UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE GENÉTICA, ECOLOGIA E EVOLUÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

Alison Pelri Albuquerque Menezes

ARCADE (*ARChaeplastida Annotation DatabasE*): um banco de dados para estudos genômicos comparativos sobre a evolução de fenótipos complexos em Archaeplastida

Belo Horizonte
2022

Alison Pelri Albuquerque Menezes

ARCADE (ARChaeplastida Annotation DatabasE): um banco de dados para estudos genômicos comparativos sobre a evolução de fenótipos complexos em Archaeplastida

Tese apresentada ao Programa de Pós-Graduação em Genética do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Genética.

Orientador: Prof. Dr. Francisco Pereira Lobo
Co-orientador: Prof. Dr. Luiz-Eduardo Vieira Del-Bem

Belo Horizonte
2022

043 Menezes, Alison Pelri Albuquerque.
ARCADE (AR*Chaeplastida Annotation Databas*E) [manuscrito] : um banco de dados para estudos genômicos comparativos sobre a evolução de fenótipos complexos em Archaeplastida / Alison Pelri Albuquerque Menezes. – 2022.
280 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Francisco Pereira Lobo. Coorientador: Prof. Dr. Luiz Eduardo Vieira Del-Bem.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Genética. 2. Genômica. 3. Tamanho do genoma. I. Lobo, Francisco Pereira. II. Del-Bem, Luiz Eduardo Vieira . III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 575

Ficha catalográfica elaborada pela bibliotecária Jéssica Patrícia Silva de Sá – CRB 6 – 3430

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Programa de Pós-Graduação em Genética
Instituto de Ciências Biológicas
ATA DE DEFESA DE TESE

ATA DA DEFESA DE TESE	**163/2022**
	Entrada 1º/2018
ALISON PELRI ALBUQUERQUE MENEZES	**CPF: 095.040.916-28**

Às nove horas do dia **20 de setembro de 2022**, reuniu-se, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"ARCADE (ARChaeplastida Annotation DatabasE): um banco de dados para estudos genômicos comparativos sobre a evolução de fenótipos complexos em Archaeplastida"**, requisito para obtenção do grau de Doutor em **Genética.** Abrindo a sessão, o Presidente da Comissão, **Francisco Pereira Lobo**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	**Instituição**	**CPF**	**Indicação**
Francisco Pereira Lobo	UFMG	012.273.736-94	APROVADO
Luiz Eduardo Vieira Del Bem	UFMG	326.143.688-30	APROVADO
Laila Alves Nahum	FIOCRUZ	666.211.786-20	APROVADO
Jurandir Vieira de Magalhães	EMBRAPA	656.980.886-91	APROVADO
Wellington Ronildo Clarindo	UFV	049.437.196-00	APROVADO

Douglas Silva Domingues	USP	297. 659.208-06	APROVADO

Pelas indicações, o candidato foi considerado: APROVADO

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 20 de setembro de 2022.

Francisco Pereira Lobo - Orientador

Luiz Eduardo Vieira Del Bem - Coorientador

Laila Alves Nahum

Jurandir Vieira de Magalhães

Wellington Ronildo Clarindo

Douglas Silva Domingues

Assinatura dos membros da banca examinadora:

Documento assinado eletronicamente por **Douglas Silva Domingues**, **Usuário Externo**, em 20/09/2022, às 13:52, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Laila Alves Nahum**, **Usuário Externo**, em 20/09/2022, às 17:41, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Jurandir Vieira de Magalhaes**, **Usuário Externo**, em 21/09/2022, às 13:32, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Luiz Eduardo Vieira Del Bem**, **Professor do Magistério Superior**, em 21/09/2022, às 16:06, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Francisco Pereira Lobo**, **Professor do Magistério Superior**, em 21/09/2022, às 19:13, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Wellington Ronildo Clarindo**, **Usuário Externo**, em 22/09/2022, às 13:43, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1769664** e o código CRC **79218764**.

Referência: Processo nº 23072.255399/2022-30 SEI nº 1769664

RESUMO

A abundância de dados genômicos de plantas fruto da diminuição dos custos de sequenciamento contrasta com a falta de bancos de dados que integrem estes dados com anotação genômica, taxonomia e fenótipos para produzir conhecimento estatisticamente sólido e biologicamente relevante. Aqui apresentamos o ARCADE (*ARChaeplastida Annotation DatabasE*), um banco de dados de 171 proteomas não redundantes de Archaeplastida de alta qualidade coletados de seis fontes primárias diferentes, juntamente com métricas de qualidade de proteoma e um número crescente de metadados associados. Como estudos de caso para demonstrar a utilidade do ARCADE , investigamos três cenários evolutivos contrastantes em termos filogenéticos e fenotípicos: 1) a expansão e contração de domínios proteicos associados à evolução do tamanho do genoma (TG) em plantas terrestres; 2) a evolução da altura máxima em angiospermas; 3) e a origem e evolução da família de genes DELAY OF GERMINATION1 (DELAY OF GERMINATION1 Gene Family, DGF) em Archaeplastida. Integramos as anotações genômicas e informações filogenéticas disponíveis no ARCADE juntamente com dados fenotípicos disponíveis publicamente para investigar dois fenótipos vegetais complexos e altamente variáveis (TG e altura). TG parece estar diminuindo ao longo da evolução, exceto por alguns ramos que podem ter sofrido aumentos independentes de TG. Descobrimos que a variação de TG em plantas terrestres está relacionada principalmente ao metabolismo de nucleotídeos, reparo de DNA e organização do genoma. Também vimos que em genomas maiores há maior frequência da superfamília de histonas 2A, responsável por diversas funções, incluindo a formação de nucleossomos e silenciamento de elementos transponíveis. Nossos resultados indicam que pode haver uma associação entre a variação do tamanho do genoma em plantas terrestres e a preservação da estabilidade do genoma, sugerindo a evolução de mecanismos para que auxiliem plantas terrestres a lidarem com a variação no TG. Sobre a evolução da altura em angiospermas, destacamos a detecção de expansões independentes do sistema de autoincompatibilidade em angiospermas mais altas, mecanismo molecular que diminui a endogamia e aumenta a diversidade genética. As angiospermas mais altas possuem menores taxas evolutivas, uma vez que também possuem ciclos de vida maiores do que plantas menores, usualmente anuais. A expansão dos sistemas de auto-incompatibilidade nas angiospermas mais altas pode ser um importante fator causando um aumento da variabilidade genética nessas espécies,

contrabalanceando suas menores taxas evolutivas.A família DGF é um componente chave na regulação de muitos processos em angiospermas, como germinação e floração. No entanto, pode ser encontrado em plantas terrestres não-angiospermas. Nossa busca em 171 espécies dos principais clados de Archaeplastida detectou a presença de genes desta família em 6 espécies de Charophyta. Este resultado é evidência de uma origem mais antiga para esta família de genes do que se pensava anteriormente e contribui para a discussão da evolução dos DGFs. Em conjunto, os resultados que obtivemos nesses estudos de caso demonstram a inovacao e relevancia cientifica de ARCADE, um recurso para estudos genômicos comparativos da evolução de fenótipos complexos em plantas.

Palavras chave: base de dados; genômica comparativa; evolução; famílias genicas; tamanho do genoma.

ABSTRACT

The abundance of plant genomic data as a result of decreasing sequencing costs contrasts with the lack of databases that integrate these data with genomic annotation, taxonomy and phenotypes to produce statistically solid and biologically relevant knowledge. Here we present ARCADE (ARChaeplastida Annotation DatabaseE), a database of 171 high quality non-redundant Archaeplastida proteomes collected from six different primary sources, along with proteome quality metrics and an increasing number of associated metadata. As case studies to demonstrate the usefulness of ARCADE, we investigated three contrasting evolutionary scenarios in phylogenetic and phenotypic terms: 1) the expansion and contraction of protein domains associated with genome size (GS) evolution in land plants; 2) the evolution of maximum height in angiosperms; 3) and the origin and evolution of the DELAY OF GERMINATION1 gene family (DELAY OF GERMINATION1 Gene Family, DGF) in Archaeplastida. We integrated the genomic annotations and phylogenetic information available in ARCADE together with publicly available phenotypic data to investigate two complex and highly variable plant phenotypes (GS and height). GS appears to be decreasing throughout evolution, except for a few branches that may have undergone independent GS increases. We found that GS variation in land plants is mainly related to nucleotide metabolism, DNA repair, and genome organization. We also saw that in larger genomes there is a higher frequency of the histone 2A superfamily, responsible for several functions, including the formation of nucleosomes and silencing of transposable elements (though epigenetic modifications). Our results indicate that there may be an association between genome size variation in land plants and the preservation of genome stability, suggesting the evolution of mechanisms to help land plants deal with GS variation. Regarding the evolution of height in angiosperms, we highlight the detection of independent expansions of the self-incompatibility system in taller angiosperms, a molecular mechanism that reduces inbreeding and increases genetic diversity. Taller angiosperms have lower evolutionary rates, as they also have longer life cycles than smaller plants, usually annuals. The expansion of self-incompatibility systems in taller angiosperms may be an important factor causing an increase in genetic variability in these species, counterbalancing their lower evolutionary rates. The DGF is a key component in the regulation of many processes in angiosperms, such as germination and flowering. However, it can be found in non-angiosperm land plants. Our

search in 171 species of the main Archaeplastida clades detected the presence of genes of this family in 6 species of Charophyta. This result is evidence of an older origin for this gene family than previously thought and contributes to the discussion of the evolution of DGFs. Taken together, the results we obtained from these case studies demonstrate the innovation and scientific relevance of ARCADE, a resource for comparative genomic studies of the evolution of complex phenotypes in plants.

Key words: database; comparative genomics; evolution; gene families; genome size.

SUMÁRIO

INTRODUÇÃO

Archaeplastida é um grupo monofilético formado por Viridiplantae (plantas terrestres e algas verdes. Chlorophyta e Charophyta), Rodophyta (algas vermelhas) e Glaucophyta (Ball *et al.,* 2011; Adl *et al.,* 2012). Esse grupo originou-se a partir da relação estabelecida no evento de endossimbiose primária, quando um ancestral das cianobactérias foi internalizado por um eucarioto unicelular. Da célula formada nesse evento descendem todos os cloroplastos atuais. As plantas representam cerca de 80% da biomassa do planeta Terra e desde o seu surgimento tem sido um dos principais fatores na formação da atmosfera e solo (Niklas, Kutschera, 2010; Bar-On, Phillips, Milo 2018). Além disso, as plantas também influenciaram na origem, manutenção e evolução de comunidades bióticas, fornecendo habitat e fonte de alimento (Sørensen *et al.,* 2011). É seguro afirmarmos que sem as plantas não seria possível a existência de vida como nós a conhecemos hoje. Em contrapartida, à medida que as plantas passaram a ocupar habitats sob as mais diversas condições ambientais, uma nova série de pressões evolutivas selecionaram características vantajosas para os novos estresses bióticos e abióticos (Vitti, Grossman, Sabeti, 2013). Ao longo da evolução, forças evolutivas moldaram a grande diversidade fenotípica (característica expressa de um organismo) e genotípica (composição genética de um organismo) que observamos hoje no diverso grupo de Archaeplastida.

Por décadas, cientistas têm se interessado pelos mistérios que envolvem os mecanismos por trás da evolução e diversificação em Archaeplastida. O início dos projetos de sequenciamento que montaram os primeiros genomas de eucariotos, incluindo a planta modelo *Arabidopsis thaliana* no final da década de 1990 (Goffeau *et al.,* 1996; Initiative, 2000; IHGSC *et al.,* 2001) criou altas expectativas na comunidade científica e sociedade em geral. Esperava-se que as informações contidas na sequência completa de DNA de um organismo pudessem responder às grandes perguntas da ciência. Quais as diferenças entre os genomas de organismos? E quais as diferenças entre indivíduos de uma mesma espécie? O que faz com que diferentes organismos possuam diferentes níveis de complexidade? Como uma célula indiferenciada dá origem a dezenas de tipos celulares? Muitas perguntas foram respondidas, outras apenas parcialmente e diversas perguntas novas surgiram. Por exemplo, uma das grandes perguntas que emergem nesse momento (e ainda não completamente respondida) foi qual a função e origem das sequências não codificantes do genoma.

De fato, aprendemos muito com cada genoma completo que foi sequenciado. A partir

de um genoma completo podemos descobrir quais genes estão presentes naquela sequência, qual a porção codificante e não codificante, como os genes se organizam no genoma, vias metabólicas presentes no organismo em questão. Aprendemos muito sobre estrutura e organização de genomas com as primeiras sequências genômicas publicadas. Esses primeiros estudos contribuíram bastante para subsequentes estudos de genômica funcional, fornecendo embasamento e algo de interesse para a investigação da função e como interagem diferentes genes em um organismo. Entretanto, para responder muitas dessas perguntas precisamos estudar fenótipos complexos, características desenvolvidas a partir de uma cadeia de fatores que interagem entre si, podendo ser tanto genéticos (RNA, DNA, proteínas), quanto ambientais (bióticos e abióticos) (Kopriva, Weber, 2021). A evolução de fenótipos complexos é um dos problemas centrais na biologia evolutiva. Em plantas, por exemplo, germinação, floração e altura são alguns fenótipos complexos de grande interesse. Compreender como esses fenótipos são regulados também tem fortes impactos na produção de alimentos e insumos industriais (Schneider, Persson, 2015).

Para entender a evolução dos diferentes fenótipos é necessário conhecer suas bases genéticas e as suas formas de herança (Roff, 1997). Um fenótipo só evolui se for herdável, ou seja, puder ser traduzido a partir de um genótipo, de um ou mais genes. Esse genótipo herdável pode sofrer modificações (mutações) e estará sujeito a sofrer os efeitos de forças evolutivas. Desse modo pode contribuir para a história evolutiva da sua espécie ou população. Grande parte dos avanços nos estudos da evolução fenotípica se deram graças a estudos quantitativos focados em fenótipos de interesse ou estudos populacionais focados nas alterações das composições alélicas de populações (, Wolf, 2002). Apesar da grande contribuição dada por essas duas linhas de pesquisa, nenhuma das duas se propõe a integrar dados de características fenotípicas e genotípicos. A biologia do desenvolvimento foi um dos primeiros campos da ciência a integrar informação fenotípica e genotípica buscando entender como genes se relacionam com fatores ambientes e morfológicos para construir fenótipos ao longo do desenvolvido dos organismos (Minelli, 2018). Entretanto, são recentes os esforços de integração de informação fenotípica em dentro de estudos de genômica comparativa (Dunn, Munro, 2016). Com o crescente aumento da quantidade de genomas disponíveis para uma grande variedade de espécies (incluindo espécies não modelo), é importante o desenvolvimento e uso de métodos que integrem dados genômicos (sequências e anotações genômicas, por exemplo) com informações fenotípicas (Nagy *et al.,* 2020).

Rapidamente descobrimos que, individualmente, um genoma não responde tantas perguntas quanto um conjunto de genomas pode responder ao serem comparados utilizando as

devidas ferramentas. Ao compararmos organismos ou grupos de organismos filogeneticamente próximos, mas que diferem no estado de algum fenótipo de interesse - seja tempo de germinação, tipo de dormência da semestre ou altura máxima alcançada -, podemos identificar genes candidatos para estudos de biologia do desenvolvimento que expliquem a relação entre o genótipo de um organismo e o fenótipo em questão. Além disso, comparando genomas sabemos cada vez mais sobre a evolução dos genomas e espécies. Sabemos, por exemplo, que o grau de complexidade biológica (comumente associado ao número de tipos celulares diferentes em um organismo) de um organismo não está diretamente relacionado com o tamanho do genoma ou a quantidade de genes codificantes de proteínas (Greilhuber, Leitch, 2013).

Conhecemos a quantidade de DNA armazenado nas células de diversos organismos muito antes do sequenciamento de DNA (Leitch, 2005). A partir dos primeiros genomas completos publicados notamos que a variação no tamanho do genoma de eucariotos se devia principalmente à região não codificante, a qual foi chamada por muito tempo de "DNA lixo". Com o avanço dos estudos de genômica funcional, entretanto, o que antes era "DNA lixo "passou a ter cada vez mais funções e importância. A porção não codificante do genoma é constituída por sequências reguladoras da expressão gênica (como sequências promotoras e acentuadoras - *enhancers* -, por exemplo), íntrons, regiões intergênicas, entre outros elementos genômicos diretamente associados à porção codificante. Além disso, uma porção altamente variável, porém significativa do genoma dos eucariotos, é formado por regiões repetitivas, como os DNA satélites (repetições de DNA em cadeia) e os elementos transponíveis. No caso das plantas, os elementos transponíveis são os principais responsáveis pela variação no tamanho do genoma (Leitch, 2005). Os DNA satélites aumentam sua abundância no genoma geralmente por erros no processo de replicação do DNA. Enquanto isso, os elementos transponíveis são capazes de se amplificarem e deslocarem no genoma independentemente da replicação do DNA. Desse modo, elementos transponíveis conseguem aumentar rapidamente seu número no genoma, caso não sejam silenciados por mutações, mecanismos transcricionais, pós-transcricionais ou epigenéticos especializados (como modificação de histonas para alterações do estado da cromatina) (Granzotto, Cruz, 2015).

A partir de estudos genômicos e funcionais compreendemos melhor a origem, evolução e diversidade das regiões repetitivas na porção não codificante dos genomas (Lisch, 2013). Esses e muitos outros avanços receberam contribuições da genômica comparativa, que faz uso da abundância de dados genômicos de alta qualidade para traduzir as informações contidas em genomas individuais em conhecimento aplicável a uma faixa maior de

organismos. Entretanto, ao buscarmos por correlações entre organismos filogeneticamente relacionados ferimos um importante pressuposto do círculo de correlações, a independência dos dados (Harvey *et al.,* 1991). Uma vez que todas as espécies possuem uma história evolutiva comum seus dados não são independentes. Espécies mais próximas filogeneticamente são mais semelhantes do que aquelas distantes entre si. Por isso, métodos de genômica comparativa levam em consideração as relações filogenéticas dos organismos estudados para mitigar o viés filogenético dos dados e trazer maior confiabilidade aos estudos evolutivos. A maioria das ferramentas atuais que se preocupam com a correção desse viés implementam o método de contrastes filogeneticamente independentes, que identifica quais comparações entre espécies são estatisticamente independentes (Felsenstein, 1985). A genômica comparativa é um tópico que vem crescendo rapidamente dentro da biologia evolutiva. Esse crescimento só tem sido possível devido à grande quantidade de dados genômicos que vêm sendo produzidos e disponibilizados nas últimas décadas. O surgimento de técnicas de sequenciamento massivo de DNA tem tornado o sequenciamento e montagem de novos genomas cada vez mais acessíveis. Desde década de 1960 as técnicas de sequenciamento de trechos de DNA têm sido aplicadas em estudos genéticos (Heather, Chain, 2016). Sanger e Coulson (1975) desenvolveram um dos mais populares métodos de sequenciamento de primeira geração, baseado em eletroforese capilar. Esse método permitiu em estudos genômicos, inclusive o sequenciamento do genoma humano (IHGSC *et al.,* 2001). Desde então, o uso de sequenciamento de DNA aumentou exponencialmente, pressionando o surgimento de tecnologias mais rápidas e menos dispendiosas. Para atender a essa demanda surgiram os métodos de sequenciamento da segunda geração (*High Throughput Sequencing* – HTS). O primeiro método a surgir foi a técnica de sequenciamento por síntese, na qual, o DNA da amostra fornecida é lido a medida que cada nova base nitrogenada adicionada durante a síntese de novas moléculas de DNA. A cada nova base adicionada é liberado um sinal que pode ser captado e interpretado pelo sequenciador (Schuster, 2008). Atualmente existem diversas plataformas de HTS com técnicas diferentes, mas a maioria delas mantém o princípio básico que foi o grande avanço desse novo método em relação ao de Sanger e Coulson (1975), a utilização de sequenciamento por síntese a partir de clonagem in vitro.

O sequenciamento por síntese de moléculas aderidas a um suporte sólido, permite sequenciar um número maior de moléculas ao mesmo tempo, reduzindo tempo e custos com sequenciamento, tornando-o muito mais acessível (Heather, Chain, 2016). Desde então houve, não só um aumento na quantidade de projetos de sequenciamento de montagem de novos genomas, mas também uma maior diversificação taxonômica das espécies alvos. Apesar de

ainda termos um viés evidente a favor de espécies modelo e de interesse agrícola e biotecnológico, temos cada vez mais linhagens sendo representadas nesse conjunto de dados (Nishiyama *et al.,* 2021; Zhang *et al.,* 2020; Li *et al.,* 2020; Nagy *et al.,* 2020). Essa extensa quantidade de dados genômicos disponíveis alimenta estudos de genômica comparativa evolutiva que são capazes de extrair informação de dados genômicos e produzir conhecimento evolutivo a partir da grande quantidade de dados (Nagy *et al.,* 2020). Desse modo, o acesso a técnicas e ferramentas de sequenciamento tem sido substituído, enquanto fator limitante em estudos genômicos, pelo acesso e domínio do uso de ferramentas de bioinformática para montagem, manipulação e análise desses genomas. Sendo assim, é importante chamar atenção para a relevância de desenvolver bancos de dados e ferramentas capazes de integrar e analisar dados genômicos em conjunto com dados fenotípicos e filogenéticos. No atual cenário, extrair conhecimento dos dados já produzidos é um caminho importante para o avanço do conhecimento na área (Nagy *et al.,* 2020).

Apesar da grande quantidade de genomas disponíveis, ainda existem poucas bases de dados que centralizem e organizem a informação genômica, taxonômica e de anotação. O maior deles constituídos pelas bases de dados do *National Center for Biotechnology Information* (NCBI). Outros bancos de dados foram criados na tentativa de integrar genomas sequenciados de especies de plantas gerados por diferentes fontes, como *Phytozome* (Goodstein *et al.,* 2011), PLAZA (Proost *et al.,* 2014) e *Fernbase* (F.-W. Li *et al.,* 2018). Entretanto, existem ainda menos iniciativas que organizam e disponibilizam informações refinadas extraídas ou associadas a esses genomas, como anotações genômicas ou informação filogenética, funcional e fenotípica. As bases de dados mais consolidadas contendo informações mais complexas e refinadas sobre genomas geralmente são dedicadas a espécies ou linhagens específicas, como *Arabidopsis thaliana* (Berardini *et al.,* 2015), *Oryza sativa* (Kawahara *et al.,* 2013) e *Brassica* (Chen *et al.,* 2021).

A existência de bancos de dados que integrem genomas, suas anotações genômicas e metadados e informação fenotípica é fundamental na extração novos conhecimentos biológicos relevantes a partir de genomas montados e anotados. A fim de preencher essa lacuna, nós produzimos ARCADE (*ARChaeplastida Annotation DatabasE*), um banco de dados que integra informação filogenética e anotações genômicas de alta qualidade em Archaeplastida, priorizando a representatividade taxonômica de linhagens não-modelo. Para tanto, realizamos uma investigação exaustiva nas bases de dados de genomas disponíveis publicamente e em seguida buscamos metadados associados tanto aos genomas diretamente, quanto à espécie. Esse conjunto de dados de alta qualidade e filogeneticamente diverso foi

utilizado para investigar a evolução de duas características complexas, a altura máxima de plantas com flores (angiospermas) e o tamanho do genoma nuclear em plantas terrestres, bem como a evolução do domínio Delay of Germination1 (DOG1) em Archaeplastida.

1 DISCUSSÃO GERAL

Existe uma quantidade muito grande de sequências de DNA disponíveis publicamente, porém há poucas e recentes iniciativas que se dedicam a organizar, refinar e integrar os metadados de toda essa informação (GenomeHubs, 2022; Ma *et al.,* 2022). Pensando nessa demanda, nós criamos visando preencher essa lacuna na organização do conhecimento científico nós criamos ARCADE, um recurso que integra proteomas preditos não-redundantes de alta qualidade, seus respectivos metadados e anotações genômicas, bem como dados fenotípicos das espécies estudadas. As bases de dados genômicos existentes para plantas até o momento focam em categorias específicas de elementos anotadores, em linhagens específicas ou possuem um conjunto de espécies filogeneticamente limitado e focado em espécies modelo ou de interesse agrícola (Berardini *et al.,* 2015; Chen *et al.,* 2021; Xue *et al.,* 2021; Ma *et al.,* 2022). Nossa base de dados oferece uma seleção mais ampla e filogeneticamente diversa de 171 proteomas preditos não-redundantes de alta qualidade do que outras plataformas similares.

Qualidade e diversidade filogenética são essenciais para a realização de bons estudos de genômica comparativa. Visando isso, integramos os proteomas disponíveis em seis bancos de dados diferentes (NCBI, PLAZA, CNGB, FernBase, Phytozome e DRYAD). Grande parte dessas bases de dados possuem uma grande proporção de seu conjunto de dados dedicada a espécies modelo e de interesse econômico, o que favorece espécies representantes das angiospermas. Entretanto, como citado anteriormente, a diversidade filogenética é essencial dentro de estudos de genômica comparativa. Quanto mais diversos e abrangente for o conjunto de dados maior será o poder estatístico de uma análise (Zwickl, Hillis, 2002; Plazzi, Ferrucci, Passamonti, 2010). Além da maior representatividade de angiospermas, a maior proporção de espécies modelo e cultivadas estudadas também cria um viés nas sequências. Cada gene codificante de proteína pode possuir uma ou mais sequências alternativas de proteínas produzidas a partir do mesmo gene. Espécies mais estudadas possuem mais isoformas conhecidas para cada gene, devido à maior abundância de estudos experimentais. Esse viés aumenta o número de sequências em um determinado proteoma para esses genes, que aparecerão numa análise como uma duplicação ou expansão, mas não por ação da evolução e sim como fruto de um viés de estudo. Nosso controle de qualidade levou em consideração e corrigiu esses vieses para que mantivéssemos apenas uma sequência proteica representativa de cada locus genômico codificante de proteína.

Embora seja possível realizar análises comparativas a nível de gene, investigando presença e ausência, duplicação ou deleção de sequências, é interessante adicionar uma camada de informação funcional a esses dados. Para adicionar informações biologicamente significativas aos genes nos disponibilizamos também dados de anotação genômica de novo via InterProScan (Jones *et al.,* 2014) para cada um dos proteomas presentes no nosso conjunto de dados. Esse tipo de análise é computacionalmente custosa, requer grande poder e tempo de processamento e os resultados podem agora ser usados em futuras análises comparativas a nível genômico, como fizemos investigando a evolução do tamanho do genoma e altura máxima; ou utilizando anotações específicas individualmente, como fizemos estudando a evolução de um domínio específico, o Delay of Germination 1 (DOG1). Juntamente com os proteomas preditos com seus respectivos dados de anotação, disponibilizamos também informações taxonômicas das espécies, o que pode guiar a amostragem de espécies da nossa base de dados em futuros estudos. Também fornecemos metadados sobre a qualidade, proveniência e protocolos de processamento dos proteomas. Incluímos também neste conjunto de dados uma tabela rica e crescente de dados fenotípicos associados a cada espécie. Esperamos que nosso banco de dados seja amplamente usado em estudos de associação entre fenótipos complexos e as mais diversas anotações genômicas que estamos fornecendo.

Com os dados organizados dentro desse banco de dados nós pudemos investigar a evolução de duas características complexas dentro do reino das plantas, a altura máxima das angiospermas e o tamanho do genoma das plantas terrestres. Nosso estudo sobre a evolução do tamanho do genoma indica que, de modo geral o tamanho do genoma nas plantas terrestres esta diminuindo, exceto por alguns clados cujo tamanho do genoma tem aumentado independente. Observamos que, à medida que o tamanho do genoma de plantas terrestres aumenta, também aumenta a frequência de domínios proteicos dedicados ao metabolismo de nucleotídeos, reparo de DNA e organização e manutenção do DNA.

Encontramos em genomas maiores uma maior frequência da superfamília de histonas 2A, responsável por diversas funções, incluindo a formação de nucleossomos e silenciamento de elementos transponíveis. Possivelmente uma resposta à presença e movimentação de elementos transponíveis, principais responsáveis pela grande variação de tamanho de genoma em eucariotos (Leitch, 2005). Essas funções moleculares que encontramos correlacionadas com a variação do tamanho do genoma podem estar associadas à preservação da estabilidade do genoma e podem indicar a evolução dos mecanismos para as plantas terrestres lidarem com a variação do GS.

Além disso, a partir dos resultados do nosso estudo sobre a evolução da altura em

angiospermas, nos propusemos que plantas mais altas investem mais recursos em reprodução cruzada. Nossos resultados indicam uma expansão de genes do locus S, responsável por mecanismos de autoincompatibilidade em plantas (J. Nasrallah, M. Nasrallah, 2014). Até onde sabemos, nenhum outro estudo de genômica comparativa sobre a altura de plantas demonstrou essa associação, reforçando a importância de uma base de dados abrangentes e integrativos, como esse que criamos. Esses dois estudos demonstram que nossa base de dados pode ser facilmente usada em analises genômicas comparativas, auxiliando na geração de hipóteses a serem testadas experimentalmente.

Além de características complexas, os proteomas preditos não redundantes gerados por esse trabalho também nos permitiram investigar a evolução da família de domínios proteicos DOG1, um componente chave para diversas redes transcricionais envolvidas com diversos fenótipos de interesse evolutivo e biotecnológico, como germinação e floração (Nishimura *et al.,* 2018; Sall *et al.,* 2019). Usando sequências do domínio DOG1 recuperadas do nosso banco de dados por um metodo baseado em Modelo Oculto de Markov (Hidden Markov Models - HMM) (Eddy, 2011), mais sensível para a busca por homólogos, nós pudemos reconstruir uma filogenia para as duas superfamílias de domínios proteicos encontradas em Viridiplantae. Nossos resultados confirmam a expansão observada por Nishiyama *et al.,* 2021) em angiospermas, bem como a monofilia e a divisão da superfamília de genes DOG1 (DOG1 gene family - DGF) em 4 famílias menores.

A partir dos proteínas contendo DOG1 recuperadas pela nossa busca usando HMM pudemos detectar pela primeira vez sequências contendo o domínio proteico DOG1 em 2 especies de algas Trebouxiophyceae (Chlorophyta) e nas 6 algas Charophyta estudadas: *C. braunii*, *C. irregularis*, *K. nitens*, *S. omearii*, *C. cushleckae* e *M. caldariorum*. Investigamos 4 classes de Chlorphytta (Chlorophyceae, Chloropicophyceae, Mamiellophyceae e Trebouxiophyceae) buscando por proteínas com o domínio DOG1 e encontramos duas sequencias da superfamília de fatores de transcrição que se ligam ao motivo TGACG (TGACG motif-binding transcription factor proteins - TGA) apenas em espécies de Trebouxiophyceae. Esse resultado, junto com a nossa análise filogenética (Anexo C, Fig. 1B) sugere que a origem desse domínio em Chlorophyta ocorreu por dois eventos independentes de transferência horizontal gênica. Evidentemente, mais estudos incluindo novos genomas desses grupos são necessários para determinar o cenário mais parcimonioso que explique a presença de genes TGA em Trebouxyophyceae. Além disso, encontramos uma cópia da família DGF na carófita *K. nitens*, o que ainda não havia sido reportado e demonstra uma origem mais ainda das DGF do que ja havia sido observado por outros trabalhos. Além disso

recuperamos também uma nova linhagem de homólogos das proteínas DGF, as proteínas INAPERTURATE POLLEN1 e 2 (INP1 e INP2). As proteínas INP foram isoladas de *Arabidopsis thaliana* e são importantes atores no desenvolvimento do pólen (Dobritsa, Coerper, 2012; Lee *et al.,* 2021). Esses resultados contribuem para a discussão sobre a origem e evolução das superfamílias DGF e TGA (Nishiyama *et al.,* 2021).

Até o momento não temos conhecimento de nenhuma outra base de dados que entregue os mesmos recursos disponibilizados pela nossa. PlantGDB, Ensembl Plants e Phytozome, por exemplo, são bases de dados que disponibilizam dados genômicos de plantas e estão consolidadas dentro da comunidade cientifica (Dong, Schlueter, Brendel, 2004; Goodstein *et al.,* 2011; Cunningham *et al.,* 2021). Entretanto, todas essas bases de dados focam seus esforços em agrupar dados sobre espécies modelos e cultivadas. Ao analisar essas bases de dados notamos uma alta representatividade de angiospermas, principalmente monocots e eudicots. Baseado nessa observação, dedicamos nossos esforços em reunir, a partir de outros bancos de dados, genomas de espécies de clados sub-representados. Desse modo, nós reunimos em um só local, genomas filogeneticamente diversos e de alta qualidade disponíveis para a utilização em análises de genômica comparativas. Nosso banco de dados integra genomas, anotações genômicas e informações filogenéticas e fenotípicas para 171 especies de Archaeplastida. Esse é um recurso útil e capaz de contribuir solidamente com a comunidade científica. No momento, estamos desenvolvendo uma ferramenta web que facilite o acesso e mineração dos dados disponibilizados por este trabalho.

2 CONCLUSÃO GERAL

Discutimos ao longo deste trabalho a relevância de estudos sobre a evolução de fenótipos complexos e suas bases genéticas. Fenótipos complexos são determinados por uma serie de fatores ambientes e genéticos que interagem entre si. Por isso, é necessário o desenvolvimento de métodos capazes de integrar dados de diferentes naturezas e origens em uma mesma análise, sobretudo, a integração de dados fenotípicos e genotípicos com as informações filogenéticas disponíveis. A partir de análises comparativas é possível gerar hipóteses sobre um fenótipo de interesse, indicando potenciais candidatos para estudos experimentais mais minuciosos sobre o fenótipo em questão. Entretanto, o desenvolvimento de estudos de genômica comparativa necessita, em primeiro lugar, da disponibilidade de dados genômicos de alta qualidade.

Nosso conjunto de dados contribui para o avanço da ciência nesse campo, fornecendo um recurso novo em relação ao que já foi publicado. Estamos disponibilizando um conjunto de proteomas preditos não-redundantes de alta qualidade, junto com metadados taxonômicos, fenotípicos e genômicos para 171 espécies filogeneticamente diversas, representando todos os principais clados de Archaeplastida.

A partir do conjunto de dados produzidos por esse trabalho, nossos estudos foram capazes de investigar a evolução de uma família de domínios proteicos e duas características complexas de grande importância evolutiva e biotecnológica. Obtivemos resultados inéditos e biologicamente relevantes, atestando a utilidade e relevância do recurso que estamos disponibilizando. Esses estudos foram realizados com base nos proteomas produzidos por esse trabalho e só foram possíveis de serem realizados dada a existência e tais dados de forma estruturada e organizada. Desse modo, nós esperamos que o recurso que nós produzimos e estamos disponibilizando para comunidade científica possa continuar contribuindo para o avanço de estudos de genômica comparativa evolutiva de espécies de plantas.

REFERÊNCIAS BIBLIOGRÁFICAS

ADAMS, Dean C.; COLLYER, Michael L. Multivariate phylogenetic comparative: evaluations, comparisons, and recommendations. Systematic Biology, v. 67, n. 1, p. 14-31, 2017. ISSN 1063-5157. DOI: 10.1093/sysbio/syx055.

ADL, Sina M. et al. The revised classification of eukaryotes. *Journal of Eukaryotic Microbiolog*y, v. 59, n. 5, p. 429-514, 2012. DOI: 10.1111/j.1550-7408.2012.00644.x.

ARABIDOPSIS GENOME INITIATIVE. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, v. 408, n. 6814, p. 796-815, 2000. ISSN 1476-4687. DOI: 10.1038/35048692.

ARNDT, David et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research*, v. 44, n. W1, p. W16-W21, 2016. ISSN 0305-1048. DOI: 10.1093/nar/gkw387.

BALL, Steven et al. The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis. *Journal of Experimental Botany*, v. 62, n. 6, p. 1775-1801, 2011. ISSN 0022-0957. DOI: 10.1093/jxb/erq411.

BAR-ON, Yinon M.; PHILLIPS, Rob; MILO, Ron. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*, v. 115, n. 25, p. 6506-6511, 2018. DOI: 10.1073/pnas.1711842115.

BARR, Jeremy J. et al. Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proceedings of the National Academy of Sciences*, v. 110, n. 26, p. 10771-10776, 2013. DOI: 10.1073/pnas.1305923110.

BEAULIEU, Jeremy M. et al. Correlated evolution of genome size and seed mass. *New Phytologist*, v. 173, n. 2, p. 422-437, 2007. DOI: 10.1111/j.1469-8137.2006.01919.x.

BENNETT, Michael D. Variation in genomic form in plants and its ecological implications. *New Phytologist*, v. 106, n. s1, p. 177-200, 1987. DOI: 10.1111/j.1469-8137.1987.tb04689.x.

BENTSINK, Leónie et al. Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis. *Proceedings of the National Academy of Sciences*, v. 103, n. 45, p. 17042-17047, 2006. DOI: 10.1073/pnas.0607877103.

CARLSON, Marc. GO.db: A set of annotation maps describing the entire Gene Ontology. 2019. R package version 3.8.2.

CROOKS, Gavin E. et al. WebLogo: a sequence logo generator. *Genome Research*, v. 14, n. 6, p. 1188-1190, 2004. DOI: 10.1101/gr.849004.

CUNNINGHAM, Fiona et al. Ensembl 2022. *Nucleic Acids Research*, v. 50, n. D1, p. D988-D995, 2021. DOI: 10.1093/nar/gkab1049.

DEDRICK, Rebekah M. et al. Prophage-mediated defence against viral attack and viral counter-defence. *Nature Microbiology*, v. 2, n. 3, p. 16251, 2017. DOI: 10.1038/nmicrobiol.2016.251.

DEKKERS, Bas J. W. et al. The Arabidopsis DELAY OF GERMINATION 1 gene affects ABSCISIC ACID INSENSITIVE 5 (ABI5) expression and genetically interacts with ABI3 during Arabidopsis seed development. *The Plant Journal*, v. 85, n. 4, p. 451-465, 2016. DOI: 10.1111/tpj.13118.

DOBRITSA, Anna A.; COERPER, Daniel. The novel plant protein INAPERTURATE POLLEN1 marks distinct cellular domains and controls formation of apertures in the Arabidopsis pollen exine. *The Plant Cell*, v. 24, n. 11, p. 4452-4464, 2012. DOI: 10.1105/tpc.112.101220.

DONG, Qunfeng; SCHLUETER, Shannon D.; BRENDEL, Volker. PlantGDB, plant genome database and analysis tools. *Nucleic Acids Research*, v. 32, supl. 1, p. D354-D359, 2004. DOI: 10.1093/nar/gkh046.

DUBOIS, Emeline et al. Homologous recombination is stimulated by a decrease in dUTPase in Arabidopsis. *PLOS ONE*, v. 6, n. 4, p. 1-8, 2011. DOI: 10.1371/journal.pone.0018658.

DUFAYARD, Jean-François et al. New insights on leucine-rich repeats receptor-like kinase orthologous relationships in angiosperms. *Frontiers in Plant Science*, v. 8, p. 381, 2017. DOI: 10.3389/fpls.2017.00381.

DUNN, Casey W.; MUNRO, Catriona. Comparative genomics and the diversity of life. *Zoologica Scripta*, v. 45, supl. 1, p. 5-13, 2016. DOI: 10.1111/zsc.12211.

DURAND, Eléonore et al. Evolution of self-incompatibility in the Brassicaceae: lessons from a textbook example of natural selection. *Evolutionary Applications*, v. 13, n. 6, p. 1279-1297, 2020. DOI: 10.1111/eva.12933.

EDDY, Sean R. Accelerated profile HMM searches. *PLOS Computational Biology*, v. 7, n. 10, p. 1-16, 2011. DOI: 10.1371/journal.pcbi.1002195.

EDGAR, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, v. 32, n. 5, p. 1792-1797, 2004. DOI: 10.1093/nar/gkh340.

EHRBAR, Kristin; HARDT, Wolf-Dietrich. Bacteriophage-encoded type III effectors in *Salmonella enterica* subspecies 1 serovar Typhimurium. *Infection, Genetics and Evolution*, v. 5, n. 1, p. 1-9, 2005. DOI: 10.1016/j.meegid.2004.07.004.

EKSTROM, Alexander et al. PlantCAZyme: a database for plant carbohydrate-active enzymes. *Database*, v. 2014, bau079, 2014. DOI: 10.1093/database/bau079.

EL-GEBALI, Sara et al. The Pfam protein families database in 2019. *Nucleic Acids Research*, v. 47, n. D1, p. D427-D432, 2018. ISSN 0305-1048. DOI: 10.1093/nar/gky995.

FALSTER, Daniel S.; WESTOBY, Mark. Plant height and evolutionary games. *Trends in Ecology & Evolution*, v. 18, n. 7, p. 337-343, 2003. DOI: 10.1016/S0169-5347(03)00061-2.

FEDAK, Halina et al. Control of seed dormancy in *Arabidopsis* by a cis-acting noncoding antisense transcript. *Proceedings of the National Academy of Sciences*, v. 113, n. 48, p. E7846-E7855, 2016. DOI: 10.1073/pnas.1608827113.

FELSENSTEIN, Joseph. Phylogenies and the comparative method. *The American Naturalist*, v. 125, n. 1, p. 1-15, 1985.

FERNÁNDEZ, Lucía; RODRÍGUEZ, Ana; GARCÍA, Pilar. Phage or foe: an insight into the impact of viral predation on microbial communities. *The ISME Journal*, v. 12, n. 5, p. 1171-1179, 2018. DOI: 10.1038/s41396-018-0049-5.

