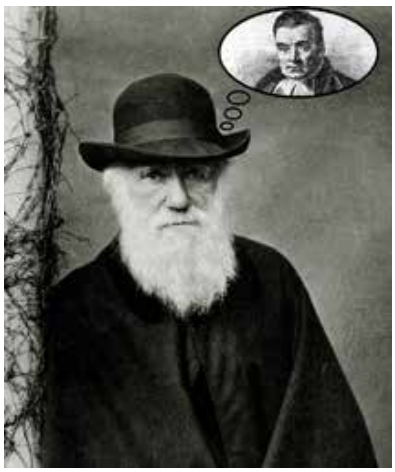
Filogenómica y genómica de la adaptación

Rafael Zardoya
Profesor de Investigación del Museo Nacional de Ciencias Naturales-CSIC, Madrid
Alejandro Sánchez-Gracia
Profesor Asociado del Departament de Genètica, Microbiologia i Estadística e Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona.
Julio Rozas
Catedrático del Departament de Genètica, Microbiologia i Estadística e Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona.

La posibilidad técnica actual de secuenciar genomas y transcriptomas completos de cualquier especie abre la puerta a cumplir la predicción de Darwin de poder reconstruir de forma rigurosa el árbol de la vida (las relaciones filogenéticas de todos los seres vivos actuales) mediante la filogenómica. El árbol de la vida sirve como marco robusto para realizar estudios de genómica evolutiva de la adaptación; es decir, identificar los genes o regiones del genoma moldeados por la acción de la selección natural y responsables de la adaptación fenotípica.



Darwin describió el origen de las especies como "el misterio de los misterios" y más de 150 años después de publicar su magna obra, comprender los mecanismos y procesos evolutivos que han generado y mantienen la biodiversidad que nos rodea sigue siendo el principal reto con el que se enfrenta el biólogo evolutivo. Aún así, hemos avanzado enormemente, en particular gracias a la aportación de la evolución molecular. Hoy sabemos que la variabilidad genética generada por mutaciones al azar en el genoma es en su mayoría neutra en términos selectivos y que existe una presión selectiva muy fuerte para evitar mutaciones deletéreas (aquellas que conllevan una pérdida de aptitud o eficacia biológica). Solo una parte ínfima de las mutaciones son beneficiosas, contribuyen a mejorar la adaptación al medio y son objeto de la selección natural positiva (o darwiniana). La diversidad biológica se genera inicialmente por mutación y se moldea por diversos mecanismos evolutivos que alteran la frecuencia de alelos en una población. De estos, el principal es la selección natural. Los otros mecanismos de diversificación son la deriva génica o selección al azar de alelos en una población, la recombinación y el flujo génico asociado a la importación de alelos procedentes de otras poblaciones por migración. Para que se produzca la especiación es necesario, además, el concurso de mecanismos específicos de interrupción del flujo génico (barreras reproductivas) que permitan la fijación diferencial de las mutaciones. Existiendo consenso en que todos estos mecanismos evolutivos están implicados en la especiación, el debate se centra en discernir el papel relativo de cada uno de ellos en la generación de la diversidad que observamos en los diferentes grupos de organismos.

La evolución trabaja de forma sutil y gradual y sus resultados son, en general, observables solo a largo plazo. Por ello, el estudio de los procesos evolutivos no es nada sencillo y normalmente se recurre a sistemas modelo en los que se observa una aceleración de las tasas evolutivas como es el caso, por ejemplo, de las radiaciones (aparición casi simultánea de nuevos linajes) en islas oceánicas, la domesticación por el hombre de especies o la evolución experimental con organismos con ciclos vitales rápidos como bacterias y virus. En estos sistemas, se pueden realizar aproximaciones a nivel ecosistémico, organísmico o genómico, siendo este último el foco del presente artículo.

