. Este aumento puede atribuirse en parte a que *T. gondii* presenta aproximadamente un 13% más de genes anotados respecto a *N. caninum*. También se observa que la distribución de tamaños de los clusters (Figura 3.3A y 3.4A) así como la cantidad total de genes en estos (Figura 3.3B y 3.4B) son similares entre ambos parásitos. Lo cual sugiere que ambos parásitos presentan mayormente familias multigénicas de similar tamaño.

La distribución de genes en los clusters obtenidos por la clusterización conjunta de ambos parásitos refleja la estrecha relación evolutiva entre *T. gondii* y *N. caninum*, lo cual es consistente con los resultados reportados por investigaciones previas (Monteiro et al., 2007; Reid et al., 2012). Esto se desprende del hecho de que casi la totalidad de los cluster observados (~98%) presentan genes de ambos organismos, sugiriendo que la gran mayoría de los genes presentes en *T. gondii* y *N. caninum* son homólogos entre sí.

La identificación de clusters que contienen exclusivamente genes de uno de los parásitos ofrece una oportunidad para investigar las diferencias fenotípicas entre *T. gondii* y *N. caninum*. Específicamente, sólo el 1.8% de los genes de *T. gondii* y el 0.5% de los genes de *N. caninum* forman parte de este tipo de clusters, totalizando 229 genes, lo que facilitará su análisis en futuros estudios. Es probable que los genes en estos clusters no compartan ortólogos cercanos entre los parásitos, sugiriendo que podrían desempeñar funciones biológicas específicas en cada especie y contribuir a sus diferencias fenotípicas. Dentro de las especificidades son de especial interés los distintos modos de transmisión (vertical y horizontal) y las variaciones en el rango de hospedadores de cada parásito (Barber & Trees, 1998; Dubey, 2007; Dubey & Schares, 2011). Sin embargo, antes de profundizar en el estudio de este grupo de genes, es necesario corroborar que tanto los genomas como las anotaciones genómicas de ambos parásitos sean de buena calidad, y estén completas, de manera de asegurarse que la ausencia de genes de un organismo en un cluster no se deba a este tipo de problemas.

Análisis de las búsquedas de las familias multigénicas en *T. gondii* y *N. caninum*

Comparación de las estrategías de clusterización

La búsquedas efectuadas sobre clusterización conjunta de los genomas fueron más completas ya que se encontraron múltiples clusters para todas las familias multigénicas buscadas (Tabla 3.7). A diferencia de las búsquedas realizadas sobre las clusterizaciones independientes, en donde casi no se obtuvieron resultados para las familias ROP y GRA de *N. caninum* (Tabla 3.6).

Además se observa para ambos parásitos, que las búsquedas de las distintas familias multigénicas realizada sobre las clusterizaciones en instancias separadas, presentaron una mayor proporción de genes sin anotar o con anotaciones no relacionadas a las familias multigénicas buscadas. Siendo más prominente esta diferencia en las familias multigénicas GRA y ROP de *T. gondii* (Tabla 3.5 y 3.7). Esto se debe a la incorporación de genes más distantes (más divergentes) a las respectivas familias, mostrando que el proceso de clusterización fue muy permisivo en este caso, incorporando en los clusters genes que no presentan suficiente homología para ser considerados de la misma familia. Por lo tanto, las búsquedas de las familias multigénicas realizadas sobre la clusterización conjunta también arrojó resultados más acertados.

La permisividad del proceso de clusterización es regulada por el valor del parámetro de inflación. Sin embargo, a pesar que se utilizó siempre el mismo valor de inflación, como se mencionó anteriormente se obtuvieron resultados con diferentes niveles de permisividad. Esto se debe al distinto número de genes utilizado en cada caso y a las distintas relaciones de divergencia entre ellos. Al realizar la clusterización conjunta, se incorpora el doble de genes y cambian a su vez las medidas de distancia y la identificación de clusters dada una misma inflación. Estos resultados dan a entender que el valor del parámetro de inflación para regular el proceso de clusterización debe ser evaluado caso a caso, ya que este es un proceso dependiente de la cantidad de proteínas a clusterizar y de las relaciones de similaridad existentes entre ellas.

Efecto de las anotaciones sobre la búsqueda de familias multigénicas

Se observa que la falta de componentes en las familias ROP y GRA a partir de la clusterización de forma independiente de *N. caninum* es una consecuencia de la anotación incompleta del genoma de este parásito. Al analizar detenidamente los resultados de la búsqueda de estas familias en la clusterización conjunta con *T. gondii* se encontraron múltiples genes de *N. caninum* sin anotación dentro de clusters con miembros ROP y GRA de *T. gondii*.

Por lo tanto, la clusterización conjunta mostró que los problemas de anotación pueden resolverse usando de forma complementaria las anotaciones de otro genoma. De esta forma, se pueden asociar genes que presentan similitud en secuencia pero que no están

anotados en un organismo, a familias génicas en el otro organismo donde sí están anotados.