FISCHER, Steve et al. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Current Protocols in Bioinformatics*, v. 35, n. 1, p. 6.12.1-6.12.19, 2011. DOI: 10.1002/0471250953.bi0612s35.

FU, Limin et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, v. 28, n. 23, p. 3150-3152, 2012. ISSN 1367-4803. DOI: 10.1093/bioinformatics/bts565.

GALILI, Tal et al. heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics*, 2017. DOI: 10.1093/bioinformatics/btx657.

GALILI, Tal. dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*, 2015. DOI: 10.1093/bioinformatics/btv428.

GENOMEHUBS 2.0. GoaT - Genomes on a Tree. 2022. Disponível em: http://genomehubs.org. Acesso em: 19 ago. 2022.

GOFFEAU, A. et al. Life with 6000 Genes. *Science*, v. 274, n. 5287, p. 546-567, 1996. DOI: 10.1126/science.274.5287.546.

GONZÁLEZ-MORALES, Sandra Isabel et al. Regulatory network analysis reveals novel regulators of seed desiccation tolerance in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, v. 113, n. 35, p. E5232-E5241, 2016. DOI: 10.1073/pnas.1610985113.

GOODSTEIN, David M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, v. 40, n. D1, p. D1178-D1186, 2011. ISSN 0305-1048. DOI: 10.1093/nar/gkr944.

GORDILLO ALTAMIRANO, Fernando et al. Bacteriophage-resistant Acinetobacter baumannii are resensitized to antimicrobials. *Nature Microbiology*, v. 6, n. 2, p. 157-161, 2021. ISSN 2058-5276. DOI: 10.1038/s41564-020-00830-7.

GRANZOTTO, Adriana; CRUZ, Guilherme Marcello Queiroga. Regulação de Elementos de Transposição: Mecanismos Epigenéticos de Silenciamento, Autorregulação e Ativação por Estresse. *In:* CARARETO, Claudia Marcia Aparecida; VITORELLO, Claudia Barros Monteiro; VAN SLUYS, Marie-Anne (Ed.). Elementos de transposição: diversidade, evolução, aplicações e impacto nos genomas dos seres vivos. São José do Rio Preto: Editora FIOCRUZ, 2015. p. 91-113. ISBN 978-85-7541-462-0. DOI: 10.7476/9788575415672.

GREILHUBER, Johann; LEITCH, Ilia J. Genome Size and the Phenotype. *In:* GREILHUBER, Johann; DOLEZEL, Jaroslav; WENDEL, Jonathan F. (Ed.). Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes. Vienna: Springer, 2013. p. 323-344. ISBN 978-3-7091-1160-4. DOI: 10.1007/978-3-7091-1160-4_20.

GROTH, Philip et al. PhenomicDB: a new cross-species genotype/phenotype resource. *Nucleic Acids Research*, v. 35, n. supl. 1, p. D696-D699, 2006. ISSN 0305-1048. DOI: 10.1093/nar/gkl662.

HARVEY, Paul H.; PAGEL, Mark D. The comparative method in evolutionary biology. Oxford: Oxford University Press, 1991. v. 239.

HAYNES, Winston A.; TOMCZAK, Aurelie; KHATRI, Purvesh. Gene annotation bias impedes biomedical research. *Scientific Reports*, v. 8, n. 1, p. 1362, 2018. ISSN 2045-2322. DOI: 10.1038/s41598-018-19333-x.

HEATHER, James M.; CHAIN, Benjamin. The sequence of sequencers: The history of

sequencing DNA. *Genomics*, v. 107, n. 1, p. 1-8, 2016. ISSN 0888-7543. DOI: 10.1016/j.ygeno.2015.11.003.

HIDALGO, Oriane et al. Is There an Upper Limit to Genome Size? *Trends in Plant Science*, v. 22, n. 7, p. 567-573, 2017. ISSN 1360-1385. DOI: 10.1016/j.tplants.2017.04.005.

HONGO, Jorge Augusto et al. CALANGO: an annotation-based, phylogeny-aware comparative genomics framework for exploring and interpreting complex genotypes and phenotypes. *bioRxiv*, 2021. DOI: 10.1101/2021.08.25.457574.

HUNG, Jui-Hung et al. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, v. 13, n. 3, p. 281-291, 2011. ISSN 1467-5463. DOI: 10.1093/bib/bbr049.

HUO, Heqiang; WEI, Shouhui; BRADFORD, Kent J. DELAY OF GERMINATION1 (DOG1) regulates both seed dormancy and flowering time through microRNA pathways. *Proceedings of the National Academy of Sciences*, v. 113, n. 15, p. E2199-E2206, 2016. DOI: 10.1073/pnas.1600558113.

IHGSC. Initial sequencing and analysis of the human genome. *Nature*, v. 409, n. 6822, p. 860-921, 2001. ISSN 1476-4687. DOI: 10.1038/35057062.

JONES, Philip et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, v. 30, n. 9, p. 1236-1240, 2014. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btu031.

KANG, Ming et al. Adaptive and nonadaptive genome size evolution in Karst endemic flora of China. *New Phytologist*, v. 202, n. 4, p. 1371-1381, 2014. DOI: 10.1111/nph.12726.

KATTGE, Jens et al. TRY plant trait database – enhanced coverage and open access. *Global Change Biology*, v. 26, n. 1, p. 119-188, 2020. DOI: 10.1111/gcb.14904.

KAWAHARA, Yoshihiro et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, v. 6, n. 1, p. 4, 2013. ISSN 1939-8433. DOI: 10.1186/1939-8433-6-4.

KAWASHIMA, Tomokazu et al. Diversification of histone H2A variants during plant evolution. *Trends in Plant Science*, v. 20, n. 7, p. 419-425, 2015. ISSN 1360-1385. DOI: 10.1016/j.tplants.2015.04.005.

KNIGHT, Charles A.; MOLINARI, Nicole A.; PETROV, Dmitri A. The Large Genome Constraint Hypothesis: Evolution, Ecology and Phenotype. *Annals of Botany*, v. 95, n. 1, p. 177-190, 2005. ISSN 0305-7364. DOI: 10.1093/aob/mci011.

KOORNEEF, Maarten; BENTSINK, Leónie; HILHORST, Henk. Seed dormancy and germination. *Current Opinion in Plant Biology*, v. 5, n. 1, p. 33-36, 2002. ISSN 1369-5266. DOI: 10.1016/S1369-5266(01)00219-9.

KOPRIVA, Stanislav; WEBER, Andreas P. M. Genetic encoding of complex traits. *Journal of Experimental Botny*, v. 72, n. 1, p. 1-3, 2021. ISSN 0022-0957. DOI: 10.1093/jxb/eraa498.

KRISHNAKUMAR, Vivek et al. Araport: the *Arabidopsis* Information Portal. *Nucleic Acids Research*, v. 43, n. D1, p. D1003-D1009, 2014. ISSN 0305-1048. DOI: 10.1093/nar/gku1200.

KUANG, Kevin; KONG, Quyu; NAPOLITANO, Francesco. pbmcapply: Tracking the Progress of Mc*pply with Progress Bar. 2022. R package version 1.5.1.

KUMAR, Sudhir et al. MEGA X: Molecular Evolutionary Genetics Analysis across Computing

Platforms. *Molecular Biology and Evolution*, v. 35, n. 6, p. 1547-1549, 2018. ISSN 0737-4038. DOI: 10.1093/molbev/msy096.

KUMAR, Sudhir et al. TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, v. 39, n. 8, msac174, 2022. ISSN 1537-1719. DOI: 10.1093/molbev/msac174.

KUMAR, Vikash et al. Genome-Wide Identification of Populus Malectin/Malectin-Like Domain-Containing Proteins and Expression Analyses Reveal Novel Candidates for Signaling and Regulation of Wood Development. *Frontiers in Plant Science*, v. 11, p. 588846, 2020. DOI: 10.3389/fpls.2020.588846.

KUMAR, Vikash et al. Poplar carbohydrate-active enzymes: whole-genome annotation and functional analyses based on RNA expression data. *The Plant Journal*, v. 99, n. 4, p. 589-609, 2019. DOI: 10.1111/tpj.14417.

LANFEAR, Robert et al. Taller plants have lower rates of molecular evolution. N*ature Communications*, v. 4, n. 1, p. 1879, 2013. ISSN 2041-1723. DOI: 10.1038/ncomms2836.

LEE, Byung Ha et al. A species-specific functional module controls formation of pollen apertures. *Nature Plants*, v. 7, n. 7, p. 966-978, 2021. ISSN 2055-0278. DOI: 10.1038/s41477-021-00951-9.

LEE, Heewook et al. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences*, v. 109, n. 41, p. E2774-E2783, 2012. DOI: 10.1073/pnas.1210309109.

LEI, Bingkun; BERGER, Frédéric. H2A Variants in Arabidopsis: Versatile Regulators of Genome Activity. *Plant Communications*, v. 1, n. 1, p. 100015, 2020. ISSN 2590-3462. DOI: 10.1016/j.xplc.2019.100015.

LEITCH, Andrew R.; LEITCH, Ilia J. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytologist*, v. 194, n. 3, p. 629-646, 2012. DOI: 10.1111/j.1469-8137.2012.04105.x.

LEITCH, Ilia J. et al. Evolution of DNA Amounts Across Land Plants (Embryophyta). *Annals of Botany*, v. 95, n. 1, p. 207-217, 2005. ISSN 0305-7364. DOI: 10.1093/aob/mci014.

LEITCH, Ilia J.; CHASE, Mark W.; BENNETT, Michael D. Phylogenetic Analysis of DNA C-values Provides Evidence for a Small Ancestral Genome Size in Flowering Plants. *Annals of Botany*, v. 82, n. supl. 1, p. 85-94, 1998. ISSN 0305-7364. DOI: 10.1006/anbo.1998.0783.

LEÓN, M.; BASTÍAS, R. Virulence reduction in bacteriophage resistant bacteria. *Frontiers in Microbiology*, v. 6, p. 343, 2015. DOI: 10.3389/fmicb.2015.00343.

LI, Fay-Wei et al. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature Plants*, v. 4, n. 7, p. 460-472, 2018. ISSN 2055-0278. DOI: 10.1038/s41477-018-0188-8.

LI, Linzhou et al. The genome of Prasinoderma coloniale unveils the existence of a third phylum within green plants. *Nature Ecology & Evolution*, v. 4, n. 9, p. 1220-1231, 2020. ISSN 2397-334X. DOI: 10.1038/s41559-020-1221-7.

LI, Weizhong; GODZIK, Adam. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, v. 22, n. 13, p. 1658-1659, 2006. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btl158.

LIOLIOS, Konstantinos et al. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, v. 38,

n. supl. 1, p. D346-D354, 2009. ISSN 0305-1048. DOI: 10.1093/nar/gkp848.

LISCH, Damon. How important are transposons for plant evolution? *Nature Reviews Genetics*, v. 14, n. 1, p. 49-61, 2013. ISSN 1471-0064. DOI: 10.1038/nrg3374.

LIU, Jian-Zhong; WHITHAM, Steven A. Overexpression of a soybean nuclear localized type–III DnaJ domain-containing HSP40 reveals its roles in cell death and disease resistance. *The Plant Journal*, v. 74, n. 1, p. 110-121, 2013. DOI: 10.1111/tpj.12108.

MA, Xuelian et al. PlantGSAD: a comprehensive gene set annotation database for plant species. *Nucleic Acids Research*, v. 50, n. D1, p. D1456-D1467, 2022. ISSN 0305-1048. DOI: 10.1093/nar/gkab794.

MANNI, Mosè et al. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, v. 38, n. 10, p. 4647-4654, 2021. ISSN 1537-1719. DOI: 10.1093/molbev/msab199.

MARKS, Rose A. et al. Representation and participation across 20 years of plant genome sequencing. *Nature Plants*, v. 7, n. 12, p. 1571-1578, 2021. ISSN 2055-0278. DOI: 10.1038/s41477-021-01031-8.

MASHAU, Aluoneswi C. et al. Plant height and lifespan predict range size in southern African grasses. *Journal of Biogeography*, v. 48, n. 12, p. 3047-3059, 2021. DOI: 10.1111/jbi.14261.

MASLOV, Sergei; SNEPPEN, Kim. Population cycles and species diversity in dynamic Kill-the-Winner model of microbial ecosystems. *Scientific Reports*, v. 7, n. 1, p. 39642, 2017. ISSN 2045-2322. DOI: 10.1038/srep39642.

MAZUECOS-AGUILERA, Ismael et al. The Role of INAPERTURATE POLLEN1 as a Pollen Aperture Factor Is Conserved in the Basal Eudicot *Eschscholzia californica* (Papaveraceae). *Frontiers in Plant Science*, v. 12, 2021. ISSN 1664-462X. DOI: 10.3389/fpls.2021.701286.

MINELLI, Alessandro. Introducing Plant Evo-Devo. *In:* Plant Evolutionary Developmental Biology: The Evolvability of the Phenotype. Cambridge: Cambridge University Press, 2018. p. 1-29. DOI: 10.1017/9781139542364.002.

MINH, Bui Quang et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, v. 37, n. 5, p. 1530-1534, 2020. ISSN 0737-4038. DOI: 10.1093/molbev/msaa015.

MOLES, Angela T. et al. Global patterns in plant height. *Journal of Ecology*, v. 97, n. 5, p. 923-932, 2009. DOI: 10.1111/j.1365-2745.2009.01526.x.

MORGAN, Martin. BiocManager: Access the Bioconductor Project Package Repository. 2022. R package version 1.30.18.

MOSAVI, Leila K. et al. The ankyrin repeat as molecular architecture for protein recognition. *Protein Science*, v. 13, n. 6, p. 1435-1448, 2004. DOI: 10.1110/ps.03554604.

MUKHERJEE, Supratim et al. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Research*, v. 45, n. D1, p. D446-D456, 2016. ISSN 0305-1048. DOI: 10.1093/nar/gkw992.

NAGY, László G. et al. Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. *Nucleic Acids Research*, v. 48, n. 5, p. 2209-2219, 2020. ISSN 0305-1048. DOI: 10.1093/nar/gkz1241.

NAKABAYASHI, Kazumi et al. The Time Required for Dormancy Release in Arabidopsis Is Determined by DELAY OF GERMINATION1 Protein Levels in Freshly Harvested Seeds. *The Plant Cell*, v. 24, n. 7, p. 2826-2838, 2012. ISSN 1040-4651. DOI: 10.1105/tpc.112.100214.

NASRALLAH, June B.; NASRALLAH, Mikhail E. S-locus receptor kinase signalling. *Biochemical Society Transactions*, v. 42, n. 2, p. 313-319, 2014. ISSN 0300-5127. DOI: 10.1042/BST20130222.

NIKLAS, Karl J.; KUTSCHERA, Ulrich. The evolution of the land plant life cycle. *New Phytologist*, v. 185, n. 1, p. 27-41, 2010. DOI: 10.1111/j.1469-8137.2009.03054.x.

NISHIMURA, Noriyuki et al. Control of seed dormancy and germination by DOG1-AHG1 PP2C phosphatase complex via binding to heme. *Nature Communications*, v. 9, n. 1, p. 2132, 2018. ISSN 2041-1723. DOI: 10.1038/s41467-018-04437-9.

NISHIYAMA, Eri et al. Ancient and recent gene duplications as evolutionary drivers of the seed maturation regulators DELAY OF GERMINATION1 family genes. *New Phytologist*, v. 230, n. 3, p. 889-901, 2021. DOI: 10.1111/nph.17201.

NISHIYAMA, Takashi et al. The structure of the deacetylase domain of *Escherichia coli* PgaB, an enzyme required for biofilm formation: a circularly permuted member of the carbohydrate esterase 4 family. *Acta Crystallographica Section D*, v. 69, n. 1, p. 44-51, 2013. DOI: 10.1107/S0907444912042059.

O'LEARY, Nuala A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, v. 44, n. D1, p. D733-D745, 2015. ISSN 0305-1048. DOI: 10.1093/nar/gkv1189.

PAGÈS, H. et al. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. 2022. R package version 1.58.0.

PANG, Shuai et al. GPU MrBayes V3.1: MrBayes on Graphics Processing Units for Protein Sequence Data. *Molecular Biology and Evolution*, v. 32, n. 9, p. 2496-2497, 2015. ISSN 0737-4038. DOI: 10.1093/molbev/msv129.

PARADIS, E.; SCHLIEP, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, v. 35, p. 526-528, 2019.

PARK, Beom Seok; LEE, Jie-Oh. Recognition of lipopolysaccharide pattern by TLR4 complexes. *Experimental & Molecular Medicine*, v. 45, n. 12, p. e66, 2013. ISSN 2092-6413. DOI: 10.1038/emm.2013.97.

PASHA, Asher et al. Araport Lives: An Updated Framework for Arabidopsis Bioinformatics. *The Plant Cell*, v. 32, n. 9, p. 2683-2686, 2020. ISSN 1040-4651. DOI: 10.1105/tpc.20.00358.

PAWLUK, April et al. A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa. *mBio*, v. 5, n. 2, p. e00896-14, 2014. DOI: 10.1128/mBio.00896-14.

PEIFFER, Jason A. et al. The Genetic Architecture Of Maize Height. *Genetics*, v. 196, n. 4, p. 1337-1356, 2014. ISSN 1943-2631. DOI: 10.1534/genetics.113.159152.

PELLICER, Jaume et al. Genome Size Diversity and Its Impact on the Evolution of Land Plants. *Genes*, v. 9, n. 2, 2018. ISSN 2073-4425. DOI: 10.3390/genes9020088.

PELLICER, Jaume; FAY, Michael F.; LEITCH, Ilia J. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, v. 164, n. 1, p. 10-15, 2010. ISSN 0024-4074. DOI:

10.1111/j.1095-8339.2010.01072.x.

PELLICER, Jaume; LEITCH, Ilia J. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytologist*, v. 226, n. 2, p. 301-305, 2020. DOI: 10.1111/nph.16261.

PETROV, Dmitri A. Evolution of genome size: new approaches to an old problem. *Trends in Genetics*, v. 17, n. 1, p. 23-28, 2001. ISSN 0168-9525. DOI: 10.1016/S0168-9525(00)02157-0.

PETROV, Dmitri A. Mutational Equilibrium Model of Genome Size Evolution. *Theoretical Population Biology*, v. 61, n. 4, p. 531-544, 2002. ISSN 0040-5809. DOI: 10.1006/tpbi.2002.1605.

PINARD, Desre et al. Comparative analysis of plant carbohydrate active enZymes and their role in xylogenesis. *BMC Genomics*, v. 16, n. 1, p. 402, 2015. ISSN 1471-2164. DOI: 10.1186/s12864-015-1571-8.

PINHEIRO, José; BATES, Douglas; R CORE TEAM. nlme: Linear and Nonlinear Mixed Effects Models. 2022. R package version 3.1-157.

PLAZZI, Federico; FERRUCCI, Ronald R.; PASSAMONTI, Marco. Phylogenetic representativeness: a new method for evaluating taxon sampling in evolutionary studies. *BMC Bioinformatics*, v. 11, n. 1, p. 209, 2010. ISSN 1471-2105. DOI: 10.1186/1471-2105-11-209.

PROOST, Sebastian et al. PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Research*, v. 43, n. D1, p. D974-D981, 2014. ISSN 0305-1048. DOI: 10.1093/nar/gku986.

PULIDO, Pablo; LEISTER, Dario. Novel DNAJ-related proteins in Arabidopsis thaliana. *New Phytologist*, v. 217, n. 2, p. 480-490, 2018. ISSN 0028-646X.

PULKKINEN, W. S.; MILLER, S. I. A Salmonella typhimurium virulence protein is similar to a Yersinia enterocolitica invasion protein and a bacteriophage lambda outer membrane protein. *Journal of Bacteriology*, v. 173, n. 1, p. 86-93, 1991. DOI: 10.1128/jb.173.1.86-93.1991.

PUTTICK, Mark N. et al. Size is not everything: rates of genome size evolution, not C-value, correlate with speciation in angiosperms. *Proceedings of the Royal Society B: Biological Sciences*, v. 282, n. 1820, p. 20152289, 2015. DOI: 10.1098/rspb.2015.2289.

RAMBAUT, Andrew et al. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, v. 67, n. 5, p. 901-904, 2018. ISSN 1063-5157. DOI: 10.1093/sysbio/syy032.

RAMISETTY, Bhaskar Chandra Mohan; SUDHAKARI, Pavithra Anantharaman. Bacterial 'Grounded' Prophages: Hotspots for Genetic Renovation and Innovation. *Frontiers in Genetics*, v. 10, 2019. ISSN 1664-8021. DOI: 10.3389/fgene.2019.00065.

REN, Ren et al. Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms. *Molecular Plant*, v. 11, n. 3, p. 414-428, 2018. ISSN 1674-2052. DOI: 10.1016/j.molp.2018.01.002.

REVELL, Liam J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, v. 3, n. 2, p. 217-223, 2012. DOI: 10.1111/j.2041-210X.2011.00169.x.

ROFF, Derek A. Evolutionary Quantitative Genetics. New York: Springer, 1997. ISBN 978-1-4615-4080-9. DOI: 10.1007/978-1-4615-4080-9.

SALL, Khadidiatou et al. DELAY OF GERMINATION 1-LIKE 4 acts as an inducer of seed

reserve accumulation. *The Plant Journal*, v. 100, n. 1, p. 7-19, 2019. DOI: 10.1111/tpj.14485.

SALZBERG, Steven L. Next-generation genome annotation: we still struggle to get it right. *Genome Biology*, v. 20, n. 1, p. 92, 2019. ISSN 1474-760X. DOI: 10.1186/s13059-019-1715-2.

SANDOVAL, Francisco J.; ZHANG, Yi; ROJE, Sanja. Flavin Nucleotide Metabolism in Plants: MONOFUNCTIONAL ENZYMES SYNTHESIZE FAD IN PLASTIDS. *Journal of Biological Chemistr*y, v. 283, n. 45, p. 30890-30900, 2008. ISSN 0021-9258. DOI: 10.1074/jbc.M803416200.

SANGER, F.; COULSON, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, v. 94, n. 3, p. 441-448, 1975. ISSN 0022-2836. DOI: 10.1016/0022-2836(75)90213-2.

SAYERS, Eric W. et al. GenBank. *Nucleic Acids Research*, v. 48, n. D1, p. D84-D86, 2019. ISSN 0305-1048. DOI: 10.1093/nar/gkz956.

SCHÄFFER, Alejandro A. et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, v. 29, n. 14, p. 2994-3005, 2001. ISSN 0305-1048. DOI: 10.1093/nar/29.14.2994.

SCHALLUS, Thomas et al. Malect*In:* A Novel Carbohydrate-binding Protein of the Endoplasmic Reticulum and a Candidate Player in the Early Steps of Protein N-Glycosylation. *Molecular Biology of the Cell*, v. 19, n. 8, p. 3404-3414, 2008. DOI: 10.1091/mbc.e08-04-0354.

SCHNEIDER, René; PERSSON, Staffan. Another brick in the wall. *Science*, v. 350, n. 6257, p. 156-157, 2015. DOI: 10.1126/science.aad3200.

SCHUSTER, Stephan C. Next-generation sequencing transforms today's biology. *Nature Methods*, v. 5, n. 1, p. 16-18, 2008. ISSN 1548-7105. DOI: 10.1038/nmeth1156.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, v. 52, n. 3-4, p. 591-611, 1965. DOI: 10.1093/biomet/52.3-4.591.

SILVEIRA, Cynthia B.; ROHWER, Forest L. Piggyback-the-Winner in host-associated microbial communities. *NPJ Biofilms and Microbiomes*, v. 2, n. 1, p. 16010, 2016. ISSN 2055-5008. DOI: 10.1038/npjbiofilms.2016.10.

SIMMONS, Emilia L. et al. Biofilm Structure Promotes Coexistence of Phage-Resistant and Phage-Susceptible Bacteria. *mSystems*, v. 5, n. 3, p. e00877-19, 2020. DOI: 10.1128/mSystems.00877-19.

SORENSEN, Iben et al. The charophycean green algae provide insights into the early origins of plant cell walls. *The Plant Journal*, v. 68, n. 2, p. 201-211, 2011. DOI: 10.1111/j.1365-313X.2011.04686.x.

STEYERT, Susan R.; KAPER, James B. Contribution of Urease to Colonization by Shiga Toxin-Producing Escherichia coli. *Infection and Immunity*, v. 80, n. 8, p. 2589-2600, 2012. DOI: 10.1128/IAI.00210-12.

SUBBURAJ, Saminathan et al. Phylogenetic Analysis, Lineage-Specific Expansion and Functional Divergence of seed dormancy 4-Like Genes in Plants. *PLOS ONE*, v. 11, n. 6, p. 1-24, 2016. DOI: 10.1371/journal.pone.0153717.

TELLO-RUIZ, Marcela K. et al. Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. *Nucleic Acids Research*, v. 49, n. D1, p. D1452-D1463, 2020. ISSN 0305-1048. DOI: 10.1093/nar/gkaa979.

TENENBAUM, D.; MAINTAINER, B. KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). 2022. R package version 1.36.3.

TOMAŽ, Špela; GRUDEN, Kristina; COLL, Anna. TGA transcription factors—Structural characteristics as basis for functional variability. *Frontiers in Plant Science*, v. 13, 2022. ISSN 1664-462X. DOI: 10.3389/fpls.2022.935819.

TONG, Chao et al. Comparative Genomics Identifies Putative Signatures of Sociality in Spiders. *Genome Biology and Evolution*, v. 12, n. 3, p. 122-133, 2020. ISSN 1759-6653. DOI: 10.1093/gbe/evaa007.

TOUCHON, Marie et al. Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. *PLOS Genetics*, v. 5, n. 1, p. 1-25, 2009. DOI: 10.1371/journal.pgen.1000344.

TUSKAN, G. A. et al. The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, v. 313, n. 5793, p. 1596-1604, 2006. DOI: 10.1126/science.1128691.

UNG, Huoi; MOEDER, Wolfgang; YOSHIOKA, Keiko. Arabidopsis Triphosphate Tunnel Metalloenzyme2 Is a Negative Regulator of the Salicylic Acid-Mediated Feedback Amplification Loop for Defense Responses. *Plant Physiology*, v. 166, n. 2, p. 1009-1021, 2014. ISSN 0032-0889. DOI: 10.1104/pp.114.248757.

VAIDYA, Gaurav; LOHMAN, David J.; MEIER, Rudolf. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics*, v. 27, n. 2, p. 171-180, 2011. DOI: 10.1111/j.1096-0031.2010.00329.x.

VAIDYANATHAN, Ramnath et al. htmlwidgets: HTML Widgets for R. 2021. R package version 1.5.4.

VANDECRAEN, Joachim et al. The impact of insertion sequences on bacterial genome plasticity and adaptability. *Critical Reviews in Microbiology*, v. 43, n. 6, p. 709-730, 2017. DOI: 10.1080/1040841X.2017.1303661.

VESELÝ, Pavel; BUREŠ, Petr; ŠMARDA, Petr. Nutrient reserves may allow for genome size increase: evidence from comparison of geophytes and their sister non-geophytic relatives. *Annals of Botany*, v. 112, n. 6, p. 1193-1200, 2013. ISSN 0305-7364. DOI: 10.1093/aob/mct185.

VINOGRADOV, Alexander E. Selfish DNA is maladaptive: evidence from the plant Red List. *Trends in Genetics*, v. 19, n. 11, p. 609-614, 2003. ISSN 0168-9525. DOI: 10.1016/j.tig.2003.09.010.

VITTI, Joseph J.; GROSSMAN, Sharon R.; SABETI, Pardis C. Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, v. 47, n. 1, p. 97-120, 2013. DOI: 10.1146/annurev-genet-111212-133526.

VOGEL, Christine; CHOTHIA, Cyrus. Protein Family Expansions and Biological Complexity. *PLOS Computational Biology*, v. 2, n. 5, p. 1-13, 2006. DOI: 10.1371/journal.pcbi.0020048.

WANG, B. et al. [The China National GeneBank owned by all, completed by all and shared by all]. Yi Chuan, v. 41, n. 20, p. 761-772, 2019. DOI: 10.16288/j.yczz.

WANG, Dandan et al. Which factors contribute most to genome size variation within angiosperms? *Ecology and Evolution*, v. 11, n. 6, p. 2660-2668, 2021. DOI: 10.1002/ece3.7222.

WANG, Xiaoxue et al. Cryptic prophages help bacteria cope with adverse environments. *Nature Communications*, v. 1, n. 1, p. 147, 2010. ISSN 2041-1723. DOI: 10.1038/ncomms1146.

WATERHOUSE, Robert M. et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, v. 35, n. 3, p. 543-548, 2017. ISSN 0737-4038. DOI: 10.1093/molbev/msx319.

WENDEL, Jonathan F. et al. Feast and famine in plant genomes. *Genetica*, v. 115, n. 1, p. 37-47, 2002. ISSN 1573-6857. DOI: 10.1023/A:1016020030189.

WICKHAM, Hadley. assertthat: Easy Pre and Post Assertions. 2019. R package version 0.2.1.

WICKHAM, Hadley. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag, 2016. ISBN 978-3-319-24277-4.

WICKHAM, Hadley. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman & Hall, 2020. ISBN 9781138331457.

WICKHAM, Hadley; HESSELBERTH, Jay; SALMON, Maëlle. pkgdown: Make Static HTML Documentation for a Package. 2022. R package version 2.0.3.

WILLI, Yvone; HOFFMAN, Ary A. Demographic factors and genetic variation influence population persistence under environmental change. *Journal of Evolutionary Biology*, v. 22, n. 1, p. 124-133, 2009. DOI: 10.1111/j.1420-9101.2008.01631.x.

WOLF, Andrea J.; UNDERHILL, David M. Peptidoglycan recognition by the innate immune system. *Nature Reviews Immunology*, v. 18, n. 4, p. 243-254, 2018. ISSN 1474-1741. DOI: 10.1038/nri.2017.136.

WOLF, Jason B. The geometry of phenotypic evolution in developmental hyperspace. *Proceedings of the National Academy of Sciences*, v. 99, n. 25, p. 15849-15851, 2002. DOI: 10.1073/pnas.012686699.

XIAO, Yu et al. Mechanisms of RALF peptide perception by a heterotypic receptor complex. *Nature*, v. 572, n. 7768, p. 270-274, 2019. ISSN 1476-4687. DOI: 10.1038/s41586-019-1409-7.

XIE, Yihui; CHENG, Joe; TAN, Xianying. DT: A Wrapper of the JavaScript Library 'DataTables'. 2022. R package version 0.23.

XUE, Han et al. qPTMplants: an integrative database of quantitative post-translational modifications in plants. *Nucleic Acids Research*, v. 50, n. D1, p. D1491-D1499, 2021. ISSN 0305-1048. DOI: 10.1093/nar/gkab945.

YANG, He et al. Malectin/Malectin-like domain-containing proteins: A repertoire of cell surface molecules with broad functional potential. *The Cell Surface*, v. 7, p. 100056, 2021. ISSN 2468-2330. DOI: 10.1016/j.tcsw.2021.100056.

YANG, Xiaohan et al. Comparative genomics can provide new insights into the evolutionary mechanisms and gene function in CAM plants. *Journal of Experimental Botany*, v. 70, n. 22, p. 6539-6547, 2019. ISSN 0022-0957. DOI: 10.1093/jxb/erz408.

YELAGANDULA, Ramesh et al. The Histone Variant H2A.W Defines Heterochromatin and Promotes Chromatin Condensation in Arabidopsis. *Cell*, v. 158, n. 1, p. 98-109, 2014. ISSN 0092-8674. DOI: 10.1016/j.cell.2014.06.006.

ZHANG, Jian et al. The hornwort genome and early land plant evolution. *Nature Plants*, v. 6, n. 2, p. 107-118, 2020. ISSN 2055-0278. DOI: 10.1038/s41477-019-0588-4.

ZU, Pengjuan; SCHIESTL, Florian P. The effects of becoming taller: direct and pleiotropic effects of artificial selection on plant height in Brassica rapa. *The Plant Journal*, v. 89, n. 5, p.

1009-1019, 2017. DOI: 10.1111/tpj.13440.

ZWICKL, Derrick J.; HILLIS, David M. Increased Taxon Sampling Greatly Reduces Phylogenetic Error. *Systematic Biology*, v. 51, n. 4, p. 588-598, 2002. ISSN 1063-5157. DOI: 10.1080/10635150290102339.

ANEXO A - USING ARCADE (ARCHAEPLASTIDA ANNOTATION DATABASE) TO UNDERSTAND THE EVOLUTION OF GENOME SIZE IN LAND PLANTS

Using ARCADE (ARChaeplastida Annotation DatabasE) to understand the evolution of genome size in land plants

Alison Pelri Albuquerque Menezes[1], João Victor dos Anjos Almeida[2], Luiz-Eduardo Del-Bem[3], and Francisco Pereira Lobo[1,*]

[1]*Departamento de Genética, Ecologia e Evolução, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.*

[2]*Universidade Estadual Paulista Júlio de Mesquita Filho, Jaboticabal, São Paulo*

[3]*Departamento de Botânica, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.*

[*]*To whom correspondence should be addressed. Tel: +55 31 34093072; Fax: +55 31 34092567; Email: franciscolobo@ufmg.br, franciscolobo@gmail.com*

Abstract

The abundance of plant genomic information caused by the decrease of sequencing costs contrasts with the lack of databases that integrate genome annotation, taxonomy and phenotypes to produce statistically sound, biologically meaningful knowledge. Here we present ARCADE (ARChaeplastida Annotation DatabasE), a database of 171 high-quality archaeplastidian non-redundant proteomes gathered from six primary genomic databases, together with proteome quality metrics and a growing number of associated metadata. As a case study to demonstrate the usefulness of ARCADE, we used it to investigate the expansion and contraction of protein domains associated with the evolution of genome size (hereafter GS). GS varies greatly among land plants and the maintenance of large genomes can be costly to cells. Although GS has been studied extensively for decades, the molecular mechanisms involved in the adaptations of plants to the increase in GS are still poorly understood. We used the annotation and phylogenetic information available in ARCADE, together with estimated GS values available for 83 land plant species, to search for associations between the abundance of protein domain families in these species and GS variation through phylogenetic-aware methods. Additionally, we estimated the GS for the ancestral nodes of the extant land plant species. GS seems to

be decreasing along the course of evolution, except for a few branches that might have undergone independent GS increases. We found 7 Pfam domains correlated with the variation in GS in land plants, mainly related to nucleotide metabolism, DNA repair and genome organization. We found larger genomes to have a greater frequency of the Histone 2A superfamily, responsible for diverse functions, including the nucleosome formation and silencing of transposable elements. These molecular functions we found correlated to GS variation suggests they may be associated with preserving genome stability in larger genomes, and might indicate the evolution of mechanisms to cope with the variation in GS in land plants.

ARCADE is available at https://bit.ly/ARCADE_OSF.

KEYWORDS: genome size, C-value, genome evoluttion, genome annotation, comparative phylogenetic methods.

1.1 INTRODUCTION

Comparative genomics has extensively contributed to elucidate genotype-phenotype associations and understand evolutionary processes at the genome level (X. Yang et al. 2019). A major conceptual data structure needed for any comparative genomics study is a standardized set of genomic elements (e.g. all protein-coding genes for all species under analysis), each of them annotated to a common dictionary of biologically meaningful annotation terms (e.g. groups of homologous regions shared across gene sets) (Tello-Ruiz et al. 2020). Furthermore, as cellular species are all descendant from the last universal common ancestor and, consequently, are non-independent from a statistical point of view, common association statistics are not suitable to infer genotype-phenotype associations across species (Nagy et al. 2020). For that, a tree-like structure describing the relationships among them is of uttermost importance for model generation, as it allows several phylogeny-aware methods to be applied when searching for genotype-phenotype associations (Adams e Collyer 2017).

Most Archaeplastida members are oxygenic photosynthetic eukaryotes comprising the red algae (Rhodophyta), green algae (Chlorophyta and Charophyta), land plants (Embryophyta), and the freshwater microscopic algae Glaucophyta. Contemporary research in evolutionary plant genomics has been highly impacted by the massive decrease in DNA

sequencing costs. The availability of genomic information from early-branching plant lineages, together with comparative genomics analyses, provides the molecular comprehension of major evolutionary phenotypic novelties in this group, such as the vascular system and the flower (Chanderbali et al. 2016; Blázquez, Nelson e Weijers 2020). These phenotypes are but the tip of the iceberg of a large number of quantitative and qualitative phenotypes readily available for comparative analyses across plants (Kattge et al. 2020).

However, the abundance of plant genomic data contrasts with the challenges associated with the gathering of high-quality, standardized genomic data needed for comparative genomics studies. Several aspects contribute to this paradox. The first and widely known issue is the excess of high-quality, experimentally validated annotation information available for genes of model organisms when compared with non-model species (Haynes, Tomczak e Khatri 2018). Another source of bias is the huge variation observed in the quality of genome assemblies due to technical issues, such as distinct sequencing technologies and assembly algorithms, but also caused by true biological facts, such as genome size variation caused by the increase of repetitive elements, domestication events, and lineage-specific whole-genome duplications (Marks et al. 2021). Together, these facts can significantly bias downstream comparative genomic analyses.

Not surprisingly, several specialized databases provide high-quality genomic and annotation data for distinct groups of plants. However, these are heavily focused on a few data-rich model organisms comprising mostly angiosperm species of commercial or scientific interest (Marks et al. 2021). This fact limits their usage to provide the genomic and phylogenetic data to survey the evolution of key phenotypic traits in Archaeplastida.

Here we present the ARChaeplastida Annotation DatabasE (ARCADE), a database of high-quality, non-redundant annotated proteomes for 171 Archaeplastida species, together with the available phylogenetic metadata. ARCADE was gathered from six major genome databases and annotated through a common pipeline to provide a rich set of homologous regions as defined by InterProScan, together with Gene Ontology (GO) and pathway annotation, when available (Jones et al. 2014). We also provide phylogenetic information for 142 species as an ultrametric species tree. The integration of phylogenetic and annotation data as provided by ARCADE allows the development of phylogeny-aware models needed for properly comparing species data. Furthermore, instead of focusing on providing user-friendly access to several layers of metadata for a few model organisms

and economically relevant angiosperms, as is the case for virtually all plant annotation resources available, we actively searched for high-quality genomes from all known Archaeplastida taxa with an emphasis on underrepresented, early branching lineages, together with their annotation and phylogenetic metadata, when available. This data is available as defined tabular text files and other file formats commonly used in bioinformatics pipelines, therefore providing an annotation and evolutionary scaffold to perform comparative analysis of plants at several taxonomic levels (Nagy et al. 2020).

To demonstrate how our database provides readily available, biologically meaningful knowledge to understand the evolution of complex traits, we used ARCADE to search for homologous regions in protein-coding genes associated with the variation of genome size (or C-value) in land plants, whose variation has been largely studied and yet is one of the greatest mysteries of genomics and evolutionary biology. The associations of genome size variation and several plant traits have been repeatedly addressed. However, the molecular mechanisms involved in the adaptations of plants to the increase in genome size are still poorly understood. Thus, we aim at using this case study to demonstrate ARCADE's usefulness by investigating which protein domain famlies are correlated with the increase in genome size in Embryophyta (land plants). There might be molecular mechanisms that allow genome size variation in nature, considering that the increase in genome size might also impose physiological pressures on their organisms (Knight, Molinari e Petrov 2005). Through analysis of comparative genomics on 83 species of land plants, we found that protein domains associated with key components of DNA biology, such as DNA repair, nucleotide metabolism and genome maintenance are enriched in plants with larger genomes when compared to the smaller ones. These results might show the way to mechanisms through which plants cope with the impact of genome increase, and comprise a compelling case study of the usefulness of our database.

1.2 METHODS

1.2.1 Building ARCADE

Survey for Archaeplastida complete genomes

We aimed at building a comprehensive and phylogenetic diverse genomic dataset for Archaeplastida species, with emphasis on early-branching lineages. Specifically, we downloaded the assembled genomic data for all Embryophyta species available in the NCBI databases and complemented our dataset with the predicted proteomes belonging to relevant and underrepresented lineages with data from the other public databases. We screened six primary genomic databases for plant genomes: 1) National Center for Biotechnology Information (NCBI) databases – RefSeq (O'Leary et al. 2015) and GenBank (Sayers et al. 2019); 2) Phytozome (Goodstein et al. 2011); 3) Gymno PLAZA (Proost et al. 2014); 4) FernBase (F.-W. Li et al. 2018); 5) CNGB (X. Wang et al. 2010); 6) and the DRYAD repository (J. Zhang et al. 2020).

Obtention of high-quality, non-redundant annotated proteomes and phylogenetic data

To reduce the proteome redundancy generated by the distinct number of isoforms of the same gene and avoid a possible bias towards model organisms, we applied two different proteome summarization protocols to our dataset. The first protocol, used on NCBI data (*in-house* pipeline), selects the longest protein sequence of each protein-coding locus based on the "locus_tag" or "gene_id" fields. However, fasta proteomes from other databases do not contain these fields in their metadata, and could not be submitted to our pipeline. To reduce redundancy of those proteomes we used the CD-HIT software (W. Li e Godzik 2006; Fu et al. 2012) with the threshold set to 1, this way we expected to keep only the longest protein isoform to represent each genomic locus. We evaluated the assembling quality of each proteome using the BUSCO software (Manni et al. 2021) to assess their gene completeness based on the set of 255 Viridiplantae BUSCOs (odb10). We kept in our dataset proteomes with completeness higher than 70% and rates lower than 20% for duplicated and fragmented genes, using the results obtained for *Arabidopsis thaliana* as a proxy for the lower-bound cutoffs for quality (Supplementary Table S-1).