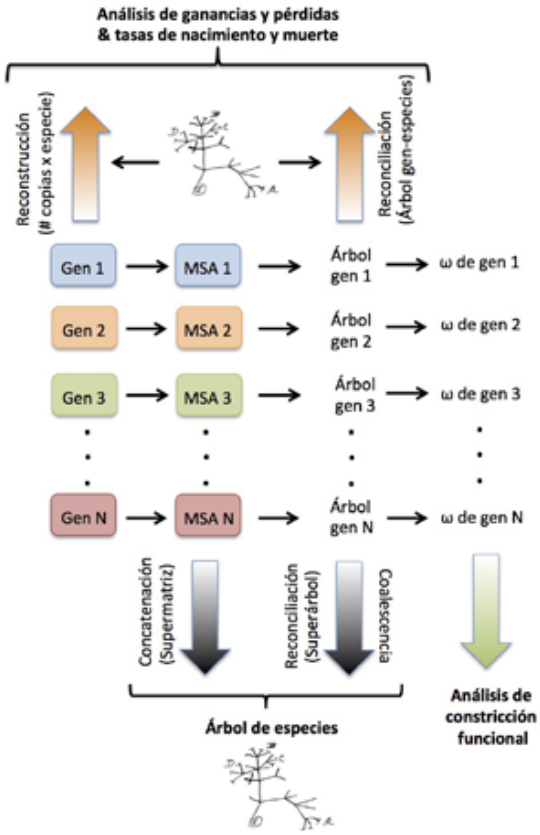
El desarrollo de las técnicas de secuenciación masiva en la última década ha permitido, por primera vez, testar hipótesis sobre el efecto de los procesos adaptativos a

›››› nivel genómico y su papel en la generación de la biodiversidad. Estas nuevas técnicas, en constante mejora, permiten obtener datos del genoma y del transcriptoma de organismos no modelo, abriendo así la puerta a los estudios evolutivos. Dentro de un marco filogenético robusto, se comparan las diferencias en la secuencia de los diferentes genes y regiones no codificantes en el genoma entre especies que comparten un antepasado común reciente, para poder asociar los cambios genéticos observados con cambios fenotípicos sujetos a selección y adaptación. De esta manera, se pueden detectar genes o regiones del genoma que pueden ser potenciales candidatos a ser los agentes causales del proceso de especiación, que con posterioridad deberán ser objeto de validación experimental con las técnicas de biología molecular apropiadas. La filogenómica es la encargada de reconstruir las relaciones filogenéticas de las especies a partir de datos genómicos y proporcionar el marco necesario a la genómica comparada y evolutiva que permite analizar la variación a nivel estructural en el genoma, el origen y la evolución de las familias génicas o detectar qué genes presentan selección positiva y a qué dominios proteicos afecta. A continuación, repasamos brevemente las metodologías utilizadas en filogenómica y genómica de la adaptación, y presentamos un ejemplo de este tipo de estudios.

FILOGENÓMICA

Darwin definió la evolución como la descendencia con modificación y él mismo se dio cuenta de que el proceso evolutivo implica que los organismos vivos presentan rela-

ESQUEMA DE FLUJO DE TRABAJO EN FILOGENÓMICA Y GENÓMICA DE LA ADAPTACIÓN



GENÓMICA DE LA ADAPTACIÓN A LA DIETA CARNÍVORA EN PLANTAS

Fukushima y colaboradores han realizado recientemente un estudio genómico en la planta carnívora *Cephalotus follicularis* (planta de jarra), con objeto de investigar las bases moleculares de la adaptación a la dieta carnívora (*figura 1*). Esta adaptación, que permite a las plantas sobrevivir en suelos pobres en nutrientes, se ha repetido a lo largo de la evolución de forma independiente (convergencia evolutiva) en varias especies filogenéticamente no relacionadas. *Cephalotus* es una planta que tiene bien diferenciadas las hojas insectívoras (unas trampas en forma de jarra para atrapar insectos) de las hojas no insectívoras (como las del resto de plantas). La coexistencia de estos dos tipos de hojas en un mismo individuo ofrece una oportunidad única para conocer las bases genéticas de la adaptación a la dieta carnívora. Con este fin, se ha secuenciado el genoma, de 1,6 Gbp, se han identificado los aproximadamente 36.000 genes que contiene, así como el transcriptoma específico de los dos tipos de hojas. Además, el estudio se ha complementado con una comparativa del proteoma (proteínas del fluido digestivo) de *Cephalotus* y otras tres especies de plantas carnívoras con un origen independiente.