Diversidad de las principales familias en base a su rol biológico

A partir de las búsquedas realizadas sobre la clusterización conjunta de ambos parásitos se obtuvieron múltiples clusters para cada una de las familias multigénicas buscadas, siendo esto concordante con la diversidad funcional que estas presentan. De todos modos, para evaluar la diversidad de cada familia con mayor profundidad, sería propicio incluir en el análisis diferentes estadísticos que se puedan asociar con la diversidad de cada cluster o grupo de clusters. Ejemplo de esto pueden ser la diversidad filogenética (PD) y la distancia filogenética media entre pares (MPD), que son usadas comúnmente para evaluar la diversidad de comunidades ecológicas en base a su filogenia (Faith, 1992; Webb et al., 2002). Los cuales pueden ser incluidos en futuras versiones del pipeline. Pero en este caso, el usuario debería tener la opción de no calcularlos, ya que sería necesario realizar una filogenia a cada uno de los clusters de los que deseen obtener estos estadísticos, pudiendo llevar un tiempo considerable en clusters de gran tamaño.

Variabilidad de largos proteicos

Al comparar entre *T. gondii* y *N. caninum*, se observa mayormente que para una misma familia multigénica, los largos proteicos promedios de sus clusters asociados a estas difieren entre ambos parásitos. Para todas las familias esta diferencia se acentúa al comparar los clusters obtenidos por las búsquedas efectuadas sobre clusterización en instancias separadas (Tablas 3.3 y 3.4). Esta variabilidad puede ser asociada a la mayor permisividad en los clusters generados en instancias independientes, los cuales presentaron abundancia de genes sin anotaciones que posiblemente tengan relaciones de homología más distantes o que no pertenezcan a las familias multigénicas buscadas.

Esto puede ser evitado, mediante la utilización de un parámetro de clusterización más restrictivo. De igual forma, la diferencia en los promedios de los largos entre ambos parásitos puede ser consecuencia de la propia variación de la familia, ejemplo de esto son la familia VIR presente en *Plasmodium* y la familia FAINT de *Theileria*, que presentan proteínas de diverso tamaño producto de la variación en su número de dominios (Bateman et al., 2004; Pain et al., 2005; Singh et al., 2014).

Análisis de la familia multigénica SRS en *T. gondii* y *N. caninum*

Los resultados obtenidos del análisis de la familia multigénica SRS en *T. gondii* y *N. caninum* coinciden con estudios previos realizados por otros autores (Adomako-Ankomah et al., 2014; Reid et al., 2012; Wasmuth et al., 2012). Se constató la expansión de esta familia en *N. caninum* respecto a *T. gondii*, existiendo una relación 2 a 1 en el número de integrantes de la familia SRS entre estos organismos. Además se encontró que los genes

de esta familia están presentes en todos los cromosomas de ambos parásitos, mayormente agrupados en tandems (Adomako-Ankomah et al., 2014; Reid et al., 2012; Wasmuth et al., 2012).

La estructura de la filogenia obtenida en conjunto con la ubicación cromosómica de los miembros de la familia multigénica SRS apoya la hipótesis de que estos genes son el resultado de eventos locales de duplicación génica (Reid et al., 2012). Esto es evidenciado a partir de la estructura del árbol en conjunto con su ubicación cromosómica de los integrantes de esta familia en cada una de las especies de parásitos. En donde genes ubicados bajo misma rama, formando un grupo monofilético, también tienen una ubicación espacial cercana en el cromosoma, formando un tándem. Siendo compatible con la hipótesis de que estos grupos de genes se originaron a partir de un gen ancestral el cual sufrió uno o varios eventos locales de duplicación. Esto se alinea con los resultados encontrados en estudios previos de Wasmuth et al., 2012, donde se ha observado que los genes de esta familia que pertenecen a un mismo tándem suelen parecerse más entre ellos que con genes de otros tándems. Sin embargo, esto no siempre ocurre en otras familias multigénicas como la familia pirs de Plasmodium (Otto et al., 2014) y los genes subteloméricos Sfil Theileria (Weir et al. 2010), posiblemente a causa de recombinación entre tándems (Reid, 2015).

Este trabajo también respalda la hipótesis de que la expansión de la familia multigénica SRS en *N. caninum* en comparación con *T. gondii* es impulsada principalmente por eventos de duplicación génica (Reid et al., 2012). En *N. caninum,* los tandems en promedio se componen de una mayor cantidad de genes que en *T. gondii.* La conservación de las secuencias en los tandems de cada parásito podría ser el resultado de eventos locales de duplicación y conversión génica (Reid, 2015).

A pesar de que la mayoría de los genes de la familia multigénica SRS se encuentren agrupados en tandems, se identificaron varios genes de copia única en *T. gondii* y *N. caninum*. De los genes analizados, 62 no están agrupados en tandems, y 28 de ellos pertenecen a clusters que se componen únicamente por dos secuencias, donde una pertenece a *T. gondii* y la otra a *N. caninum*. Estos clusters sugieren que son genes ortólogos especializados de ‘copia única’ dentro de la familia SRS. Además, se encontraron 3 genes pertenecientes a *N. caninum* (Ncaninum_LIV_000621700.1, Ncaninum_LIV_000571600.1, Ncaninum_LIV_000162600.1) anotados como hipotéticos que tienen homólogos en *T. gondii* anotados como SRS. Por lo tanto, estos resultados sugieren que estos tres genes hipotéticos son integrantes de la familia multigénica SRS.

Conclusiones

Se desarrolló un pipeline automático y flexible para la clusterización de genes a partir de un genoma anotado. El mismo fue utilizado en los genomas de *T. gondii* y *N. caninum*, pero potencialmente puede ser utilizado en cualquier grupo de genomas que tengan disponibles su ensamblaje y anotación. A pesar de los desafíos encontrados (diferencias internas de los archivos de anotación, variedad de librerías usadas y dependencia de programas elaborados por terceros) el pipeline es plenamente funcional, pero sería conveniente

implementar estrategías de garantía de calidad (software QA) y usar ‘containers’ para facilitar su uso para otros usuarios.

