Mathias

26 Octubre, 2025

Orígenes evolutivos de familias génicas en parásitos kinetoplástidos

Introduccion

La historia de la vida está marcada por transiciones en diferentes momentos de la evolucion entre formas de vida libres y parasitarias, cada transicion es acompañada por profundas transformaciones genómicas y fenotípicas. Los parásitos representan una porción significativa de la biodiversidad eucariota, con su aparición se han generado cambios en la evolución y la ecología de prácticamente todos los demás organismos.

El parasitismo anteriormente se asociaba a una evolución reductiva, es decir pérdida génica, simplificación metabólica y dependencia del hospedador. Sin embargo, el parasitismo no implican únicamente pérdida de genes, sino también reorganización y expansión de ciertos genes. Al depender energéticamente del hospedador, los parásitos pueden enfrentar menos restricciones en cuanto al contenido genomico, lo que posibilita la retención y diversificación de familias génicas asociadas a la interacción hospedador-parásito.

Los Kinetoplástidos son un grupo de eucariotas que representan un buen modelo para estudiar estos procesos evolutivos, dado que incluye tanto organismos de vida libre (*Bodo saltans, Neobodo designs*) asi como parasitos de gran relevancia médica (*Trypanosoma y Leishmania*) y agrícola (*Phytomonas*). Esta diversidad en un unico grupo permite establecer comparaciones directas a nivel evolutivo sobre ganancia y perdidas de genes.

Los genomas de *Trypanosoma brucei*, *Trypanosoma cruzi*, *Leishmania major* y *Pythomonas* tienen una organización genómica compleja, caracterizada por una alta proporción familias génicas implicadas en la interacción con el hospedador, la evasión inmune y la adhesión celular. Entre las familias con mayor representacion se ecuentran trans-sialidasas, mucinas, MASP, amastinas y GP63, con distinto grados de divergencia entre e intra especies.

El origen evolutivo de estas familias genicas asociadas al estilo de vida parasitario continúa siendo un tema en discusion. Una interrogante clave para entender los procesos evolutivos es determinar si estas familias génicas son innovaciones que surgieron junto con la transición hacia el parasitismo, o si derivan de genes de especies ancestrales preexistentes que posteriormente pudieron diversificarse y expandirse en respuesta a presiones selectivas ambientales como la interacción con el hospedador.

Para abordar esta pregunta se requiere un enfoque comparativo, en el cual se integre información genómica, filogenética y funcional, abarcando tanto parásitos como organismos de vida libre.

Objetivos

Objetivo general

Investigar los diferentes orígenes y la historia evolutiva de familias multigénicas en Kinetoplástidos parásitos, con el fin de identificar patrones de expansión, conservación y diversificacion asociados al parasitismo a traves de la evolucion.

Objetivos específicos

Identificar las familias génicas conservadas en cada especie parasitaria y aquellas compartidas entre especies parásitas y de vida libre.

Reconstruir las relaciones evolutivas y los patrones de duplicación y pérdida de las familias génicas transsialidasas, mucinas, MASP, amastinas y GP63.

Analizar la organización genómica y la distribución cromosómica de dichas familias en cada especie.

Metodologia

Obtención y preparación de datos genómicos

Los datos genómicos se obtendrán de la base de datos de Ensembl Protists (https://protists.ensembl.org/), seleccionando a las especies de Kinetoplástidos parásitos (*Trypanosoma cruzi, T. brucei, Leishmania major*) y de vida libre (como *Bodo saltans y Neobodo designis*), ademas se integraran genomas de Euglenozoa (*Euglena gracilis, Diplonema papillatum, Perkinsela spp. y Naegleria gruberi*) con el fin de abarcar la diversidad filogenética del grupo y de organismos ancestrales.

Desde el entorno online de Ensembl Protists se descargan los datos genomicos para cada especie, archivos formato FASTA (CDS y proteínas), archivos en formato GFF3 con las anotaciones y coordenadas. Ademas,archivos en formato FASTA del genoma completo.