Para este estudio se han utilizado diferentes aproximaciones analíticas de la genómica comparada y evolutiva, incluyendo i) métodos filogenómicos (para determinar con precisión las relaciones evolutivas entre *Cephalotus* y otras ocho especies de angiospermas dicotiledóneas con el genoma ya secuenciado), ii) métodos para identificar y agrupar las familias de genes (identificar los ortogrupos) en las nueve especies de plantas com-

ciones de ascendencia-descendencia entre sí, lo cual puede ser visualizado en forma de lo que él denominó el árbol de la vida. La reconstrucción de las relaciones filogenéticas entre las diferentes especies es el primer paso necesario de cualquier estudio evolutivo, como el del impacto de la selección natural en la adaptación. Las especies actuales forman las puntas del árbol filogenético y en su reconstrucción se trata de establecer las denominadas relaciones de grupo hermano mediante la identificación de ancestros comunes recientes, en definitiva, de establecer clados o grupos monofiléticos que incluyen un ancestro y todos sus descendientes. Los primeros árboles filogenéticos se reconstruyeron a partir de la comparación de caracteres morfológicos. Sin embargo, a partir de los años 60 del pasado siglo se empezó a utilizar la información acumulada (desde la separación a partir de ancestros comunes) en las secuencias de ADN o de proteínas de las diferentes especies para reconstruir las relaciones filogenéticas. Para ello, se desarrollaron los métodos de máxima verosimilitud (ML) que utilizan un modelo probabilístico de evolución de la sustitución nucleotídica o aminoacídica para estimar el árbol que hace más verosímiles las secuencias observadas. Sin embargo, no fue hasta los años 90 con la implementación de la técnica de PCR que permitía amplificar y secuenciar un fragmento ortólogo (adquirido por descendencia) en diferentes especies no modelo, y la mejora en la capacidad computacional que aceleraba las búsquedas heurísticas de los valores de los parámetros del modelo que maximizan la verosimilitud del árbol, cuando los estudios filogenéticos se generalizaron, ofreciendo hipótesis robustas desde el punto de vista estadístico de las relaciones filogenéticas. A finales del siglo pasado, los métodos de inferencia filogenética se enriquecieron con el uso de la estadística bayesiana que propone el cálculo de la probabilidad posterior de un árbol; es decir, la probabilidad de que el árbol sea correcto dados unos datos y un modelo evolutivo. La probabilidad posterior de un árbol se calcula mediante el Teorema de Bayes que relaciona la probabilidad previa del árbol con la verosimilitud (probabilidad de los datos dados un árbol y modelo evolutivo).

Aunque supuso un gran avance y la resolución de muchas relaciones filogenéticas dentro del árbol de la vida, el uso de fragmentos cortos (típicamente se secuenciaban y analizaban unos 1000 pb) tiene sus limitaciones a la hora de inferir filogenias. Por un lado, pueden constituir un muestreo no representativo de la variación genómica (error muestral) y por otro pueden no acumular la variabilidad apropiada para dilucidar las relaciones filogenéticas a un nivel taxonómico concreto, dando lugar a árboles sin resolver o incluso relaciones incorrectas. Estas limitaciones han sido superadas en la actualidad por el uso generalizado de la secuenciación masiva que permiten obtener el genoma o el transcriptoma de una especie y por lo tanto construir alineamientos al menos uno o dos órdenes de magnitud más largos (10-100k pb). El problema al que se enfrenta la filogenómica es que existe una alta heterogeneidad entre diferentes regiones del genoma, afectando por ejemplo a la composición nucleotídica y las tasas evolutivas. Además, es importante tener en cuenta que la historia evolutiva de un gen o una parte del genoma (árbol de gen) no siempre se tiene que corresponder con la historia evolutiva de la especie (árbol de especies) debido a fenómenos evolutivos como la transferencia horizontal de genes (relacio- ›››

paradas, iii) métodos para determinar las familias de genes que se han expandido específicamente en el genoma de *Cephalotus* (ya que algunas de ellas pueden estar asociadas al cambio de dieta) y iv) métodos para el análisis comparativo de los genes que se están expresando diferencialmente en la hojas insectívoras.

Los resultados de la investigación han permitido identificar algunos de los cambios genéticos asociados con la emergencia de la dieta carnívora en plantas. Por ejemplo, se ha determinado que las hojas que atrapan insectos han adquirido nuevas funciones enzimáticas mediante coopción (cambio de función en genes ya existentes): la quitinasa básica, que rompe la quitina (el principal componente del exoesqueleto de los insectos), y la fosfatasa ácida púrpura, que libera los grupos fosfato de las moléculas y ayuda a movilizar el fósforo de las presas entre otras. Los resultados del estudio proteómico comparativo sugieren, además, que el proceso evolutivo se ha repetido de forma independiente en varias especies, a partir de un conjunto de genes y proteínas con función ancestral similar. Esto es, la capacidad de las plantas carnívoras para digerir animales en suelos empobrecidos ha sido el resultado de la acción de la selección natural que ha promovido varios cambios genéticos recurrentes sobre un mismo conjunto de genes de un número reducido de rutas metabólicas.