El análisis comparativo preliminar de la clusterización de los genomas de *T. gondii* y *N. caninum* muestra similitudes congruentes con el estrecho parentesco evolutivo que presentan estos organismos. Los clusters obtenidos presentan una cantidad y distribución de tamaños similar y en su gran mayoría presentan genes de ambos parásitos. Los clusters con genes exclusivos de uno de los parásitos representan una oportunidad para investigar las diferencias específicas que existen entre ellos, como sus principales formas de transmisión y el rango de hospedadores. Sin embargo, es necesario confirmar la completitud de las anotaciones genómicas antes de estudiar estos genes en profundidad

Usando diferentes estrategias se lograron analizar e identificar clusters pertenecientes a las familias SRS, ROP, GRA y MIC en ambos parásitos. La clusterización conjunta fue la estrategia más exitosa ya que a partir de ella se obtuvieron datos más completos y consistentes, en cambio la clusterización independiente, principalmente sobre *N. caninum* presentó problemas relacionados con sus anotaciones genómicas. Esta experiencia demuestra que la clusterización conjunta es una estrategia valiosa para identificar genes y clusters, ya que permite el uso de forma complementaria de las anotaciones de los distintos genomas.

A partir del análisis de la familia multigénica SRS, este trabajo logró reproducir resultados obtenidos anteriormente por otros estudios, entre los que se destacan: la expansión de esta familia en *N. caninum* respecto a *T. gondii*; que sus genes se distribuyen por todos los cromosomas de estos parásitos, estando comúnmente agrupados en tandems; y la mayor similitud entre los genes de un mismo tandems que los pertenecientes de otros (Adomako-Ankomah et al., 2014; Reid et al., 2012; Wasmuth et al., 2012). Además se produjo evidencia para anotar tres genes hipotéticos pertenecientes a *N. caninum* como miembros de la familia SRS.

Finalmente, se propone utilizar el pipeline junto con la estrategia de clusterización conjunta para identificar nuevos miembros en otras familias multigénicas de estas especies, o incluso en otras con anotaciones genómicas menos precisas. Además, este enfoque podría emplearse para identificar genes que influyen en las diferencias biológicas específicas de cada especie, lo que ayudaría a comprender la diversidad y adaptaciones de estos parásitos.