Se aplicarán filtros en el buscador de Ensembl Protists para incluir únicamente genes anotados como "protein_coding" y excluir elementos hipotéticos o sin evidencia de transcripción.

Posteriormente, los datos se analizaran en un entorno local. Asimismo, se verificara que toda anotacion y formatos se mantenga uniforme en las diferentes especies todo esto se hara a traves de comandos de bash y comandos de awk. Las anotaciones se estandarizarán mediante scripts propios en R y Python, garantizando la compatibilidad entre especies y el correcto enlace entre las secuencias y sus metadatos (posición genómica, longitud, producto génico y tipo de anotación). Sobre todo normalizar las etiquetas de anotacion que por lo general difieren en cambios minimos pudiendo generar errores de analisis posteriores.

Identificación y clasificación de familias multigénicas

Dada que las anotaciones entre especie varian ya que son anotados en base a diferentes programas reanotaremos estas familias multigenicas a partir de un set de secuencias curadas de forma manual. Se realizará una búsqueda basandonos en homología de secuencias mediante la herramienta BLAST con sus aplicaciones. Las secuencias identificadas serán agrupadas en familias génicas en base a criterios de similitud y se constuiran perfiles de hmm a partir de los alineamientos para labusqueda de homologos remotos en especes ancestrales.

Análisis filogenético

Los alineamientos de cada familia se haran utilizando MAFFT, seguidos de la construcción filogenética mediante IQ-TREE con modelos de sustitución, seleccionando el mejor modelo que se ajuste. Se inferirán los patrones de duplicación y pérdida génica mediante superposicion en un árboles de especies. Permitiendo identificar eventos de expansión o contracción asociados a la transición hacia el parasitismo.

Caracterización genómica y distribución cromosómica

Para los cálculos de longitud de secuencias, contenido de GC y frecuencias de dinucleótidos y trinucleótidos se realizarán en R utilizando el paquete seqinR. Este paquete permite manipular y analizar secuencias en formato FASTA de forma sencilla. Obteniendo para cada gen o familia un conjunto de caracteristicas genómicas.

Las posiciones genómicas de cada familia se analizarán para evaluar su organización y distribución cromosómica, haciendo énfasis en la posible asociación de el entorno genomico, regiones subteloméricas y regiones centroméricas.

Estos resultados serán guardados en tablas que integrarán información de varias especies, familias génicas y caracteristicas. Posteriormente, los valores obtenidos se someterán a análisis multivariantes, como el Análisis de Componentes Principales (PCA) haciendo uso de la herramienta scikit-learn de python, explorando patrones globales de estas caracteristicas en cada una de las familias.

Aprendizaje automático

Se integrarán características genómicas, filogenéticas y estructurales de las familias multigénicas para realizar una clusterización basada en esta informacion. Los análisis se haran en Python, haciendo uso de la librería scikit-learn para aplicar métodos de reducción de dimensionalidad (PCA, t-SNE) y técnicas de clustering (k-means, DBSCAN). Permitira identificar clusters entre familias que tengan caracteristicas similares en las diferentes especies o las que no compartan nunguna de las caracteristicas relevadas.

Desarrollo de un flujo de trabajo reproducible

Con el objetivo de asegurar la trazabilidad y reproducibilidad de los resultados, se implementará un pipeline automatizado en Nextflow, integrando todas las etapas del análisis (descarga de datos, filtrado, alineamiento, reconstrucción filogenética, cálculos genómicos y análisis multivariante). El código y los resultados intermedios serán versionados mediante Git y alojados en un repositorio GitHub público, siguiendo buenas prácticas de ciencia abierta y documentación transparente.

Visualización e integración de resultados

Los resultados se presentarán mediante representaciones gráficas y filogenéticas que integren la distribución evolutiva y genómica de las familias analizadas. Se espera obtener mapas filogenéticos que reflejen los patrones de expansión génica, la relación con regiones genómicas particulares y los posibles vínculos entre estructura y función. Finalmente, los datos integrados permitirán proponer hipótesis sobre el origen y diversificación de las familias multigénicas en los Kinetoplástidos, aportando un marco comparativo para comprender la evolución del parasitismo dentro del grupo.