We used InterProScan 5 (Jones et al. 2014) to perform a *de novo* annotation of the 171

non-redundant proteomes that fulfill the previous BUSCO cutoffs (from now on referred as high-quality non-redundant proteomes - NRPs). Specifically, all non-redundant proteomes were annotated to 15 distinct databases that are integrated into InterProScan that provide homology and functional annotation information for all proteomes, if available.

The ARCADE database comprises a tabular file containing curated metadata for all 171 species, such as species binomial names, NCBI TaxonIDs, database of origin and proteome quality metrics, among others. We also provide annotation information, including raw output files from InterProScan for each species, as well as parsed data for all 15 distinct annotation databases members of InterProScan. Finally, we also gathered a newick file containing a species tree for 142 species gathered from the TimeTree web tool (S. Kumar, Suleski et al. 2022). The initial tree produced by this tool was enriched by including species from closely related taxa as placeholders for the original species. Example: if one species from an order could not be placed on the tree and 1) it is the only species from this order present in ARCADE and 2) another species from the same order is available, we replaced the original species by the one present in TimeTree to include the original species in the tree.

1.2.2 Case study – evolution of genome size in Embryophyta

Computation environment

The analysis done in this project was executed in a Dell server with 2 processors Intel Xeon E5-4610 v2 2.3GHz totalizing 64 threads; 128GB of RAM and operational system CentOS Linux release 7.5.1804, with support to the programming languages PERL5 v. 7.5 and R v. 3.0.0.

Data Collection and Functional Annotation

As ARCADE contains species' scientific names as its primary key to integrate distinct data types, it is trivial to gather information from other databases and integrate with the genomic annotations we provide. To investigate the evolution of genome size in land plants, we obtained experimentally measured 1C-value data for all the studied species in the Plant C-values Database (Pellicer e I J. Leitch 2020) and in the GoaT database (Challis et al. 2017) (Supplementary Table S-1). Here we are using 1C-value and genome size convertibly, although the genome size would equal to 2C-value divided by ploidy level

(I. J. Leitch, Chase e Bennett 1998). Then, we converted the 1C-value from picogram (pg) to megabase pairs (Mb), where 1 pg equals 980 Mb and log10-transformed those values. To evaluate the distribution of GS values in our sample we performed the Shapiro-Wilk test (SHAPIRO e WILK 1965). In the end, we gathered genotypic, phylogenetic, and genotypic data for 83 species of Embryophyta (Supplementary Table S-1, Fig. 1) to perform our phylogenetic comparative analysis.

Comparative analysis

We studied the evolution of genome size in land plants in two steps. First, we estimated the ancestral genome size and their respective 95% confidence interval (CI) for all the internal nodes of the obtained phylogeny based on the genome size of 86 extant species of land plants. That was made by the method of maximum likelihood implemented in the "*fastAnc*" function of the *phytools* R package (Revell 2012). Then we mapped the estimated ancestral states (*fastAnc* function) on the phylogeny obtained from the Time Tree of Life database (S. Kumar, Suleski et al. 2022) using ggplot2 (Sievert 2016).

Second, we investigated protein domains' abundance and frequency correlated with the increase in genome size in land plants. To do so, we used CALANGO (Hongo et al. 2021), which integrates protein domain annotation (both frequency and absolute abundance in each genome), genotypic (genome size), and phylogenetic data. To avoid bias from the phylogenetic interdependencies within the data, the CALANGO applies the method of Phylogenetically Independent Contrasts (PIC) (Felsenstein 1985) using a phylogenetic tree, from the Time Tree of Life database (S. Kumar, Suleski et al. 2022).

To summarize results and facilitate data interpretation, the Pfam domains obtained from the results of the CALANGO analysis were used to find, in the InterProScan annotation files, the *Arabidopsis thaliana* genes containing them. Then we searched for information about those genes using the ThaleMine (Krishnakumar et al. 2014; Pasha et al. 2020) data mining tool to gather relevant functional information.

1.3 RESULTS AND DISCUSSION

1.3.1 The ARCADE database: an overview

We created ARCADE as a resource to organize and integrate high-quality non-redundant predicted proteomes from Archaeplastida species, together with their respective phylogenetic metadata and genomic annotations, when available. For that, we surveyed six major genome databases that host plant genomes and, starting from 1381 species, we gathered 171 high-quality, non-redundant proteomes of Archaeoplastida organisms from the following ten major groups (number of species between parenthesis): Angiosperm (98), Gymnosperm (8), Bryophyta (24), Monilophyta (14), Charophyta (6), Chlorophyta (16), Glaucophyta (1), Lycophyta (3), Prasinodermophyta (1), Rhodophyta (1) (Figure 1A; see also section “Comparison of ARCADE and other plant genome databases” for a deeper discussion on individual database contribution for the final list of species available in ARCADE). We could confidently produce a species tree for 142 species found in ARCADE (84.2%) that represent all major taxa but Prasinodermophyta, therefore providing a taxonomically diverse phylogenetic scaffold for phylogeny-aware studies, as we demonstrate in our case study (Figure 1B).

Model organisms are data-rich species and expected to have more genomic information available, including the number of known isoforms, and these may be a significant source of bias (L. Chen et al. 2014), and non-redundant proteomes are a mainstream source of unbiased genomic components for comparative genomic studies (Vogel e Chothia 2006). We used locus information available at NCBI to generate non-redundant proteomes for all species. However, as genomes from all other databases do not provide isoform-level information, we proceed by using CD-HIT to remove redundancy from the predicted proteomes and evaluated whether this strategy also introduced any undesirable downstream bias. For that we compared the non-redundant proteomes of four major groups summarized by both methods (in house and CD-HIT): angiosperms, bryophytes, charophytes and chlorophytes. Specifically, we evaluated these groups for the number of protein sequences found in each proteome, together with BUSCO metrics for proteome completeness (Fig. S-1 B-G). We found CD-HIT summarization to produce, on average, smaller proteomes for three out of the four groups (angiosperms, bryophytes and chlorophytes), with an opposite trend observed for charophytes (Fig. S-1B). The BUSCO results for completeness

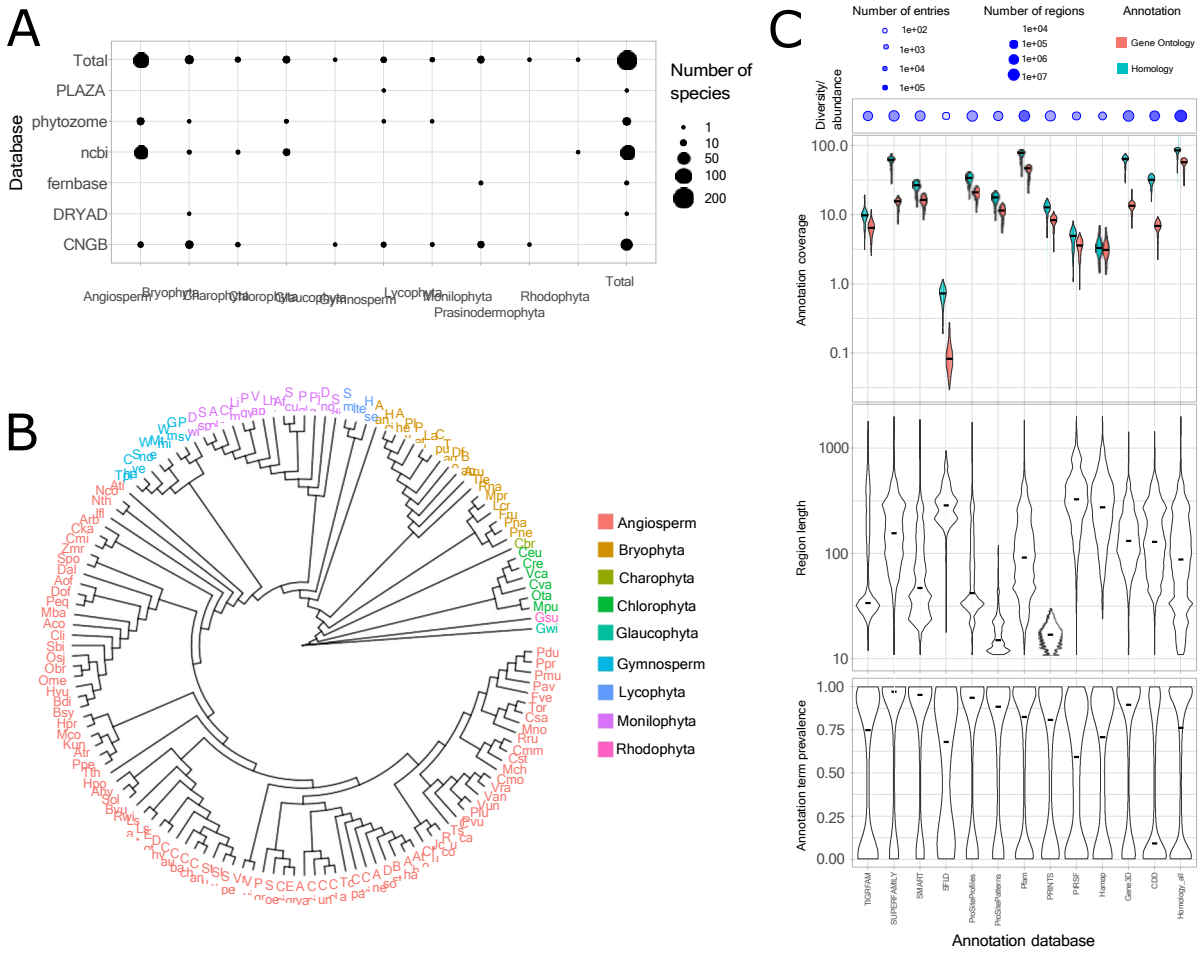


Figura 1: Description of ARCADE database. A) Relative species richness and taxonomic diversity of the plant genome databases used to build ARCADE. B) Species tree for 142 species found in ARCADE (extracted from the TimeTree web tool (S. Kumar, Suleski et al. 2022) as described in "Methods" (species abbreviations can be found in supplementary materials). C) Characterization of the 15 InterPro databases used to annotate ARCADE sequence data. Plots represent information as follows: Diversity/abundance – the number of entries.

and single-copy BUSCOs displayed a similar trend (Fig. S-1C-D).

The number of duplicated BUSCOs, which are a widespread proxy to detect genome assemblies with duplications introduced by genome assembly issues, and is also expected to be higher in proteomes with more than one isoform per locus, showed similar distributions for all four groups but bryophytes, where CD-HIT summarization strategy appears to produce a higher number of duplicated genes (Supplementary Figure S-5). Up to this point, one could conclude that the CD-HIT may be introducing bias in at least the non-

redundant proteomes from bryophyte species. However, the analysis of fragmented and missing BUSCOs, which are likely not to be influenced by any summarization protocol, found that proteomes summarized by CD-HIT more fragmented and have more missing BUSCOs than their *in-house* counterparts, especially for the bryophytes (Supplementary S-6-S-7). Taken together, we conclude that the CD-HIT summarization strategy does not introduce any significant bias during the production of non-redundant proteomes. The differences observed are likely due to other factors, such as true natural variation or genome assembly issues. Although databases other than NCBI were very important to increase diversity in our dataset, our results for BUSCO values suggest that in general, their assembly quality is lower than the NCBI ones.

The annotation of all 171 high-quality non-redundant proteomes using the 15 databases as provided by InterProScan predicted a total of 23,173,794 homologous regions within non-redundant proteomes annotated by one of the 23,514 distinct annotation terms, defined as each distinct homologous region from a specific annotation database. We found these databases to provide highly variable sets of annotation terms for sequence length, taxonomic prevalence and functional diversity (Figure 1C; group "Homology_all" represents all entries from all databases). There is a considerable variation in annotation term abundance (regions annotated by an annotation term) and diversity (number of distinct annotation terms) contributed by each database. Two databases (Pfam and CDD) have approximately 25% and 50% of the total annotation abundance and diversity, respectively (group Figure 1C, "Diversity/abundance" plot).

We also found distinct databases to have large differences in the annotation coverage of non-redundant proteomes, defined as the fraction of non-redundant proteomes annotated by at least one annotation term. Three databases (Pfam, SUPERFAMILY and GENE3D) provide the broadest values of proteome annotations. Importantly, we found considerable fraction of the proteomes found in ARCADE are annotated by at least one annotation term (Figure 1C, "Annotation coverage" plot, "Homology_all" group). In contrast with all other annotation databases, Pfam also contributes with the largest fraction of Gene Ontology (GO) annotation. The annotation databases are also highly variable in terms of the length of each homologous region, ranging from databases enriched in short sequences (e.g. PRINTS, ProSitePatterns) to databases with median length homologous regions around 300 amino acid residues (e.g. PIRSF).

Finally, we evaluated annotation term prevalence across proteomes, defined as the fraction of proteomes where an annotation term was found (Figure 1C, "Annotation term prevalence" plot). We observed databases to possess a bimodal distribution with annotation terms that have either low prevalence, being observed in few proteomes and corresponding to lineage-specific homologous regions, to annotation terms highly prevalent, comprising homologous regions found in the majority of the species under analysis. The vast majority of databases are enriched in highly prevalent homologous regions, with CDD being the only database enriched in homologous genes with low prevalence across species. Taken together, we conclude that ARCADE represents a considerable effort towards providing standardized and high-quality phylogenetic and annotation data, where individual annotation databases provide a highly variable set of homologous regions in terms of abundance, diversity, phylogenetic distribution, annotation coverage and functional characterization.

1.3.2 Comparison of ARCADE and other plant genome databases

The genome databases that host data for plant species, such as the ones we used to mine data from, are developed for scientists with little or no computational background. As so, they provide rich graphical user interfaces to allow data access and analysis through genome browsers and the integration of several layers of genomic metadata. However, these databases mostly host data from specific lineages, or have a phylogenetically limited and biased dataset focused on model species or of agricultural interest (Ma et al. 2022). ARCADE, in contrast, was specifically developed to provide a taxonomically diverse selection of high-quality, non-redundant predicted proteomes, together with its phylogenetic and annotation data, if available. Our target audience is the community of plant computational biologists that may take advantage of having a taxonomically diverse set of species and their fundamental data types needed for the study of complex genotype-phenotype associations. More importantly, the data provided in ARCADE is not found in any single plant genome database, as we demonstrate below.

NCBI, CNGB, and Phytozome are some of the richest resources in plant genomic data. NCBI databases are arguably the most commonly used resources for genetic and genomic studies in the western world, thus it gathers a very wide range of different types of data, very well curated and organized. CNGB is a large-scale database with an enormous

number and diversity of genomes, but from thousands of predicted proteomes, only 54 unique species were approved by our quality control. Furthermore, we found BUSCO results to be consistently worst from genomes summarized by CD-HIT, as is the case for all CNGB entries (Supplementary Figures S-3-S-7).

Even though a single database (NCBI) contributes with the majority of entries of ARCADE (94 spp, 60,0%), no individual genomic database was found to contribute with species for all ten major Archaeplastida groups, with CNGB and NCBI being the most taxonomically rich databases and contributing with 8 and 5 groups, respectively. Furthermore, no single group was found in all genomic databases, with Bryophyta being the most prevalent group (4 out of 6 databases), and with three major groups observed in a single database (Rhodophyta – NCBI, Prasinodermophyta and Glaucophyta – CNGB).

Among all the databases we searched, Phytozome is the only one containing functional annotations for their predicted proteomes and the most phylogenetically diverse as well. The databases that bring functional annotation to their protein sequences are more specific, focusing on tRNA (Chan e Lowe 2015) or specific enzyme families (Ekstrom et al. 2014), for example. Phytozome is well-curated and offers comprehensive and uniform annotation for the genomes of its 139 Archaeplastida species, while our datasets encompass 171 plant species. Although Phytozome includes more phylogenetic diverse species, the number of representants of non-flowering plants is insufficient, even lacking species of the major clade of charophytes.

We also used genomes from lineage-specific databases, such as FernBase and Gymno PLAZA, to increase the diversity of ferns and gymnosperms in our dataset. Only *Anthoceros angustus* did not come from a genomic database; this proteome was obtained from the data repository DRYAD where it was uploaded by the research group responsible for sequencing its genome. Taken together, we consider that our effort to obtain a diverse set of high-quality and annotated non-redundant proteomes for Archaeplastida was successful, as these would not be obtainable from any single database we evaluated.

1.3.3 Case study: using ARCADE to study the evolution of genome size in land plants

The nuclear genome size varies about 2,400-fold in land plants, ranging from 61 Mb in *Genlisea tuberosa* (the smallest genome known yet) (Fleischmann et al. 2014) to 152,000

Mb in *Paris japonica* (Pellicer, Fay e I. J. Leitch 2010). Since genome sizes are readily available since long before the beginning of genomic studies, variations on this trait have been largely evaluated for associations (andILeitch 2005). Previous studies have found correlations between genome size and many plants' phenotypic and ecological traits, such as cell and seed size, habitat, and distribution (Bennett 1987; Knight, Molinari e Petrov 2005; Beaulieu et al. 2007; Kang et al. 2014).

In land plants, large genomes emerged mainly due to repeated events of polyploidy and the imbalance between the amplification and removal of transposable elements from DNA (Pellicer, Hidalgo et al. 2018; Petrov 2001). Genome size imposes restrictions on compatible life strategies and ecological options, as it will affect plant development and fitness (Knight, Molinari e Petrov 2005; A. R. Leitch e I. J. Leitch 2012). Due to the costs of maintaining their genome (e.g. nutrient and water supply), plants with larger genomes will be limited to more stable environments, where selective pressures are more relaxed (Knight, Molinari e Petrov 2005; Veselý, Bureš e Šmarda 2013; Hidalgo et al. 2017). Many authors have proposed different evolutionary models that place genome size as a trait that responds to selective pressures, adaptively or not, or as an evolutionary neutral trait. Therefore, it is fair to assume that genome size variation holds biological importance (Petrov 2002; Vinogradov 2003).

To demonstrate how ARCADE provides the phylogenetic and annotation scaffold needed to survey the evolution of complex traits, we used our dataset to survey the influence of genome size (C-value) variation in land plants and the abundance of homologous genes as defined by the Pfam annotation (Fig. S-2 contains the workflow for this analysis). We started by gathering the C-values for 86 species of land plants found in ARCADE. At this point, it is worth noting that five out of the six genome databases used to generate ARCADE contributed with species for this case study, therefore demonstrating the unfeasibility of using any individual database to reproduce our study case study and, consequently, the usefulness of ARCADE.

Integrating phylogenetic and trait data to estimate ancestral of genome size

Shapiro-Wilk test showed that the distribution of C-values is not normal, even after log10-transformation ($W = 0.89674$, $p < 0.0001$), with a long-tail towards larger genomes (Fig. 2). The range of C-values varies from 86.24 Mb (for the lycophyte *S. moellendorffii*)

to 24,792.30 Mb (for gymnosperm *Cephalotaxus harringtonia*). The genome size average in our dataset is 2,394.33 Mb, but 67 species (of 86 in total) have genomes smaller than the average, dragging the median down to 618.63 Mb. Most plants with large genomes are gymnosperms (14,992.32 Mb on average) or monilophytes (4,258.59 Mb on average) (Figure 2). But there are also representatives of the plants with large genomes within angiosperms (i.e. *Hordeum vulgare* with 5,379 Mb) and lycophytes (i.e. *Huperzia selago* with 4,938.9 Mb). We used current C-values, together with the species tree provided by ARCADE, to estimate the genome size for the ancestral nodes of the land plant phylogeny (Figure 2A contains the estimations; Figures 2B-C contains the actual values for species and major groups, respectively).

The estimated ancestral genome size of the land plants was 1088.98 Mb (95% CI: 212.93 Mb to 5,569.38 Mb), smaller than the average genome size of the extant plants. From that point in evolution, it shows bidirectional evolution. The GS in most clades are decreasing, but we can notice that in few clades GS is increasing independently along the phylogeny of land plants, as previously observed (Bennett 1987; Wendel et al. 2002).

The last common ancestor of all spermatophytes is estimated to have a genome of 2916.91 Mb (95% CI: 789.24 Mb to 10780.43 Mb), following an increase in genome size in Gymnosperms, with the largest genomes, as mentioned before and a last common ancestor with a genome size of 5265.78 Mb (95% CI: 1535 Mb to 18064.16 Mb). Then, there is a pattern of decrease in genome size within the Angiosperm clade, with a genome of 958.67 Mb (95% CI: 324.80 Mb to 2829.53 Mb) in their common ancestor and an average of 2457.63 Mb among the extant species of angiosperms. However, we can notice that a few lineages of angiosperms went through an increase in genome size, such as the ancestral state of the common ancestor of Asteraceae, Orchidaceae, and Solanaceae, for which the estimated genome size of their common ancestor was 2742.47 Mb (95% CI: 2575.76 Mb to 2919.96 Mb), 1675.07 Mb (95% CI: 686.5 Mb to 4086.64), and 1607.76 Mb (95% CI: 944.53 Mb to 2736.68), respectively. On the other hand, in Mosses, we observe an average of 376.53 Mb, and the common ancestor genome size estimative is 417.36 Mb (95% CI: 129.85 Mb to 1341.49 Mb). Similarly, in Liverworts the average genome size is 587.97 Mb and the estimative for the genome size of the last common ancestor is 842.49 Mb (95% CI: 131.84 Mb to 5383.54 Mb). We can notice close species from other lineages with both increasing and decreasing genome size. There are two representatives in the

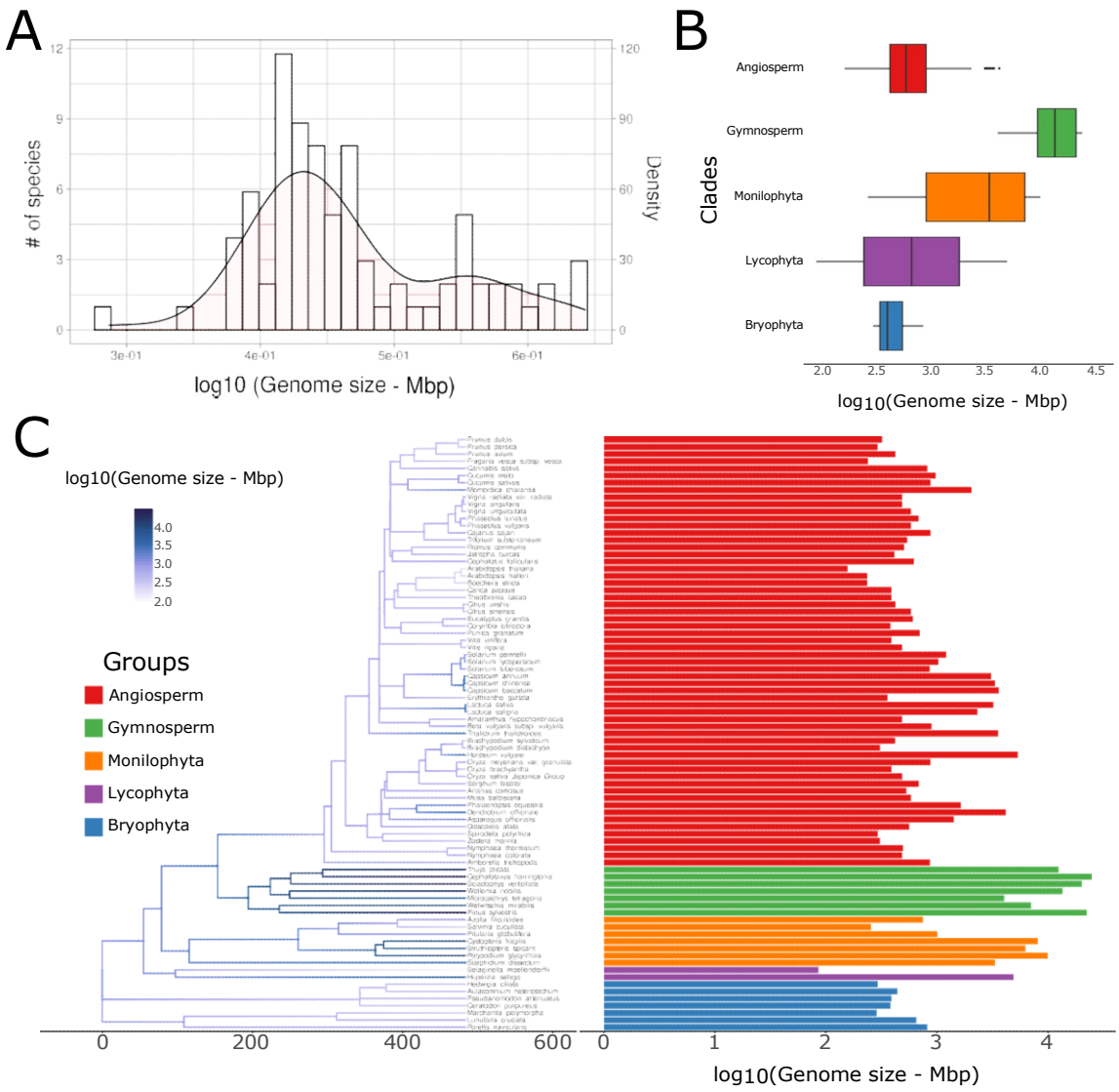


Figura 2: Evolution of genome size in land plants. A) Overall distribution of genome size in land plants. B) Variation of genome size in groups of land plants. C) Estimation of ancestral genome size values.

Lycophyta branch, *S. moellendorffii* with the smallest genome in our dataset, 86.24 Mb, and *H. selago* with 4938.9 Mb of genome size, among the largest ones.

Almost every node and branch tip in the bryophyte clades (mosses and liverworts) present a decrease in genome size, except for the *Aulacomnium heterostichum*, which has a genome of 655.26 Mb and descends from a node estimated to have a genome size of 393.17 Mb. However, both clades maintain their genome sizes very small, as described before for bryophytes (andILeitch 2005).

Every other clade has examples of branches both increasing and decreasing the genome

size, which is consistent with the literature (andILeitch 2005). Lycophytes and ferns are both very diverse clades concerning genome sizes (D. Wang et al. 2021), thus great cases to observe the variation of genome size within a clade. Although we have a small sample of lycophyte representatives, they interestingly are *Selaginella moellendorffii* (the smallest genome in our dataset) and *H. selago* (among the largest ones in our dataset). On the other hand, in ferns, it is easy to observe the different patterns of genome size evolution, while the Salvinales show a clear decrease in genome size, we can see the genome size of Polypodiales increasing, some of the largest in the dataset.

The estimated genome size for the ancestor of Spermatophyta is 2916.91 Mb and, again, we notice two different pattern behaviors. In gymnosperms, the genomes sizes are uniform within the clade (Puttick, Clark e Donoghue 2015) and increase along the phylogeny, except for *Welwitschia mirabilis* and *Microcachrys tetragona*. andILeitch (2005) also demonstrated this genome downsizing in the Gnetales, as the *W. mirabilis*, suggesting a contraction of genome size in the order (andILeitch 2005). Displaying a different trend from gymnosperms, the angiosperms show a high diversity of genome size, but mostly very small genome sizes, from its common ancestor to most of the nodes and branches within the clades (I. J. Leitch, Soltis et al. 2005). Although most genomes are very small, few derived clades evolved independently intermediate genomes, specially Solanaceae, Orchidaceae, and Asteraceae (andILeitch 2005). The patterns of genome size distribution and diversity in angiosperms could explain the success and diversity of angiosperm species, as genome size is known to be correlated (directly or indirectly) to several functional and adaptive traits (I. J. Leitch, Chase e Bennett 1998; andILeitch 2005; Puttick, Clark e Donoghue 2015; Carta et al. 2022).

Protein domains associated with genome size in land plants

We used CALANGO, an *in-house* comparative genomics tool that integrates genotypic, phenotypic and annotation data, to build phylogeny-aware linear models and search for phenotype-genotype associations, to investigate possible homologous regions associated with the variation of genome size in land plants (Hongo et al. 2021). Specifically, we used 1) the the log10(C-value) and 2) either the absolute counts or the relative frequencies (the ratio of each Pfam annotation count to the total number Pfam annotation in each proteome) of annotations terms defined by Pfam as data vectors for phenotypes

and genotypes, respectively. From a total of 5,721 Pfam IDs that annotate at least one protein-coding gene from one of the 83 genomes of land plants under analysis, we found that 7 of them (0.12%) are significantly correlated with the increase in genome size in land plants (q-value < 0.1) (Table 1, Figure 3). All of them were found in at least 90% of the studied species across the whole phylogeny of land plants (prevalence > 0.9), suggesting these are conserved homologous regions shared across most land plants in this analysis.

Tabela 1: Protein domains found significantly correlated to the genome size evolution in land plants and their respective annotations, number of copies in the two species with shortest (*Selaginella moellendorffii*, *Arabidopsis thaliana*) and largest (*Pinus sylvestris*, *Cephalotaxus harringtonia*) genomes in our dataset, the signal of the correlation, and whether the correlation is regarding the Pfam domain's abundance or frequency.

Pfam ID	**Annotation**	**Selaginella moellendorffii**	**Arabidopsis thaliana**	**Pinus sylvestris**	**Cephalotaxus harringtonia**	**Correlation**	**Experiment**
PF00684	DnaJ central domain	8	8	8	8	positive	Abundance
PF00692	dUTPase	1	1	3	2	positive	Frequency
PF01113	Dihydrodipicolinate reductase, N-terminus	2	3	1	2	negative	Abundance
PF01872	RibD C-terminal domain	2	1	2	4	positive	Abundance
PF01928	CYTH domain	4	3	5	3	positive	
PF02867	Ribonucleotide reductase, barrel domain	3	1	2	2	positive	Frequency/ Abundance
PF16211	C-terminus of histone H2A	7	12	29	8	positive	Frequency

The absolute count of 3 Pfam domains was positively correlated with the increase of genome size in land plants. We also found 4 Pfam domains with relative frequencies positively associated with the increase in genome size in land plants, and a single Pfam domain (PF02867) found to be associated when considering both absolute and relative abundancies (Supplementary Table S-2). Based on information mined from the manual curation of *A. thaliana* genes (Pasha et al. 2020), we noticed that as genome size increases the abundance and frequency of four Pfam domains found in proteins playing critical roles in DNA biology, such as nucleotide metabolism, DNA repair, and chromatin organization. Two out of the six Pfam domains with positive associations are key enzymes of nucleo-

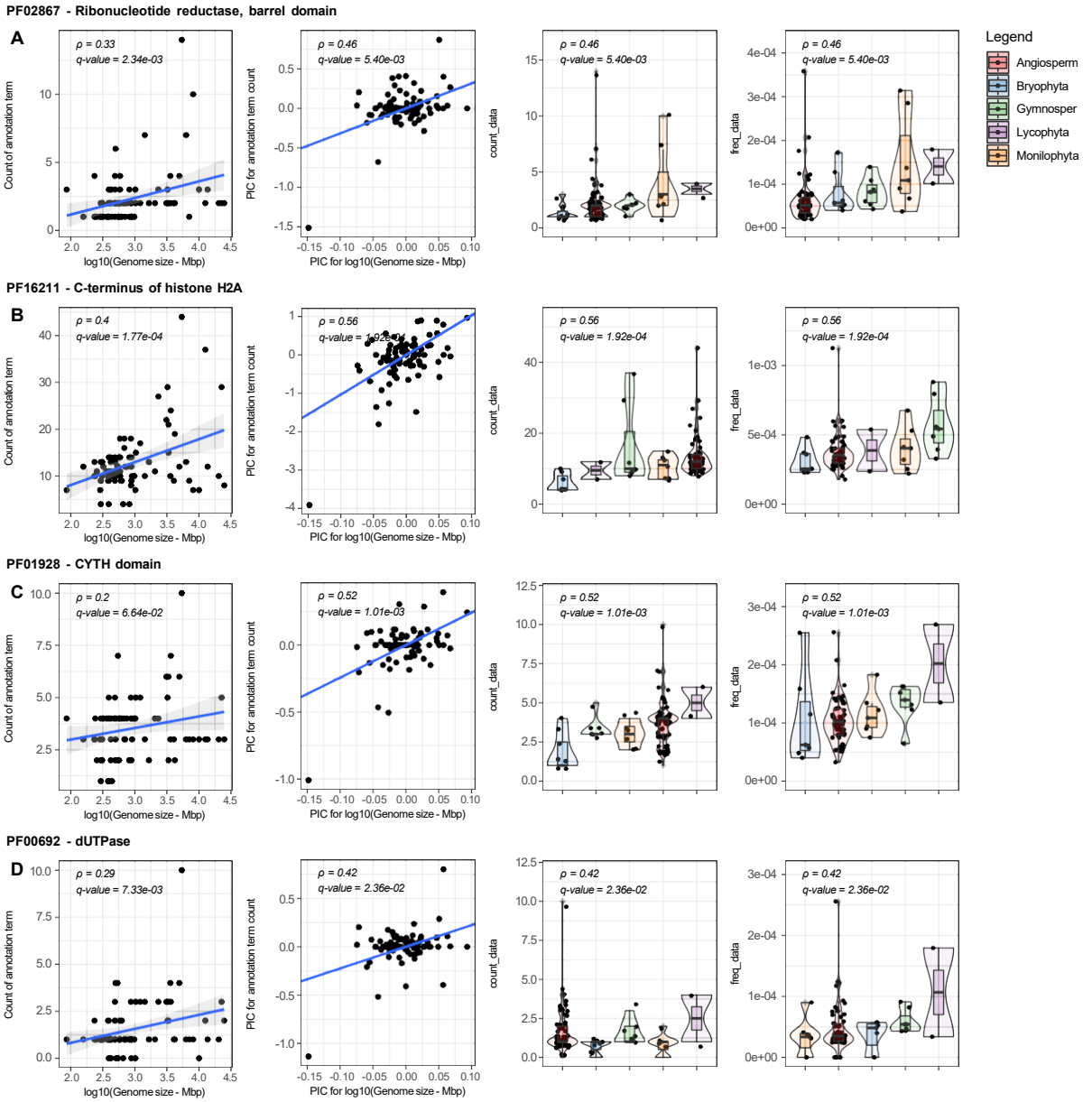


Figura 3: Protein domains with absolute/relative abundances in proteomes associated with genome size variation in land plants. A) Ribonucleotide reductase, barrel domain (PF02867); B) C-terminus of histone H2A (PF16211); C) CUTH domain (PF01928); D) dUTPase (PF00692); E) RibD C-terminal domain (PF01872); F) Dihydrodipicolinate reductase, N-terminus (PF01113); G) and DnaJ central domain (PF00684).

tide metabolism: PF02867 (Ribonucleotide reductase, barrel domain) and PF00692 (dUTPase). Ribonucleotide reductases are key enzymes for DNA synthesis as they produce deoxyribonucleotides through the reduction of ribonucleotides. Deoxyuridine triphosphatases (dUTPases) are a group of enzymes that hydrolyze dUTP (deoxyuridine monophosphate) to dUMP (deoxyuridine monophosphate) and pyrophosphate. Through this reaction,

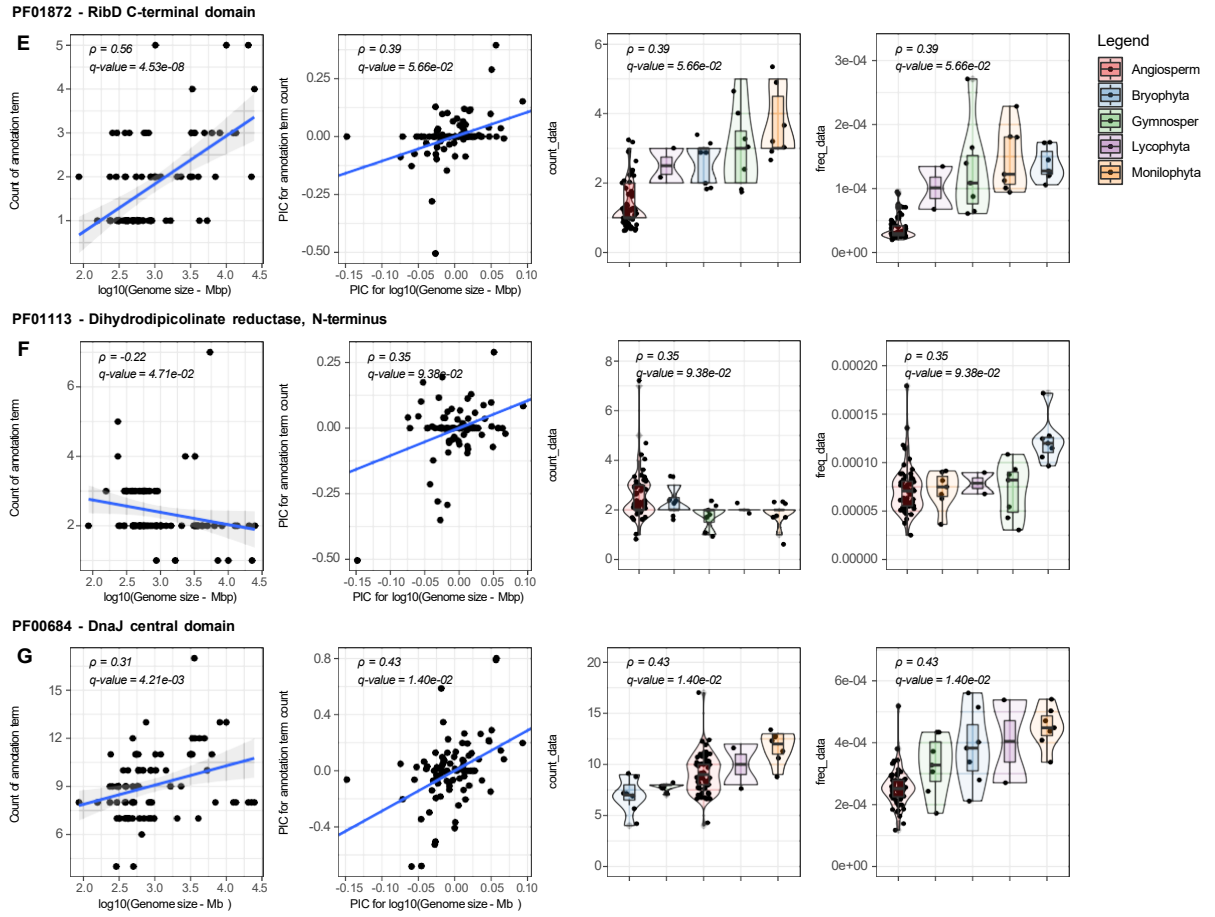


Figura 3: **CONT.** Protein domains with absolute/relative abundances in proteomes associated with genome size variation in land plants. A) Ribonucleotide reductase, barrel domain (PF02867); B) C-terminus of histone H2A (PF16211); C) CUTH domain (PF01928); D) dUTPase (PF00692); E) RibD C-terminal domain (PF01872); F) Dihydrodipicolinate reductase, N-terminus (PF01113); G) and DnaJ central domain (PF00684).

dUTPases perform two key functions for DNA metabolism and repair: 1) keep a low dUTP level in the cell, therefore preventing the misincorporation of dUTP instead of dTTP by DNA polymerase; 2) provides dUMP, a component for dTTP synthesis (Dubois et al. 2011). Therefore, the positive association of both enzymatic activities and genome size suggests these critical components of DNA metabolism and repair may play important roles for the stability of larger genomes.

The domain PF01872 (RibD C-terminal domain), also positively associated with genome size, is found in components of the riboflavin biosynthesis pathway. Riboflavin is a starting molecular scaffold for the sequential production of Flavin adenine dinucleotide (FAD) and Flavin mononucleotide (FNM). In plants, FAD and DNM are essential co-

factors required for DNA repair and also for several central metabolic pathways such as photosynthesis, mitochondrial electron transport, fatty acid oxidation and photoreception (Sandoval, Y. Zhang e Roje 2008). We hypothesize that the role of these cofactors on DNA repair may provide more genomic stability for larger genomes. Their central roles in several other major biochemical pathways may

Our results indicate that as the genome size of land plants is also associated with an increase in the frequency of the C-terminus domain of the histone H2A superfamily (PF16211). The protein variants in this superfamily contain histones responsible for nucleosome formation, regulation of DNA transcription and replication, and repair of DNA double-strand breaks (Kawashima et al. 2015). In *A. thaliana*, H2A variants are also responsible for the silencing of transposable elements (TEs) (Lei e Berger 2020). It is well accepted that divergence in genome size between species is mainly due to the abundance of TEs (andILeitch 2005). Thus, the association of H2A protein frequency and genome size could signal the evolution of a mechanism to protect genome integrity from TEs amplification and translocation. To the best of our knowledge, these associations have not been reported elsewhere, and provide compelling evidence of how ARCADE can be used to produce new, biologically relevant knowledge.