Material Suplementario

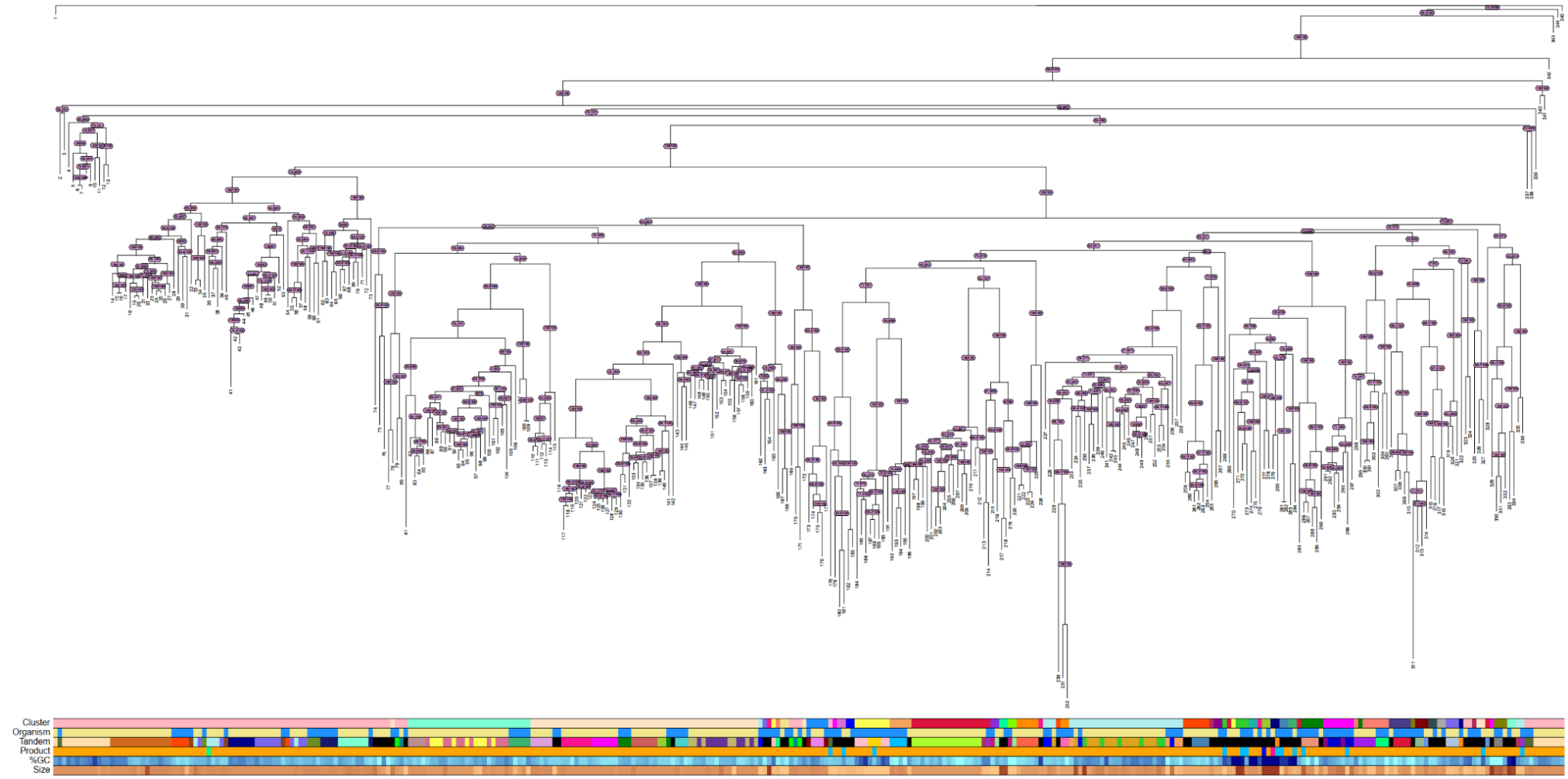


Figura S1. Árbol filogenético de proteínas SRS en *T. gondii* y *N. caninum*. Inferencia filogenética obtenida a partir de todos los integrantes de la familia SRS. Se utilizó el método de máxima verosimilitud. En la parte superior de cada nodo se encuentra su apoyo estadístico obtenido por 5000 iteraciones de UF-bootstrap y sh-aLRT. En la parte

inferior se encuentra una barra de colores que indica para cada integrante cluster, organismo (Organism) y tándem al que pertenecen; así como su, producto proteico (Product), porcentaje de GC (%GC) y largo proteico en aminoácidos (Size). Cada integrante de la filogenia está numerado del 1 al 345, siguiendo el orden de las filas de la tabla usada para generar la figura (véase https://github.com/JPereira199/GenomicClusterExplorer/blob/main/Supplementary_Data/TableS1.tab).

Referencias

Adomako-Ankomah, Y., Wier, G. M., Borges, A. L., Wand, H. E., & Boyle, J. P. (2014). Differential Locus Expansion Distinguishes *Toxoplasmatinae* Species and Closely Related Strains of Toxoplasma gondii. *MBio*, *5*(1), e01003-13. https://doi.org/10.1128/mBio.01003-13

Al-Qassab, S. E., Reichel, M. P., & Ellis, J. T. (2010). On the Biological and Genetic Diversity in Neospora caninum. *Diversity*, *2*(3), 411-438. https://doi.org/10.3390/d2030411

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389-3402. https://doi.org/10.1093/nar/25.17.3389

Atkinson, R., Harper, P. A. W., Ryce, C., Morrison, D. A., & Ellis, J. T. (1999). Comparison of the biological characteristics of two isolates of *Neospora caninum*. *Parasitology*, *118*(4), 363-370. https://doi.org/10.1017/S0031182098003898

Barber, J. S., Gasser, R. B., Ellis, J., Reichel, M. P., McMillan, D., & Trees, A. J. (1997). Prevalence of Antibodies to Neospora caninum in Different Canid Populations. *The Journal of Parasitology*, *83*(6), 1056. https://doi.org/10.2307/3284361

Barber, J. S., & Trees, A. J. (1998). Naturally occurring vertical transmission of Neospora caninum in dogs. *International Journal for Parasitology*, *28*(1), 57-64. https://doi.org/10.1016/S0020-7519(97)00171-9

Barragan, A., & Sibley, L. D. (2002). Transepithelial Migration of Toxoplasma gondii Is Linked to Parasite Motility and Virulence. *Journal of Experimental Medicine*, *195*(12), 1625-1633. https://doi.org/10.1084/jem.20020258

Barragan, A., & Sibley, L. D. (2003). Migration of Toxoplasma gondii across biological barriers. *Trends in Microbiology*, *11*(9), 426-430.

https://doi.org/10.1016/s0966-842x(03)00205-1

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., & Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Research*, *32*(Database issue), D138-141. https://doi.org/10.1093/nar/gkh121

Beckers, C. J., Dubremetz, J. F., Mercereau-Puijalon, O., & Joiner, K. A. (1994). The Toxoplasma gondii rhoptry protein ROP 2 is inserted into the parasitophorous vacuole membrane, surrounding the intracellular parasite, and is exposed to the host cell cytoplasm. *Journal of Cell Biology*, *127*(4), 947-961. https://doi.org/10.1083/jcb.127.4.947

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2008). GenBank. *Nucleic Acids Research*, *36*(Database issue), D25-D30. https://doi.org/10.1093/nar/gkm929

Berná, L., Marquez, P., Cabrera, A., Greif, G., Francia, M. E., & Robello, C. (2021). Reevaluation of the Toxoplasma gondii and Neospora caninum genomes reveals misassembly, karyotype differences, and chromosomal rearrangements. *Genome Research*, *31*(5), 823-833. https://doi.org/10.1101/gr.262832.120

Blank, M. L., & Boyle, J. P. (2018). Effector variation at tandem gene arrays in tissue-dwelling coccidia: Who needs antigenic variation anyway? *Current Opinion in Microbiology*, *46*, 86-92. https://doi.org/10.1016/j.mib.2018.09.001

Boothroyd, J. C., & Dubremetz, J.-F. (2008). Kiss and spit: The dual roles of Toxoplasma rhoptries. *Nature Reviews Microbiology*, *6*(1), Article 1. https://doi.org/10.1038/nrmicro1800