Figura 1.
Imagen de los dos tipos de hojas (en forma de jarra y planas) de la planta de jarra de Australia (o jarrito enano), *Cephalotus follicularis*. Estas hojas, en forma de jarra, están adaptadas para la captura y digestión de insectos. La planta es originaria de Australia Occidental.
Crédito: Mitsuyasu Hasebe.

›››› nes xenólogas), la duplicación de genes (relaciones parálogas) o la separación incompleta de linajes causada por polimorfismos ancestrales, anteriores al proceso de especiación (ya que esta no es instantánea), y la especiación híbrida que se produce cuando dos especies dan lugar a una tercera.

Una vez obtenidos los datos de secuencia crudos, es necesario ensamblarlos, anotarlos (si es posible) y de forma obligatoria definir si se cumplen o no las relaciones de ortología. A partir de aquí, existen tres posibilidades a la hora de analizar los datos de secuenciación masiva. La primera y más extendida en la actualidad es realizar un alineamiento múltiple de secuencias (MSA) ortólogas de cada especie de forma individual (por ejemplo con *MAFFT*, *T-COFFEE*) y posteriormente concatenarlas. Cada secuencia ortóloga constituye una partición del alineamiento concatenado (supermatriz) y se calculan los valores de los parámetros del modelo evolutivo para cada partición de forma independiente. A continuación bien por máxima verosimilitud (con *RAxML* o *IQ tree*) o mediante inferencia Bayesiana (con *MrBayes* o *PhyloBayes*) se obtiene un árbol de especies a partir de la información conjunta de todas las particiones. Esta metodología asume que los diferentes ortólogos de una especie comparten una historia común que prevalece al concatenar sobre el ruido individual (conflictivo) de cada partición. Una segunda aproximación consiste en generar con máxima verosimilitud o inferencia Bayesiana un árbol de gen para cada partición y a continuación combinarlos mediante métodos de consenso, concordancia o reconciliación (con *BUCKy* o *ASTRALII*) obtener un superárbol que representa el árbol de especies. Finalmente, y para la reconstrucción filogenética de especies cercanas, se han desarrollado métodos de coalescencia que tienen en cuenta la diferenciación entre las secuencia ancestral y las derivadas y que directamente estiman el árbol de especies a partir de cada una de las particiones teniendo en cuenta la posible señal conflictiva de cada partición (con *STARBEAST2* o *PHYLDOG*). Estos últimos métodos requieren mayor capacidad computacional y están actualmente en continua mejora. En muchos casos, los tres tipos de aproximaciones convergen en las mismas hipótesis filogenéticas, o si hay diferencias están no tienen soporte estadístico. Un caso distinto, y no resuelto, ocurre cuando las diferentes metodologías resuelven de forma diferente ciertas relaciones filogenéticas (típicamente en conflicto) y además lo hacen con máximo apoyo estadístico. Se trata precisamente de aquellas partes del árbol de la vida que siempre han sido fuente de controversia y se corresponden con nodos cortos (los diferentes linajes se separaron en relativamente poco tiempo) en los que la señal filogenética es reducida. Esta es la última frontera para conseguir cumplir la visión de Darwin de obtener un árbol de la vida completamente resuelto que sirva como marco robusto y estable para los estudios comparativos.

ANÁLISIS DE GANANCIAS Y PÉRDIDAS DE GENES

Una familia multigénica es un conjunto de genes relacionados filogenéticamente (homólogos) que derivan de una serie de eventos de duplicación (genes parálogos). Uno de los ejemplos más conocidos es el de las alfa y beta globinas. El análisis comparativo del número de miembros de una determinada familia en varias especies puede servir para inferir el papel de la selección natural y de la adaptación. Actualmente, y con la disponibilidad de la secuencia genómica completa de varias especies filogenéticamente relacionadas, se pueden realizar estos análisis a escala genómica, es decir analizando todas las familias multigénicas. Este análisis permite identificar y localizar en la filogenia todas las ganancias y pérdidas de genes, a partir de las cuales se pueden estimar las tasas de nacimiento y muerte (típicamente expresado por gen y por millón de años). Y lo que es más importante, el análisis de la función de aquellos genes que se han expandido o contraído de forma significativa en linajes particulares (dentro de un marco estadístico riguroso), puede dar pistas de qué caracteres (genes particulares y funciones concretas) han afectado al proceso adaptativo.