The three remaining associations of homologous regions and genome size in land plants lack clear interpretations, and may indicate previously unknown biological processes associated with genome size variation. The CYTH domain (PF01928) converts ATP to 3',5'-cyclic AMP and pyrophosphate. In *A. thaliana*, this domain is found in three triphosphate tunnel metalloenzymes, a poorly characterized group of homologs with roles in senescence and in defense response to pathogens (Ung, Moeder e Yoshioka 2014). Proteins containing Pfam PF00684 (DnaJ central domain) are co-chaperones that act together with Hsp70 proteins regulating protein folding and homeostasis in response to several biotic and abiotic stresses (Liu e Whitham 2013; Pulido e Leister 2018). The only negative association was between genome size increase and PF01113 (Dihydrodipicolinate reductase, N-terminus). This domain annotates three *A. thaliana* genes involved in the diaminopimelate biosynthetic process, a crucial molecule linking the mitochondrial electron transport chain, a key component of energetic metabolism, to amino acid catabolism and the tricarboxylic acid cycle (Cavalcanti et al. 2018).

1.4 CONCLUSION

As the number of high-quality annotated genomes for major cellular lineages increases due to both the reduction of sequencing costs and the improvements of DNA sequencing technologies, genome assembly and annotation algorithms, the challenge to extract biologically meaningful, statistically sound knowledge changes from data production to data curation and modelling (Nagy et al. 2020). ARCADE was developed to address both issues by specifically providing two major data types needed for comparative genomics: sets of genomic elements shared across species' genomes of interest and annotated to common dictionaries of biological terms and a species tree to generate phylogeny-aware models. As we demonstrated, ARCADE is the first database to provide such data types for a rich set of Archaeplastida groups not available anywhere else.

Our case study of the evolution of genome size in land plants found that species with large genome sizes have in common the presence of a higher density of genes dedicated to DNA biology. These functions may be key to preserving the stability of larger genomes. Protein variants of the histone superfamily 2A, which we found correlated with genome size increase, mark heterochromatic regions of the genome for DNA methylation silencing of TEs (Yelagandula et al. 2014). These proteins are not likely to be the reason why some plant genomes are larger, but might be a mechanism selected in lineages with larger genomes through which they cope with genome size changes. Further studies should focus on better understanding how these proteins relate to the evolution and function of plant genomes, including TEs.

1.5 DATA AVAILABILITY

All processed data needed to fully reproduce the case study (genome annotation files, phylogenetic tree, phenotypic information and CALANGO configuration files) are available at https://bit.ly/ARCADE_OSF. All raw data used in case studies (genome IDs and sources for phenotypic data) are available as supplementary tables.

1.6 ACKNOWLEDGEMENT

We would like to thank Dr. Anderson Vieira Chaves for the insightful discussions during the elaboration of this research, and Dr. Ilia J Leitch for sharing the genome size data.

1.7 FUNDING

We are grateful to the Graduate Program in Genetics, the Graduate Program in Bioinformatics, the "Centro de Processamento de Alto Desempenho/ICB"(CEPAD/SAGARANA cluster), and the Vice Dean for Research from Universidade Federal de Minas Gerais, Brazil, for the financial and computational support for this research. This work was partially funded by CAPES/Brazil (Grant 001).

1.8 CONFLICT OF INTEREST

All authors declare no conflict of interest for this publication

1.9 REFERENCES

2.0, GenomeHubs (2022). *GoaT - Genomes on a Tree*. Last Accessed: August 19th 2022.

Adams, Dean C. e Michael L. Collyer (jul. de 2017). “Multivariate Phylogenetic Comparative: Evaluations, Comparisons, and Recommendations”. Em: *Systematic Biology* 67.1, pp. 14–31. issn: 1063-5157. doi: 10.1093/sysbio/syx055.

Adl, Sina M. et al. (2012). “The Revised Classification of Eukaryotes”. Em: *Journal of Eukaryotic Microbiology* 59.5, pp. 429–514. doi: https://doi.org/10.1111/j.1550-7408.2012.00644.x.

Allaire, JJ et al. (2022). *rmarkdown: Dynamic Documents for R*. R package version 2.14.

andILeitch, Michael Bennet (2005). “CHAPTER 2 - Genome Size Evolution in Plants”. Em: *The Evolution of the Genome*. Ed. por T. Ryan Gregory. Burlington: Academic Press, pp. 89–162. isBn: 978-0-12-301463-4. doi: https://doi.org/10.1016/B978-012301463-4/50004-8.

Arndt, David et al. (mai. de 2016). “PHASTER: a better, faster version of the PHAST phage search tool”. Em: *Nucleic Acids Research* 44.W1, W16–W21. issn: 0305-1048. doi: 10.1093/nar/gkw387.

Ball, Steven et al. (jan. de 2011). “The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis”. Em: *Journal of Experimental Botany* 62.6, pp. 1775–1801. issn: 0022-0957. doi: 10.1093/jxb/erq411.

Bar-On, Yinon M., Rob Phillips e Ron Milo (2018). “The biomass distribution on Earth”. Em: *Proceedings of the National Academy of Sciences* 115.25, pp. 6506–6511. doi: 10.1073/pnas.1711842115.

Barr, Jeremy J. et al. (2013). “Bacteriophage adhering to mucus provide a non-host-derived immunity”. Em: *Proceedings of the National Academy of Sciences* 110.26, pp. 10771–10776. doi: 10.1073/pnas.1305923110.

Beaulieu, Jeremy M. et al. (2007). “Correlated evolution of genome size and seed mass”. Em: *New Phytologist* 173.2, pp. 422–437. doi: https://doi.org/10.1111/j.1469-8137.2006.01919.x.

Bennett, Michael D. (1987). “VARIATION IN GENOMIC FORM IN PLANTS AND ITS ECOLOGICAL IMPLICATIONS”. Em: *New Phytologist* 106.s1, pp. 177–200. doi: https://doi.org/10.1111/j.1469-8137.1987.tb04689.x.

Bentsink, Leónie et al. (2006). “Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis”. Em: *Proceedings of the National Academy of Sciences* 103.45, pp. 17042–17047. doi: 10.1073/pnas.0607877103.

Berardini, Tanya Z. et al. (2015). “The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome”. Em: *genesis* 53.8, pp. 474–485. doi: https://doi.org/10.1002/dvg.22877.

Bertozzi Silva, Juliano, Zachary Storms e Dominic Sauvageau (jan. de 2016). “Host receptors for bacteriophage adsorption”. Em: *FEMS Microbiology Letters* 363.4. fnw002. issn: 0378-1097. doi: 10.1093/femsle/fnw002.

Blázquez, Miguel A., David C. Nelson e Dolf Weijers (2020). “Evolution of Plant Hormone Response Pathways”. Em: *Annual Review of Plant Biology* 71.1. PMID: 32017604, pp. 327–353. doi: 10.1146/annurev-arplant-050718-100309.

Buchfink, Benjamin, Chao Xie e Daniel H. Huson (jan. de 2015). “Fast and sensitive protein alignment using DIAMOND”. Em: *Nature Methods* 12.1, pp. 59–60. issn: 1548-7105. doi: 10.1038/nmeth.3176.

Carrillo-Barral, Néstor, María del Carmen Rodríguez-Gacio e Angel Jesús Matilla (2020). “Delay of Germination-1 (DOG1): A Key to Understanding Seed Dormancy”. Em: *Plants* 9.4. issn: 2223-7747. doi: 10.3390/plants9040480.

Carta, Angelino et al. (2022). “Correlated evolution of seed mass and genome size varies among life forms in flowering plants”. Em: *Seed Science Research* 32.1, pp. 46–52. doi: 10.1017/S0960258522000071.

Cavalcanti, João Henrique F et al. (set. de 2018). “An L,L-diaminopimelate aminotransferase mutation leads to metabolic shifts and growth inhibition in Arabidopsis”. Em: *Journal of Experimental Botany* 69.22, pp. 5489–5506. issn: 0022-0957. doi: 10.1093/jxb/ery325.

Challis, Richard J. et al. (mai. de 2017). “GenomeHubs: simple containerized setup of a custom Ensembl database and web server for any species”. Em: *Database* 2017. bax039. issn: 1758-0463. doi: 10.1093/database/bax039.

Chamberlain, Scott A. e Eduard Szöcs (2013). “taxize: taxonomic search and retrieval in R”. Em: *F1000 Research* 2, p. 191. doi: 10.12688/f1000research.2-191.v1.

Chan, Patricia P. e Todd M. Lowe (dez. de 2015). “GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes”. Em: *Nucleic Acids Research* 44.D1, pp. D184–D189. issn: 0305-1048. doi: 10.1093/nar/gkv1309.

Chanderbali, Andre S et al. (mar. de 2016). “Evolving Ideas on the Origin and Evolution of Flowers: New Perspectives in the Genomic Era”. Em: *Genetics* 202.4, pp. 1255–1265. issn: 1943-2631. doi: 10.1534/genetics.115.182964.

Chaudhuri, Roy R. e Ian R. Henderson (2012). “The evolution of the Escherichia coli phylogeny”. Em: *Infection, Genetics and Evolution* 12.2, pp. 214–226. issn: 1567-1348. doi: https://doi.org/10.1016/j.meegid.2012.01.005.

Chen, Haixu et al. (nov. de 2021). “BRAD V3.0: an upgraded Brassicaceae database”. Em: *Nucleic Acids Research* 50.D1, pp. D1432–D1441. issn: 0305-1048. doi: 10.1093/nar/gkab1057.

Chen, Lu et al. (mar. de 2014). “Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity”. Em:

Molecular Biology and Evolution 31.6, pp. 1402–1413. issn: 0737-4038. doi: 10.1093/molbev/msu083.

Cheng, Joe et al. (2021). *htmltools: Tools for HTML*. R package version 0.5.2.

Cirillo, D M et al. (1996). "Identification of a domain in Rck, a product of the Salmonella typhimurium virulence plasmid, required for both serum resistance and cell invasion". Em: *Infection and Immunity* 64.6, pp. 2019–2023. doi: 10.1128/iai.64.6.2019-2023.1996.

Coghlan, Avril et al. (jan. de 2019). "Comparative genomics of the major parasitic worms". Em: *Nature Genetics* 51.1, pp. 163–174. issn: 1546-1718. doi: 10.1038/s41588-018-0262-1.

Cornwell, Will e Shinichi Nakagawa (2017). "Phylogenetic comparative methods". Em: *Current Biology* 27.9, R333–R336. issn: 0960-9822. doi: https://doi.org/10.1016/j.cub.2017.03.049.

Correa, Adrienne M. S. et al. (ago. de 2021). "Revisiting the rules of life for viruses of microorganisms". Em: *Nature Reviews Microbiology* 19.8, pp. 501–513. issn: 1740-1534. doi: 10.1038/s41579-021-00530-x.

Cotter, Paul D., R. Paul Ross e Colin Hill (fev. de 2013). "Bacteriocins — a viable alternative to antibiotics?" Em: *Nature Reviews Microbiology* 11.2, pp. 95–105. issn: 1740-1534. doi: 10.1038/nrmicro2937.

Crooks, Gavin E. et al. (2004). "WebLogo: A Sequence Logo Generator". Em: *Genome Research* 14.6, pp. 1188–1190. doi: 10.1101/gr.849004.

Cunningham, Fiona et al. (nov. de 2021). "Ensembl 2022". Em: *Nucleic Acids Research* 50.D1, pp. D988–D995. issn: 0305-1048. doi: 10.1093/nar/gkab1049.

Dedrick, Rebekah M. et al. (jan. de 2017). "Prophage-mediated defence against viral attack and viral counter-defence". Em: *Nature Microbiology* 2.3, p. 16251. issn: 2058-5276. doi: 10.1038/nmicrobiol.2016.251.

Dekkers, Bas J.W. et al. (2016). "The Arabidopsis DELAY OF GERMINATION 1 gene affects ABSCISIC ACID INSENSITIVE 5 (ABI5) expression and genetically interacts with ABI3 during Arabidopsis seed development". Em: *The Plant Journal* 85.4, pp. 451–465. doi: https://doi.org/10.1111/tpj.13118.

Dobritsa, Anna A. e Daniel Coerper (nov. de 2012). "The Novel Plant Protein INAPERTURATE POLLEN1 Marks Distinct Cellular Domains and Controls Formation of

Apertures in the Arabidopsis Pollen Exine ". Em: *The Plant Cell* 24.11, pp. 4452–4464. issn: 1040-4651. doi: 10.1105/tpc.112.101220.

Dong, Qunfeng, Shannon D. Schlueter e Volker Brendel (jan. de 2004). "PlantGDB, plant genome database and analysis tools". Em: *Nucleic Acids Research* 32.suppl_1, pp. D354–D359. issn: 0305-1048. doi: 10.1093/nar/gkh046.

Dubois, Emeline et al. (abr. de 2011). "Homologous Recombination Is Stimulated by a Decrease in dUTPase in Arabidopsis". Em: *PLOS ONE* 6.4, pp. 1–8. doi: 10.1371/journal.pone.0018658.

Dufayard, Jean-François et al. (2017). "New Insights on Leucine-Rich Repeats Receptor-Like Kinase Orthologous Relationships in Angiosperms". Em: *Frontiers in Plant Science* 8, p. 381. doi: 10.3389/fpls.2017.00381.

Dunn, Casey W. e Catriona Munro (2016). "Comparative genomics and the diversity of life". Em: *Zoologica Scripta* 45.S1, pp. 5–13. doi: https://doi.org/10.1111/zsc.12211.

Durand, Eléonore et al. (2020). "Evolution of self-incompatibility in the Brassicaceae: Lessons from a textbook example of natural selection". Em: *Evolutionary Applications* 13.6, pp. 1279–1297. doi: https://doi.org/10.1111/eva.12933.

Eddy, Sean R. (out. de 2011). "Accelerated Profile HMM Searches". Em: *PLOS Computational Biology* 7.10, pp. 1–16. doi: 10.1371/journal.pcbi.1002195.

Edgar, Robert C. (mar. de 2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". Em: *Nucleic Acids Research* 32.5, pp. 1792–1797. issn: 0305-1048. doi: 10.1093/nar/gkh340.

Ehrbar, Kristin e Wolf-Dietrich Hardt (2005). "Bacteriophage-encoded type III effectors in Salmonella enterica subspecies 1 serovar Typhimurium". Em: *Infection, Genetics and Evolution* 5.1, pp. 1–9. issn: 1567-1348. doi: https://doi.org/10.1016/j.meegid.2004.07.004.

Ekstrom, Alexander et al. (ago. de 2014). "PlantCAZyme: a database for plant carbohydrate-active enzymes". Em: *Database* 2014. bau079. issn: 1758-0463. doi: 10.1093/database/bau079.

Falster, Daniel S. e Mark Westoby (2003). "Plant height and evolutionary games". Em: *Trends in Ecology & Evolution* 18.7, pp. 337–343. issn: 0169-5347. doi: https://doi.org/10.1016/S0169-5347(03)00061-2.

Fedak, Halina et al. (2016). “Control of seed dormancy in Arabidopsis by a cis-acting noncoding antisense transcript”. Em: *Proceedings of the National Academy of Sciences* 113.48, E7846–E7855. doi: 10.1073/pnas.1608827113.

Felsenstein, Joseph (1985). “Phylogenies and the Comparative Method”. Em: *The American Naturalist* 125.1, pp. 1–15. issn: 00030147, 15375323.

Fernández, Lucía, Ana Rodríguez e Pilar García (mai. de 2018). “Phage or foe: an insight into the impact of viral predation on microbial communities”. Em: *The ISME Journal* 12.5, pp. 1171–1179. issn: 1751-7370. doi: 10.1038/s41396-018-0049-5.

Fischer, Steve et al. (2011). “Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups”. Em: *Current Protocols in Bioinformatics* 35.1, pp. 6.12.1–6.12.19. doi: https://doi.org/10.1002/0471250953.bi0612s35.

Fleischmann, Andreas et al. (out. de 2014). “Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms”. Em: *Annals of Botany* 114.8, pp. 1651–1663. issn: 0305-7364. doi: 10.1093/aob/mcu189.

Fu, Limin et al. (out. de 2012). “CD-HIT: accelerated for clustering the next-generation sequencing data”. Em: *Bioinformatics* 28.23, pp. 3150–3152. issn: 1367-4803. doi: 10.1093/bioinformatics/bts565.

Galili, Tal (2015). “dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering”. Em: *Bioinformatics*. doi: 10.1093/bioinformatics/btv428.

Galili, Tal et al. (2017). “heatmaply: an R package for creating interactive cluster heatmaps for online publishing”. Em: *Bioinformatics*. doi: 10.1093/bioinformatics/btx657.

El-Gebali, Sara et al. (out. de 2018). “The Pfam protein families database in 2019”. Em: *Nucleic Acids Research* 47.D1, pp. D427–D432. issn: 0305-1048. doi: 10.1093/nar/gky995.

Goffeau, A. et al. (1996). “Life with 6000 Genes”. Em: *Science* 274.5287, pp. 546–567. doi: 10.1126/science.274.5287.546.

González-Morales, Sandra Isabel et al. (2016). “Regulatory network analysis reveals novel regulators of seed desiccation tolerance in Arabidopsis thaliana”. Em: *Proceedings*

of the National Academy of Sciences 113.35, E5232–E5241. doi: 10 .1073 / pnas . 1610985113.

Goodstein, David M. et al. (nov. de 2011). "Phytozome: a comparative platform for green plant genomics". Em: *Nucleic Acids Research* 40.D1, pp. D1178–D1186. issn: 0305-1048. doi: 10.1093/nar/gkr944.

Gordillo Altamirano, Fernando et al. (fev. de 2021). "Bacteriophage-resistant Acinetobacter baumannii are resensitized to antimicrobials". Em: *Nature Microbiology* 6.2, pp. 157–161. issn: 2058-5276. doi: 10.1038/s41564-020-00830-7.

Granzotto, Adriana e Guilherme Marcello Queiroga Cruz (2015). "Regulação de Elementos de Transposição: Mecanismos Epigenéticos de Silenciamento, Autorregulação e Ativação por Estresse". Em: *Elementos de transposição: diversidade, evolução, aplicações e impacto nos genomas dos seres vivos*. Ed. por Claudia Marcia Aparecida Carareto, Claudia Barros Monteiro-Vitorello e Marie-Anne Van Sluys. São José do Rio Preto: Editora FIOCRUZ, pp. 91–113. isBn: 978-85-7541-462-0. doi: https://doi.org/10.7476/9788575415672.

Greilhuber, Johann e I J. Leitch (2013). "Genome Size and the Phenotype". Em: *Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes*. Ed. por Johann Greilhuber, Jaroslav Dolezel e Jonathan F. Wendel. Vienna: Springer Vienna, pp. 323–344. isBn: 978-3-7091-1160-4. doi: 10.1007/978-3-7091-1160-4_20.

Groth, Philip et al. (set. de 2006). "PhenomicDB: a new cross-species genotype/phenotype resource". Em: *Nucleic Acids Research* 35.suppl_1, pp. D696–D699. issn: 0305-1048. doi: 10.1093/nar/gkl662.

Harvey, Paul H, Mark D Pagel et al. (1991). *The comparative method in evolutionary biology*. Vol. 239. Oxford university press Oxford.

Haynes, Winston A., Aurelie Tomczak e Purvesh Khatri (jan. de 2018). "Gene annotation bias impedes biomedical research". Em: *Scientific Reports* 8.1, p. 1362. issn: 2045-2322. doi: 10.1038/s41598-018-19333-x.

Heather, James M. e Benjamin Chain (2016). "The sequence of sequencers: The history of sequencing DNA". Em: *Genomics* 107.1, pp. 1–8. issn: 0888-7543. doi: https://doi.org/10.1016/j.ygeno.2015.11.003.

Hidalgo, Oriane et al. (2017). “Is There an Upper Limit to Genome Size?” Em: *Trends in Plant Science* 22.7, pp. 567–573. issn: 1360-1385. doi: https://doi.org/10.1016/j.tplants.2017.04.005.

Hongo, Jorge Augusto et al. (2021). “CALANGO: an annotation-based, phylogeny-aware comparative genomics framework for exploring and interpreting complex genotypes and phenotypes”. Em: *bioRxiv*. doi: 10.1101/2021.08.25.457574.

Hung, Jui-Hung et al. (set. de 2011). “Gene set enrichment analysis: performance evaluation and usage guidelines”. Em: *Briefings in Bioinformatics* 13.3, pp. 281–291. issn: 1467-5463. doi: 10.1093/bib/bbr049.

Huo, Heqiang, Shouhui Wei e Kent J. Bradford (2016). “DELAY OF GERMINATION1 (DOG1) regulates both seed dormancy and flowering time through microRNA pathways”. Em: *Proceedings of the National Academy of Sciences* 113.15, E2199–E2206. doi: 10.1073/pnas.1600558113.

IHGSC et al. (fev. de 2001). “Initial sequencing and analysis of the human genome”. Em: *Nature* 409.6822, pp. 860–921. issn: 1476-4687. doi: 10.1038/35057062.

Initiative, The Arabidopsis Genome (dez. de 2000). “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana”. Em: *Nature* 408.6814, pp. 796–815. issn: 1476-4687. doi: 10.1038/35048692.

Jones, Philip et al. (jan. de 2014). “InterProScan 5: genome-scale protein function classification”. Em: *Bioinformatics* 30.9, pp. 1236–1240. issn: 1367-4803. doi: 10.1093/bioinformatics/btu031.

Kang, Ming et al. (2014). “Adaptive and nonadaptive genome size evolution in Karst endemic flora of China”. Em: *New Phytologist* 202.4, pp. 1371–1381. doi: https://doi.org/10.1111/nph.12726.

Kattge, Jens et al. (2020). “TRY plant trait database – enhanced coverage and open access”. Em: *Global Change Biology* 26.1, pp. 119–188. doi: https://doi.org/10.1111/gcb.14904.

Kawahara, Yoshihiro et al. (fev. de 2013). “Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data”. Em: *Rice* 6.1, p. 4. issn: 1939-8433. doi: 10.1186/1939-8433-6-4.

Kawashima, Tomokazu et al. (jul. de 2015). “Diversification of histone H2A variants during plant evolution”. Em: *Trends in Plant Science* 20.7, pp. 419–425. issn: 1360-1385. doi: 10.1016/j.tplants.2015.04.005.

Knight, Charles A., Nicole A. Molinari e Dmitri A. Petrov (jan. de 2005). “The Large Genome Constraint Hypothesis: Evolution, Ecology and Phenotype”. Em: *Annals of Botany* 95.1, pp. 177–190. issn: 0305-7364. doi: 10.1093/aob/mci011.

Koornneef, Maarten, Leónie Bentsink e Henk Hilhorst (2002). “Seed dormancy and germination”. Em: *Current Opinion in Plant Biology* 5.1, pp. 33–36. issn: 1369-5266. doi: https://doi.org/10.1016/S1369-5266(01)00219-9.

Kopriva, Stanislav e Andreas P M Weber (jan. de 2021). “Genetic encoding of complex traits”. Em: *Journal of Experimental Botany* 72.1, pp. 1–3. issn: 0022-0957. doi: 10.1093/jxb/eraa498.

Krishnakumar, Vivek et al. (nov. de 2014). “Araport: the Arabidopsis Information Portal”. Em: *Nucleic Acids Research* 43.D1, pp. D1003–D1009. issn: 0305-1048. doi: 10.1093/nar/gku1200.

Kuang, Kevin, Quyu Kong e Francesco Napolitano (2022). *pbmcapply: Tracking the Progress of Mc*pply with Progress Bar*. R package version 1.5.1.

Kumar, Sudhir, Glen Stecher et al. (mai. de 2018). “MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms”. Em: *Molecular Biology and Evolution* 35.6, pp. 1547–1549. issn: 0737-4038. doi: 10.1093/molbev/msy096.

Kumar, Sudhir, Michael Suleski et al. (ago. de 2022). “TimeTree 5: An Expanded Resource for Species Divergence Times”. Em: *Molecular Biology and Evolution* 39.8. msac174. issn: 1537-1719. doi: 10.1093/molbev/msac174.

Kumar, Vikash, Evgeniy N. Donev et al. (2020). “Genome-Wide Identification of Populus Malectin/Malectin-Like Domain-Containing Proteins and Expression Analyses Reveal Novel Candidates for Signaling and Regulation of Wood Development”. Em: *Frontiers in Plant Science* 11, p. 588846. doi: 10.3389/fpls.2020.588846.

Kumar, Vikash, Matthieu Hainaut et al. (2019). “Poplar carbohydrate-active enzymes: whole-genome annotation and functional analyses based on RNA expression data”. Em: *The Plant Journal* 99.4, pp. 589–609. doi: https://doi.org/10.1111/tpj.14417.

Lanfear, Robert et al. (mai. de 2013). “Taller plants have lower rates of molecular evolution”. Em: *Nature Communications* 4.1, p. 1879. issn: 2041-1723. doi: 10.1038/ncomms2836.

Lee, Byung Ha et al. (jul. de 2021). “A species-specific functional module controls formation of pollen apertures”. Em: *Nature Plants* 7.7, pp. 966–978. issn: 2055-0278. doi: 10.1038/s41477-021-00951-9.

Lee, Heewook et al. (2012). “Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing”. Em: *Proceedings of the National Academy of Sciences* 109.41, E2774–E2783. doi: 10.1073/pnas.1210309109.

Lei, Bingkun e Frédéric Berger (2020). “H2A Variants in Arabidopsis: Versatile Regulators of Genome Activity”. Em: *Plant Communications* 1.1, p. 100015. issn: 2590-3462. doi: https://doi.org/10.1016/j.xplc.2019.100015.

Leitch, A. R. e I. J. Leitch (2012). “Ecological and genetic factors linked to contrasting genome dynamics in seed plants”. Em: *New Phytologist* 194.3, pp. 629–646. doi: https://doi.org/10.1111/j.1469-8137.2012.04105.x.

Leitch, I. J., Mark W. Chase e Michael D. Bennett (dez. de 1998). “Phylogenetic Analysis of DNA C-values Provides Evidence for a Small Ancestral Genome Size in Flowering Plants”. Em: *Annals of Botany* 82.suppl_1, pp. 85–94. issn: 0305-7364. doi: 10.1006/anbo.1998.0783.

Leitch, I. J., D. E. Soltis et al. (jan. de 2005). “Evolution of DNA Amounts Across Land Plants (Embryophyta)”. Em: *Annals of Botany* 95.1, pp. 207–217. issn: 0305-7364. doi: 10.1093/aob/mci014.

León, M e R Bastías (2015). “Virulence reduction in bacteriophage resistant bacteria.” Em: *Frontiers in Microbiology* 343.6. doi: http://dx.doi.org/10.3389/fmicb.2015.00343.

Li, Fay-Wei et al. (jul. de 2018). “Fern genomes elucidate land plant evolution and cyanobacterial symbioses”. Em: *Nature Plants* 4.7, pp. 460–472. issn: 2055-0278. doi: 10.1038/s41477-018-0188-8.

Li, Linzhou et al. (set. de 2020). “The genome of Prasinoderma coloniale unveils the existence of a third phylum within green plants”. Em: *Nature Ecology &, Evolution* 4.9, pp. 1220–1231. issn: 2397-334X. doi: 10.1038/s41559-020-1221-7.

Li, Weizhong e Adam Godzik (mai. de 2006). “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”. Em: *Bioinformatics* 22.13, pp. 1658–1659. issn: 1367-4803. doi: 10.1093/bioinformatics/btl158.

Liolios, Konstantinos et al. (nov. de 2009). “The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata”. Em: *Nucleic Acids Research* 38.suppl_1, pp. D346–D354. issn: 0305-1048. doi: 10.1093/nar/gkp848.

Lisch, Damon (jan. de 2013). “How important are transposons for plant evolution?” Em: *Nature Reviews Genetics* 14.1, pp. 49–61. issn: 1471-0064. doi: 10.1038/nrg3374.

Liu, Jian-Zhong e Steven A. Whitham (2013). “Overexpression of a soybean nuclear localized type–III DnaJ domain-containing HSP40 reveals its roles in cell death and disease resistance”. Em: *The Plant Journal* 74.1, pp. 110–121. doi: https:// doi.org/10.1111/tpj.12108.

M, Carlson (2019). *GO.db: A set of annotation maps describing the entire Gene Ontology*. R package version 3.8.2.

Ma, Xuelian et al. (set. de 2022). “PlantGSAD: a comprehensive gene set annotation database for plant species”. Em: *Nucleic Acids Research* 50.D1, pp. D1456–D1467. issn: 0305-1048. doi: 10.1093/nar/gkab794.

Manni, Mosè et al. (jul. de 2021). “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes”. Em: *Molecular Biology and Evolution* 38.10, pp. 4647–4654. issn: 1537-1719. doi: 10.1093/molbev/msab199.

Marks, Rose A. et al. (dez. de 2021). “Representation and participation across 20 years of plant genome sequencing”. Em: *Nature Plants* 7.12, pp. 1571–1578. issn: 2055-0278. doi: 10.1038/s41477-021-01031-8.

Mashau, Aluoneswi C. et al. (2021). “Plant height and lifespan predict range size in southern African grasses”. Em: *Journal of Biogeography* 48.12, pp. 3047–3059. doi: https://doi.org/10.1111/jbi.14261.

Maslov, Sergei e Kim Sneppen (jan. de 2017). “Population cycles and species diversity in dynamic Kill-the-Winner model of microbial ecosystems”. Em: *Scientific Reports* 7.1, p. 39642. issn: 2045-2322. doi: 10.1038/srep39642.

Mazuecos-Aguilera, Ismael et al. (2021). “The Role of INAPERTURATE POLLEN1 as a Pollen Aperture Factor Is Conserved in the Basal Eudicot Eschscholzia californica (Papaveraceae)”. Em: *Frontiers in Plant Science* 12. issn: 1664-462X. doi: 10.3389/fpls.2021.701286.

Minelli, Alessandro (2018). “Introducing Plant Evo-Devo”. Em: *Plant Evolutionary Developmental Biology: The Evolvability of the Phenotype*. Cambridge University Press, pp. 1–29. doi: 10.1017/9781139542364.002.

Minh, Bui Quang et al. (fev. de 2020). “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era”. Em: *Molecular Biology and Evolution* 37.5, pp. 1530–1534. issn: 0737-4038. doi: 10.1093/molbev/msaa015.

Moles, Angela T. et al. (2009). “Global patterns in plant height”. Em: *Journal of Ecology* 97.5, pp. 923–932. doi: https://doi.org/10.1111/j.1365-2745.2009.01526.x.

Morgan, Martin (2022). *BiocManager: Access the Bioconductor Project Package Repository*. R package version 1.30.18.

Mosavi, Leila K. et al. (2004). “The ankyrin repeat as molecular architecture for protein recognition”. Em: *Protein Science* 13.6, pp. 1435–1448. doi: https://doi.org/10.1110/ps.03554604.

Mukherjee, Supratim et al. (out. de 2016). “Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements”. Em: *Nucleic Acids Research* 45.D1, pp. D446–D456. issn: 0305-1048. doi: 10.1093/nar/gkw992.

Nagy, László G et al. (jan. de 2020). “Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing”. Em: *Nucleic Acids Research* 48.5, pp. 2209–2219. issn: 0305-1048. doi: 10.1093/nar/gkz1241.

Nakabayashi, Kazumi et al. (jul. de 2012). “The Time Required for Dormancy Release in Arabidopsis Is Determined by DELAY OF GERMINATION1 Protein Levels in Freshly Harvested Seeds”. Em: *The Plant Cell* 24.7, pp. 2826–2838. issn: 1040-4651. doi: 10.1105/tpc.112.100214.

Nasrallah, June B. e Mikhail E. Nasrallah (mar. de 2014). “S-locus receptor kinase signalling”. Em: *Biochemical Society Transactions* 42.2, pp. 313–319. issn: 0300-5127. doi: 10.1042/BST20130222.

Niklas, Karl J. e Ulrich Kutschera (2010). “The evolution of the land plant life cycle”. Em: *New Phytologist* 185.1, pp. 27–41. doi: https://doi.org/10.1111/j.1469-8137.2009.03054.x.

Nishimura, Noriyuki et al. (jun. de 2018). “Control of seed dormancy and germination by DOG1-AHG1 PP2C phosphatase complex via binding to heme”. Em: *Nature Communications* 9.1, p. 2132. issn: 2041-1723. doi: 10.1038/s41467-018-04437-9.

Nishiyama, Eri et al. (2021). “Ancient and recent gene duplications as evolutionary drivers of the seed maturation regulators DELAY OF GERMINATION1 family genes”. Em: *New Phytologist* 230.3, pp. 889–901. doi: https://doi.org/10.1111/nph.17201.

Nishiyama, Takashi et al. (jan. de 2013). “The structure of the deacetylase domain of Escherichia coli PgaB, an enzyme required for biofilm formation: a circularly permuted member of the carbohydrate esterase 4 family”. Em: *Acta Crystallographica Section D* 69.1, pp. 44–51. doi: 10.1107/S0907444912042059.

O’Leary, Nuala A. et al. (nov. de 2015). “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. Em: *Nucleic Acids Research* 44.D1, pp. D733–D745. issn: 0305-1048. doi: 10.1093/nar/gkv1189.

Pagès, H et al. (2022). *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. R package version 1.58.0.

Pang, Shuai et al. (mai. de 2015). “GPU MrBayes V3.1: MrBayes on Graphics Processing Units for Protein Sequence Data”. Em: *Molecular Biology and Evolution* 32.9, pp. 2496–2497. issn: 0737-4038. doi: 10.1093/molbev/msv129.

Paradis, E. e K. Schliep (2019). “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R”. Em: *Bioinformatics* 35, pp. 526–528.

Park, Beom Seok e Jie-Oh Lee (dez. de 2013). “Recognition of lipopolysaccharide pattern by TLR4 complexes”. Em: *Experimental & Molecular Medicine* 45.12, e66–e66. issn: 2092-6413. doi: 10.1038/emm.2013.97.

Pasha, Asher et al. (jul. de 2020). “Araport Lives: An Updated Framework for Arabidopsis Bioinformatics”. Em: *The Plant Cell* 32.9, pp. 2683–2686. issn: 1040-4651. doi: 10.1105/tpc.20.00358.

Pawluk, April et al. (2014). “A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa”. Em: *mBio* 5.2, e00896–14. doi: 10.1128/mBio.00896-14.

Peiffer, Jason A et al. (abr. de 2014). “The Genetic Architecture Of Maize Height”. Em: *Genetics* 196.4, pp. 1337–1356. issn: 1943-2631. doi: 10.1534/genetics.113.159152.

Pellicer, Jaume, Michae F. Fay e I. J. Leitch (set. de 2010). “The largest eukaryotic genome of them all?” Em: *Botanical Journal of the Linnean Society* 164.1, pp. 10–15. issn: 0024-4074. doi: 10.1111/j.1095-8339.2010.01072.x.

Pellicer, Jaume, Oriane Hidalgo et al. (2018). “Genome Size Diversity and Its Impact on the Evolution of Land Plants”. Em: *Genes* 9.2. issn: 2073-4425. doi: 10.3390/genes9020088.

Pellicer, Jaume e I J. Leitch (2020). “The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies”. Em: *New Phytologist* 226.2, pp. 301–305. doi: https://doi.org/10.1111/nph.16261.

Petrov, Dmitri A. (jan. de 2001). “Evolution of genome size: new approaches to an old problem”. Em: *Trends in Genetics* 17.1, pp. 23–28. issn: 0168-9525. doi: 10.1016/S0168-9525(00)02157-0.

— (2002). “Mutational Equilibrium Model of Genome Size Evolution”. Em: *Theoretical Population Biology* 61.4, pp. 531–544. issn: 0040-5809. doi: https://doi.org/10.1006/tpbi.2002.1605.

Pinard, Desre et al. (mai. de 2015). “Comparative analysis of plant carbohydrate active enZymes and their role in xylogenesis”. Em: *BMC Genomics* 16.1, p. 402. issn: 1471-2164. doi: 10.1186/s12864-015-1571-8.

Pinheiro, José, Douglas Bates e R Core Team (2022). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-157.

Plazzi, Federico, Ronald R. Ferrucci e Marco Passamonti (abr. de 2010). “Phylogenetic representativeness: a new method for evaluating taxon sampling in evolutionary studies”. Em: *BMC Bioinformatics* 11.1, p. 209. issn: 1471-2105. doi: 10.1186/1471-2105-11-209.

Proost, Sebastian et al. (out. de 2014). “PLAZA 3.0: an access point for plant comparative genomics”. Em: *Nucleic Acids Research* 43.D1, pp. D974–D981. issn: 0305-1048. doi: 10.1093/nar/gku986.

Pulido, Pablo e Dario Leister (2018). “Novel DNAJ-related proteins in Arabidopsis thaliana”. Em: *The New Phytologist* 217.2, pp. 480–490. issn: 0028646X, 14698137.

Pulkkinen, W S e S I Miller (1991). “A Salmonella typhimurium virulence protein is similar to a Yersinia enterocolitica invasion protein and a bacteriophage lambda outer membrane protein”. Em: *Journal of Bacteriology* 173.1, pp. 86–93. doi: 10.1128/jb.173.1.86-93.1991.

Puttick, Mark N., James Clark e Philip C. J. Donoghue (2015). “Size is not everything: rates of genome size evolution, not C-value, correlate with speciation in angiosperms”. Em: *Proceedings of the Royal Society B: Biological Sciences* 282.1820, p. 20152289. doi: 10.1098/rspb.2015.2289.

Rambaut, Andrew et al. (abr. de 2018). “Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7”. Em: *Systematic Biology* 67.5, pp. 901–904. issn: 1063-5157. doi: 10.1093/sysbio/syy032.

Ramisetty, Bhaskar Chandra Mohan e Pavithra Anantharaman Sudhakari (2019). “Bacterial ‘Grounded’ Prophages: Hotspots for Genetic Renovation and Innovation”. Em: *Frontiers in Genetics* 10. issn: 1664-8021. doi: 10.3389/fgene.2019.00065.

Ren, Ren et al. (2018). “Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms”. Em: *Molecular Plant* 11.3. Genome Biology, pp. 414–428. issn: 1674-2052. doi: https://doi.org/10.1016/j.molp.2018.01.002.

Revell, Liam J. (2012). “phytools: an R package for phylogenetic comparative biology (and other things)”. Em: *Methods in Ecology and Evolution* 3.2, pp. 217–223. doi: https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Roff, Derek A. (1997). *Evolutionary Quantitative Genetics*. New York: Springer New York. isBn: 978-1-4615-4080-9. doi: https://doi.org/10.1007/978-1-4615-4080-9.

Sall, Khadidiatou et al. (2019). “DELAY OF GERMINATION 1-LIKE 4 acts as an inducer of seed reserve accumulation”. Em: *The Plant Journal* 100.1, pp. 7–19. doi: https://doi.org/10.1111/tpj.14485.

Salzberg, Steven L. (mai. de 2019). “Next-generation genome annotation: we still struggle to get it right”. Em: *Genome Biology* 20.1, p. 92. issn: 1474-760X. doi: 10.1186/s13059-019-1715-2.

Sandoval, Francisco J., Yi Zhang e Sanja Roje (nov. de 2008). “Flavin Nucleotide Metabolism in Plants: MONOFUNCTIONAL ENZYMES SYNTHESIZE FAD IN PLASTIDS

*". Em: *Journal of Biological Chemistry* 283.45, pp. 30890–30900. issn: 0021-9258. doi: 10.1074/jbc.M803416200.

Sanger, F. e A.R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". Em: *Journal of Molecular Biology* 94.3, pp. 441–448. issn: 0022-2836. doi: https://doi.org/10.1016/0022-2836(75)90213-2.

Sayers, Eric W et al. (out. de 2019). "GenBank". Em: *Nucleic Acids Research* 48.D1, pp. D84–D86. issn: 0305-1048. doi: 10.1093/nar/gkz956.

Schäffer, Alejandro A. et al. (jul. de 2001). "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements". Em: *Nucleic Acids Research* 29.14, pp. 2994–3005. issn: 0305-1048. doi: 10.1093/nar/29.14.2994.

Schallus, Thomas et al. (2008). "Malectin: A Novel Carbohydrate-binding Protein of the Endoplasmic Reticulum and a Candidate Player in the Early Steps of Protein N-Glycosylation". Em: *Molecular Biology of the Cell* 19.8. PMID: 18524852, pp. 3404–3414. doi: 10.1091/mbc.e08-04-0354.

Schneider, Rene e Staffan Persson (2015). "Another brick in the wall". Em: *Science* 350.6257, pp. 156–157. doi: 10.1126/science.aad3200.

Schuster, Stephan C. (jan. de 2008). "Next-generation sequencing transforms today's biology". Em: *Nature Methods* 5.1, pp. 16–18. issn: 1548-7105. doi: 10.1038/nmeth1156.

SHAPIRO, S. S. e M. B. WILK (dez. de 1965). "An analysis of variance test for normality (complete samples)". Em: *Biometrika* 52.3-4, pp. 591–611. doi: 10.1093/biomet/52.3-4.591.

Sievert, Carson (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. isbn: 978-3-319-24277-4.

Silveira, Cynthia B. e Forest L. Rohwer (jul. de 2016). "Piggyback-the-Winner in host-associated microbial communities". Em: *npj Biofilms and Microbiomes* 2.1, p. 16010. issn: 2055-5008. doi: 10.1038/npjbiofilms.2016.10.

Simmons, Emilia L. et al. (2020). "Biofilm Structure Promotes Coexistence of Phage-Resistant and Phage-Susceptible Bacteria". Em: *mSystems* 5.3, e00877–19. doi: 10.1128/mSystems.00877-19.