Bradley, P. J., Ward, C., Cheng, S. J., Alexander, D. L., Coller, S., Coombs, G. H., Dunn, J. D., Ferguson, D. J., Sanderson, S. J., Wastling, J. M., & Boothroyd, J. C. (2005). Proteomic Analysis of Rhoptry Organelles Reveals Many Novel Constituents for Host-Parasite Interactions in Toxoplasma gondii. *Journal of Biological Chemistry*, *280*(40), 34245-34258. https://doi.org/10.1074/jbc.M504158200

Calarco, L., & Ellis, J. (2020). Species diversity and genome evolution of the pathogenic protozoan parasite, Neospora caninum. *Infection, Genetics and Evolution*, *84*, 104444. https://doi.org/10.1016/j.meegid.2020.104444

Carruthers, V. B. (2002). Host cell invasion by the opportunistic pathogen Toxoplasma gondii. *Acta Tropica*, *81*(2), 111-122. https://doi.org/10.1016/S0001-706X(01)00201-7

Cruz-Mirón, R., Ramírez-Flores, C. J., Lagunas-Cortés, N., Mondragón-Castelán, M., Ríos-Castro, E., González-Pozos, S., Aguirre-García, M. M., & Mondragón-Flores, R. (2021). Proteomic characterization of the pellicle of Toxoplasma gondii. *Journal of Proteomics*, *237*, 104146. https://doi.org/10.1016/j.jprot.2021.104146

Darde, M. L., Bouteille, B., & Pestre-Alexandre, M. (1988). Isoenzymic characterization of seven strains of Toxoplasma gondii by isoelectrofocusing in polyacrylamide gels. *The American Journal of Tropical Medicine and Hygiene*, *39*(6), 551-558. https://doi.org/10.4269/ajtmh.1988.39.551

Despa, M. L. (2014). Comparative study on software development methodologies. *Database Systems Journal*, *5*(3), 37-56.

Dongen, S. (2000). Graph Clustering by Flow Simulation. *PhD thesis, Center for Math and Computer Science (CWI)*.

Dubey, J. P. (2003). Review of Neospora caninum and neosporosis in animals. *The Korean Journal of Parasitology*, *41*(1), 1-16. https://doi.org/10.3347/kjp.2003.41.1.1

Dubey, J. P. (2004). Toxoplasmosis – a waterborne zoonosis. *Veterinary Parasitology*, *126*(1-2), 57-72. https://doi.org/10.1016/j.vetpar.2004.09.005

Dubey, J. P. (2007). The History and Life Cycle of Toxoplasma gondii. En *Toxoplasma Gondii* (pp. 1-17). Elsevier. https://doi.org/10.1016/B978-012369542-0/50003-9

Dubey, J. P., & Schares, G. (2011). Neosporosis in animals—The last five years. *Veterinary Parasitology*, *180*(1-2), 90-108. https://doi.org/10.1016/j.vetpar.2011.05.031

Dubey, J. P., Schares, G., & Ortega-Mora, L. M. (2007). Epidemiology and Control of Neosporosis and *Neospora caninum*. *Clinical Microbiology Reviews*, *20*(2), 323-367. https://doi.org/10.1128/CMR.00031-06

Dubremetz, J. F. (2007). Rhoptries are major players in Toxoplasma gondii invasion and host cell interaction. *Cellular Microbiology*, *9*(4), 841-848. https://doi.org/10.1111/j.1462-5822.2007.00909.x

Dubremetz, J. F., & Lebrun, M. (2012). Virulence factors of Toxoplasma gondii. *Microbes and Infection*, *14*(15), 1403-1410. https://doi.org/10.1016/j.micinf.2012.09.005

El Hajj, H., Lebrun, M., Arold, S. T., Vial, H., Labesse, G., & Dubremetz, J. F. (2007). ROP18 Is a Rhoptry Kinase Controlling the Intracellular Proliferation of Toxoplasma gondii. *PLoS Pathogens*, *3*(2), e14. https://doi.org/10.1371/journal.ppat.0030014

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, *61*(1), 1-10. https://doi.org/10.1016/0006-3207(92)91201-3

Gajria, B., Bahl, A., Brestelli, J., Dommer, J., Fischer, S., Gao, X., Heiges, M., Iodice, J., Kissinger, J. C., Mackey, A. J., Pinney, D. F., Roos, D. S., Stoeckert, C. J., Wang, H., & Brunk, B. P. (2008). ToxoDB: An integrated Toxoplasma gondii database resource. *Nucleic Acids Research*, *36*(Database issue), D553-556. https://doi.org/10.1093/nar/gkm981

Ghanem, M. E., Suzuki, T., Akita, M., & Nishibori, M. (2009). Neospora caninum and complex vertebral malformation as possible causes of bovine fetal mummification. *The Canadian Veterinary Journal*, *50*(4), 389-392.

Gold, D. A., Kaplan, A. D., Lis, A., Bett, G. C. L., Rosowski, E. E., Cirelli, K. M., Bougdour, A., Sidik, S. M., Beck, J. R., Lourido, S., Egea, P. F., Bradley, P. J., Hakimi, M.-A., Rasmusson, R. L., & Saeij, J. P. J. (2015). The Toxoplasma Dense Granule Proteins GRA17 and GRA23 Mediate the Movement of Small Molecules between the Host and the Parasitophorous Vacuole. *Cell Host & Microbe*, *17*(5), 642-652. https://doi.org/10.1016/j.chom.2015.04.003

Gubbels, M.-J., White, M., & Szatanek, T. (2008). The cell cycle and Toxoplasma gondii cell division: Tightly knit or loosely stitched? *International Journal for Parasitology*, *38*(12), 1343-1358. https://doi.org/10.1016/j.ijpara.2008.06.004

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C.