Para este tipo de análisis evolutivos se han desarrollado varios métodos basados en modelos estocásticos (típicamente metodologías de ML para estimar las tasas de nacimiento y muerte de genes. El flujo de trabajo típico para realizar este tipo de análisis comprende:

1. La determinación de todos los grupos de genes descendientes del mismo ancestro común en las especies estudiadas (u ortogrupos). En este paso se agrupan en un mismo ortogrupo tanto los miembros ortólogos de la familia (genes de copia única u ortólogos 1:1), como los genes originados por duplicación en un mismo linaje genético durante la diversificación de las especies estudiadas (relaciones N:N de co-ortólogos y co-parálogos). Esta agrupación se realiza en base a la similitud entre secuencias (típicamente en función de los "E-values" de *blast*), y aplicar al-

Tabla 1. Lista de programas más relevantes.

Software	Pagina Web	Descripción
ASTRALII	github.com/smirarab/ASTRAL	Estimación de árboles de especie mediante coalescencia.
BadiRate	www.ub.edu/softevol/badirate/	Estimación de las tasas de nacimiento y muerte de genes.
BUCKy	www.stat.wisc.edu/~ane/bucky/	Estimación de árboles de especie mediante concordancia Bayesiana.
CAFE	hahnlab.github.io/CAFE/	Estimación de las tasas de nacimiento y muerte de genes.
Hyphy	hyphy.org/w/index.php/Main_Page	Análisis del impacto de la selección natural en regiones codificadoras.
IQ tree	www.iqtree.org/	Reconstrucción filogenética con máxima verosimilitud.
MAFFT	mafft.cbrc.jp/alignment/server/	Alineamientos múltiples.
MrBayes	mrbayes.sourceforge.net/	Inferencia bayesiana.
Notung	www.cs.cmu.edu/~durand/Notung/	Estimación del número de ganancias y pérdidas de genes.
OrthoFinder	github.com/davidemms/OrthoFinder	Determinación de grupos de ortólogos.
OrthoMCL	www.orthomcl.org/	Determinación de grupos de ortólogos.
PAML	abacus.gene.ucl.ac.uk/software/paml.html	Análisis del impacto de la selección natural en regiones codificadoras.
PHYLDOG	pbil.univ-lyon1.fr/software/phyldog/	Estimación de árboles de especie en presencia de duplicaciones.
Phylobayes	www.phylobayes.org/	Inferencia bayesiana con modelos mixtos (CAT).
r8s	sourceforge.net/projects/r8s/	Estimación de árboles ultramétricos (reloj molecular).
RAxML	sco.h-its.org/exelixis/web/software/raxml/index.html	Reconstrucción filogenética con máxima verosimilitud.
STARBEAST2	github.com/genomescale/starbeast2	Estimación de árboles de especie mediante coalescencia.
T-COFFEE	tcoffee.crg.cat/	Alineamientos múltiples.

goritmos de ML optimizados para realizar un agrupamiento automático (por ejemplo con *OrthoMCL* y *OrthoFinder*).

2. Inferencia de un árbol filogenético ultramétrico (en el que la longitud de las ramas representa tiempo) de las especies. Para esta inferencia típicamente se utilizan todos los genes de copia única (ortólogos 1:1) de todos los genomas, y se procede tal como se indica en el apartado de filogenómica (MSA, concatenación de las secuencias generando una supermatriz, reconstrucción de un árbol sin raíz por ML). Posteriormente, el árbol sin raíz se linealiza (con *r8s*), y se calibra en base a dataciones externas de nodos particulares (por ejemplo con información del registro fósil).

3. Los pasos anteriores permiten estimar el número de ganancias y pérdidas de genes (y las tasas de nacimiento y muerte) en cada uno de los linajes de la filogenia, bien por metodologías de reconciliación entre el árbol de los genes de cada ortogrupo con el de las especies (con *Notung*), o directamente a partir del número de miembros de la familias presentes en las especies actuales mediante modelos de ML (con *BadiRate* o *CAFE*).