Sørensen, Iben et al. (2011). “The charophycean green algae provide insights into the early origins of plant cell walls”. Em: *The Plant Journal* 68.2, pp. 201–211. doi: https://doi.org/10.1111/j.1365-313X.2011.04686.x.

Steyert, Susan R. e James B. Kaper (2012). “Contribution of Urease to Colonization by Shiga Toxin-Producing Escherichia coli”. Em: *Infection and Immunity* 80.8, pp. 2589–2600. doi: 10.1128/IAI.00210-12.

Subburaj, Saminathan et al. (jun. de 2016). “Phylogenetic Analysis, Lineage-Specific Expansion and Functional Divergence of seed dormancy 4-Like Genes in Plants”. Em: *PLOS ONE* 11.6, pp. 1–24. doi: 10.1371/journal.pone.0153717.

Tello-Ruiz, Marcela K et al. (nov. de 2020). “Gramene 2021: harnessing the power of comparative genomics and pathways for plant research”. Em: *Nucleic Acids Research* 49.D1, pp. D1452–D1463. issn: 0305-1048. doi: 10.1093/nar/gkaa979.

Tenenbaum D, Maintainer B (2022). *KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)*. R package version 1.36.3.

Tomaž, Špela, Kristina Gruden e Anna Coll (2022). “TGA transcription factors—Structural characteristics as basis for functional variability”. Em: *Frontiers in Plant Science* 13. issn: 1664-462X. doi: 10.3389/fpls.2022.935819.

Tong, Chao et al. (jan. de 2020). “Comparative Genomics Identifies Putative Signatures of Sociality in Spiders”. Em: *Genome Biology and Evolution* 12.3, pp. 122–133. issn: 1759-6653. doi: 10.1093/gbe/evaa007.

Touchon, Marie et al. (jan. de 2009). “Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths”. Em: *PLOS Genetics* 5.1, pp. 1–25. doi: 10.1371/journal.pgen.1000344.

Tuskan, G. A. et al. (2006). “The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray)”. Em: *Science* 313.5793, pp. 1596–1604. doi: 10.1126/science.1128691.

Ung, Huoi, Wolfgang Moeder e Keiko Yoshioka (set. de 2014). “Arabidopsis Triphosphate Tunnel Metalloenzyme2 Is a Negative Regulator of the Salicylic Acid-Mediated Feedback Amplification Loop for Defense Responses ”. Em: *Plant Physiology* 166.2, pp. 1009–1021. issn: 0032-0889. doi: 10.1104/pp.114.248757.

Vaidya, Gaurav, David J. Lohman e Rudolf Meier (2011). “SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon

information". Em: *Cladistics* 27.2, pp. 171–180. doi: https://doi.org/10.1111/j.1096-0031.2010.00329.x.

Vaidyanathan, Ramnath et al. (2021). *htmlwidgets: HTML Widgets for R*. R package version 1.5.4.

Vandecraen, Joachim et al. (2017). "The impact of insertion sequences on bacterial genome plasticity and adaptability". Em: *Critical Reviews in Microbiology* 43.6. PMID: 28407717, pp. 709–730. doi: 10.1080/1040841X.2017.1303661.

Veselý, Pavel, Petr Bureš e Petr Šmarda (ago. de 2013). "Nutrient reserves may allow for genome size increase: evidence from comparison of geophytes and their sister non-geophytic relatives". Em: *Annals of Botany* 112.6, pp. 1193–1200. issn: 0305-7364. doi: 10.1093/aob/mct185.

Vinogradov, Alexander E (2003). "Selfish DNA is maladaptive: evidence from the plant Red List". Em: *Trends in Genetics* 19.11, pp. 609–614. issn: 0168-9525. doi: https://doi.org/10.1016/j.tig.2003.09.010.

Vitti, Joseph J., Sharon R. Grossman e Pardis C. Sabeti (2013). "Detecting Natural Selection in Genomic Data". Em: *Annual Review of Genetics* 47.1. PMID: 24274750, pp. 97–120. doi: 10.1146/annurev-genet-111212-133526.

Vogel, Christine e Cyrus Chothia (mai. de 2006). "Protein Family Expansions and Biological Complexity". Em: *PLOS Computational Biology* 2.5, pp. 1–13. doi: 10.1371/journal.pcbi.0020048.

Wang, B et al. (2019). "[The China National GeneBank owned by all, completed by all and shared by all]". Em: *Yi Chuan* 20.41, pp. 761–772. doi: 10.16288/j.yczz..

Wang, Dandan et al. (2021). "Which factors contribute most to genome size variation within angiosperms?" Em: *Ecology and Evolution* 11.6, pp. 2660–2668. doi: https://doi.org/10.1002/ece3.7222.

Wang, Xiaoxue et al. (dez. de 2010). "Cryptic prophages help bacteria cope with adverse environments". Em: *Nature Communications* 1.1, p. 147. issn: 2041-1723. doi: 10.1038/ncomms1146.

Waterhouse, Robert M et al. (dez. de 2017). "BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics". Em: *Molecular Biology and Evolution* 35.3, pp. 543–548. issn: 0737-4038. doi: 10.1093/molbev/msx319.

Wendel, Jonathan F. et al. (mai. de 2002). “Feast and famine in plant genomes”. Em: *Genetica* 115.1, pp. 37–47. issn: 1573-6857. doi: 10.1023/A:1016020030189.

Wickham, Hadley (2019). *assertthat: Easy Pre and Post Assertions*. R package version 0.2.1.

— (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman & Hall. isbn: 9781138331457.

Wickham, Hadley, Jay Hesselberth e Maëlle Salmon (2022). *pkgdown: Make Static HTML Documentation for a Package*. R package version 2.0.3.

Willi, Yvone e Ary A. Hoffman (2009). “Demographic factors and genetic variation influence population persistence under environmental change”. Em: *Journal of Evolutionary Biology* 22.1, pp. 124–133. doi: https : / / doi . org / 10 . 1111 / j . 1420 - 9101.2008.01631.x.

Wolf, Andrea J. e David M. Underhill (abr. de 2018). “Peptidoglycan recognition by the innate immune system”. Em: *Nature Reviews Immunology* 18.4, pp. 243–254. issn: 1474-1741. doi: 10.1038/nri.2017.136.

Wolf, Jason B. (2002). “The geometry of phenotypic evolution in developmental hyperspace”. Em: *Proceedings of the National Academy of Sciences* 99.25, pp. 15849–15851. doi: 10.1073/pnas.012686699.

Xiao, Yu et al. (mai. de 2019). “Mechanisms of RALF peptide perception by a heterotypic receptor complex”. Em: *Nature* 572.7768, pp. 270–274. issn: 1476-4687. doi: 10.1038/s41586-019-1409-7.

Xie, Yihui, Joe Cheng e Xianying Tan (2022). *DT: A Wrapper of the JavaScript Library ’DataTables’*. R package version 0.23.

Xue, Han et al. (out. de 2021). “qPTMplants: an integrative database of quantitative post-translational modifications in plants”. Em: *Nucleic Acids Research* 50.D1, pp. D1491–D1499. issn: 0305-1048. doi: 10.1093/nar/gkab945.

Yang, He et al. (2021). “Malectin/Malectin-like domain-containing proteins: A repertoire of cell surface molecules with broad functional potential”. Em: *The Cell Surface* 7, p. 100056. issn: 2468-2330. doi: https://doi.org/10.1016/j.tcsw.2021.100056.

Yang, Xiaohan et al. (set. de 2019). “Comparative genomics can provide new insights into the evolutionary mechanisms and gene function in CAM plants”. Em: *Journal*

of Experimental Botany 70.22, pp. 6539–6547. issn: 0022-0957. doi: 10.1093/jxb/erz408.

Yelagandula, Ramesh et al. (jul. de 2014). "The Histone Variant H2A.W Defines Heterochromatin and Promotes Chromatin Condensation in Arabidopsis". Em: *Cell* 158.1, pp. 98–109. issn: 0092-8674. doi: 10.1016/j.cell.2014.06.006.

Zhang, Jian et al. (fev. de 2020). "The hornwort genome and early land plant evolution". Em: *Nature Plants* 6.2, pp. 107–118. issn: 2055-0278. doi: 10.1038/s41477-019-0588-4.

Zu, Pengjuan e Florian P. Schiestl (2017). "The effects of becoming taller: direct and pleiotropic effects of artificial selection on plant height in Brassica rapa". Em: *The Plant Journal* 89.5, pp. 1009–1019. doi: https://doi.org/10.1111/tpj.13440.

Zwickl, Derrick J. e David M. Hillis (jul. de 2002). "Increased Taxon Sampling Greatly Reduces Phylogenetic Error". Em: *Systematic Biology* 51.4, pp. 588–598. issn: 1063-5157. doi: 10.1080/10635150290102339.

1.10 SUPPLEMENTARY MATERIALS

Fig. 1 full caption: Description of ARCADE database. A) Relative species richness and taxonomic diversity of the plant genome databases used to build ARCADE. B) Species tree for 142 species found in ARCADE (extracted from the TimeTree web tool (S. Kumar, Suleski et al. 2022) as described in "Methods". Species abbreviations are as follows: *Acer yangbiense* (Aya), *Akebia trifoliata* (Atr), *Amaranthus hypochondriacus* (Ahy), *Amborella trichopoda* (Ati), *Ananas comosus* (Aco), *Andreaea rupestris* (Aru), *Anthoceros angustus* (Aan), *Arabidopsis halleri* (Aha), *Arabidopsis thaliana* (Ath), *Arabis nemorensis* (Ane), *Ascarina rubricaulis* (Arb), *Asparagus officinalis* (Aof), *Asplenium platyneuron* (Apl), *Aulacomnium heterostichum* (Ahe), *Azolla filiculoides* (Afi), *Beta vulgaris subsp. vulgaris* (Bvu), *Boechera stricta* (Bst), *Brachypodium distachyon* (Bdi), *Brachypodium sylvaticum* (Bsy), *Buxbaumia aphylla* (Bap), *Cajanus cajan* (Cca), *Cannabis sativa* (Csa), *Capsicum annuum* (Can), *Capsicum baccatum* (Cba), *Capsicum chinense* (Cch), *Carex littledalei* (Cli), *Carica papaya* (Cpa), *Castanea mollissima* (Cmo), *Cephalotaxus harringtonia* (Cha), *Cephalotus follicularis* (Cfo), *Ceratodon purpureus* (Cpu), *Chara braunii* (Cbr), *Chlamydomonas eustigma* (Ceu), *Chlamydomonas reinhardtii* (Cre), *Chlorella variabilis* (Cva), *Cinnamomum kanehirae* (Cka), *Cinnamomum micranthum* (Cmi), *Citrus clementina* (Ccl), *Citrus sinensis* (Csi), *Citrus unshiu* (Cun), *Cleome violacea* (Cvi), *Coleochaete irregularis* (Cir), *Corymbia citriodora* (Cci), *Cucumis melo* (Cme), *Cucumis sativus* (Cst), *Cuscuta australis* (Cau), *Cystopteris fragilis* (Cfr), *Danaea nodosa* (Dno), *Dendrobium officinale* (Dof), *Descurainia sophioides* (Dso), *Dioscorea alata* (Dal), *Diphyscium foliosum* (Dfo), *Diplazium wichurae* (Dwi), *Dorcoceras hygrometricum* (Dhy), *Erythranthe guttata* (Egt), *Eucalyptus grandis* (Egr), *Fragaria vesca subsp. vesca* (Fve), *Frullania spp.* (Fru), *Galdieria sulphuraria* (Gsu), *Gloeochaete wittrockiana* (Gwi), *Gnetum montanum* (Gmo), *Hakea prostrata* (Hpo), *Hedwigia ciliata* (Hci), *Hordeum vulgare* (Hvu), *Huperzia selago* (Hse), *Hypecoum procumbens* (Hpr), *Illicium floridanum* (Ifl), *Isoetes tegetiformans* (Ite), *Jatropha curcas* (Jcu), *Kingdonia uniflora* (Kun), *Lactuca saligna* (Lsa), *Lactuca sativa* (Lst), *Leucobryum albidum* (Lal), *Leucostegia immersa* (Lim), *Lonchitis hirsuta* (Lhi), *Lunularia cruciata* (Lcr), *Macleaya cordata* (Mco), *Marchantia polymorpha* (Mpo), *Microcachrys tetragona* (Mte), *Micromonas pusilla CCMP1545* (Mpu), *Momordica charantia* (Mch), *Morus notabilis* (Mno), *Musa balbisiana* (Mba), *Nymphaea colorata* (Nco), *Nymphaea thermarum* (Nth), *Oryza brachyantha* (Obr), *Oryza meyeriana var. granulata*

(Ome), *Oryza sativa* (Osa), *Ostreococcus tauri* (Ota), *Pellia neesiana* (Pne), *Phalaenopsis equestris* (Peq), *Phaseolus lunatus* (Plu), *Phaseolus vulgaris* (Pvu), *Pilularia globulifera* (Pgl), *Pinus sylvestris* (Psy), *Plenasium javanicum* (Pja), *Podophyllum peltatum* (Ppe), *Polypodium glycyrrhiza* (Pgy), *Porella navicularis* (Pna), *Prunus avium* (Pav), *Prunus dulcis* (Pdu), *Prunus mume* (Pmu), *Prunus persica* (Ppr), *Pseudanomodon attenuatus* (Pat), *Pulvigera lyellii* (Ply), *Punica granatum* (Pgr), *Rhamnella rubrinervis* (Rru), *Rhododendron williamsianum* (Rwi), *Ricciocarpos natans* (Rna), *Ricinus communis* (Rco), *Salvinia cucullata* (Scu), *Sceptridium dissectum* (Sdi), *Sciadopitys verticillata* (Sve), *Selaginella moellendorffii* (Smo), *Solanum lycopersicum* (Sly), *Solanum pennellii* (Spe), *Solanum tuberosum* (Stu), *Sorghum bicolor* (Sbi), *Spinacia oleracea* (Sol), *Spirodela polyrhiza* (Spo), *Struthiopteris spicant* (Ssp), *Syzygium oleosum* (Soe), *Takakia lepidozioides* (Tle), *Thalictrum thalictroides* (Tth), *Theobroma cacao* (Tca), *Thuja plicata* (Tpl), *Timmia austriaca* (Tau), *Trema orientale* (Tor), *Trifolium subterraneum* (Tsu), *Vigna angularis* (Van), *Vigna radiata var. radiata* (Vra), *Vigna unguiculata* (Vun), *Vitis riparia* (Vri), *Vitis vinifera* (Vvi), *Vittaria appalachiana* (Vap), *Volvox carteri f. nagariensis* (Vca), *Welwitschia mirabilis* (Wmi), *Wollemia nobilis* (Wno), and *Zostera marina* (Zmr). C) Characterization of the 15 InterPro databases used to annotate ARCADE sequence data. Plots represent information as follows: Diversity/abundance – the number of entries

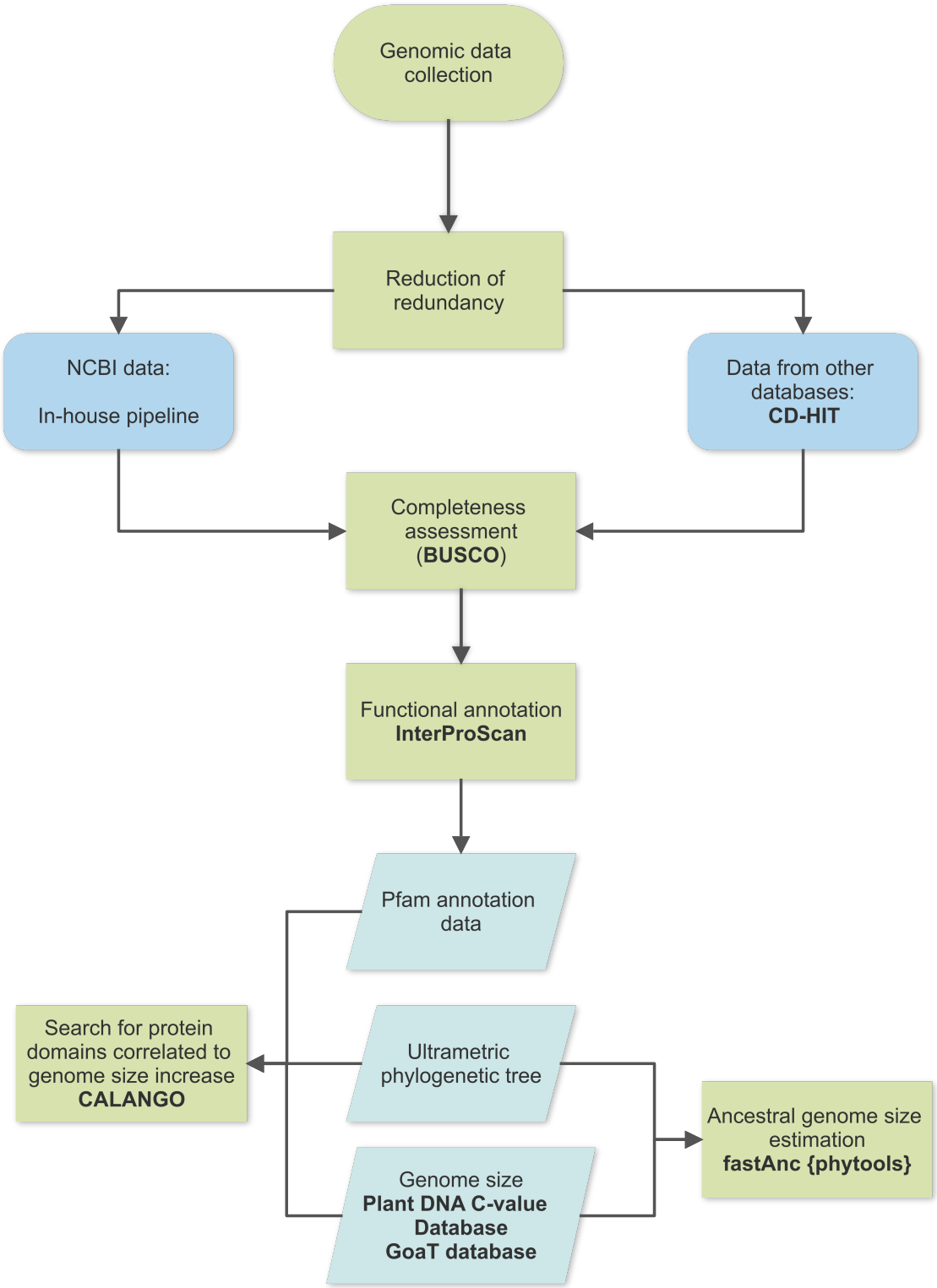


Figura S-1: Flowchart illustrating a summarized view of the procedures performed in this project.

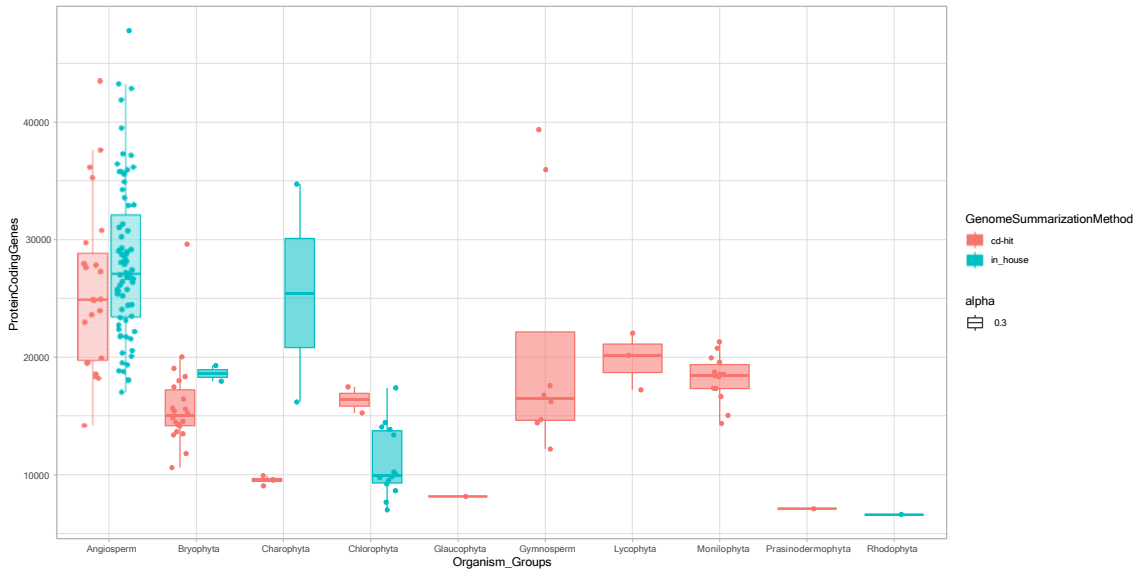


Figura S-2: Distribution of protein coding genes in the major Archaeplastida clades after each one of the methods we use for redundancy reductions (CD-HIT (W. Li e Godzik 2006; Fu et al. 2012) and our in-house pipeline).

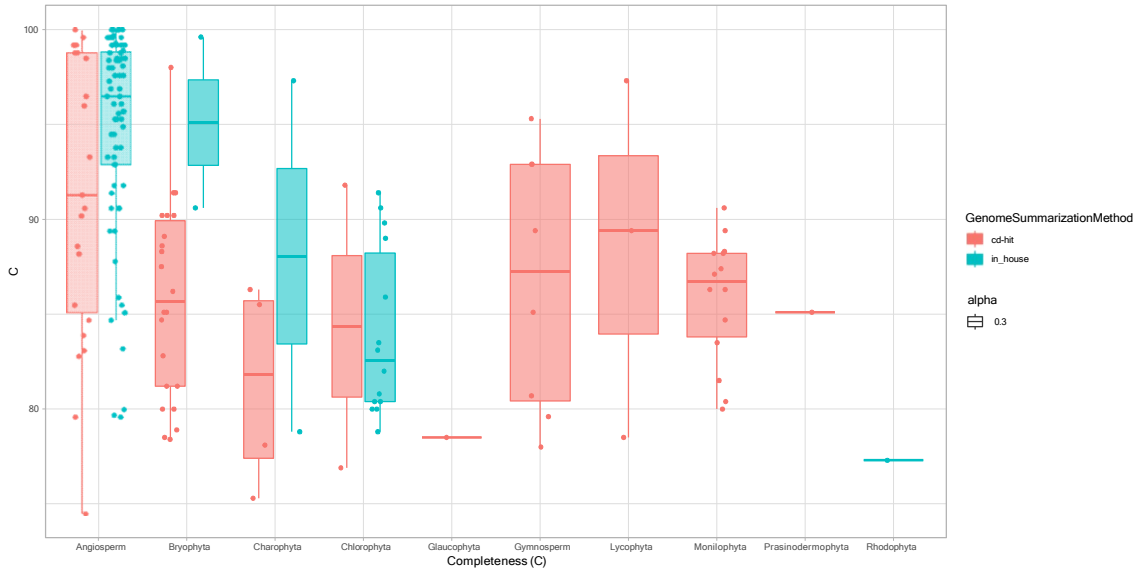


Figura S-3: Comparison of CD-HIT (W. Li e Godzik 2006; Fu et al. 2012) and our in-house method for redundancy reduction according to BUSCO's (Manni et al. 2021) gene completeness (C).

ANEXO B - CALANGO: A PHYLOGENY-AWARE COMPARATIVE GENOMICS TOOL FOR DISCOVERING QUANTITATIVE GENOTYPE-PHENOTYPE ASSOCIATIONS

CALANGO: a phylogeny-aware comparative genomics tool for discovering quantitative genotype-phenotype associations

Jorge Augusto Hongo[1,a] Giovanni Marques de Castro[2,a]
Alison Pelri Albuquerque Menezes[2,a] Agnello César Rios Picorelli[2] Thieres Tayroni Martins da Silva[2] Eddie Luidy Imada[3] Luigi Marchionni[3] Luiz-Eduardo Del-Bem[2] Anderson Vieira Chaves[2] Gabriel Magno de Freitas Almeida[4] Felipe Campelo [5] Francisco Pereira Lobo[2,*]

[1] *Instituto de Computação, Universidade Estadual de Campinas, Campinas, SP, 13083-872, Brazil.*

[2] *Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil.*

[3] *Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, 10021, USA.*

[4] *Faculty of Biosciences, Fisheries and Economics, Norwegian College of Fishery Science, UiT The Arctic University of Norway, Tromsø, 9019, Norway.*

[5] *As*ton University, Birmingham, B4 7ET, UK .

[a] *Joint authors. * To whom correspondence should be addressed. Tel: +55 31 34093072; Fax: +55 31 34092567; Email: franciscolobo@ufmg.br, franciscolobo@gmail.com*

The increasing availability of genomic, annotation, evolutionary and phenotypic data for species contrasts with the lack of studies that adequately integrate these heterogeneous data sources to produce biologically meaningful knowledge. Here, we present CALANGO, a phylogeny-aware comparative genomics tool that uncovers functional molecular convergences and homologous regions associated with quantitative genotypes and phenotypes across species, enabling the fast discovery of novel statistically sound, biologically relevant phenotype-genotype associations. We demonstrate the usefulness of CALANGO in two case studies. The first one unveils potential causal links between prophage density and the pathogenicity phenotype in *Escherichia coli*, and confidently demonstrates how

CALANGO supports the investigation of basic causal relationships by enabling a level of counterfactual investigation of observed associations in the data. As a second case study, we used our tool to search for homologous regions associated with a complex phenotypic trait in a major group of eukaryotes: the evolution of maximum height in angiosperms. We confidently identify a previously unknown association between maximum plant height and the expansion of the self-incompatibility system, a molecular mechanism that prevents inbreeding and increases genetic diversity. Taller species also have lower rates of molecular evolution due to their longer generation times, a critical concern for their long-term viability. The new mechanism we report could counterbalance this fact, and have far-reaching consequences for fields as diverse as conservation biology and agriculture. CALANGO is provided as a fully operational R package that can be freely installed from CRAN.

KEYWORDS: comparative genomics, evolution of quantitative phenotypes, genotype-phenotype association, comparative methods, molecular functional convergence.

ABBREVIATIONS: QVAL - quantitative values across lineages, DUF – domain of unknown function, SRK - S-locus receptor kinase, SCR - S-locus cysteine-rich protein.

2.1 INTRODUCTION

Living species display a wide range of quantitative variation across both their phenotypes and genomic content. Vast amounts of data describing phenotypic variation and high-quality genomes are available as databases and other structured and unstructured data sources (Groth et al. 2006; Liolios et al. 2009). The main bottleneck for the extraction of biologically meaningful knowledge from phenotypic and genomic differences across species, therefore, no longer lies in obtaining such data, but instead on analyzing these heterogeneous data types in a biologically meaningful manner and under a comparative and evolutionary genomics framework to gain insights into the putative genomic mechanisms associated with the evolution of complex quantitative phenotypes (Nagy et al. 2020).

A major challenge when developing data-modeling schema and statistical workflows to aggregate these data sources is to account for controllable sources of biases and errors, such as data dependencies arising from common ancestry (Cornwell e Nakagawa 2017). Furthermore, most comparative genomics strategies rely exclusively on patterns

of variation of homologous genomic elements across genomes as the basic unit of comparison. However, this approach fails to capture molecular functional convergences of non-homologous genomic elements fulfilling the same biological function (Coghlan et al. 2019; Tong et al. 2020). From the computational and statistical perspectives, there is a clear need of a tool capable of not only correcting for multiple hypothesis testing (Hung et al. 2011), but also mitigating frequent biases in genomic data arising from usual bioinformatics procedures such as genome assembly, gene prediction and annotation (Waterhouse et al. 2017).

In this article we present CALANGO (*Comparative AnaLysis with ANnotation-based Genomic cOmponents*), a first-principles, general comparative genomics tool designed to account for the aforementioned issues while searching for association between quantitative phenotype/genotypes in distinct species or lineages, and the abundance of annotation terms of their genomic components (Fig. 1). These annotations may reflect both homologous regions, as in traditional comparative genomics studies, or molecular convergences based on, *e.g.*, Gene Ontology (GO) terms, which can be used to detect molecular functional convergences in the emergence of complex phenotypes such as sociality and parasitism (Coghlan et al. 2019; Tong et al. 2020).

We validated CALANGO using two complementary case studies that differ in major aspects, such as evolutionary time, taxonomy, and biological phenomena under analysis. The first one comprises the analysis of the biological interaction of integrated bacteriophages (prophages) and *Escherichia coli* lineages, using the density of prophages as a proxy variable. The second evaluated the variation of a complex phenotype in a major group of eukaryotes, namely the evolution of plant height in angiosperms, a key trait for the ecology, physiology and evolution of this group (Mashau et al. 2021; Zu e Schiestl 2017).

CALANGO is available as an open-source R package, which can be installed directly from CRAN and also from the project website (https://labpackages.github.io/CALANGO/), where usage examples and long-format documentation can be found. CALANGO outputs interactive HTML5 reports, which facilitate sharing and fast communication of results and integration with existing bioinformatics pipelines.

2.2 MATERIAL AND METHODS

2.2.1 Genomic data modeling

CALANGO is designed to process distinct classes of genomic components, such as protein-coding genes, protein domains and promoter regions, as well as their respective annotations, to extract biologically meaningful patterns of variation (Fig. S-1A). Our tool explicitly models a genome and the biological functions coded within it by using two associated concepts. The first is a genomic component, defined as a common element type that can be discriminated in the group of genomes under analysis (Fig. S-1A; genomic component IDs in this case could be protein-coding genes, their translated protein sequences, promoters, or protein domains coded within a genome). The second is a set of controlled terms used to annotate the genomic components, which are expected to reflect some specific biological aspect thought to help answer the biological question under investigation, *e.g.*, Pfam domain IDs or GO terms (Fig. S-1A, annotation term IDs). By explicitly dissociating genomic components from their annotation terms, CALANGO makes it possible to use annotation schemas designed to represent the functional similarities of genomic components that may not share a common ancestor, enabling comparative genomics at the function level.

2.2.2 Input data

Genomic features, annotation, and dictionary files are simple tabular files containing textual information used to describe genomic elements and their annotations (Fig. 1A).

- **Annotation/dictionary data**: a Perl script (*calanguize_genomes.pl*) that parses Genbank files into high-quality genomic annotation data compatible with CALANGO input data (Fig. 1B) is distributed as part of the package. The script performs the following steps:

A) Downloads genomic data for the species/individual to be analyzed; B) Extracts the protein-coding genes described; C) Provides a single coding sequence per locus, reporting only the longest coding sequence per locus to avoid possible biases introduced due to the larger number of isoforms described for model organisms (Vogel e Chothia 2006); D) Executes BUSCO for genome completeness evaluation (Waterhouse et al. 2017); E) Annotates all valid protein sequences using InterProScan (Jones et al. 2014); F) Generates

CALANGO-compatible files for annotated genomes and dictionaries of annotation terms;

- **Phylogenetic tree data**: CALANGO currently supports fully dichotomous, ultrametric trees in the newick or nexus formats (Fig. 1C). Trees with multichotomies are converted into a dichotomous tree with branches of length zero.

- **Metadata**: CALANGO expects a metadata file containing QVAL values, groups for heatmap and boxplot visualization, normalization factors, and other information needed for proper execution (Fig. 1A).

2.2.3 Analyzing data using CALANGO

CALANGO starts its analysis with (i) a set of genomic components from distinct genomes annotated using a common, controlled set of annotation terms; (ii) a dictionary file defining each annotation term in a biologically meaningful way; and (iii) a metadata file containing genome-centered information, such as values for optional normalization of annotation count values in each genome (*e.g.* total count of annotation terms per genome), and the quantitative phenotype/genotype used to rank genomes (Fig. 1A). Users must also provide (iv) an ultrametric phylogenetic tree containing all lineages in a given analysis, which allows CALANGO to correct for phylogeny-related dependencies in the values of annotation terms in distinct genomes and QVAL values (Felsenstein 1985).

Based on these input files, CALANGO computes different classes of association statistics between the phenotype and annotation terms: three commonly used correlation statistics (Pearson's, Spearman's, and Kendall's correlation values and corresponding p-values) and a phylogeny-aware linear model constructed using Phylogenetic Independent Contrasts (PICs). To account for the multiple hypothesis scenario of simultaneously searching for associations between QVAL and thousands of annotation terms, CALANGO reports FDR-corrected q-values for each association statistic.

CALANGO additionally computes the variance / standard deviation of annotation term counts and two customized statistics that summarize how abundant an annotation term is and how frequently it is observed in distinct genomes: the sum of annotation terms and their prevalence (fraction of genomes where an annotation term is observed).

Two main output structures are provided (Fig. 1C). The first is a list-type object containing all computed results, which can be used to survey specific downstream hypotheses. This object also contains all input parameters used to generate the results,

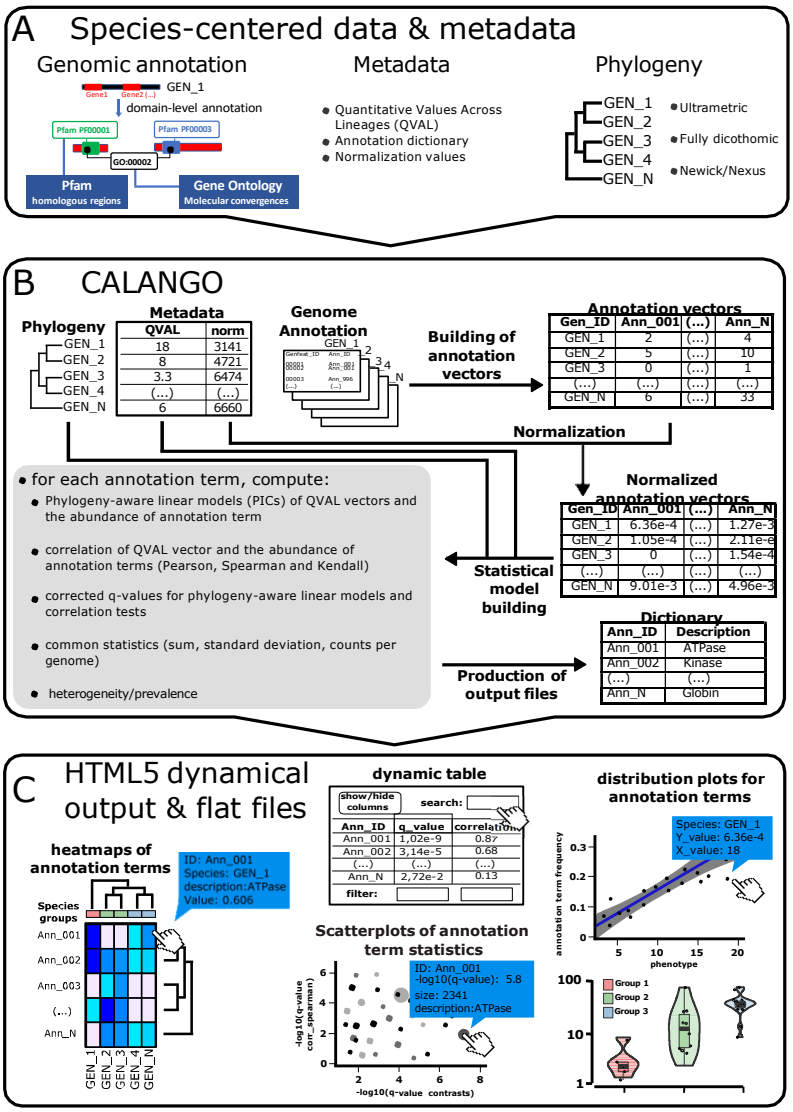


Figura 1: General structure of CALANGO. A) Species-centric input data types needed to run CALANGO: a set of genomic elements from distinct species and their associated annotation data, phylogenetic relationships across species, and the quantitative phenotype/genotype to be surveyed. B) CALANGO execution starts by reading input data and building of annotation vectors (number of occurrences of annotation terms per genome). It can perform optional data normalization to account for relevant variations in genomic content (e.g., genomes with vastly distinct numbers of protein-coding genes). The annotation vectors are then used, together with the quantitative phenotype/genotype, to build phylogeny-aware statistical models and other useful quantities to evaluate possible associations. C) CALANGO outputs dynamical HTML5 reports, which can be accessed using any HTML5-enabled web browser and contain graphical and interactive data representations and summaries, including tables, heatmaps, scatterplots and boxplots highlighting specific aspects of the associations detected.

therefore providing a simple and convenient way to share results as well as all necessary parameters required for full reproducibility.

The second output is a full interactive HTML5 report / website that can be easily shared, hosted online or browsed locally using any modern web browser. The CALANGO outputs were designed to facilitate more transparent reporting of results and sharing of raw data and code. This user-friendly output facilitates the critical evaluation of all statistics provided by CALANGO in a dynamic tabular and graphical manner.

Four kinds of interactive results are provided by the tool. The first is a biclustered heatmap based on annotation terms (clustered based on their values) and species under analysis (clustered according to the user-provided phylogenetic tree), which allows easy inspection of annotation term distribution across phylogenetic groups and refinement of questions based on interactive exploration of the graph (Fig. 1C). The second comprises interactive scatterplots of annotation terms as distributed by their corrected q-values arising from phylogeny-aware models and from common correlation tests. Dot size and transparency are used to highlight interesting annotation terms (both highly frequent and variable across species).

The third type of output is a dynamical table where users may further explore and filter results. Each line contains several computed statistics (*e.g.*, correlation values, q-values for PIC linear model and correlation tests, sum, prevalence, and coefficient of variation) related to a single annotation term, as well as the individual counts of that annotation term in each genome. This table allows users to filter results based on any data column, selecting data slices for further inspection. The dynamic table also contains links to the fourth type of interactive output, namely individual plots of annotation terms results, which includes scatterplots, linear model trend lines and confidence bands for actual data values, ranked data and phylogenetic-aware linear models, and violin plots with superimposed raw data, allowing users to visually inspect how the frequency of annotation terms is distributed in the distinct user-defined groups. The heatmap in Fig. 2A and the association scatterplots in Figures 2C and 3C are direct examples of the CALANGO graphical output.

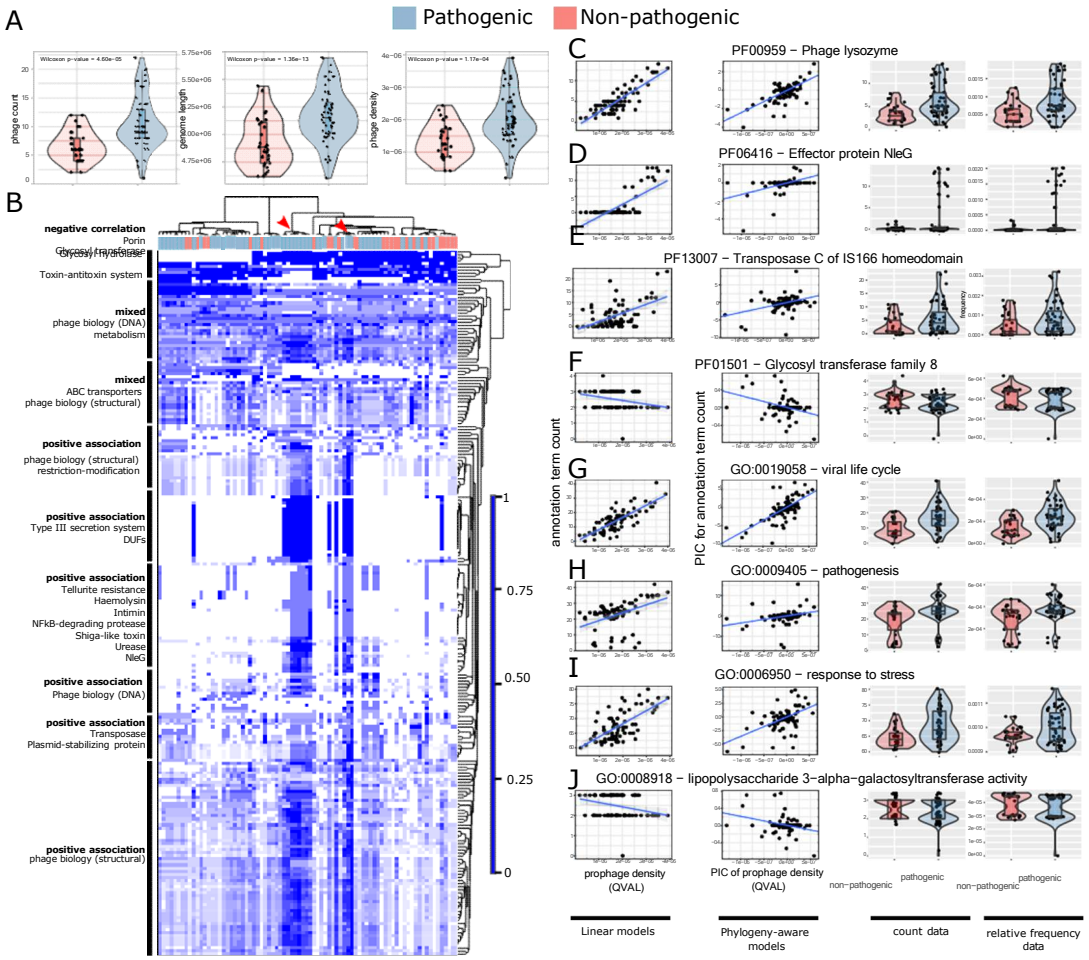


Figura 2: *Escherichia coli* data. A) Comparison of the number of prophages, genome length, and prophage density in pathogenic (blue) and non-pathogenic (pink) E. coli. B) Heatmap as produced by CALANGO integrates phylogenetic, phenotypic and annotation data. Species clustering are based on phylogeny, and classes are the same as "A". Each square contains additional information, and can be accessed through "on mouse over" events. C) Examples of graphical outputs displaying additional information of annotation terms associated with prophage density. From left to right: common association statistics (linear models with Pearson's correlation data), phylogeny-aware linear models (PIC values for QVAL and annotation data counts) and boxplots with absolute and relative values of annotation terms in the user-defined groups. From top to bottom: examples of annotation terms playing roles in (i) viral life cycle (single domain); (ii) virulence mechanisms (single domain); (iii) transposases (single domain); (iv) LPS biosynthesis (single domain); (v) viral process (GO annotation, multiple domains, functional annotation); (vi) pathogenicity (functional annotation); (vii) stress response mechanisms (functional annotation); (viii) LPS biosynthesis (functional annotation).