R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

He, X., Grigg, M. E., Boothroyd, J. C., & Garcia, K. C. (2002). Structure of the immunodominant surface antigen from the Toxoplasma gondii SRS superfamily. *Nature Structural Biology*, *9*(8), 606-611. https://doi.org/10.1038/nsb819

Hosseini, S. A., Amouei, A., Sharif, M., Sarvi, Sh., Galal, L., Javidnia, J., Pagheh, A. S., Gholami, S., Mizani, A., & Daryani, A. (2019). Human toxoplasmosis: A systematic review for genetic diversity of *Toxoplasma gondii* in clinical samples. *Epidemiology and Infection*, *147*, e36. https://doi.org/10.1017/S0950268818002947

Howe, D. K., & Sibley, L. D. (1995). Toxoplasma gondii Comprises Three Clonal Lineages: Correlation of Parasite Genotype with Human Disease. *The Journal of Infectious Diseases*, *Volume 172,*(Issue 6), 1561-15668. https://doi.org/10.1093/infdis/172.6.1561

Jerome, M. E., Radke, J. R., Bohne, W., Roos, D. S., & White, M. W. (1998). Toxoplasma Gondii Bradyzoites Form Spontaneously during Sporozoite-Initiated Development. *Infection and Immunity*, *66*(10), 4838-4844.

Jung, C., Lee, C. Y.-F., & Grigg, M. E. (2004). The SRS superfamily of Toxoplasma surface proteins. *International Journal for Parasitology*, *34*(3), 285-296. https://doi.org/10.1016/j.ijpara.2003.12.004

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772-780. https://doi.org/10.1093/molbev/mst010

Kemp, L. E., Yamamoto, M., & Soldati-Favre, D. (2013). Subversion of host cellular functions by the apicomplexan parasites. *FEMS Microbiology Reviews*, *37*(4), 607-631. https://doi.org/10.1111/1574-6976.12013

Khan, A., Shaik, J. S., Sikorski, P., Dubey, J. P., & Grigg, M. E. (2020). Neosporosis: An Overview of Its Molecular Epidemiology and Pathogenesis. *Engineering*, *6*(1), 10-19.

https://doi.org/10.1016/j.eng.2019.02.010

Lebrun, M., Michelin, A., El Hajj, H., Poncet, J., Bradley, P. J., Vial, H., & Dubremetz, J. F. (2005). The rhoptry neck protein RON4 re-localizes at the moving junction during Toxoplasma gondii invasion. *Cellular Microbiology*, *7*(12), 1823-1833. https://doi.org/10.1111/j.1462-5822.2005.00646.x

Lehmann, T., Graham, D., Dahl, E., Bahiaoliveira, L., Gennari, S., & Dubey, J. (2004). Variation in the structure of Toxoplasma gondii and the roles of selfing, drift, and epistatic selection in maintaining linkage disequilibria. *Infection, Genetics and Evolution*, *4*(2), 107-114. https://doi.org/10.1016/j.meegid.2004.01.007

Lekutis, C., Ferguson, D. J. P., Grigg, M. E., Camps, M., & Boothroyd, J. C. (2001). Surface antigens of Toxoplasma gondii: Variations on a theme. *International Journal for Parasitology*, *31*(12), 1285-1292. https://doi.org/10.1016/S0020-7519(01)00261-2

Lorenzi, H., Khan, A., Behnke, M. S., Namasivayam, S., Swapna, L. S., Hadjithomas, M., Karamycheva, S., Pinney, D., Brunk, B. P., Ajioka, J. W., Ajzenberg, D., Boothroyd, J. C., Boyle, J. P., Dardé, M. L., Diaz-Miranda, M. A., Dubey, J. P., Fritz, H. M., Gennari, S. M., Gregory, B. D., … Sibley, L. D. (2016). Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic Toxoplasma gondii genomes. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms10147

McCann, C. M., Vyse, A. J., Salmon, R. L., Thomas, D., Williams, D. J. L., McGarry, J. W., Pebody, R., & Trees, A. J. (2008). Lack of Serologic Evidence of Neospora caninum in Humans, England. *Emerging Infectious Diseases*, *14*(6), 978-980. https://doi.org/10.3201/eid1406.071128

McInnes, L. M., Irwin, P., Palmer, D. G., & Ryan, U. M. (2006). In vitro isolation and characterisation of the first canine Neospora caninum isolate in Australia. *Veterinary Parasitology*, *137*(3-4), 355-363. https://doi.org/10.1016/j.vetpar.2006.01.018

McInnes, L. M., Ryan, U. M., O'Handley, R., Sager, H., Forshaw, D., & Palmer, D. G. (2006). Diagnostic significance of Neospora caninum DNA detected by PCR in cattle serum.