MODELOS DE EVOLUCIÓN DE CODONES Y ADAPTACIÓN A NIVEL PROTEICO

La variación en el nivel de constricción funcional entre genes homólogos de distintas especies es una fuente muy importante de información sobre el proceso adaptativo. Unos de los métodos más potentes y ampliamente usados consiste en comparar las tasas de sustitución sinónima (d_S, o el número de cambios sinónimos por posición sinónima) y no-sinónima (d_N, número de cambios no sinónimos por posición no-sinónima); es decir, aquellas sustituciones que no ›››

›››› provocan cambios de aminoácido en la proteína codificada y aquellas que sí, respectivamente. Si todos los cambios nucleotídicos que se produjeran en un gen fueran estrictamente neutros (sin ningún efecto sobre la eficacia biológica de los individuos), se esperaría que d_S fuera igual a d_N, y por lo tanto su ratio ($\omega = d_N/d_S$) sería igual a 1. La principal fuerza evolutiva que moldea los cambios evolutivos en la región codificadora de un gen es la selección purificadora (o negativa) que elimina las mutaciones (principalmente no-sinónimas) con efectos deletéreos de las poblaciones, por lo que el valor de d_N será menor que en el caso de neutralidad, y ω será típicamente inferior a 1 (la selección natural evitará su fijación de la población). Solo las mutaciones beneficiosas se podrán fijar por selección natural (selección positiva o darwiniana) de forma más rápida que las sinónimas neutras generando valores ω mayores que 1. Es decir, valores significativamente distintos de $\omega = 1$ ($\omega < 1$ o $\omega > 1$), indican la acción recurrente de la selección natural. En genómica de la adaptación los segundos ($\omega > 1$) son los más importantes ya que indican cambios de aminoácido que han sido promovidos por la selección natural como respuesta adaptativa a un cambio geológico, ambiental, biótico, etcétera. Este análisis, al realizarse bajo un marco estadístico riguroso, permite determinar qué genes, funciones génicas, procesos biológicos, etcétera, han evolucionado por selección positiva; es decir, han participado de forma activa en un proceso de adaptación a nivel molecular.

De los varios programas que existen para estimar ω los más populares son PAML (el más usado) y HyPhy (más versátil, pero más complejo de utilizar). Aunque existen diferencias en los modelos evolutivos que implementan estos programas, ambos permiten estimaciones por ML. Un estudio genómico típico usando modelos de codones comprende de los siguientes pasos:

1. Seleccionar el conjunto de genes (y sus MSA) a estudiar. Un análisis particular incluye el estudio de un MSA de la zona codificadora de los genes homólogos de varias especies (típicamente ortólogos 1:1, aunque también pueden incluirse parálogos). A nivel genómico se podrían estudiar los MSA de todos o un grupo de genes compartidos por todas las especies.

2. Incluir el árbol filogenético aceptado de los genes a estudiar (ver apartado de filogenómica).

3. Realizar el análisis con PAML o HyPhy. Entre otras opciones, se tiene que indicar un modelo evolutivo de codones concreto (que dependerá de nuestra hipótesis de estudio). Entre los modelos más comunes está el denominado "M0", que asume que todos los codones y todos los linajes del árbol filogenético comparten un mismo valor de ω), y el modelo "FR", que asume que cada linaje de la filogenia pueda tener su propio valor de ω (compartido por todos los codones en ese linaje). Modelos más complejos que permiten determinar estadísticamente si existe selección positiva en algún linaje, o incluso en posiciones concretas de linajes específicos son también aplicados habitualmente. ■

PARA LEER MÁS

Bleidorn C (2017). Phylogenomics. An introduction. 222 pp. Springer International Publishing: Basel, Switzerland.

Demuth JP and Hahn MW (2009). The life and death of gene families. *Bioessays* 31: 29-39.

Fukushima K, et al. (2017). The pitcher plant *Cephalotus* genome reveals genetic changes associated with carnivory. *Nature Ecol. Evol.* 1: 0059.

Vargas P, Zardoya R. (Eds) (2014). The Tree of Life. 713 pp. Sinauer Associates. Sunderland, MA, USA.

Yang Z & Bielawski JP (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*, 15: 496-503.