2.2.4 Experimental design for case studies

Two annotation schemas (Fig. S-1B) were used to annotate the same genomic elements in each case study, allowing us to contrast the results and evaluate the influence of annotation schema in the analyses:

A (*domain2Pfam*) A domain-level analysis, in which individual predicted protein domains were annotated using a traditional homology-based annotation schema (Pfam IDs) to emulate the results obtained through standard comparative genomics analyses;

B (*domain2GO*) The same domain-level genomic components were annotated using the GO term annotation for the Pfam domains, to highlight how CALANGO may identify functional convergences of non-homologous sequences not detectable in the first analysis.

Finally, as we are performing exploratory data analyses aimed at both validating CALANGO and finding associations that are potentially useful for downstream hypothesis generation, we used a q-value cutoff of < 0.1 for the selection of significative associations.

2.2.5 Acquiring data for *Escherichia coli*

A thorough literature review was performed to select *E. coli* genomes that could be unambiguously classified as pathogenic or non-pathogenic according to GOLD metadata (Supplementary Table 1) (Mukherjee et al. 2016). We evaluated genome assembly quality based on the expected gene content using BUSCO 3.0 and the Enterobacteriales order (Waterhouse et al. 2017). All genomes have BUSCO values of single-copy orthologs greater than 0.95. We predicted the numbers and location of prophages in *E. coli* chromosomes using PHASTER with standard parameters, and considered all classes of prophages (Arndt et al. 2016).

To obtain an ultrametric phylogenetic tree for the 80 *E. coli* genomes we initially computed groups of homologous genes using OrthoMCL v2 (Fischer et al. 2011) with default parameters and DIAMOND (Buchfink, C. Xie e Huson 2015) for the sequence alignment step. At this point, we included two species from the same family as external groups: *Escherichia fergusonii* (ATCC 35469, NC_011740) and *Shigella dysenteriae* (Sd197, NC_007606) (Chaudhuri e Henderson 2012). We proceeded by randomly selecting 25 universal 1-1 orthologs from OrthoMCL output file to be used for downstream phylogenetic analyses (Supplementary Table 1).

The global alignment of coding regions was done using MUSCLE (Edgar, 2004), fol-

lowed by the manual curation of alignments using MEGA X (S. Kumar, Stecher et al. 2018). We used SequenceMatrix (Vaidya, Lohman e Meier 2011) to concatenate individual alignments in a single matrix using, and MEGA X to select the best nucleotide substitution models for individual gene alignments. We used MRBAYES (Pang et al. 2015) as provided by the CIPRES SCIENCE platform to estimate phylogenetic relationships based on the Bayesian information criterion (BIC) and assuming a minimal mutation rate (μ) of 2x10-10 mutations per nucleotide site per cell division, following the work of H. Lee et al. (2012). The parameters of the posterior probability model, tree branch lengths, and tree topology were obtained using the *Metropolis-coupled Markov Chain Monte Carlo* (MCMCMC) algorithm. Two independent runs of two simultaneous chains for 20 million generations were performed for each analysis, collecting tree samples every 500 generations and discarding the 20% first runs as a burn-in step. Tracer v1.6 was used to evaluate MCMCMC convergence and the conjoint distributions of parameters and trees (Rambaut et al. 2018). The final summarized tree topology was saved in *newick* format and used as input for our analyses.

2.2.6 Removal of genes of viral origin in *E. coli* genomes

For each protein-coding gene in all 80 *E. coli* genomes, we generated a BED file containing the IDs of coded proteins and the genomic coordinates of their corresponding genes. We proceeded by generating BED files for each predicted integrated prophage from the output of PHASTER. We then used bedtools (Quinlan, 2014) to remove all protein-coding genes located within prophage coordinates, generating simulated bacterial genomes where genes of viral origin were removed and blocking out any effect of these genes in the CALANGO output while holding all other variables constant, which allowed us to use CALANGO as a support tool to investigate potential causal relationships.

2.2.7 Manual curation of Pfam domains associated with prophage density

We used the individual description of Pfam IDs provided by InterProScan together with the Pfam ID entries from the Pfam website (El-Gebali et al. 2018) to classify all Pfam IDs associated with prophage density to the broadest categories that functionally encom-

pass them (Supplementary Table 2, sheet “domain2PfamCount”, column “major_role”).

2.2.8 Acquiring data for Angiosperms

We downloaded the predicted proteomes of species of flowering plants from NCBI (O’Leary et al. 2015) and Phytozome (Goodstein et al. 2011). We detected a high level of duplication in most proteomes due to the presence of protein isoform sequences produced by variant mRNA splicing. Two methods were applied to reduce the redundancy arising from the presence of isoforms. The first was to summarize the NCBI proteomes per genomic locus, keeping only the longest protein sequence for each gene based on the “locus_tag” or “gene_id” information. As proteomes from other databases do not contain these fields in their sequence headers, we used the CD-HIT software with the identity threshold set to 1.0 to reduce the redundancy level for those proteomes (W. Li e Godzik 2006).

We used BUSCO to assess the assembly quality of the proteomes based on the expected content of nearly universal ortholog genes (Embryophita db10) (Waterhouse et al. 2017). Only proteomes with completeness higher than 80% and the rates of duplicated and fragmented genes lower than 15% were kept. We obtained the species phylogenetic relationship from the Time Tree of Life database (S. Kumar, Suleski et al. 2022), and used InterProScan (Jones et al. 2014) to perform a *de novo* annotation using Pfam data on the predicted proteomes. A total of 54 species of Angiosperms were selected as having the three data types needed to execute CALANGO: high-quality non-redundant proteomes, available quantitative phenotype (maximum plant height) and phylogenetic mapping (Supplementary Table S-3).

A considerable fraction of angiosperms is polyploid, and may also have been subjected to previous whole-duplication events (Ren et al. 2018). These facts may introduce biases when using count data alone, as we have done for the *E. coli* data, since a much more variable genomic content is expected in the non-redundant proteomes of the flowering plants (even though our filtering pipeline removed genomes with high gene duplication values). CALANGO may use a normalizing value (*e.g.* the total number of domains found in a non-redundant proteome) to compute relative frequencies of annotation terms. However, as we also demonstrated, this metric may introduce spurious associations (See Supplementary File 1, section “Evaluating annotation term frequencies and counts” and

Figure S-2 A-B for a deeper discussion on the biases of using relative frequencies and absolute counts)).

To account for all these facts, annotation terms were considered to be associated with the maximum plant height only if both the relative frequency and absolute count were found to be significantly associated. We used CALANGO to search for significant associations with the following cutoffs: 1) q-value for corrected contrasts and Pearson's correlation < 0.1; and 2) annotation terms with total occurrence count > 50 (to search for major expansions and prevent spurious associations due to the common errors and pitfalls found in eukaryotic genomes).

For the comparative genomics analysis, we cross-referenced genomic annotation data, phylogenetic information and the logarithm of the plants' maximum height values to search for Pfam domains and GO terms associated with this phenotype using the *domain2Pfam* and *domain2GO* annotation schemas (Fig. S-1B).

2.2.9 Estimation of ancestor states for height in Angiosperms

We used the method of maximum likelihood implemented in the "fastAnc" function of the phytools package to estimate the ancestral height state for all internal nodes of the phylogeny, based on the height values for the 54 extant species under analysis (Revell 2012).

2.3 RESULTS

2.3.1 CALANGO - a brief overview

CALANGO provides a flexible platform for investigating the variation of quantitative phenotypes/genotypes in a phylogeny – from now on referred as Quantitative Values Across Lineages (QVALs) – using current genomic knowledge (Fig. 1). By dissociating genomic components from their annotation schemas, CALANGO allows comparative genomics analyses at several levels, such as promoters, domains or genes; and, by using distinct annotation schemas, CALANGO enables the search for associations of both homologous regions and functional molecular convergences as provided by, *e.g.*, GO annotation (Fig. S-1A; Methods, section "Genomic data modeling").

Starting with a set of species/linages and their proper data types, CALANGO searches for associations of annotation terms and QVALs using comparative methods to deal with the data dependencies that emerge when analyzing phylogenetically related organisms, together with traditional statistical models (Fig. 1B; Methods, section "Analyzing data using CALANGO"). It also allows normalizing the relative abundance of annotation terms in different genomes by, *e.g.*, genome length or total number of protein-coding genes, to account for the high variability in genome sizes and contents (Supplementary File 1, "Evaluating annotation term frequencies and counts").

CALANGO outputs a dynamic HTML5 report containing graphical elements and tables with association statistics and other useful quantities, intended to instigate users to interact and actively interpret the results. The full set of user-defined input parameters is also returned alongside all computed results as a *list* object, which can be easily integrated into other bioinformatics pipelines (Fig. 1C).

2.3.2 Case study 1: coevolution of *Escherichia coli* lineages and their integrated bacteriophages

Escherichia coli have a remarkable genomic variability, with a considerable fraction of this variation comprising horizontally transferred genes through integrated bacteriophages (prophages) (Touchon et al. 2009). This genomic diversity is reflected in the distinct ecological niches occupied by this bacterium, which is found in several body niches of animal hosts as a commensal or pathogen. Bacteriophage infections are not always deleterious to their bacterial hosts. While obligate lytic phages represent agents of cell death and population control, persistent lysogenic phages are responsible for gene transfer and mutualism. In a microbial population the lysis-lysogeny events are dynamic and extremes of a continuum comprising antagonistic and beneficial biological interactions (Correa et al. 2021).

Virulence factors are an archetypal example of bacteriophage-mediated horizontal gene transfer that can result in fitness increase in new bacterial hosts, including pathogenic *E. coli* (Steyert e Kaper 2012). Despite being a well-known phenomenon, we are not aware of any systematic evaluation of the association between prophage occurrence and the abundance of non-homologous virulence factors. As prophages are themselves genomic elements with specific coordinates, this case study also allows us to selectively remove

the effect of viral genes on CALANGO's results, enabling the potential investigation of associations of causal origin. Therefore, we consider this as an interesting scenario to evaluate our tool, as it has expected causal associations and represents a complex biological interaction, likely to contain previously undescribed biological phenomena.

We performed a thorough literature review to select 80 *E. coli* lineages with both gapless genomes (plus plasmids, when available) and reliable information regarding its pathogenicity status, and computed the annotation, phylogenetic and QVAL data needed to run CALANGO (Supplementary Table 1). Pathogenic *E. coli* were found to have a significantly higher number of prophages, even after accounting for variations in genome size (Fig. 2A; Supplementary File 1, section "Integrated bacteriophages in pathogenic and non-pathogenic *E. coli* "). We proceeded by using the prophage density as a proxy QVAL to further survey this biological interaction.

To objectively evaluate GO usefulness to detect molecular functional convergences, we annotated the same set of protein domains found in each bacterial genome by using either their Pfam IDs (*Pfam2domain*) or the GO terms associated with them (*Pfam2GO*) (Fig. S-1B, Methods, section "Experimental design for case studies"). We found GO annotations to be more abundant and more prevalent than our *Pfam2domain* annotation, a scenario compatible with the integration of biological functions from non-homologous domains at the GO space (Supplementary File 1, section "Comparison of functional and homology-based annotation schemas"; Fig. S-1C).

Homologous regions and biological roles associated with prophage density in *E. coli* We found 230 out of the 3,335 Pfam domains observed in our *E. coli* dataset (6,8%) to be significantly associated with prophage density (corrected q-values for both phylogeny-aware models and Pearson's correlation < 0.1). Fig. 2B is a heatmap of these terms as produced by CALANGO together with our manual annotation. Of these, 207 domains presented positive correlation (from 0.28 to 0.85) and 23 showed negative correlation (from -0.26 to -0.47) (Supplementary Table 2, sheet "domain2PfamCount"; see also Supplementary File 1, section "Evaluating annotation term frequencies and counts" and Fig. S-2A-B for the rationale of working with raw count data instead of normalized frequencies).

We found the majority of positively associated domains (125/207 domains, 60.4%) to have clear roles in viral life cycle, such as lysozymes and integrases (Fig. 2B). Fig. 2C illustrates a typical CALANGO output for one of these domains (with additional examples

in Fig. S-3A-C). The second largest category encompasses several classes of virulence factors (58/207 domains, 28.0%) (Fig. 2B; Fig. 2D; Fig. S-3D-E). Some virulence factors, such as Shiga-like toxins and effectors of the Type III secretion system, are known to be commonly horizontally transferred by bacteriophages in specific *E. coli* pathotypes (Ehrbar e Hardt 2005; Steyert e Kaper 2012), an association also detected by CALANGO (Supplementary File 1, section "Homologous genes and biological roles associated with prophage density in *E. coli*"). These two categories provide a compelling example of how CALANGO can uncover known associations of causal origin. Some domains of unknown function (DUFs) have a distribution pattern similar to that of virulence factors (Fig. 2B), suggesting these DUFs may be uncharacterized pathogenicity domains and demonstrating how CALANGO can be used to prioritize targets for experimental characterization.

CALANGO also detected positive associations that unveil novel biological interactions between immune genes found in bacterial genomes, prophages, and other classes of mobile elements such as transposases and plasmids. Furthermore, several of these homologous regions are located outside prophage regions, suggesting a complex interplay of symbiosis and competition between them (Fig. 2B; Fig. 2E; Sup. Figures S-3F-H; see also Supplementary File 1 for a deeper discussion on these associations). The 23 negative associations suggest that *E. coli* lineages with fewer integrated prophages – which are also less likely to be non-pathogenic – have a set of genes enabling a greater diversity of lifestyles at several levels, ranging from metabolic pathways and membrane transport to community-level processes, such as biofilm formation (Fig. 2B; Fig. S-3I-L).

Interestingly, we observed negative associations of prophage density and components of the cell wall and of the LPS biosynthesis pathway (Fig. 2B; Fig. 2F). Both classes of molecules are receptors of bacteriophages for cellular infection, but also major activators of the vertebrate immune system (Park e J.-O. Lee 2013; Bertozzi Silva, Storms e Sauvageau 2016; A. J. Wolf e Underhill 2018). These undocumented negative associations may be a consequence of the selective pressure against prophage infection resulting in the loss of these components. However, these losses can also confer an advantage to pathogenic bacteria when infecting vertebrate hosts, as they may be less likely to trigger host's immune responses, and could represent a previously unknown aspect of the emergence of a virulence phenotype in this species.

The results provided by the *domain2GO* schema largely support the same conclusions

found by our manual curation of the *domain2Pfam* results for both positive and negative associations, highlighting how GO annotation provides an interpretability that is at least qualitatively equivalent to human curation (Fig. 2G-H; Fig. S-4A-N; Supplementary Table 2, sheet "domain2GOCount"). Additionally, as several of these biological roles are performed by non-homologous domains (Supplementary File 1, section "GO annotation provides curation-level of biological knowledge"), these results further demonstrate how CALANGO, together with a GO annotation, allows comparative genomics analysis at the function level.

The prophage-mediated horizontal transfer of virulence factors is a known mechanism for fitness increase of pathogenic *E. coli* (Steyert e Kaper 2012). We also found a positive association of the term GO:0006950 (*response to stress*, Pearson's correlation of 0.69, q-value for contrasts of 4.65e-05), which may represent a new example of virus-mediated transfer of non-homologous fitness genes that fulfill a common biological role (Fig. 2I; Supplementary File 1, section "Stress response genes are associated with prophage density in *E. coli* "; Supplementary Table 2). We hypothesize that integrated prophages could also contribute to fitness increase of bacteria specifically under conditions of stress, such as the host's immune response against pathogenic bacterial lineages (X. Wang et al. 2010), providing yet another dimension of this complex biological interaction and providing additional evidence of how GO-level annotation can support the detection of previously unknown associations.

Since prophages are genomic elements with defined genomic coordinates within bacterial genomes, this case study allows us to perform an *in silico* controlled experiment to evaluate the ability of CALANGO to support the investigation of a potential causal relationship, namely that annotation terms are associated with prophage density *because* they annotate protein-coding genes of viral origin located within prophage coordinates (Supplementary File 1, section "Associations with prophage density after removal of genes of viral origin").

This experiment consisted of re-executing CALANGO holding the QVAL values fixed (i.e., as calculated previously) and removing all genes of viral origin. This effectively blocks out possible effects of prophage genes on the output of CALANGO, allowing us to test whether the significant associations detected earlier between the annotation terms and the QVAL are indeed due to genes of viral origin. As expected, the vast majo-

rity of protein domains annotated as being of viral origin were no longer significantly associated with prophage density after the removal of genes of viral origin (123/125, 98.4%). A similar pattern was found for GO terms (Supplementary Table 2, sheets "domain2PfamCountLessPhages" and "domain2GOCountLessPhages"). Interestingly, several classes of virulence factor domains were still significantly associated with prophage density after the removal of genes of viral origin, a scenario compatible with bacteriophage-mediated horizontal gene transfer followed by prophage degeneration. However, several other classes virulence domains that were totally or mostly located within detectable prophage genomes were not found to be associated after blocking the effect of viral genes, which suggests a synergistic interaction between virulence factors acquired from different evolutionary origins (Supplementary Table 2, sheet "virulence_factors").

This *in silico* manipulative experiment showcases the ability of CALANGO to support the investigation of basic causal relationships by enabling a level of counterfactual investigation of observed associations in the data. While this is still short of a fully developed causal inference package for genomic data, the ability to uncover some causal relationships from data by *in silico* isolation and testing of the influence of possible confounders can provide valuable insights into biologically meaningful phenomena, as illustrated in this case study.

2.3.3 Case study 2: Homologous regions associated with maximum height in Angiosperms

Maximum height is a key trait in the ecology, physiology, and evolution of land plants, and the understanding of the molecular mechanisms associated with the emergence of complex phenotype has consequences for fields as diverse as conservation biology and agriculture (Falster e Westoby 2003; Moles et al. 2009). Angiosperms, or the flowering plants, are the largest and most diverse group of land plants, displaying a remarkable phenotypic variation, including in plant height. Artificial selection experiments within single species supports the notion that plant height is a trait strongly controlled by genes that can evolve fast under phenotypic selection, with h^2 values ranging from 0.5 to 0.9 (Peiffer et al. 2014; Zu e Schiestl 2017). However, there is a lack of studies associating the evolution of this phenotype across species under a comparative genomics perspective.

Increases in height seems to be associated with reproductive success, a relationship

thought to be caused by several factors, including increased pollination, seed dispersal mechanisms, and access to light. Thus, tall plants that emerged from short ancestors likely experienced positive selection for height, a trait that is potentially under selection in natural populations (Zu e Schiestl 2017). Shorter plants, on the other hand, have smaller generation times and, consequently, higher rates of evolution, which provides a higher capacity of phenotypical change in distinct environments, and is a concern for the long-term viability of taller species and the ecosystems that depends on them (Lanfear et al. 2013).

The evolution of a complex phenotype like height is likely coupled with many other plants traits, such as rates of mitosis in meristematic tissues, cell expansion, development of leaves and reproductive organs, pollination syndrome, community composition, among others, and plays important roles in the success of establishment of distinct species (Mashau et al. 2021; Moles et al. 2009). As such, the evolution of height in Angiosperms represents a compelling case study to further evaluate CALANGO and demonstrate its usefulness to reveal new biological knowledge.

2.3.4 Protein domains associated with maximum height in flowering plants unveil independent expansions of reproductive and developmental processes in taller species

We surveyed the specialized literature and sequence databases, together with our *in-house* annotation pipeline, to gather the annotation, phylogenetic and QVAL information for 54 angiosperms species with high-quality non-redundant proteomes available (Supplementary Table S-3, see Methods, section “Acquiring genomic, phylogenetic, and phenotypic data for Angiosperms”). Our dataset has species with maximum height varying from 20 cm (the wild strawberry *Fragaria vesca*, Rosaceae) to 55 meters (the tree *Eucalyptus grandis*, Myrtaceae), more than two orders of magnitude (Fig. 3A, Fig. S-5A). We found the ancestor states of height in Angiosperms to be highly uniform, with internal nodes having mostly average values and phenotypic extremes occurring multiple times, a pattern compatible with independent emergence of this trait (Fig. 3A, Fig. S-5B).

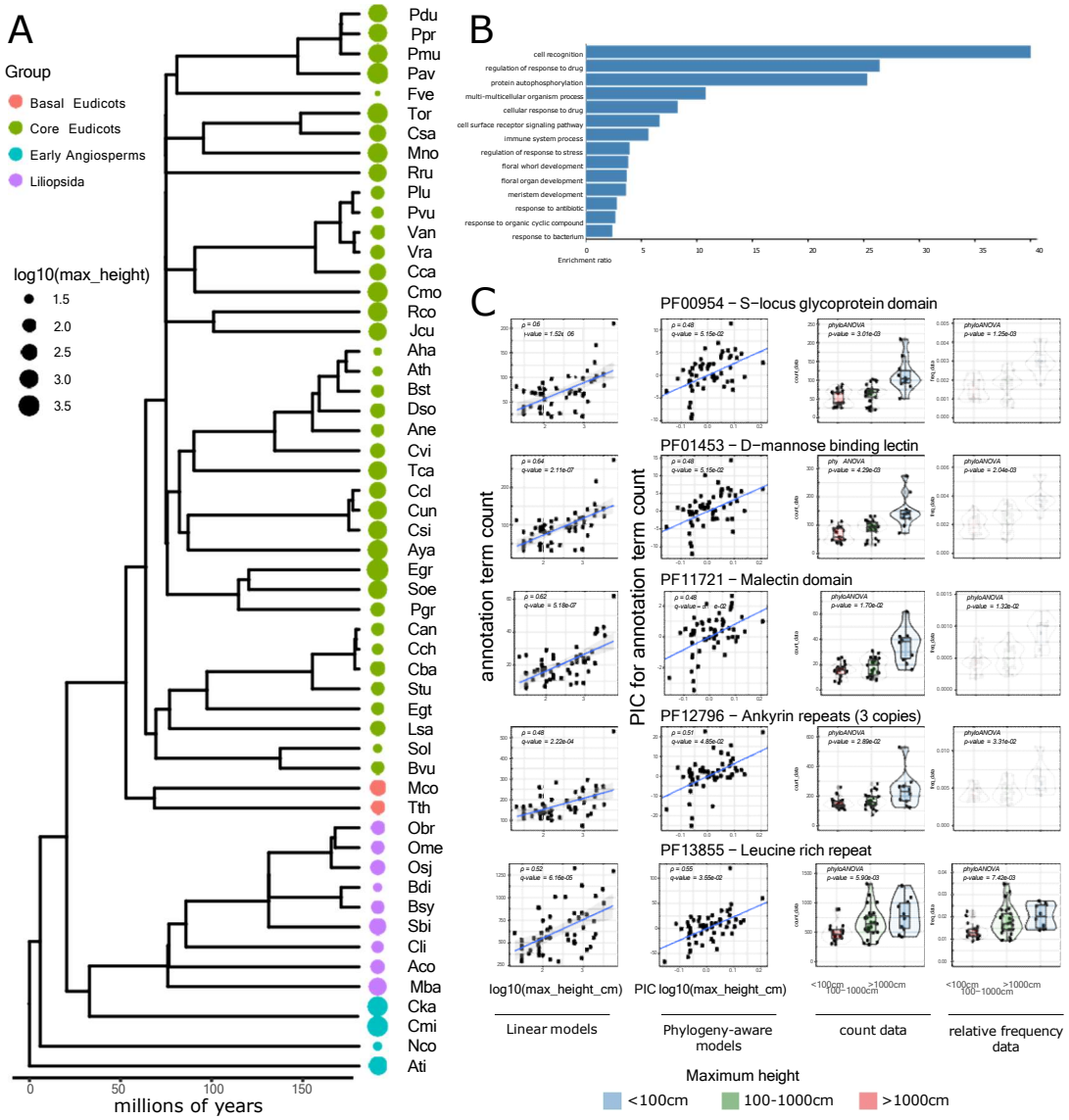


Figura 3: Angiosperm data. A) Maximum height variation across the phylogeny of 54 angiosperms species with high-quality proteomes available. Full caption including all species codes can be found in supplementary materials section; B) Biological processes significantly enriched in protein-coding genes containing domains associated with maximum height in angiosperms. C) Examples of protein domains significantly expanded in taller plants fulfilling different biological roles. From top to bottom: (i) and (ii) increase of genetic diversity through cross-pollination; (iii) cell wall biology; (iv) embryogenesis; (v) immunity and stress response mechanisms. From left to right: linear models from raw count data, phylogeny-aware linear models; boxplot with raw counts; boxplot with normalized counts.

We again used the *domain2Pfam* and *domain2GO* annotation schemas to search for

homologous regions and biological roles associated with maximum height. From a total of 4066 domains with at least 50 copies when considering all species, we found seven of them to be positively associated with increase in the plants' maximum height (Supplementary Table 4, sheet "associated_domains", see also Methods, section "Experimental design for case studies" for the parameters, experimental design and rationale).

Even though no GO term was found to be significantly associated with the QVAL, an enrichment analysis using the 457 *Arabidopsis thaliana* genes annotated to these Pfam domains found an overrepresentation of genes belonging to reproduction, embryogenesis pathways, including secondary growth, immune system, and stress response mechanisms (Fig. 3B-C; Supplementary Table 4, sheet "Arabidopsis_genes"). The increase in the number of genes coding for some of these processes, such as immunity and wood tissue development, has already been reported for *Populus trichocarpa* - a model organism for tree biology - when compared with *A. thaliana* (Tuskan et al. 2006). The expansions of immune and stress response genes are likely to represent adaptions required for the longer lifespan of taller species, which results in exposure to long-term infections and a myriad of stress sources. Importantly, *P. trichocarpa* had not been included in our analysis due to a large amount of gene duplication events detected by our pre-processing pipeline. Therefore, CALANGO independently provides evidence supporting the association of genes fulfilling these biological roles and the emergence of taller species.

Taller plants have lower rates of molecular evolution, presumably due to longer generation times and slower long-term rates of mitosis in their apical meristems. This fact is a concern for the long-term viability of these organisms and, consequently, for the large number of ecosystems where they play critical roles (Lanfear et al. 2013).

Self-Incompatibility (SI) systems are non-homologous molecular mechanisms that prevent inbreeding and promotes out-crossing in flowering plants (Durand et al. 2020). Three of the domains associated with maximum height in angiosperms are components of the most well-characterized SI system (Fig. 3C, "PF00954 – *S-locus glycoprotein domain*"; Supplementary Figure S-6A-B; Supplementary Table S-4, sheet "associated_domains"). The SI system found by CALANGO (from now on referred simply as "SI") has been described in exquisite molecular details in the Brassicaceae family, even though it is widely distributed in flowering plants, and is an archetypal example of natural (balancing) selection maintaining genetic variation over long evolutionary times through inbreeding

avoidance and rare-allele advantage (Durand et al. 2020).

The SI system is a mating barrier controlled by a single highly polymorphic locus (S-locus), which codes for two genes in closely-linked genes: 1) the S-locus receptor kinase (SRK), a glycoprotein with kinase activity that allows stigma cells to discriminate between pollen from the same organism or from genetically related individuals (all four domains are located in these genes); 2) the S-locus cysteine-rich protein (SCR), expressed in pollen coat and the ligand of SRK (J. B. Nasrallah e M. E. Nasrallah 2014). In the case of self-fertilization, the SCR protein in pollen is structurally complementary to the SRK protein found in the same S locus haplotype, activating a signaling pathway that inhibits pollen tube development. Not surprisingly, the S-locus is highly polymorphic, and has been extensively characterized in several populations across multiple species and in distinct ecological contexts (Durand et al. 2020).

We found 37 copies of PF00954 – the signature domain of SRK genes – in *A. thaliana* (maximum height of 0.30 meters, the eighth smallest plant in our dataset), while *E. grandis* (maximum height of 55 meters, the tallest plant in our dataset) has 210 copies of this domain (a 5.68-fold increase). The remaining three components of the SI system have similar expansion profiles (Fig. S-4, Supplementary Table S-4, sheet "associated_domains").

Our findings describe a previously unknown landscape of the total genetic variability in the SI system, as the combinatorial possibilities of multiple loci coding for components of the SI mechanism, each locus being itself potentially highly variable, greatly expands the total number of alleles hosted in each genome and, consequently, the fraction of possible incompatible individuals in the populations of taller plants. In this scenario, successful fertilization events would have a greater chance of outcrossing in species that can potentially host a greater number of S-locus haplotypes, therefore allowing taller plants to increase their evolutionary rates through cross-pollination with unrelated individuals.

CALANGO revealed a previously unknown molecular mechanism that can promote outcrossing in taller species with longer generation times and may counterbalance their lower evolutionary rates. This observation may have profound consequences for our understanding of the evolutionary future of taller plants and the critical roles they play in their ecosystems and in human economic activities, as the ability of a species to adapt to changing environments fundamentally depends on their underlying mutation rates (Willi

e Hoffman 2009).

2.3.5 Comparison of CALANGO with conceptually similar software

As recently reviewed by Nagy et al. (2020), several tools are already available to search for associations between the patterns of occurrence of homologous genomic components (mostly secondary losses of homologous regions in specific lineages) and a binary phenotype of interest (Nagy et al. 2020). One important distinction between CALANGO and these tools is the class of phenotypic data accepted by them. Most of the methods for searching gene losses associated with categorical phenotypes (presence/absence) which, while useful to describe several types of biological variation, cannot be used to survey quantitative phenotypic data without the usage of *ad hoc* thresholds to define classes. Our tool, therefore, considerably expands the strategies currently available to search for associations between genetic and phenotypic data across species.

Furthermore, despite being successful in detecting gene losses associated with the emergence of complex phenotypes, none of these methods incorporate current genomic knowledge at the function level as provided by GO annotation. Instead, they exclusively evaluate the associations between sets of homologous elements across genomes and phenotypic data. GO terms have been shown to capture patterns of functional convergence and to provide a deeper biological comprehension of the genomic evolution of complex phenotypes, such as parasitism and sociality, and can provide a functional landscape for comparative genomics at the molecular function level (Coghlan et al. 2019; Tong et al. 2020).

By dissociating the genomic components from their functional annotation, CALANGO provides flexibility to survey the distribution of several classes of genomic components, such as protein domains or entire genes. Furthermore, as we demonstrated, distinct annotation schemas allow both the emulation of classic comparative genomics analysis (by using an homology-based annotation dictionary) and a pathway- or function-based comparison (by using a GO-based dictionary), a feature that, to the best of our knowledge, is not present in any existing tools. As illustrated in our case studies, the combination of both strategies delivers a richer, more biologically meaningful interpretation of the results, including the detection of functional molecular convergences which could not be

discovered using homology-based annotation.

Also, in contrast with virtually all the software reviewed by Nagy et al. (2020), which mostly provide text files as main output, CALANGO produces a rich set of dynamical HTML5 result files containing several statistics and other useful quantities, together with their visual representations. For advanced users, CALANGO provides all results as a *list* of standard R objects, therefore allowing easy integration with other computational pipelines. The installation procedure for our tool is straightforward, as it is available as a fully operational R package.

2.4 DISCUSSION

The post-genomic era has brought a plethora of high-quality sequenced genomes, ranging from previously underrepresented early-branching lineages of cellular organisms to thousands of genomes of a single bacterial species. In contrast to this abundance of genomic data, there are currently no standardized methods in computational statistics for extracting genomic properties associated with a quantitative genotype or phenotype of interest across genomes. CALANGO addresses this gap in the comparative genomics field by integrating phylogenetic, genomic, annotation and phenotypic data together in order to perform this task.

Our two case studies comprise datasets that are highly contrasting in terms of evolutionary time, taxonomic diversity and the quantitative phenotype/genotype under analysis. The first evaluated the biological roles associated with the change of a complex genotype (the density of prophages) in a single bacterial species. We found, as expected, a considerable association with genes of viral origin. By removing these genes and blocking their effect from the analysis, this case study allowed us to demonstrate how CALANGO can support the investigation of causal associations. We also observed several unknown associations at the function level that point to a much richer scenario of the biological interaction between bacteria, their prophages and other classes of mobile elements. We emphasize how the horizontal acquisition of adaptive genes, such as virulence factors and stress response genes, may allow bacteria to thrive in distinct environments.

The second case study detected domain expansions associated with a complex phenotype (maximum height) in the flowering plants, a major group of multicellular eu-

karyotes. Tall plants have morphological and physiological adaptions to the challenges of growing vertically (Falster e Westoby 2003), and concomitantly harbor several advantages in dispersal and establishment success rates (Mashau et al. 2021). Our case study described several mechanisms that greatly improve our understanding of the genomic regions and molecular mechanisms associated with emergence and maintenance of this complex phenotype. Of special interest, we described a previously unknown reproductive strategy that may allow taller plants the increase their genetic diversity through the independent expansion of the S-locus and the allocation of more resources to cross-pollination, a fact with long-reaching consequences for fields as diverse as biotechnology, agriculture and conservation biology. More importantly, our tool also indicated several testable hypotheses in both case studies, indicating how it can be used to prioritize downstream targets for experimental characterization.

CALANGO represents a considerable step towards the establishment of an annotation-based, phylogeny-aware comparative genomics framework to survey genomic data beyond sequence level, and to search for associations between quantitative variables across lineages sharing a common ancestor and the multiple layers of biological knowledge coded in their genomes.

2.5 DATA AVAILABILITY

All processed data needed to fully reproduce the two case studies (genome annotation files, phylogenetic tree, phenotypic information and CALANGO configuration files) are available at https://labpackages.github.io/CALANGO/. All raw data used in case studies (genome IDs and sources for phenotypic data) are available as supplementary tables.

2.6 SUPPLEMENTARY DATA

Supplementary data will be available at Bioinformatics online.

2.7 ACKNOWLEDGEMENT

We would like to thank Prof. Daniel dos Santos Mansur and Prof. Glória Regina Franco for the insightful discussions during the elaboration of this research. We also would like to thank them, for the financial support for this research and publication fees.

2.8 FUNDING

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior /Brazil [Grant 001]. Funding for open access charge: Graduate Program in Genetics, the Graduate Program in Bioinformatics, and the Vice Dean for Research from Universidade Federal de Minas Gerais, Brazil.

2.9 CONFLICT OF INTEREST

All authors declare no conflict of interest for this publication

2.10 REFERENCES

2.0, GenomeHubs (2022). *GoaT - Genomes on a Tree*. Last Accessed: August 19th 2022.

Adams, Dean C. e Michael L. Collyer (jul. de 2017). "Multivariate Phylogenetic Comparative: Evaluations, Comparisons, and Recommendations". Em: *Systematic Biology* 67.1, pp. 14–31. issn: 1063-5157. doi: 10.1093/sysbio/syx055.

Adl, Sina M. et al. (2012). "The Revised Classification of Eukaryotes". Em: *Journal of Eukaryotic Microbiology* 59.5, pp. 429–514. doi: https://doi.org/10.1111/j.1550-7408.2012.00644.x.

Allaire, JJ et al. (2022). *rmarkdown: Dynamic Documents for R*. R package version 2.14.

andILeitch, Michael Bennet (2005). "CHAPTER 2 - Genome Size Evolution in Plants". Em: *The Evolution of the Genome*. Ed. por T. Ryan Gregory. Burlington: Academic Press, pp. 89–162. isBn: 978-0-12-301463-4. doi: https://doi.org/10.1016/B978-012301463-4/50004-8.

Arndt, David et al. (mai. de 2016). “PHASTER: a better, faster version of the PHAST phage search tool”. Em: *Nucleic Acids Research* 44.W1, W16–W21. issn: 0305-1048. doi: 10.1093/nar/gkw387.

Ball, Steven et al. (jan. de 2011). “The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis”. Em: *Journal of Experimental Botany* 62.6, pp. 1775–1801. issn: 0022-0957. doi: 10.1093/jxb/erq411.

Bar-On, Yinon M., Rob Phillips e Ron Milo (2018). “The biomass distribution on Earth”. Em: *Proceedings of the National Academy of Sciences* 115.25, pp. 6506–6511. doi: 10.1073/pnas.1711842115.

Barr, Jeremy J. et al. (2013). “Bacteriophage adhering to mucus provide a non-host-derived immunity”. Em: *Proceedings of the National Academy of Sciences* 110.26, pp. 10771–10776. doi: 10.1073/pnas.1305923110.

Beaulieu, Jeremy M. et al. (2007). “Correlated evolution of genome size and seed mass”. Em: *New Phytologist* 173.2, pp. 422–437. doi: https://doi.org/10.1111/j.1469-8137.2006.01919.x.

Bennett, Michael D. (1987). “VARIATION IN GENOMIC FORM IN PLANTS AND ITS ECOLOGICAL IMPLICATIONS”. Em: *New Phytologist* 106.s1, pp. 177–200. doi: https://doi.org/10.1111/j.1469-8137.1987.tb04689.x.

Bentsink, Leónie et al. (2006). “Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis”. Em: *Proceedings of the National Academy of Sciences* 103.45, pp. 17042–17047. doi: 10.1073/pnas.0607877103.

Berardini, Tanya Z. et al. (2015). “The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome”. Em: *genesis* 53.8, pp. 474–485. doi: https://doi.org/10.1002/dvg.22877.

Bertozzi Silva, Juliano, Zachary Storms e Dominic Sauvageau (jan. de 2016). “Host receptors for bacteriophage adsorption”. Em: *FEMS Microbiology Letters* 363.4. fnw002. issn: 0378-1097. doi: 10.1093/femsle/fnw002.

Blázquez, Miguel A., David C. Nelson e Dolf Weijers (2020). “Evolution of Plant Hormone Response Pathways”. Em: *Annual Review of Plant Biology* 71.1. PMID: 32017604, pp. 327–353. doi: 10.1146/annurev-arplant-050718-100309.

Buchfink, Benjamin, Chao Xie e Daniel H. Huson (jan. de 2015). “Fast and sensitive protein alignment using DIAMOND”. Em: *Nature Methods* 12.1, pp. 59–60. issn: 1548-7105. doi: 10.1038/nmeth.3176.

Carrillo-Barral, Néstor, María del Carmen Rodríguez-Gacio e Angel Jesús Matilla (2020). “Delay of Germination-1 (DOG1): A Key to Understanding Seed Dormancy”. Em: *Plants* 9.4. issn: 2223-7747. doi: 10.3390/plants9040480.

Carta, Angelino et al. (2022). “Correlated evolution of seed mass and genome size varies among life forms in flowering plants”. Em: *Seed Science Research* 32.1, pp. 46–52. doi: 10.1017/S0960258522000071.

Cavalcanti, João Henrique F et al. (set. de 2018). “An L,L-diaminopimelate aminotransferase mutation leads to metabolic shifts and growth inhibition in Arabidopsis”. Em: *Journal of Experimental Botany* 69.22, pp. 5489–5506. issn: 0022-0957. doi: 10.1093/jxb/ery325.

Challis, Richard J. et al. (mai. de 2017). “GenomeHubs: simple containerized setup of a custom Ensembl database and web server for any species”. Em: *Database* 2017. bax039. issn: 1758-0463. doi: 10.1093/database/bax039.

Chamberlain, Scott A. e Eduard Szöcs (2013). “taxize: taxonomic search and retrieval in R”. Em: *F1000 Research* 2, p. 191. doi: 10.12688/f1000research.2-191.v1.

Chan, Patricia P. e Todd M. Lowe (dez. de 2015). “GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes”. Em: *Nucleic Acids Research* 44.D1, pp. D184–D189. issn: 0305-1048. doi: 10.1093/nar/gkv1309.

Chanderbali, Andre S et al. (mar. de 2016). “Evolving Ideas on the Origin and Evolution of Flowers: New Perspectives in the Genomic Era”. Em: *Genetics* 202.4, pp. 1255–1265. issn: 1943-2631. doi: 10.1534/genetics.115.182964.

Chaudhuri, Roy R. e Ian R. Henderson (2012). “The evolution of the Escherichia coli phylogeny”. Em: *Infection, Genetics and Evolution* 12.2, pp. 214–226. issn: 1567-1348. doi: https://doi.org/10.1016/j.meegid.2012.01.005.

Chen, Haixu et al. (nov. de 2021). “BRAD V3.0: an upgraded Brassicaceae database”. Em: *Nucleic Acids Research* 50.D1, pp. D1432–D1441. issn: 0305-1048. doi: 10.1093/nar/gkab1057.

Chen, Lu et al. (mar. de 2014). “Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity”. Em:

Molecular Biology and Evolution 31.6, pp. 1402–1413. issn: 0737-4038. doi: 10.1093/molbev/msu083.

Cheng, Joe et al. (2021). *htmltools: Tools for HTML*. R package version 0.5.2.