Veterinary Parasitology, *142*(3), 207-213. https://doi.org/10.1016/j.vetpar.2006.07.013

Michelin, A., Bittame, A., Bordat, Y., Travier, L., Mercier, C., Dubremetz, J.-F., & Lebrun, M. (2009). GRA12, a Toxoplasma dense granule protein associated with the intravacuolar membranous nanotubular network. *International Journal for Parasitology*, *39*(3), 299-306. https://doi.org/10.1016/j.ijpara.2008.07.011

Miller, C., Quinn, H., Windsor, P., & Ellis, J. (2002). Characterisation of the first Australian isolate of Neospora caninum from cattle. *Australian Veterinary Journal*, *80*(10), 620-625. https://doi.org/10.1111/j.1751-0813.2002.tb10967.x

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530-1534. https://doi.org/10.1093/molbev/msaa015

Monteiro, R. M., Richtzenhain, L. J., Pena, H. F. J., Souza, S. L. P., Funada, M. R., Gennari, S. M., Dubey, J. P., Sreekumar, C., Keid, L. B., & Soares, R. M. (2007). Molecular phylogenetic analysis in Hammondia-like organisms based on partial Hsp70 coding sequences. *Parasitology*, *134*(Pt 9), 1195-1203. https://doi.org/10.1017/S0031182007002612

Morrison, D. (2009). Evolution of the Apicomplexa: Where are we now? *Trends in parasitology*, *25*, 375-382. https://doi.org/10.1016/j.pt.2009.05.010

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., … Venter, J. C. (2000). A whole-genome assembly of Drosophila. *Science (New York, N.Y.)*, *287*(5461), 2196-2204. https://doi.org/10.1126/science.287.5461.2196

Nkumama, I. N., O'Meara, W. P., & Osier, F. H. A. (2017). Changes in Malaria Epidemiology in Africa and New Challenges for Elimination. *Trends in Parasitology*, *33*(2), 128-140.

https://doi.org/10.1016/j.pt.2016.11.006

Otto, T. D., Böhme, U., Jackson, A. P., Hunt, M., Franke-Fayard, B., Hoeijmakers, W. A. M., Religa, A. A., Robertson, L., Sanders, M., Ogun, S. A., Cunningham, D., Erhart, A., Billker, O., Khan, S. M., Stunnenberg, H. G., Langhorne, J., Holder, A. A., Waters, A. P., Newbold, C. I., … Janse, C. J. (2014). A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biology*, *12*(1), 86. https://doi.org/10.1186/s12915-014-0086-0

Pain, A., Renauld, H., Berriman, M., Murphy, L., Yeats, C. A., Weir, W., Kerhornou, A., Aslett, M., Bishop, R., Bouchier, C., Cochet, M., Coulson, R. M. R., Cronin, A., de Villiers, E. P., Fraser, A., Fosker, N., Gardner, M., Goble, A., Griffiths-Jones, S., … Hall, N. (2005). Genome of the Host-Cell Transforming Parasite Theileria annulata Compared with T. parva. *Science*, *309*(5731), 131-133. https://doi.org/10.1126/science.1110418

Paraiso, F., Challita, S., Al-Dhuraibi, Y., & Merle, P. (2016). Model-Driven Management of Docker Containers. *2016 IEEE 9th International Conference on Cloud Computing (CLOUD)*, 718-725. https://doi.org/10.1109/CLOUD.2016.0100

Pontius, J. U., Mullikin, J. C., Smith, D. R., Team, A. S., Lindblad-Toh, K., Gnerre, S., Clamp, M., Chang, J., Stephens, R., Neelam, B., Volfovsky, N., Schäffer, A. A., Agarwala, R., Narfström, K., Murphy, W. J., Giger, U., Roca, A. L., Antunes, A., Menotti-Raymond, M., … O'Brien, S. J. (2007). Initial sequence and comparative analysis of the cat genome. *Genome Research*, *17*(11), 1675-1689. https://doi.org/10.1101/gr.6380007

Quinn, H. E., Ellis, J. T., & Smith, N. C. (2002). Neospora caninum: A cause of immune-mediated failure of pregnancy? *Trends in Parasitology*, *18*(9), 391-394. https://doi.org/10.1016/S1471-4922(02)02324-3

Reichel, M. P., Alejandra Ayanegui-Alcérreca, M., Gondim, L. F. P., & Ellis, J. T. (2013). What is the global economic impact of Neospora caninum in cattle—The billion dollar question. *International Journal for Parasitology*, *43*(2), 133-142. https://doi.org/10.1016/j.ijpara.2012.10.022

Reichel, M. P., Ellis, J. T., & Dubey, J. P. (2007). Neosporosis and hammondiosis in dogs.

The Journal of Small Animal Practice, *48*(6), 308-312. https://doi.org/10.1111/j.1748-5827.2006.00236.x

Reid, A. J. (2015). Large, rapidly evolving gene families are at the forefront of host–parasite interactions in *Apicomplexa*. *Parasitology*, *142*(S1), S57-S70. https://doi.org/10.1017/S0031182014001528

Reid, A. J., Vermont, S. J., Cotton, J. A., Harris, D., Hill-Cawthorne, G. A., Könen-Waisman, S., Latham, S. M., Mourier, T., Norton, R., Quail, M. A., Sanders, M., Shanmugam, D., Sohal, A., Wasmuth, J. D., Brunk, B., Grigg, M. E., Howard, J. C., Parkinson, J., Roos, D. S., … Wastling, J. M. (2012). Comparative Genomics of the Apicomplexan Parasites Toxoplasma gondii and Neospora caninum: Coccidia Differing in Host Range and Transmission Strategy. *PLOS Pathogens*, *8*(3), e1002567. https://doi.org/10.1371/journal.ppat.1002567

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276-277.