Cirillo, D M et al. (1996). "Identification of a domain in Rck, a product of the Salmonella typhimurium virulence plasmid, required for both serum resistance and cell invasion". Em: *Infection and Immunity* 64.6, pp. 2019–2023. doi: 10.1128/iai.64.6.2019-2023.1996.

Coghlan, Avril et al. (jan. de 2019). "Comparative genomics of the major parasitic worms". Em: *Nature Genetics* 51.1, pp. 163–174. issn: 1546-1718. doi: 10.1038/s41588-018-0262-1.

Cornwell, Will e Shinichi Nakagawa (2017). "Phylogenetic comparative methods". Em: *Current Biology* 27.9, R333–R336. issn: 0960-9822. doi: https://doi.org/10.1016/j.cub.2017.03.049.

Correa, Adrienne M. S. et al. (ago. de 2021). "Revisiting the rules of life for viruses of microorganisms". Em: *Nature Reviews Microbiology* 19.8, pp. 501–513. issn: 1740-1534. doi: 10.1038/s41579-021-00530-x.

Cotter, Paul D., R. Paul Ross e Colin Hill (fev. de 2013). "Bacteriocins — a viable alternative to antibiotics?" Em: *Nature Reviews Microbiology* 11.2, pp. 95–105. issn: 1740-1534. doi: 10.1038/nrmicro2937.

Crooks, Gavin E. et al. (2004). "WebLogo: A Sequence Logo Generator". Em: *Genome Research* 14.6, pp. 1188–1190. doi: 10.1101/gr.849004.

Cunningham, Fiona et al. (nov. de 2021). "Ensembl 2022". Em: *Nucleic Acids Research* 50.D1, pp. D988–D995. issn: 0305-1048. doi: 10.1093/nar/gkab1049.

Dedrick, Rebekah M. et al. (jan. de 2017). "Prophage-mediated defence against viral attack and viral counter-defence". Em: *Nature Microbiology* 2.3, p. 16251. issn: 2058-5276. doi: 10.1038/nmicrobiol.2016.251.

Dekkers, Bas J.W. et al. (2016). "The Arabidopsis DELAY OF GERMINATION 1 gene affects ABSCISIC ACID INSENSITIVE 5 (ABI5) expression and genetically interacts with ABI3 during Arabidopsis seed development". Em: *The Plant Journal* 85.4, pp. 451–465. doi: https://doi.org/10.1111/tpj.13118.

Dobritsa, Anna A. e Daniel Coerper (nov. de 2012). "The Novel Plant Protein INAPERTURATE POLLEN1 Marks Distinct Cellular Domains and Controls Formation of

Apertures in the Arabidopsis Pollen Exine ". Em: *The Plant Cell* 24.11, pp. 4452–4464. issn: 1040-4651. doi: 10.1105/tpc.112.101220.

Dong, Qunfeng, Shannon D. Schlueter e Volker Brendel (jan. de 2004). "PlantGDB, plant genome database and analysis tools". Em: *Nucleic Acids Research* 32.suppl_1, pp. D354–D359. issn: 0305-1048. doi: 10.1093/nar/gkh046.

Dubois, Emeline et al. (abr. de 2011). "Homologous Recombination Is Stimulated by a Decrease in dUTPase in Arabidopsis". Em: *PLOS ONE* 6.4, pp. 1–8. doi: 10.1371/journal.pone.0018658.

Dufayard, Jean-François et al. (2017). "New Insights on Leucine-Rich Repeats Receptor-Like Kinase Orthologous Relationships in Angiosperms". Em: *Frontiers in Plant Science* 8, p. 381. doi: 10.3389/fpls.2017.00381.

Dunn, Casey W. e Catriona Munro (2016). "Comparative genomics and the diversity of life". Em: *Zoologica Scripta* 45.S1, pp. 5–13. doi: https://doi.org/10.1111/zsc.12211.

Durand, Eléonore et al. (2020). "Evolution of self-incompatibility in the Brassicaceae: Lessons from a textbook example of natural selection". Em: *Evolutionary Applications* 13.6, pp. 1279–1297. doi: https://doi.org/10.1111/eva.12933.

Eddy, Sean R. (out. de 2011). "Accelerated Profile HMM Searches". Em: *PLOS Computational Biology* 7.10, pp. 1–16. doi: 10.1371/journal.pcbi.1002195.

Edgar, Robert C. (mar. de 2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". Em: *Nucleic Acids Research* 32.5, pp. 1792–1797. issn: 0305-1048. doi: 10.1093/nar/gkh340.

Ehrbar, Kristin e Wolf-Dietrich Hardt (2005). "Bacteriophage-encoded type III effectors in Salmonella enterica subspecies 1 serovar Typhimurium". Em: *Infection, Genetics and Evolution* 5.1, pp. 1–9. issn: 1567-1348. doi: https://doi.org/10.1016/j.meegid.2004.07.004.

Ekstrom, Alexander et al. (ago. de 2014). "PlantCAZyme: a database for plant carbohydrate-active enzymes". Em: *Database* 2014. bau079. issn: 1758-0463. doi: 10.1093/database/bau079.

Falster, Daniel S. e Mark Westoby (2003). "Plant height and evolutionary games". Em: *Trends in Ecology & Evolution* 18.7, pp. 337–343. issn: 0169-5347. doi: https://doi.org/10.1016/S0169-5347(03)00061-2.

Fedak, Halina et al. (2016). “Control of seed dormancy in Arabidopsis by a cis-acting noncoding antisense transcript”. Em: *Proceedings of the National Academy of Sciences* 113.48, E7846–E7855. doi: 10.1073/pnas.1608827113.

Felsenstein, Joseph (1985). “Phylogenies and the Comparative Method”. Em: *The American Naturalist* 125.1, pp. 1–15. issn: 00030147, 15375323.

Fernández, Lucía, Ana Rodríguez e Pilar García (mai. de 2018). “Phage or foe: an insight into the impact of viral predation on microbial communities”. Em: *The ISME Journal* 12.5, pp. 1171–1179. issn: 1751-7370. doi: 10.1038/s41396-018-0049-5.

Fischer, Steve et al. (2011). “Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups”. Em: *Current Protocols in Bioinformatics* 35.1, pp. 6.12.1–6.12.19. doi: https://doi.org/10.1002/0471250953.bi0612s35.

Fleischmann, Andreas et al. (out. de 2014). “Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms”. Em: *Annals of Botany* 114.8, pp. 1651–1663. issn: 0305-7364. doi: 10.1093/aob/mcu189.

Fu, Limin et al. (out. de 2012). “CD-HIT: accelerated for clustering the next-generation sequencing data”. Em: *Bioinformatics* 28.23, pp. 3150–3152. issn: 1367-4803. doi: 10.1093/bioinformatics/bts565.

Galili, Tal (2015). “dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering”. Em: *Bioinformatics*. doi: 10.1093/bioinformatics/btv428.

Galili, Tal et al. (2017). “heatmaply: an R package for creating interactive cluster heatmaps for online publishing”. Em: *Bioinformatics*. doi: 10.1093/bioinformatics/btx657.

El-Gebali, Sara et al. (out. de 2018). “The Pfam protein families database in 2019”. Em: *Nucleic Acids Research* 47.D1, pp. D427–D432. issn: 0305-1048. doi: 10.1093/nar/gky995.

Goffeau, A. et al. (1996). “Life with 6000 Genes”. Em: *Science* 274.5287, pp. 546–567. doi: 10.1126/science.274.5287.546.

González-Morales, Sandra Isabel et al. (2016). “Regulatory network analysis reveals novel regulators of seed desiccation tolerance in Arabidopsis thaliana”. Em: *Proceedings*

of the National Academy of Sciences 113.35, E5232–E5241. doi: 10 .1073 / pnas . 1610985113.

Goodstein, David M. et al. (nov. de 2011). "Phytozome: a comparative platform for green plant genomics". Em: *Nucleic Acids Research* 40.D1, pp. D1178–D1186. issn: 0305-1048. doi: 10.1093/nar/gkr944.

Gordillo Altamirano, Fernando et al. (fev. de 2021). "Bacteriophage-resistant Acinetobacter baumannii are resensitized to antimicrobials". Em: *Nature Microbiology* 6.2, pp. 157–161. issn: 2058-5276. doi: 10.1038/s41564-020-00830-7.

Granzotto, Adriana e Guilherme Marcello Queiroga Cruz (2015). "Regulação de Elementos de Transposição: Mecanismos Epigenéticos de Silenciamento, Autorregulação e Ativação por Estresse". Em: *Elementos de transposição: diversidade, evolução, aplicações e impacto nos genomas dos seres vivos*. Ed. por Claudia Marcia Aparecida Carareto, Claudia Barros Monteiro-Vitorello e Marie-Anne Van Sluys. São José do Rio Preto: Editora FIOCRUZ, pp. 91–113. isBn: 978-85-7541-462-0. doi: https://doi.org/10.7476/9788575415672.

Greilhuber, Johann e I J. Leitch (2013). "Genome Size and the Phenotype". Em: *Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes*. Ed. por Johann Greilhuber, Jaroslav Dolezel e Jonathan F. Wendel. Vienna: Springer Vienna, pp. 323–344. isBn: 978-3-7091-1160-4. doi: 10.1007/978-3-7091-1160-4_20.

Groth, Philip et al. (set. de 2006). "PhenomicDB: a new cross-species genotype/phenotype resource". Em: *Nucleic Acids Research* 35.suppl_1, pp. D696–D699. issn: 0305-1048. doi: 10.1093/nar/gkl662.

Harvey, Paul H, Mark D Pagel et al. (1991). *The comparative method in evolutionary biology*. Vol. 239. Oxford university press Oxford.

Haynes, Winston A., Aurelie Tomczak e Purvesh Khatri (jan. de 2018). "Gene annotation bias impedes biomedical research". Em: *Scientific Reports* 8.1, p. 1362. issn: 2045-2322. doi: 10.1038/s41598-018-19333-x.

Heather, James M. e Benjamin Chain (2016). "The sequence of sequencers: The history of sequencing DNA". Em: *Genomics* 107.1, pp. 1–8. issn: 0888-7543. doi: https://doi.org/10.1016/j.ygeno.2015.11.003.

Hidalgo, Oriane et al. (2017). “Is There an Upper Limit to Genome Size?” Em: *Trends in Plant Science* 22.7, pp. 567–573. issn: 1360-1385. doi: https://doi.org/10.1016/j.tplants.2017.04.005.

Hongo, Jorge Augusto et al. (2021). “CALANGO: an annotation-based, phylogeny-aware comparative genomics framework for exploring and interpreting complex genotypes and phenotypes”. Em: *bioRxiv*. doi: 10.1101/2021.08.25.457574.

Hung, Jui-Hung et al. (set. de 2011). “Gene set enrichment analysis: performance evaluation and usage guidelines”. Em: *Briefings in Bioinformatics* 13.3, pp. 281–291. issn: 1467-5463. doi: 10.1093/bib/bbr049.

Huo, Heqiang, Shouhui Wei e Kent J. Bradford (2016). “DELAY OF GERMINATION1 (DOG1) regulates both seed dormancy and flowering time through microRNA pathways”. Em: *Proceedings of the National Academy of Sciences* 113.15, E2199–E2206. doi: 10.1073/pnas.1600558113.

IHGSC et al. (fev. de 2001). “Initial sequencing and analysis of the human genome”. Em: *Nature* 409.6822, pp. 860–921. issn: 1476-4687. doi: 10.1038/35057062.

Initiative, The Arabidopsis Genome (dez. de 2000). “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana”. Em: *Nature* 408.6814, pp. 796–815. issn: 1476-4687. doi: 10.1038/35048692.

Jones, Philip et al. (jan. de 2014). “InterProScan 5: genome-scale protein function classification”. Em: *Bioinformatics* 30.9, pp. 1236–1240. issn: 1367-4803. doi: 10.1093/bioinformatics/btu031.

Kang, Ming et al. (2014). “Adaptive and nonadaptive genome size evolution in Karst endemic flora of China”. Em: *New Phytologist* 202.4, pp. 1371–1381. doi: https://doi.org/10.1111/nph.12726.

Kattge, Jens et al. (2020). “TRY plant trait database – enhanced coverage and open access”. Em: *Global Change Biology* 26.1, pp. 119–188. doi: https://doi.org/10.1111/gcb.14904.

Kawahara, Yoshihiro et al. (fev. de 2013). “Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data”. Em: *Rice* 6.1, p. 4. issn: 1939-8433. doi: 10.1186/1939-8433-6-4.

Kawashima, Tomokazu et al. (jul. de 2015). “Diversification of histone H2A variants during plant evolution”. Em: *Trends in Plant Science* 20.7, pp. 419–425. issn: 1360-1385. doi: 10.1016/j.tplants.2015.04.005.

Knight, Charles A., Nicole A. Molinari e Dmitri A. Petrov (jan. de 2005). “The Large Genome Constraint Hypothesis: Evolution, Ecology and Phenotype”. Em: *Annals of Botany* 95.1, pp. 177–190. issn: 0305-7364. doi: 10.1093/aob/mci011.

Koornneef, Maarten, Leónie Bentsink e Henk Hilhorst (2002). “Seed dormancy and germination”. Em: *Current Opinion in Plant Biology* 5.1, pp. 33–36. issn: 1369-5266. doi: https://doi.org/10.1016/S1369-5266(01)00219-9.

Kopriva, Stanislav e Andreas P M Weber (jan. de 2021). “Genetic encoding of complex traits”. Em: *Journal of Experimental Botany* 72.1, pp. 1–3. issn: 0022-0957. doi: 10.1093/jxb/eraa498.

Krishnakumar, Vivek et al. (nov. de 2014). “Araport: the Arabidopsis Information Portal”. Em: *Nucleic Acids Research* 43.D1, pp. D1003–D1009. issn: 0305-1048. doi: 10.1093/nar/gku1200.

Kuang, Kevin, Quyu Kong e Francesco Napolitano (2022). *pbmcapply: Tracking the Progress of Mc*pply with Progress Bar*. R package version 1.5.1.

Kumar, Sudhir, Glen Stecher et al. (mai. de 2018). “MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms”. Em: *Molecular Biology and Evolution* 35.6, pp. 1547–1549. issn: 0737-4038. doi: 10.1093/molbev/msy096.

Kumar, Sudhir, Michael Suleski et al. (ago. de 2022). “TimeTree 5: An Expanded Resource for Species Divergence Times”. Em: *Molecular Biology and Evolution* 39.8. msac174. issn: 1537-1719. doi: 10.1093/molbev/msac174.

Kumar, Vikash, Evgeniy N. Donev et al. (2020). “Genome-Wide Identification of Populus Malectin/Malectin-Like Domain-Containing Proteins and Expression Analyses Reveal Novel Candidates for Signaling and Regulation of Wood Development”. Em: *Frontiers in Plant Science* 11, p. 588846. doi: 10.3389/fpls.2020.588846.

Kumar, Vikash, Matthieu Hainaut et al. (2019). “Poplar carbohydrate-active enzymes: whole-genome annotation and functional analyses based on RNA expression data”. Em: *The Plant Journal* 99.4, pp. 589–609. doi: https://doi.org/10.1111/tpj.14417.

Lanfear, Robert et al. (mai. de 2013). “Taller plants have lower rates of molecular evolution”. Em: *Nature Communications* 4.1, p. 1879. issn: 2041-1723. doi: 10.1038/ncomms2836.

Lee, Byung Ha et al. (jul. de 2021). “A species-specific functional module controls formation of pollen apertures”. Em: *Nature Plants* 7.7, pp. 966–978. issn: 2055-0278. doi: 10.1038/s41477-021-00951-9.

Lee, Heewook et al. (2012). “Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing”. Em: *Proceedings of the National Academy of Sciences* 109.41, E2774–E2783. doi: 10.1073/pnas.1210309109.

Lei, Bingkun e Frédéric Berger (2020). “H2A Variants in Arabidopsis: Versatile Regulators of Genome Activity”. Em: *Plant Communications* 1.1, p. 100015. issn: 2590-3462. doi: https://doi.org/10.1016/j.xplc.2019.100015.

Leitch, A. R. e I. J. Leitch (2012). “Ecological and genetic factors linked to contrasting genome dynamics in seed plants”. Em: *New Phytologist* 194.3, pp. 629–646. doi: https://doi.org/10.1111/j.1469-8137.2012.04105.x.

Leitch, I. J., Mark W. Chase e Michael D. Bennett (dez. de 1998). “Phylogenetic Analysis of DNA C-values Provides Evidence for a Small Ancestral Genome Size in Flowering Plants”. Em: *Annals of Botany* 82.suppl_1, pp. 85–94. issn: 0305-7364. doi: 10.1006/anbo.1998.0783.

Leitch, I. J., D. E. Soltis et al. (jan. de 2005). “Evolution of DNA Amounts Across Land Plants (Embryophyta)”. Em: *Annals of Botany* 95.1, pp. 207–217. issn: 0305-7364. doi: 10.1093/aob/mci014.

León, M e R Bastías (2015). “Virulence reduction in bacteriophage resistant bacteria.” Em: *Frontiers in Microbiology* 343.6. doi: http://dx.doi.org/10.3389/fmicb.2015.00343.

Li, Fay-Wei et al. (jul. de 2018). “Fern genomes elucidate land plant evolution and cyanobacterial symbioses”. Em: *Nature Plants* 4.7, pp. 460–472. issn: 2055-0278. doi: 10.1038/s41477-018-0188-8.

Li, Linzhou et al. (set. de 2020). “The genome of Prasinoderma coloniale unveils the existence of a third phylum within green plants”. Em: *Nature Ecology &, Evolution* 4.9, pp. 1220–1231. issn: 2397-334X. doi: 10.1038/s41559-020-1221-7.

Li, Weizhong e Adam Godzik (mai. de 2006). “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”. Em: *Bioinformatics* 22.13, pp. 1658–1659. issn: 1367-4803. doi: 10.1093/bioinformatics/btl158.

Liolios, Konstantinos et al. (nov. de 2009). “The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata”. Em: *Nucleic Acids Research* 38.suppl_1, pp. D346–D354. issn: 0305-1048. doi: 10.1093/nar/gkp848.

Lisch, Damon (jan. de 2013). “How important are transposons for plant evolution?” Em: *Nature Reviews Genetics* 14.1, pp. 49–61. issn: 1471-0064. doi: 10.1038/nrg3374.

Liu, Jian-Zhong e Steven A. Whitham (2013). “Overexpression of a soybean nuclear localized type–III DnaJ domain-containing HSP40 reveals its roles in cell death and disease resistance”. Em: *The Plant Journal* 74.1, pp. 110–121. doi: https:// doi.org/10.1111/tpj.12108.

M, Carlson (2019). *GO.db: A set of annotation maps describing the entire Gene Ontology*. R package version 3.8.2.

Ma, Xuelian et al. (set. de 2022). “PlantGSAD: a comprehensive gene set annotation database for plant species”. Em: *Nucleic Acids Research* 50.D1, pp. D1456–D1467. issn: 0305-1048. doi: 10.1093/nar/gkab794.

Manni, Mosè et al. (jul. de 2021). “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes”. Em: *Molecular Biology and Evolution* 38.10, pp. 4647–4654. issn: 1537-1719. doi: 10.1093/molbev/msab199.

Marks, Rose A. et al. (dez. de 2021). “Representation and participation across 20 years of plant genome sequencing”. Em: *Nature Plants* 7.12, pp. 1571–1578. issn: 2055-0278. doi: 10.1038/s41477-021-01031-8.

Mashau, Aluoneswi C. et al. (2021). “Plant height and lifespan predict range size in southern African grasses”. Em: *Journal of Biogeography* 48.12, pp. 3047–3059. doi: https://doi.org/10.1111/jbi.14261.

Maslov, Sergei e Kim Sneppen (jan. de 2017). “Population cycles and species diversity in dynamic Kill-the-Winner model of microbial ecosystems”. Em: *Scientific Reports* 7.1, p. 39642. issn: 2045-2322. doi: 10.1038/srep39642.

Mazuecos-Aguilera, Ismael et al. (2021). “The Role of INAPERTURATE POLLEN1 as a Pollen Aperture Factor Is Conserved in the Basal Eudicot Eschscholzia californica (Papaveraceae)”. Em: *Frontiers in Plant Science* 12. issn: 1664-462X. doi: 10.3389/fpls.2021.701286.

Minelli, Alessandro (2018). “Introducing Plant Evo-Devo”. Em: *Plant Evolutionary Developmental Biology: The Evolvability of the Phenotype*. Cambridge University Press, pp. 1–29. doi: 10.1017/9781139542364.002.

Minh, Bui Quang et al. (fev. de 2020). “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era”. Em: *Molecular Biology and Evolution* 37.5, pp. 1530–1534. issn: 0737-4038. doi: 10.1093/molbev/msaa015.

Moles, Angela T. et al. (2009). “Global patterns in plant height”. Em: *Journal of Ecology* 97.5, pp. 923–932. doi: https://doi.org/10.1111/j.1365-2745.2009.01526.x.

Morgan, Martin (2022). *BiocManager: Access the Bioconductor Project Package Repository*. R package version 1.30.18.

Mosavi, Leila K. et al. (2004). “The ankyrin repeat as molecular architecture for protein recognition”. Em: *Protein Science* 13.6, pp. 1435–1448. doi: https://doi.org/10.1110/ps.03554604.

Mukherjee, Supratim et al. (out. de 2016). “Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements”. Em: *Nucleic Acids Research* 45.D1, pp. D446–D456. issn: 0305-1048. doi: 10.1093/nar/gkw992.

Nagy, László G et al. (jan. de 2020). “Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing”. Em: *Nucleic Acids Research* 48.5, pp. 2209–2219. issn: 0305-1048. doi: 10.1093/nar/gkz1241.

Nakabayashi, Kazumi et al. (jul. de 2012). “The Time Required for Dormancy Release in Arabidopsis Is Determined by DELAY OF GERMINATION1 Protein Levels in Freshly Harvested Seeds”. Em: *The Plant Cell* 24.7, pp. 2826–2838. issn: 1040-4651. doi: 10.1105/tpc.112.100214.

Nasrallah, June B. e Mikhail E. Nasrallah (mar. de 2014). “S-locus receptor kinase signalling”. Em: *Biochemical Society Transactions* 42.2, pp. 313–319. issn: 0300-5127. doi: 10.1042/BST20130222.

Niklas, Karl J. e Ulrich Kutschera (2010). “The evolution of the land plant life cycle”. Em: *New Phytologist* 185.1, pp. 27–41. doi: https://doi.org/10.1111/j.1469-8137.2009.03054.x.

Nishimura, Noriyuki et al. (jun. de 2018). “Control of seed dormancy and germination by DOG1-AHG1 PP2C phosphatase complex via binding to heme”. Em: *Nature Communications* 9.1, p. 2132. issn: 2041-1723. doi: 10.1038/s41467-018-04437-9.

Nishiyama, Eri et al. (2021). “Ancient and recent gene duplications as evolutionary drivers of the seed maturation regulators DELAY OF GERMINATION1 family genes”. Em: *New Phytologist* 230.3, pp. 889–901. doi: https://doi.org/10.1111/nph.17201.

Nishiyama, Takashi et al. (jan. de 2013). “The structure of the deacetylase domain of Escherichia coli PgaB, an enzyme required for biofilm formation: a circularly permuted member of the carbohydrate esterase 4 family”. Em: *Acta Crystallographica Section D* 69.1, pp. 44–51. doi: 10.1107/S0907444912042059.

O’Leary, Nuala A. et al. (nov. de 2015). “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. Em: *Nucleic Acids Research* 44.D1, pp. D733–D745. issn: 0305-1048. doi: 10.1093/nar/gkv1189.

Pagès, H et al. (2022). *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. R package version 1.58.0.

Pang, Shuai et al. (mai. de 2015). “GPU MrBayes V3.1: MrBayes on Graphics Processing Units for Protein Sequence Data”. Em: *Molecular Biology and Evolution* 32.9, pp. 2496–2497. issn: 0737-4038. doi: 10.1093/molbev/msv129.

Paradis, E. e K. Schliep (2019). “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R”. Em: *Bioinformatics* 35, pp. 526–528.

Park, Beom Seok e Jie-Oh Lee (dez. de 2013). “Recognition of lipopolysaccharide pattern by TLR4 complexes”. Em: *Experimental & Molecular Medicine* 45.12, e66–e66. issn: 2092-6413. doi: 10.1038/emm.2013.97.

Pasha, Asher et al. (jul. de 2020). “Araport Lives: An Updated Framework for Arabidopsis Bioinformatics”. Em: *The Plant Cell* 32.9, pp. 2683–2686. issn: 1040-4651. doi: 10.1105/tpc.20.00358.

Pawluk, April et al. (2014). “A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa”. Em: *mBio* 5.2, e00896–14. doi: 10.1128/mBio.00896-14.

Peiffer, Jason A et al. (abr. de 2014). “The Genetic Architecture Of Maize Height”. Em: *Genetics* 196.4, pp. 1337–1356. issn: 1943-2631. doi: 10.1534/genetics.113.159152.

Pellicer, Jaume, Michae F. Fay e I. J. Leitch (set. de 2010). “The largest eukaryotic genome of them all?” Em: *Botanical Journal of the Linnean Society* 164.1, pp. 10–15. issn: 0024-4074. doi: 10.1111/j.1095-8339.2010.01072.x.

Pellicer, Jaume, Oriane Hidalgo et al. (2018). “Genome Size Diversity and Its Impact on the Evolution of Land Plants”. Em: *Genes* 9.2. issn: 2073-4425. doi: 10.3390/genes9020088.

Pellicer, Jaume e I J. Leitch (2020). “The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies”. Em: *New Phytologist* 226.2, pp. 301–305. doi: https://doi.org/10.1111/nph.16261.

Petrov, Dmitri A. (jan. de 2001). “Evolution of genome size: new approaches to an old problem”. Em: *Trends in Genetics* 17.1, pp. 23–28. issn: 0168-9525. doi: 10.1016/S0168-9525(00)02157-0.

— (2002). “Mutational Equilibrium Model of Genome Size Evolution”. Em: *Theoretical Population Biology* 61.4, pp. 531–544. issn: 0040-5809. doi: https://doi.org/10.1006/tpbi.2002.1605.

Pinard, Desre et al. (mai. de 2015). “Comparative analysis of plant carbohydrate active enZymes and their role in xylogenesis”. Em: *BMC Genomics* 16.1, p. 402. issn: 1471-2164. doi: 10.1186/s12864-015-1571-8.

Pinheiro, José, Douglas Bates e R Core Team (2022). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-157.

Plazzi, Federico, Ronald R. Ferrucci e Marco Passamonti (abr. de 2010). “Phylogenetic representativeness: a new method for evaluating taxon sampling in evolutionary studies”. Em: *BMC Bioinformatics* 11.1, p. 209. issn: 1471-2105. doi: 10.1186/1471-2105-11-209.

Proost, Sebastian et al. (out. de 2014). “PLAZA 3.0: an access point for plant comparative genomics”. Em: *Nucleic Acids Research* 43.D1, pp. D974–D981. issn: 0305-1048. doi: 10.1093/nar/gku986.

Pulido, Pablo e Dario Leister (2018). “Novel DNAJ-related proteins in Arabidopsis thaliana”. Em: *The New Phytologist* 217.2, pp. 480–490. issn: 0028646X, 14698137.

Pulkkinen, W S e S I Miller (1991). “A Salmonella typhimurium virulence protein is similar to a Yersinia enterocolitica invasion protein and a bacteriophage lambda outer membrane protein”. Em: *Journal of Bacteriology* 173.1, pp. 86–93. doi: 10.1128/jb.173.1.86-93.1991.

Puttick, Mark N., James Clark e Philip C. J. Donoghue (2015). “Size is not everything: rates of genome size evolution, not C-value, correlate with speciation in angiosperms”. Em: *Proceedings of the Royal Society B: Biological Sciences* 282.1820, p. 20152289. doi: 10.1098/rspb.2015.2289.

Rambaut, Andrew et al. (abr. de 2018). “Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7”. Em: *Systematic Biology* 67.5, pp. 901–904. issn: 1063-5157. doi: 10.1093/sysbio/syy032.

Ramisetty, Bhaskar Chandra Mohan e Pavithra Anantharaman Sudhakari (2019). “Bacterial ‘Grounded’ Prophages: Hotspots for Genetic Renovation and Innovation”. Em: *Frontiers in Genetics* 10. issn: 1664-8021. doi: 10.3389/fgene.2019.00065.

Ren, Ren et al. (2018). “Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms”. Em: *Molecular Plant* 11.3. Genome Biology, pp. 414–428. issn: 1674-2052. doi: https://doi.org/10.1016/j.molp.2018.01.002.

Revell, Liam J. (2012). “phytools: an R package for phylogenetic comparative biology (and other things)”. Em: *Methods in Ecology and Evolution* 3.2, pp. 217–223. doi: https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Roff, Derek A. (1997). *Evolutionary Quantitative Genetics*. New York: Springer New York. isBn: 978-1-4615-4080-9. doi: https://doi.org/10.1007/978-1-4615-4080-9.

Sall, Khadidiatou et al. (2019). “DELAY OF GERMINATION 1-LIKE 4 acts as an inducer of seed reserve accumulation”. Em: *The Plant Journal* 100.1, pp. 7–19. doi: https://doi.org/10.1111/tpj.14485.

Salzberg, Steven L. (mai. de 2019). “Next-generation genome annotation: we still struggle to get it right”. Em: *Genome Biology* 20.1, p. 92. issn: 1474-760X. doi: 10.1186/s13059-019-1715-2.

Sandoval, Francisco J., Yi Zhang e Sanja Roje (nov. de 2008). “Flavin Nucleotide Metabolism in Plants: MONOFUNCTIONAL ENZYMES SYNTHESIZE FAD IN PLASTIDS

*". Em: *Journal of Biological Chemistry* 283.45, pp. 30890–30900. issn: 0021-9258. doi: 10.1074/jbc.M803416200.

Sanger, F. e A.R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". Em: *Journal of Molecular Biology* 94.3, pp. 441–448. issn: 0022-2836. doi: https://doi.org/10.1016/0022-2836(75)90213-2.

Sayers, Eric W et al. (out. de 2019). "GenBank". Em: *Nucleic Acids Research* 48.D1, pp. D84–D86. issn: 0305-1048. doi: 10.1093/nar/gkz956.

Schäffer, Alejandro A. et al. (jul. de 2001). "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements". Em: *Nucleic Acids Research* 29.14, pp. 2994–3005. issn: 0305-1048. doi: 10.1093/nar/29.14.2994.

Schallus, Thomas et al. (2008). "Malectin: A Novel Carbohydrate-binding Protein of the Endoplasmic Reticulum and a Candidate Player in the Early Steps of Protein N-Glycosylation". Em: *Molecular Biology of the Cell* 19.8. PMID: 18524852, pp. 3404–3414. doi: 10.1091/mbc.e08-04-0354.

Schneider, Rene e Staffan Persson (2015). "Another brick in the wall". Em: *Science* 350.6257, pp. 156–157. doi: 10.1126/science.aad3200.

Schuster, Stephan C. (jan. de 2008). "Next-generation sequencing transforms today's biology". Em: *Nature Methods* 5.1, pp. 16–18. issn: 1548-7105. doi: 10.1038/nmeth1156.

SHAPIRO, S. S. e M. B. WILK (dez. de 1965). "An analysis of variance test for normality (complete samples)". Em: *Biometrika* 52.3-4, pp. 591–611. doi: 10.1093/biomet/52.3-4.591.

Sievert, Carson (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. isbn: 978-3-319-24277-4.

Silveira, Cynthia B. e Forest L. Rohwer (jul. de 2016). "Piggyback-the-Winner in host-associated microbial communities". Em: *npj Biofilms and Microbiomes* 2.1, p. 16010. issn: 2055-5008. doi: 10.1038/npjbiofilms.2016.10.

Simmons, Emilia L. et al. (2020). "Biofilm Structure Promotes Coexistence of Phage-Resistant and Phage-Susceptible Bacteria". Em: *mSystems* 5.3, e00877–19. doi: 10.1128/mSystems.00877-19.

Sørensen, Iben et al. (2011). “The charophycean green algae provide insights into the early origins of plant cell walls”. Em: *The Plant Journal* 68.2, pp. 201–211. doi: https://doi.org/10.1111/j.1365-313X.2011.04686.x.

Steyert, Susan R. e James B. Kaper (2012). “Contribution of Urease to Colonization by Shiga Toxin-Producing Escherichia coli”. Em: *Infection and Immunity* 80.8, pp. 2589–2600. doi: 10.1128/IAI.00210-12.

Subburaj, Saminathan et al. (jun. de 2016). “Phylogenetic Analysis, Lineage-Specific Expansion and Functional Divergence of seed dormancy 4-Like Genes in Plants”. Em: *PLOS ONE* 11.6, pp. 1–24. doi: 10.1371/journal.pone.0153717.

Tello-Ruiz, Marcela K et al. (nov. de 2020). “Gramene 2021: harnessing the power of comparative genomics and pathways for plant research”. Em: *Nucleic Acids Research* 49.D1, pp. D1452–D1463. issn: 0305-1048. doi: 10.1093/nar/gkaa979.

Tenenbaum D, Maintainer B (2022). *KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)*. R package version 1.36.3.

Tomaž, Špela, Kristina Gruden e Anna Coll (2022). “TGA transcription factors—Structural characteristics as basis for functional variability”. Em: *Frontiers in Plant Science* 13. issn: 1664-462X. doi: 10.3389/fpls.2022.935819.

Tong, Chao et al. (jan. de 2020). “Comparative Genomics Identifies Putative Signatures of Sociality in Spiders”. Em: *Genome Biology and Evolution* 12.3, pp. 122–133. issn: 1759-6653. doi: 10.1093/gbe/evaa007.

Touchon, Marie et al. (jan. de 2009). “Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths”. Em: *PLOS Genetics* 5.1, pp. 1–25. doi: 10.1371/journal.pgen.1000344.

Tuskan, G. A. et al. (2006). “The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray)”. Em: *Science* 313.5793, pp. 1596–1604. doi: 10.1126/science.1128691.

Ung, Huoi, Wolfgang Moeder e Keiko Yoshioka (set. de 2014). “Arabidopsis Triphosphate Tunnel Metalloenzyme2 Is a Negative Regulator of the Salicylic Acid-Mediated Feedback Amplification Loop for Defense Responses ”. Em: *Plant Physiology* 166.2, pp. 1009–1021. issn: 0032-0889. doi: 10.1104/pp.114.248757.

Vaidya, Gaurav, David J. Lohman e Rudolf Meier (2011). “SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon

information". Em: *Cladistics* 27.2, pp. 171–180. doi: https://doi.org/10.1111/j.1096-0031.2010.00329.x.

Vaidyanathan, Ramnath et al. (2021). *htmlwidgets: HTML Widgets for R*. R package version 1.5.4.

Vandecraen, Joachim et al. (2017). "The impact of insertion sequences on bacterial genome plasticity and adaptability". Em: *Critical Reviews in Microbiology* 43.6. PMID: 28407717, pp. 709–730. doi: 10.1080/1040841X.2017.1303661.

Veselý, Pavel, Petr Bureš e Petr Šmarda (ago. de 2013). "Nutrient reserves may allow for genome size increase: evidence from comparison of geophytes and their sister non-geophytic relatives". Em: *Annals of Botany* 112.6, pp. 1193–1200. issn: 0305-7364. doi: 10.1093/aob/mct185.

Vinogradov, Alexander E (2003). "Selfish DNA is maladaptive: evidence from the plant Red List". Em: *Trends in Genetics* 19.11, pp. 609–614. issn: 0168-9525. doi: https://doi.org/10.1016/j.tig.2003.09.010.

Vitti, Joseph J., Sharon R. Grossman e Pardis C. Sabeti (2013). "Detecting Natural Selection in Genomic Data". Em: *Annual Review of Genetics* 47.1. PMID: 24274750, pp. 97–120. doi: 10.1146/annurev-genet-111212-133526.

Vogel, Christine e Cyrus Chothia (mai. de 2006). "Protein Family Expansions and Biological Complexity". Em: *PLOS Computational Biology* 2.5, pp. 1–13. doi: 10.1371/journal.pcbi.0020048.

Wang, B et al. (2019). "[The China National GeneBank owned by all, completed by all and shared by all]". Em: *Yi Chuan* 20.41, pp. 761–772. doi: 10.16288/j.yczz..

Wang, Dandan et al. (2021). "Which factors contribute most to genome size variation within angiosperms?" Em: *Ecology and Evolution* 11.6, pp. 2660–2668. doi: https://doi.org/10.1002/ece3.7222.

Wang, Xiaoxue et al. (dez. de 2010). "Cryptic prophages help bacteria cope with adverse environments". Em: *Nature Communications* 1.1, p. 147. issn: 2041-1723. doi: 10.1038/ncomms1146.

Waterhouse, Robert M et al. (dez. de 2017). "BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics". Em: *Molecular Biology and Evolution* 35.3, pp. 543–548. issn: 0737-4038. doi: 10.1093/molbev/msx319.

Wendel, Jonathan F. et al. (mai. de 2002). “Feast and famine in plant genomes”. Em: *Genetica* 115.1, pp. 37–47. issn: 1573-6857. doi: 10.1023/A:1016020030189.

Wickham, Hadley (2019). *assertthat: Easy Pre and Post Assertions*. R package version 0.2.1.

— (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman & Hall. isbn: 9781138331457.

Wickham, Hadley, Jay Hesselberth e Maëlle Salmon (2022). *pkgdown: Make Static HTML Documentation for a Package*. R package version 2.0.3.

Willi, Yvone e Ary A. Hoffman (2009). “Demographic factors and genetic variation influence population persistence under environmental change”. Em: *Journal of Evolutionary Biology* 22.1, pp. 124–133. doi: https : / / doi . org / 10 . 1111 / j . 1420 - 9101.2008.01631.x.

Wolf, Andrea J. e David M. Underhill (abr. de 2018). “Peptidoglycan recognition by the innate immune system”. Em: *Nature Reviews Immunology* 18.4, pp. 243–254. issn: 1474-1741. doi: 10.1038/nri.2017.136.

Wolf, Jason B. (2002). “The geometry of phenotypic evolution in developmental hyperspace”. Em: *Proceedings of the National Academy of Sciences* 99.25, pp. 15849–15851. doi: 10.1073/pnas.012686699.

Xiao, Yu et al. (mai. de 2019). “Mechanisms of RALF peptide perception by a heterotypic receptor complex”. Em: *Nature* 572.7768, pp. 270–274. issn: 1476-4687. doi: 10.1038/s41586-019-1409-7.

Xie, Yihui, Joe Cheng e Xianying Tan (2022). *DT: A Wrapper of the JavaScript Library ’DataTables’*. R package version 0.23.

Xue, Han et al. (out. de 2021). “qPTMplants: an integrative database of quantitative post-translational modifications in plants”. Em: *Nucleic Acids Research* 50.D1, pp. D1491–D1499. issn: 0305-1048. doi: 10.1093/nar/gkab945.

Yang, He et al. (2021). “Malectin/Malectin-like domain-containing proteins: A repertoire of cell surface molecules with broad functional potential”. Em: *The Cell Surface* 7, p. 100056. issn: 2468-2330. doi: https://doi.org/10.1016/j.tcsw.2021.100056.

Yang, Xiaohan et al. (set. de 2019). “Comparative genomics can provide new insights into the evolutionary mechanisms and gene function in CAM plants”. Em: *Journal*

of Experimental Botany 70.22, pp. 6539–6547. issn: 0022-0957. doi: 10.1093/jxb/erz408.

Yelagandula, Ramesh et al. (jul. de 2014). "The Histone Variant H2A.W Defines Heterochromatin and Promotes Chromatin Condensation in Arabidopsis". Em: *Cell* 158.1, pp. 98–109. issn: 0092-8674. doi: 10.1016/j.cell.2014.06.006.

Zhang, Jian et al. (fev. de 2020). "The hornwort genome and early land plant evolution". Em: *Nature Plants* 6.2, pp. 107–118. issn: 2055-0278. doi: 10.1038/s41477-019-0588-4.

Zu, Pengjuan e Florian P. Schiestl (2017). "The effects of becoming taller: direct and pleiotropic effects of artificial selection on plant height in Brassica rapa". Em: *The Plant Journal* 89.5, pp. 1009–1019. doi: https://doi.org/10.1111/tpj.13440.

Zwickl, Derrick J. e David M. Hillis (jul. de 2002). "Increased Taxon Sampling Greatly Reduces Phylogenetic Error". Em: *Systematic Biology* 51.4, pp. 588–598. issn: 1063-5157. doi: 10.1080/10635150290102339.