Robert-Gangneux, F., & Dardé, M.-L. (2012). Epidemiology of and Diagnostic Strategies for Toxoplasmosis. *Clinical Microbiology Reviews*, *25*(2), 264-296. https://doi.org/10.1128/CMR.05013-11

Rojo-Montejo, S., Collantes-Fernández, E., Regidor-Cerrillo, J., Álvarez-García, G., Marugan-Hernández, V., Pedraza-Díaz, S., Blanco-Murcia, J., Prenafeta, A., & Ortega-Mora, L. M. (2009). Isolation and characterization of a bovine isolate of Neospora caninum with low virulence. *Veterinary Parasitology*, *159*(1), 7-16. https://doi.org/10.1016/j.vetpar.2008.10.009

Saeij, J. P. J., Boyle, J. P., Grigg, M. E., Arrizabalaga, G., & Boothroyd, J. C. (2005). Bioluminescence imaging of Toxoplasma gondii infection in living mice reveals dramatic differences between strains. *Infection and Immunity*, *73*(2), 695-702. https://doi.org/10.1128/IAI.73.2.695-702.2005

Schock, A., Innes, E. A., Yamane, I., Latham, S. M., & Wastling, J. M. (2001). Genetic and biological diversity among isolates of *Neospora caninum*. *Parasitology*, *123*(1),

13-23. https://doi.org/10.1017/S003118200100796X

Seeber, F., Dubremetz, J. F., & Boothroyd, J. C. (1998). Analysis of Toxoplasma gondii stably transfected with a transmembrane variant of its major surface protein, SAG1. *Journal of Cell Science*, *111 (Pt 1)*, 23-29. https://doi.org/10.1242/jcs.111.1.23

Shwab, E. K., Zhu, X.-Q., Majumdar, D., Pena, H. F. J., Gennari, S. M., Dubey, J. P., & Su, C. (2014). Geographical patterns of *Toxoplasma gondii* genetic diversity revealed by multilocus PCR-RFLP genotyping. *Parasitology*, *141*(4), 453-461. https://doi.org/10.1017/S0031182013001844

Sibley, L. D., & Boothroyd, J. C. (1992). Virulent strains of Toxoplasma gondii comprise a single clonal lineage. *Nature*, *359*(6390), 82-85. https://doi.org/10.1038/359082a0

Sibley, L. D., Khan, A., Ajioka, J. W., & Rosenthal, B. M. (2009). Genetic diversity of Toxoplasma gondii in animals and humans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1530), 2749-2761. https://doi.org/10.1098/rstb.2009.0087

Silva, R. C., & Machado, G. P. (2016). Canine neosporosis: Perspectives on pathogenesis and management. *Veterinary Medicine : Research and Reports*, *7*, 59-70. https://doi.org/10.2147/VMRR.S76969

Singh, V., Gupta, P., & Pande, V. (2014). Revisiting the multigene families: Plasmodium var and vir genes. *J Vector Borne Dis*.

Soete, M., Fortier, B., Camus, D., & Dubremetz, J. F. (1993). Toxoplasma gondii: Kinetics of Bradyzoite-Tachyzoite Interconversion in vitro. *Experimental Parasitology*, *76*(3), 259-264. https://doi.org/10.1006/expr.1993.1031

Speer, C. A., Dubey, J. P., McAllister, M. M., & Blixt, J. A. (1999). Comparative ultrastructure of tachyzoites, bradyzoites, and tissue cysts of Neospora caninum and Toxoplasma gondii. *International Journal for Parasitology*, *29*(10), 1509-1519. https://doi.org/10.1016/S0020-7519(99)00132-0

Su, C., Khan, A., Zhou, P., Majumdar, D., Ajzenberg, D., Dardé, M.-L., Zhu, X.-Q., Ajioka, J. W., Rosenthal, B. M., Dubey, J. P., & Sibley, L. D. (2012). Globally diverse

Toxoplasma gondii isolates comprise six major clades originating from a small number of distinct ancestral lineages. *Proceedings of the National Academy of Sciences*, *109*(15), 5844-5849. https://doi.org/10.1073/pnas.1203190109

Tenter, A. M., Heckeroth, A. R., & Weiss, L. M. (2000). Toxoplasma gondii: From animals to humans. *International journal for parasitology*, *30*(12-13), 1217-1258.

Wasmuth, J. D., Pszenny, V., Haile, S., Jansen, E. M., Gast, A. T., Sher, A., Boyle, J. P., Boulanger, M. J., Parkinson, J., & Grigg, M. E. (2012). Integrated bioinformatic and targeted deletion analyses of the SRS gene superfamily identify SRS29C as a negative regulator of Toxoplasma virulence. *MBio*, *3*(6), e00321-12. https://doi.org/10.1128/mBio.00321-12

Webb, C. O., Ackerly, D. D., McPeek, M. A., & Donoghue, M. J. (2002). Phylogenies and Community Ecology. *Annual Review of Ecology and Systematics*, *33*, 475-505.

Weber, F. H., Jackson, J. A., Sobecki, B., Choromanski, L., Olsen, M., Meinert, T., Frank, R., Reichel, M. P., & Ellis, J. T. (2013). On the Efficacy and Safety of Vaccination with Live Tachyzoites of Neospora caninum for Prevention of Neospora-Associated Fetal Loss in Cattle. *Clinical and Vaccine Immunology*, *20*(1), 99-105. https://doi.org/10.1128/CVI.00225-12

Wendte, J. M., Gibson, A. K., & Grigg, M. E. (2011). Population genetics of Toxoplasma gondii: New perspectives from parasite genotypes in wildlife. *Veterinary parasitology*, *182*(1), 96-111. https://doi.org/10.1016/j.vetpar.2011.07.018