2.11 SUPPLEMENTARY MATERIALS

Figure 3 full caption: Angiosperm data. A) Maximum height variation across the phylogeny of 54 angiosperms species with high-quality proteomes available. Species names are as follows: *Ananas comosus* (Aco), *Arabidopsis halleri* (Aha), *Arabis nemorensis* (Ane), *Arabidopsis thaliana* (Ath), *Amborella trichopoda* (Ati), *Acer yangbiense* (Aya), *Brachypodium distachyon* (Bdi), *Boechera stricta* (Bst), *Brachypodium sylvaticum* (Bsy), *Beta vulgaris* (Bvu), *Capsicum annuum* (Can), *Capsicum baccatum* (Cba), *Cajanus cajan* (Cca), *Capsicum chinense* (Cch), *Citrus clementina* (Ccl), *Cinnamomum kanehirae* (Cka), *Carex littledalei* (Cli), *Cinnamomum micranthum* (Cmi), *Castanea mollissima* (Cmo), *Cannabis sativa* (Csa), *Citrus sinensis* (Csi), *Citrus unshiu* (Cun), *Cleome violacea* (Cvi), *Descurainia sophioides* (Dso), *Eucalyptus grandis* (Egr), *Erythranthe guttata* (Egt), *Fragaria vesca* (Fve), *Jatropha curcas* (Jcu), *Lactuca saligna* (Lsa), *Musa balbisiana* (Mba), *Macleaya cordata* (Mco), *Morus notabilis* (Mno), *Nymphaea colorata* (Nco), *Oryza brachyantha* (Obr), *Oryza meyeriana* (Ome), *Oryza sativa* (Osj), *Prunus avium* (Pav), *Prunus dulcis* (Pdu), *Punica granatum* (Pgr), *Phaseolus lunatus* (Plu), *Prunus mume* (Pmu), *Prunus persica* (Ppr), *Phaseolus vulgaris* (Pvu), *Ricinus communis* (Rco),

Rhamnella rubrinervis (Rru), *Sorghum bicolor* (Sbi), *Syzygium oleosum* (Soe), *Spinacia oleracea* (Sol), *Solanum tuberosum* (Stu), *Theobroma cacao* (Tca), *Trema orientale* (Tor), *Thalictrum thalictroides* (Tth), *Vigna angularis* (Van). B) Biological processes significantly enriched in protein-coding genes containing domains associated with maximum height in angiosperms. C) Examples of protein domains significantly expanded in taller plants fulfilling different biological roles. From top to bottom: (i) and (ii) increase of genetic diversity through cross-pollination; (iii) cell wall biology; (iv) embryogenesis; (v) immunity and stress response mechanisms. From left to right: linear models from raw count data, phylogeny-aware linear models; boxplot with raw counts; boxplot with normalized counts.

2.11.1 Supplementary Results

Integrated bacteriophages in pathogenic and non-pathogenic *E. coli*

We selected 80 high-quality, gapless *E. coli* genomes (chromosomes and plasmids) that could be unambiguously classified as either pathogenic (51) or non-pathogenic (29) lineages, excluding laboratory strains (Supplementary Table 1). We predicted the numbers and location of prophages in *E. coli* chromosomes using PHASTER (Arndt et al. 2016) and found 729 prophages, with counts varying from one to 22 integrated viruses per *E. coli* genome (Fig. 2A).

We found pathogenic lineages to have a significantly higher count of prophages when compared with non-pathogenic ones (Fig. 2A, “Prophage count”). We also observed pathogenic lineages to possess genome sizes significantly larger than non-pathogenic lineages (Fig. 2A, “Genome length”). To exclude the possibility that pathogenic *E. coli* have more integrated prophages due to their larger genomes we proceeded by computing prophage densities for each genome (the number of prophages divided by genome length), which was also found to be significantly higher in pathogenic lineages (Fig. 2B, “Prophage density”). We concluded that pathogenic *E. coli* have an increased prophage occurrence in their genomes even after accounting for genome size and phylogeny as a possible source of bias. We proceeded by using phage density as a QVAL to survey the *E. coli* dataset for biological functions associated with prophage occurrence.

Comparison of functional and homology-based annotation schemas

We investigated whether a GO-based annotation effectively integrates information from non-homologous domains that fulfill the same biological role by evaluating the prevalence of annotation terms, defined as the percentage of genomes where they were observed. This is a useful statistic to evaluate whether a GO-based annotation, which is expected to capture non-homologous functional similarities, is more observed across genomes when compared with a homology-based annotation schema. The sum of annotation term occurrence in all genomes also provides a proxy for the overall abundance of annotation terms in distinct annotation schemas. If a GO-based annotation of a set of genomic components integrates the information from non-homologous coding regions that fulfill the same biological roles, we expect it to have both a greater prevalence and a greater sum value than a Pfam-based annotation of the same set of components. Additionally, we also expect these initial differences in the occurrence of annotation terms to be reflected in the annotation terms found to be significantly associated with prophage density.

The distribution of prevalence values in all the 3729 annotation terms from *domain2Pfam* analysis observed in at least one genome revealed a highly skewed distribution, where most annotation terms are observed in the vast majority of genomes (Supplementary Figure 1C, "Prevalence" chart, "domain2Pfam all"). We found that 65.7% of the Pfam IDs (2450 out of 3729) are observed in more than 90% of the *E. coli* proteomes. A minor, second distribution peak is observed in Pfam IDs with the smallest prevalence values, comprising 11.6% of the Pfam IDs with prevalence < 0.1 (432 out of 3729 Pfam IDs).

The distribution of GO terms used to annotate the same set of genomic components (*domain2GO* experiment) has an even more biased distribution regarding terms with high prevalence values, where 88.8% (1768 out of 1992 terms found annotating at least one Pfam entry) have prevalence values greater than 0.9 (Supplementary Figure 1C, "Prevalence" chart, "domain2GO_all"). Furthermore, only 2.3% of the GO terms have prevalence values smaller than 0.1 (46 out of 1992). This indicates that, even though the majority of terms from both functional- and homology-based annotation schemas are observed in most genomes, the terms from the functional-based schema are significantly more prevalent than those from a homology-based one (Wilcoxon test, p-value < 2.22e-16). The higher prevalence observed in our GO-based annotation suggests this annotation schema is indeed capable of better representing common biological themes shared across genomes than a

homology-based one.

The higher values of prevalence observed in annotation terms from *domain2GO* appear to be reflected in significantly associated terms from this annotation schema (Supplementary Figure 1C, "Prevalence" chart, "domain2GO_all" and "domain2GO_sig"), which were also found to be highly biased towards high prevalence values (80.8%, 177 out of 219). The observed difference in the median prevalence from all terms from *domain2GO* annotation and the significant terms from the same annotation was not found to be statistically significant (Wilcoxon test, p-value = 0.057).

In contrast to *domain2GO* annotation, the significant annotation terms from the *domain2Pfam* experiment have much lower prevalence values than the distribution observed in all *domain2Pfam* annotation terms (Supplementary Figure 1C, "Prevalence" chart, "domain2Pfam_all" and "domain2Pfam_sig"). Only 25,7% (59 out of 230) of the significant annotation terms have prevalence values greater than 0.9. We also found the median prevalence value for the associated terms to be significantly lower than both the one observed for all annotation terms from the *domain2Pfam* experiment and the significant terms from *domain2GO* experiment (Wilcoxon test, p-values < 2.22e-16).

We further evaluated annotation term occurrence in homology- and functional-based annotation schemas by computing the sum of occurrences of each annotation term in each annotation schema (Supplementary Figure 1C, "Sum" chart). We initially observed that the homologous regions have a significantly smaller sum of occurrences than GO terms (medians of 80 and 227, respectively; Wilcoxon test, p-value < 2.22e-16). The distribution of significant annotation terms from homology-based annotations schema was also found to be more similar to the distribution of sums of all *domain2Pfam* terms (medians of 108.5 and 80, respectively; Wilcoxon test, p-value = 3.6e-04) than the significant associations of functional-based annotations when compared with the sum of term occurrences for *domain2GO* experiment (medians of 1125 and 227, respectively for associated GO IDs and all GO terms; Wilcoxon test, p-value < 2.22e-16). Furthermore, associated GO terms have significantly higher sum values than associated Pfam IDs (medians of 1125 and 108.5, respectively; Wilcoxon test, p-value < 2.22e-16). The scenario where annotation terms from a GO-based annotation schema are both more prevalent and more abundant than terms from the Pfam-based annotation is compatible with the several classes of genes non-homologous genes of viral origin that play complementary roles in the viral life cycle found

to be associated with prophage densities. These observations are also compatible with the integration of biological knowledge at the function level, where non-homologous Pfam domains observed in distinct genomes are annotated using the same GO term, therefore increasing both the prevalence and the abundance of GO terms in comparison to Pfam IDs.

Evaluating annotation term frequencies and counts

When evaluating annotation terms associated with QVAL, CALANGO allows users to use either raw count data of annotation terms or to normalize these values to account for biologically meaningful differences across genomes. This normalization procedure may be interesting when evaluating genomes that have, for instance, considerable variation in their number of protein-coding genes, which may cause annotation term counts to be more abundant in genomes with larger numbers of genes.

On the other hand, normalization using a common factor induces dependencies in the relative frequencies for each annotation term, as the resulting values must add to one (Supplementary Figure S-2A). This causes a mathematical artifact that can produce spurious significant associations of frequency of terms having constant or near-constant count values, due to relative frequencies of other terms whose counts are associated with QVAL.

Supplementary Figure S-2A illustrates five hypothetical genomes with two annotation terms, one occurring once in each of them (*e.g.*, a universal 1-1 ortholog, represented as an orange box), and another having a variable count occurrence that is associated with a QVAL vector (blue box). When computing associations from count data, we only observe the blue term to be associated with QVAL (Supplementary Figure S-2A, "Blue count" and "Orange count" plots), as expected.

When examining relative frequencies, however, we notice the emergence of a spurious association of orange frequencies with the QVAL, in addition to the expected association of blue terms (Fig. S-2A, "Blue freq" and "Orange freq" plots). This results from the dependency structure between the blue terms and the QVAL "contaminating" the QVAL-independent orange counts via the normalization denominator, which results in a negative correlation between the orange and blue frequencies (Fig. S-2A, "Blue and Orange freq" plot).

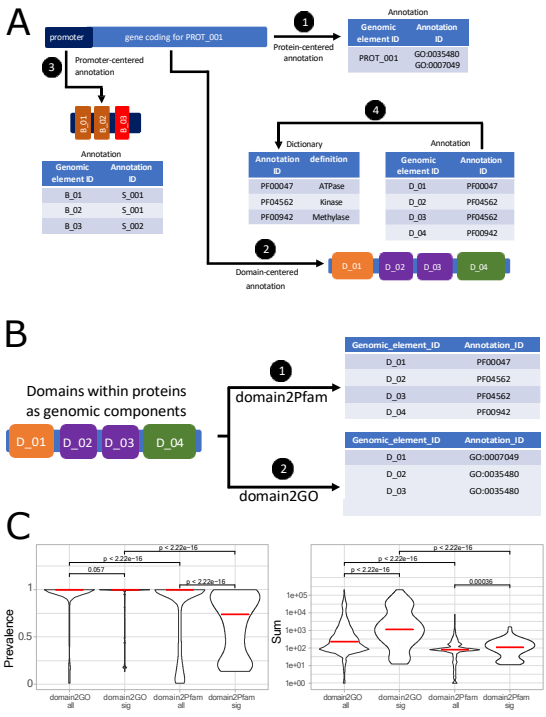


Figura S-1: Genomic components, annotation terms, and feature normalization. A) Genomic components and their annotation terms: 1) Protein-centered annotation to Gene Ontology terms. A single protein-coding gene may be annotated to as many GO terms as needed. 2) Domain-centered annotation of protein domains to domain IDs, such as the ones available in Pfam. 3) Promoter-centered annotation of distinct binding sites of specific DNA binding proteins. 4) relationship between annotation term IDs and their definitions (in this case, hypothetical protein domain IDs and their biological roles). B) Hypothetical genomes and their associated metadata for normalization. The figure represents annotation terms with either the same number of copies in all genomes (orange boxes) or with a variable number of copies in all genomes (blue boxes). The table contains QVAL for these genomes, the normalization factor for annotation counts (sum of all annotation terms), raw annotation term counts, and their relative frequencies (raw counts of each annotation term in a genome normalized by the sum of all annotation terms in a genome). C) Correlation of QVAL, count, and frequency data. Each plot contains the correlation for the hypothetical variables in Supplementary Figure 1B. It is possible to observe that, the annotation term represented by blue boxes is associated with QVAL (both count and normalized values). The orange term, however, is not associated when considering count data, as it is present as a single-copy region in all genomes. As the relative frequencies of the orange annotation terms in each genome are dependent on the relative frequency of blue annotation terms, a relative increase in the frequency of blue annotation terms causes the relative decrease of orange annotation terms in the same genome.

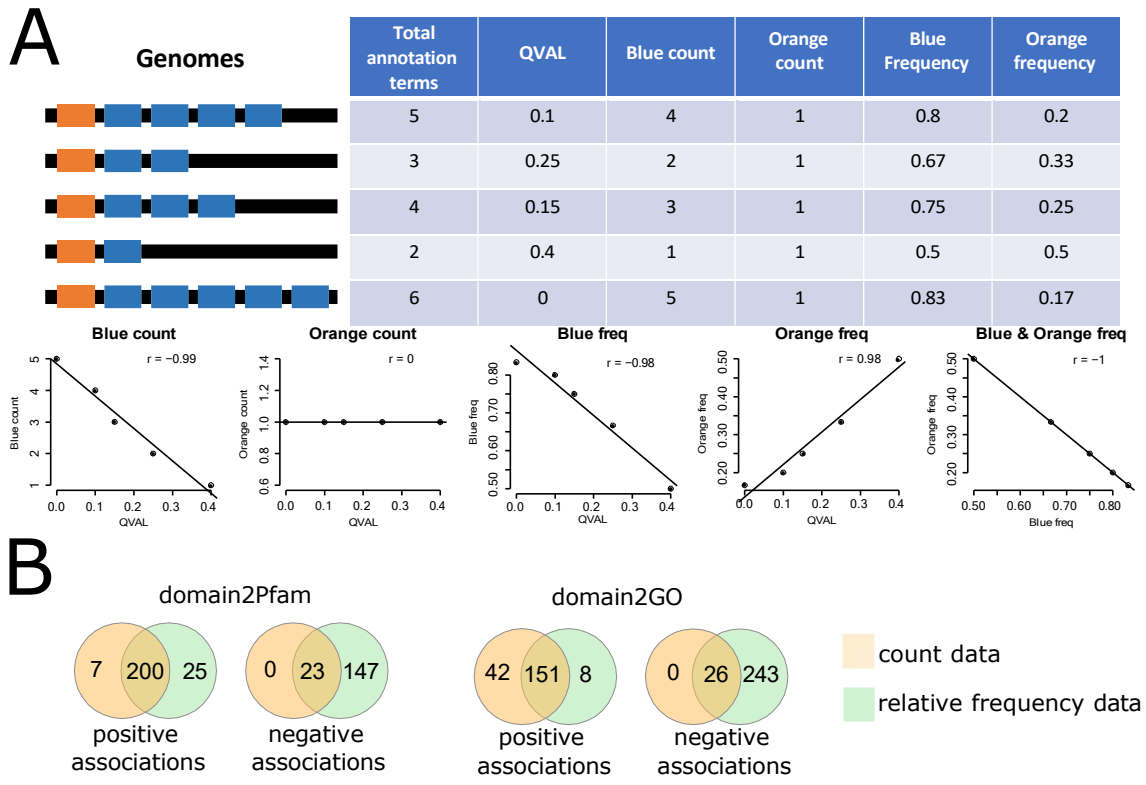


Figura S-2: Heatmap of protein domains associated with prophage density in *E. coli* after the removal of genes of viral origin as produced by CALANGO. Species clustering was performed using the E. coli phylogeny produced in this study, annotation terms clustering was performed using Manhattan distance and average clustering. Red arrows indicate *E. coli* lineages with some of the highest values of domains of viral origin (before removal of genes of viral origin) and virulence factors.

In this work we initially used CALANGO to survey both annotation term counts and relative frequencies for possible associations with prophage densities (Supplementary Table 2, Fig. S-2B). Although the majority of positively associated terms were observed to be the same when considering either count or relative frequencies, we found most of the negatively associated terms to be detected only when considering relative frequencies (Fig. S-2B).

Furthermore, manual inspection of the annotation terms detected in the second case found a considerable fraction of them to correspond to core housekeeping cellular processes (Supplementary Table 2), further suggesting that their counts are likely to be near constant across genomes, and that the negative associations found in frequency data is likely to be an artifact caused by the issue highlighted earlier. Taken together, these experiments demonstrate that relative frequencies are not always the most adequate representation of annotation term occurrence, which motivated our decision to proceed with count data alone for the *E. coli* dataset.

Homologous genes and biological roles associated with prophage density in *E. coli*

Figure 2 shows an annotated version of the heatmap produced by CALANGO using *domain2Pfam*, representing all Pfam domain IDs associated with prophage density. This figure demonstrates how our tool integrates phylogenetic, phenotypic, and annotation data to allow downstream exploratory analysis. Species clustering is done using the phylogenetic tree, together with visualization of user-defined groups (pathogenic and non-pathogenic *E. coli* lineages, in our case), allowing the detection of interesting distribution patterns of annotation terms while considering, in our case, phylogenetic and phenotypic/ecological information. Pathogenic and non-pathogenic lineages are distributed with no clear broader grouping pattern, suggesting that both phenotypes emerged and/or were lost several times during the evolution of these lineages. Two pathogenic *E. coli* groups have the highest count of most Pfam domains (Fig. 2, red arrows). These lineages comprise enterohemorrhagic (EHEC) and enteropathogenic (EPEC) pathotypes, including all O157:H7 lineages (larger group), a Shiga-like toxin-producing serotype, and an important source of foodborne disease (Steyert e Kaper 2012).

We found the domain clustering to comprise mostly two major functional classes:

groups of non-homologous regions of viral origin that play several roles in the viral life cycle, such as integrases, structural proteins, lysozymes, and DNA metabolism enzymes, together with three main classes of virulence factors known to play important roles in EHEC and EPEC disease etiology: Type III secretion systems, urease, adhesins, hemolysins, NFkB-degrading protease, attaching and effacing virulence factors, and Shiga-like toxin (2-4). The two groups of pathogenic *E. coli* with the highest count of associated Pfam IDs also contain most of the genomes where the virulence factor clusters were observed.

Positive association: virulence factors and insertion elements

The second most abundant category of protein domains found to be associated with prophage density were several classes of virulence factors. Interestingly, we found that 44 out of the 58 (75,9%) of domains annotated by us as virulence factors are still significantly associated with prophage density even after the removal of viral genes, suggesting that a considerable fraction of such genomic elements is not located within predicted prophage genomes (Supplementary Table S-2, sheet "virulence_factor", Supplementary File 1, section “Annotation terms associated with prophage density after removal of genes of viral origin”). In fact, from the 4379 domains coding for virulence factors, 3122 (71.3%) are located outside detectable regions of viral origin.

We observed that 7 out of the 14 virulence factor domains not associated with prophage density after the removal of the genes of viral origin are components of pathogenicity mechanisms long known to be horizontally transferred by bacteriophage integration, such as Shiga-like toxins and effectors of Type III secretion system (Steyert e Kaper 2012; Ehrbar e Hardt 2005) (*e.g.* all 22 copies of domain PF02258 - *Shiga-like toxin beta subunit*, Supplementary Table 2). This provides additional evidence that our strategy to remove the effect of genes of viral origin was successful.

The unexpected observation that most virulence factors are located outside predicted prophages and remain associated with the QVAL after the removal of genes of viral origin suggests a more complex relationship between virulence factors and prophage occurrence. One possibility is that the presence of these domains is a consequence of previous bacteriophage integration events that, over time, resulted in 1) prophage degeneration by the loss of genes needed for viral replication, as a result of selective pressure for smaller

bacterial genomes with lower replication times; and 2) the maintenance of genes contributing to a pathogenicity phenotype due to a fitness increase provided by such genes in pathogenic *E. coli* lineages (Ramisetty e Sudhakari 2019)(6) - a scenario compatible with the association of domain PF06316 after the removal of viral genes (Fig. S-3E). A non-excluding alternative hypothesis is that some of these virulence domains may have arrived in bacterial genomes by non-phage-mediated horizontal transfer events and, together with virulence factors acquired through bacteriophage integration, contribute to a pathogenic phenotype. More complex hypotheses are also reasonable, such as the occurrence of synergism between the presence of a prophage and the external virulence factor, resulting in increased fitness to the bacteria when both are present.

We found five domains coding for insertion elements to be positively associated with prophage density, even after the removal of genes of viral origin (Supplementary Table 2, sheet "domain2PfamCountLessPhages", PF13007 - *Transposase C of IS166 homeodomain*). Despite having been initially described as classic deleterious genomic parasites, insertion elements have been increasingly documented as important agents of bacterial genome evolution by providing the horizontal acquisition of genomic components that confer adaptation to new niches, including virulence factors, and, interestingly, may provide resistance to bacteriophage infection (Vandecraen et al. 2017).

Although there is plenty of evidence that bacteriophage integration contributes to pathogenic phenotypes in bacterial species through horizontal gene transfer of virulence factors [32, 36], we found no previous studies demonstrating a significant association between prophage density in bacterial genomes and the occurrence of virulence factors, while correcting for possible confounders and known biases, such as phylogeny and genome quality. Furthermore, our results demonstrate that the majority of virulence factors remain associated with prophage density even after the removal of genes of viral origin, therefore excluding the contribution of all known genes of viral origin for these associations. These unforeseen results showcase how CALANGO can be used to enable controlled *in silico* experiments for the investigation of biologically meaningful questions.

Positive association: other biological roles

Besides protein domains associated with phage biology, virulence factors, and insertion elements, CALANGO detected 36 additional domains positively associated with prophage

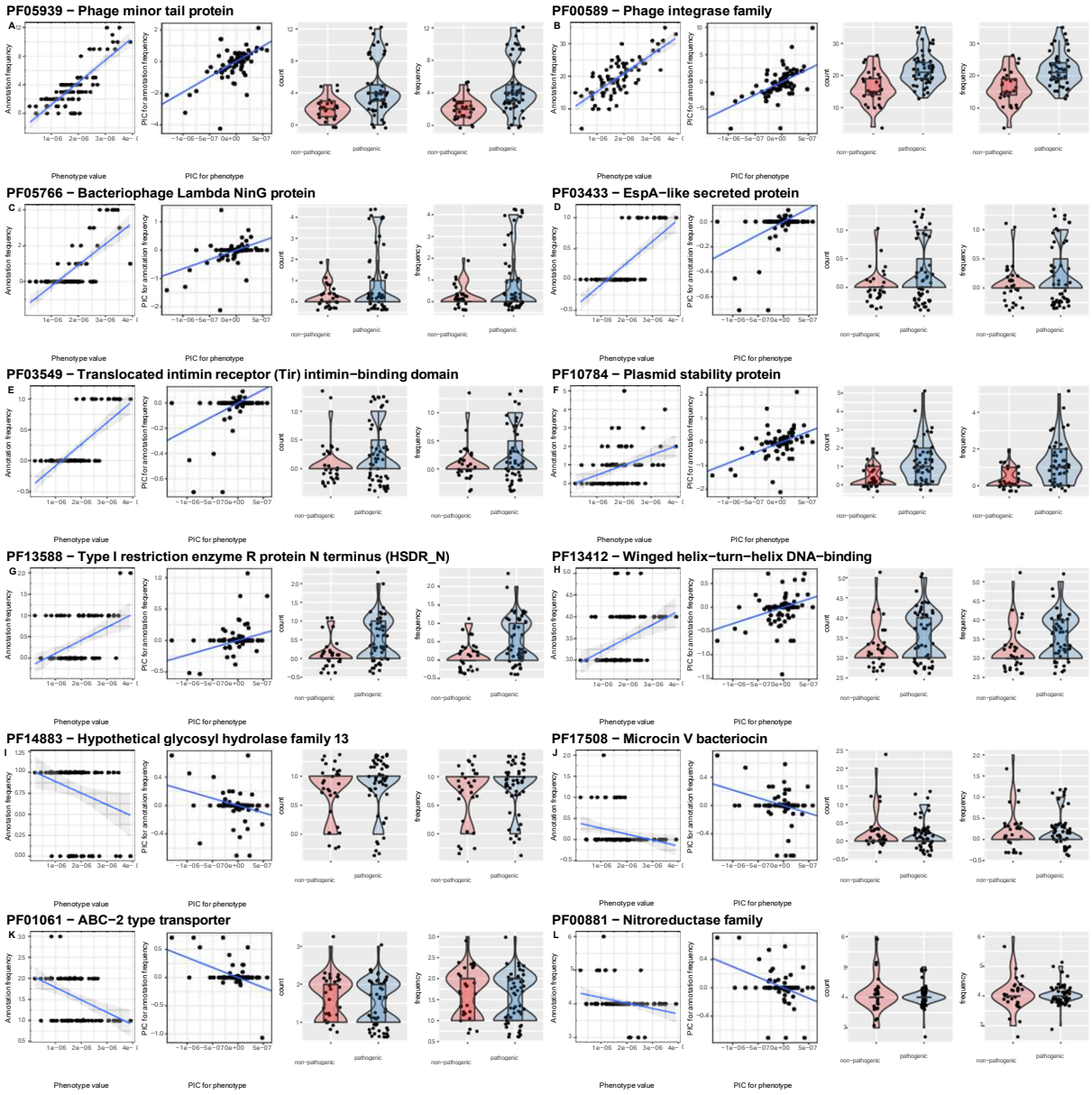


Figura S-3: Additional examples of Pfam domains associated with prophage density. For each Pfam domain ID, from left to right, CALANGO provides two plots for traditional association statistics (a scatterplot, together with a linear model, for direct data visualization, and a scatterplot of ranked data, together with a locally estimated scatterplot smoothing (LOESS)) and a plot containing the contrast data for phylogeny-aware linear regression of phylogenetically independent contrasts.

density that fulfill several biological roles. Remarkably, these include four domains that code for anti-viral components of bacterial immunity, such as putative components of the restriction-modification and CRISPR-Cas systems (*e.g.* PF13588 - *Type I restriction enzyme R protein N terminus (HSDR_N)*, Supplementary Figure S-3G). It may be tempting at this point to assume that integrated prophages constitute a selective pressure for the occurrence and maintenance of such mechanisms. However, these four domains are not associated with prophage density after removal of genes of viral origin, therefore suggesting that at least a fraction of them occur within prophage genomes. We hypothesize these could represent viral genes related to the hijacking of bacterial immunity to prevent, for instance, infection by competing viruses, but further studies are needed for confirmation. Furthermore, some components of anti-viral mechanisms were also found to be negatively associated with prophage density, as we shall see, suggesting a more complex interplay between integrated prophages and the occurrence of anti-viral mechanisms (Supplementary File 1, section "Distinct anti-viral mechanisms are positively and negatively associated with prophage density").

We also found 10 positively associated DUFs that may either comprise domains of viral origin playing roles in viral life cycles, or genes performing other biological functions found to be associated with prophage density, such as previously unknown virulence factors (Supplementary Table 2). Three out of the 10 DUFs remain significantly associated with the QVAL even after the removal of all predicted prophage genes, indicating that at least a fraction of them is located outside detectable integrated viral genomes. This property is shared with several of the known virulence factors we observed, making these domains interesting targets for downstream functional evaluation.

Biological roles negatively associated with prophage density

A total of 23 domains were found to have significant negative associations with prophage density (Fig. 2), and 20 of them are associated after the removal of genes of viral origin. Most of these domains suggest a scenario where *E. coli* lineages with fewer integrated prophages, which tend to be non-pathogenic, have a set of genes enabling a greater diversity of lifestyles at several levels. At the molecular/metabolic level, we found domains coding for metabolic functions (*e.g.* PF00881 - *Nitroreductase family*, Supplementary Figure S-3L) and transmembrane transporters involved in the uptake of ions and organic

molecules (*e.g.* PF00950 - *ABC 3 transport family* 3 transport family, Fig. S-3K).

We found two domains representing distinct facets of bacterial diversity at the population level, the first coding for a putative glycosyl hydrolase described with a known role in biofilm formation (PF14883 - *Hypothetical glycosyl hydrolase family 13*) (T. Nishiyama et al. 2013). Bacteriophages can modulate biofilm formation in several bacterial species, and certain phages are efficient biofilm destroyers (Fernández, Rodríguez e García 2018). However, inside biofilms, susceptible bacterial populations could be protected from phage infections by their resistant counterparts (Simmons et al. 2020). In this case, the presence of the PF14883 domain in the bacterium could be a protective factor against phages, and explain the negative association detected by CALANGO. The second domain (PF17508 - *Microcin V bacteriocin*, Fig. S-3J) codes for a bacteriocin, an umbrella term used to describe protein toxins produced by bacteria that inhibit the growth of closely related species (Cotter, Ross e Hill 2013). The death of related competitors could be a protective measure against phages, allowing a bacteriocin-producing population to keep in check potential viral spreaders in its surroundings during kill-the-winner dynamics (Maslov e Sneppen 2017).

Phage-bacteria interactions inside a metazoan have been shown to differ from those happening in the environment or in laboratory conditions. At the metazoan-environment interface (mucosal layers), the pressure from phages might be greater than elsewhere, considering the impact of enriched mucosal-associated phages (Barr et al. 2013). Also, the mucosal layer could influence lysis-lysogeny decisions, and prophage-containing bacteria have been speculated to be protected in this environment by superinfection exclusion (Silveira e Rohwer 2016). This correlates well to CALANGO findings that pathogenic *E. coli* tends to have more associated prophages, as the inserted viral genes could provide protection against phages colonizing the metazoans during the bacterial invasion process. This situation could also explain why some virulence factors coded by the bacteria are positively associated with prophages: the prophage-mediated protection could add up to the virulence gain, synergistically increasing the bacterial fitness in pathogenic situations.

We also observed a negative association of a component of an anti-viral mechanism, even after removing genes of viral origin (PF04313 - *Type I restriction enzyme R protein N terminus (HSDR_N)*), suggesting this domain may play some role in preventing viral infections in the lineages with a smaller prophage occurrence (Supplementary File 1,

section “Anti-viral mechanisms are positively and negatively associated with prophage density”).

Three interesting examples are the negative associations for two glycosyltransferase domains (PF08437 and PF01501) that are components of the lipopolysaccharide (LPS) biosynthesis pathway, and for one peptidoglycan binding domain (PF01471) found in genes involved in cell wall processes (Fig. 2F, Supplementary Figure S-3I). LPS and peptidoglycan are components of cell wall and have roles both as major players of *E. coli* virulence, but also as receptors of several bacteriophages during adsorption.

LPS are a major component of cell walls of gram-negative bacteria and, together with peptidoglycan, are also receptors of several bacteriophages during adsorption to the bacterial cell surface when starting the infection process (Bertozzi Silva, Storms e Sauvageau 2016). Therefore, the decrease of these domains may be a consequence of the selective pressure caused by bacteriophage infection, resulting in surface modifications leading to the loss or change of molecules used by the phages as viral receptors. Acquisition of phage resistance through surface modification is detrimental to bacterial virulence for different pathogenic bacterial species (León e Bastías 2015; Gordillo Altamirano et al. 2021).

However, LPS molecules have also been long recognized as major activators of the innate branch of the vertebrate immune system (Park e J.-O. Lee 2013). Consequently, the decrease of these domains may also represent evidence of the selective pressures of vertebrate immune systems on pathogenic bacteria, resulting in a decrease in *E. coli* immunogenicity through the loss of bacterial cell wall components. It is possible that both effects may be happening concomitantly, with the selective pressure induced by prophage infection resulting in the decrease of genes coding for components of LPS and other components of the cell wall, but also causing an exaptation scenario where this selection enables pathogenic *E. coli* to better adapt to its niche by partially evading host immune system. This illustrates another powerful aspect of our software, namely its ability to elicit relevant testable hypotheses from the data. In the particular case of this study, it would be possible to evaluate the relative fitness contribution of these domains through targeted knockout in distinct *E. coli* lineages, followed by controlled experiments investigating vertebrate immune system evasion or prophage infection rates.

Anti-viral mechanisms are positively and negatively associated with prophage density

We found five Pfam IDs associated with prophage density that are also components of bacterial immune systems to prevent viral infections, with four positive associations and one negative. Among the positive associations we found two DNA methylases (PF05063 - *MT-A70* and PF01555 - *DNA methylase*), one DNA-binding domain found in CRISPR negative transcriptional regulators (PF13412 - *Winged helix-turn-helix DNA-binding*) (Pawluk et al. 2014) and a restriction component of type I restriction-modification systems (PF13588 - *Type I restriction enzyme R protein N terminus (HSDR_N)*). All positive associations are not observed after the removal of genes of viral origin, indicating that a considerable fraction of these domains is found within prophages (Supplementary Table 2, sheet “domain2PfamCountLessPhages").

Interestingly, both the DNA methylases and the negative transcriptional regulator may confer advantages for bacteriophages to evade bacterial immune systems. As for PF13588, it is worth noting that we also found a domain described as a component of restriction-modification mechanism to be negatively associated with prophage density (PF04313 - *Type I restriction enzyme R protein N terminus (HSDR_N)*), suggesting that some restriction systems may occupy distinct biological roles in extremes of phage density. A possible hypothesis is that some restriction-modification systems may be horizontally transferred by bacteriophage genomes and confer bacterial resistance to additional bacteriophage phage infections to avoid competition (Dedrick et al. 2017). As for the negatively associated restriction system, it is a component of bacterial genomes observed outside prophage regions that may provide a more general bacteriophage infection resistance in *E. coli* with few integrated prophages. Also, the loss of such systems in lineages with greater values of prophage density may suggest such loss may be advantageous for a parasitic lifestyle. We again highlight that CALANGO output produces hypotheses that are testable through genome edition of specific genomic components followed by relative fitness evaluation in controlled environments.

GO annotation provides curation-level of biological knowledge

The *domain2Pfam* analysis required extensive curation of domain functions to provide proper biological context to our findings. Furthermore, protein domain IDs represent

homologous conserved regions within proteins and do not capture biological knowledge, such as functional similarities shared by non-homologous genomic components.

As CALANGO models genomic components independently from annotation terms, it is possible to objectively evaluate the influence of distinct annotation schemas used to annotate the same set of genomic components (Supplemenetary Figure S-1B). At this point, we are interested in evaluating if GO annotation (*domain2GO*) would detect the same major biological themes observed during our manual curation of *domain2Pfam* results. Additionally, we evaluated if the integration of biological knowledge at the function level through GO annotation allows the detection of biological functions associated with prophage density that was not immediately discernible in our *domain2Pfam* analysis.

From the set of 1963 GO terms found to annotate at least five distinct protein domains as predicted by Pfam, CALANGO found 217 (11%) to be significantly associated with prophage density, with 195 positively correlated terms (correlations between 0.25 and 0.86) and 26 negatively associated (correlations between -0.27 and -0.54) (Supplementary Table 2, sheet “domain2GOCount”).

As observed in the *domain2Pfam* analysis, we again found the majority of the positive associations to represent major aspects of bacteriophage biology and life cycle, including both general concepts and more specific components of the lytic and lysogenic cycles (*e.g.* GO:0019058 (*Viral life cycle*) and GO:0019068 (*virion assembly*) (Fig. 2G-H, Fig. S-4A-F, Supplementary Table 2, sheet “domain2GOCount”).

The protein domains annotated to these GOs comprise non-homologous sequences that play complementary roles in bacteriophage life cycles and their relationships with its bacterial host, such as in the several structural components of viral particles (*e.g.*, portal proteins, head-to-tail joining proteins, tail proteins, head proteins, and capsids). Together, these results demonstrate how CALANGO integrates information from non-homologous sequences at the function level to automatically provide the same functional roles found in *domain2Pfam* analysis through manual curation and literature review.

CALANGO also detected associated GO terms representing general and specific aspects of *E. coli* pathogenicity to be the second-largest group of annotation terms positively associated with prophage density. The more general GO term GO:0009405 (pathogenesis) annotates several non-homologous virulence factors, such as cell invasion and adhesion, toxins, hemolysins, colicins and components of secretion systems, and provides further

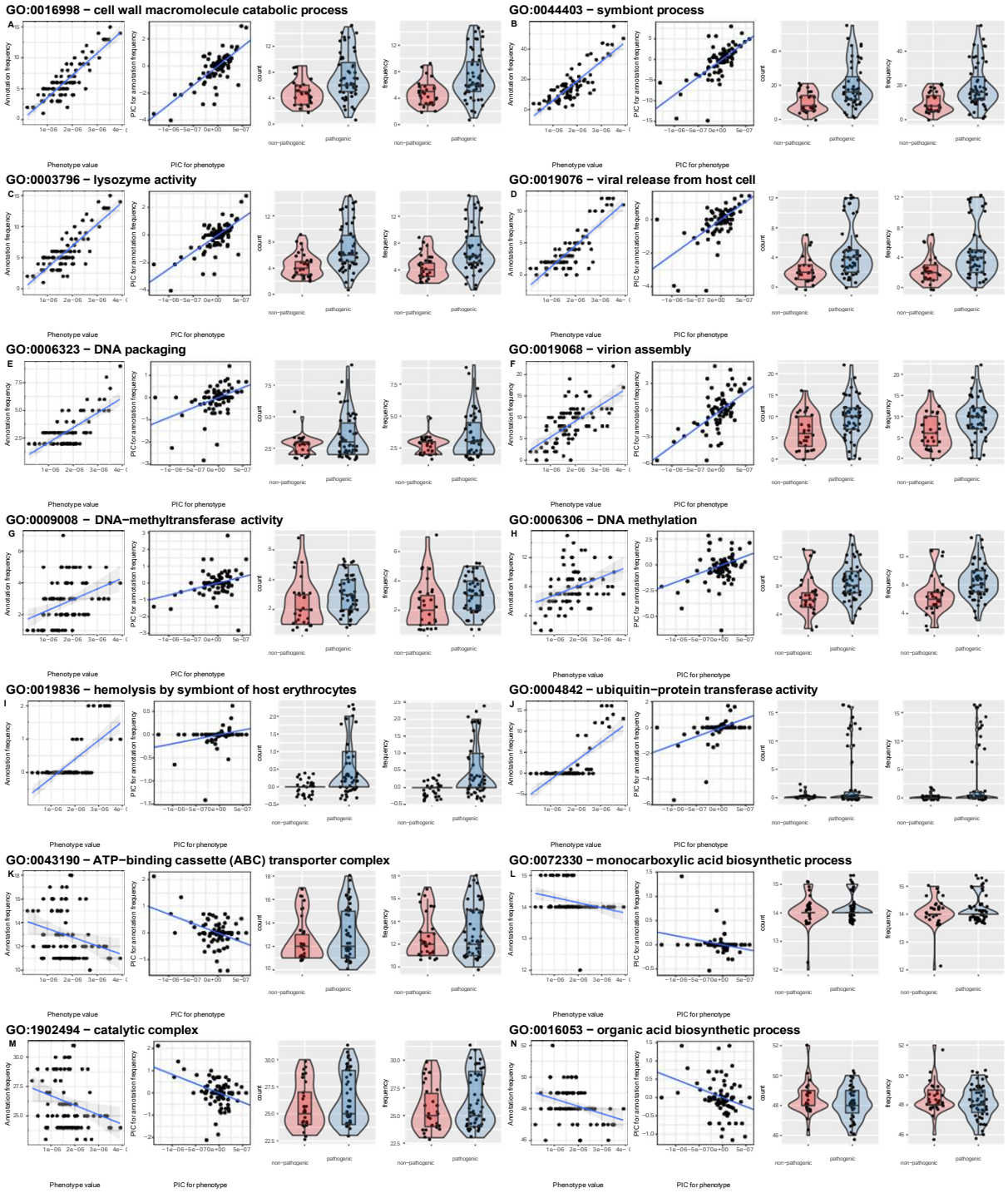


Figura S-4: Additional examples of GO terms associated with prophage density. For each GO term ID, from left to right, CALANGO provides two plots for traditional association statistics (a scatterplot, together with a linear model, for direct data visualization, and a scatterplot of ranked data, together with a locally estimated scatterplot smoothing (LOESS)) and a plot containing the contrast data for phylogeny-aware linear regression of phylogenetically independent contrasts.

evidences of how CALANGO finds associations (Figure S-5E). We also found other GO terms representing specific pathogenicity modules and mechanisms known to play important roles in distinct *E. coli* pathotypes, such as type III secretion and urease activity (Wickham 2020; Y. Xie, Cheng e Tan 2022) (Fig. S-4H-I).

The 26 GO terms negatively associated with prophage density largely reflect the conclusions achieved at the domain-level analysis (*domain2Pfam*) after extensive manual curation (Supplementary Table 2, sheet "domain2GOCount"). We again found negative associations of GO terms reflecting a higher metabolic diversity (*e.g.*, GO:0043190 - *ATP-binding cassette (ABC) transporter complex* ; GO:0072330 - *monocarboxylic acid biosynthetic process*) and also terms indicating the previous negative association between components of the LPS biosynthesis pathway (*e.g.*, GO:0008918 - *lipopolysaccharide 3-alpha-galactosyltransferase activity*, Fig. S-4K-N). Together, these results provide further evidence that, starting from the same set of genomic elements annotated to distinct dictionaries of biological roles, CALANGO can automatically detect biologically meaningful associations that are comparable to our manual curation of the results and equivalent to traditional comparative genomics analysis.

Stress response genes are associated with prophage density in *E. coli*

We found 54 Pfam domains annotated as stress response mechanisms that play roles in several stress-related biological processes (Supplementary Table 2, sheet "stress_respons_genes"). Approximately 30% these domains code for core biological functions observed mostly as single-copy universal orthologs, such as components of DNA repair pathways, oxidative stress pathways, transcription factors, chaperones and heat shock proteins. As they do not vary across genomes, they may not be the ones accounting for the association of GO:0006950 and prophage density.

The other 38 domains were observed in accessory proteins found in some *E. coli* lineages but not in others. Among them we observed components of restriction-modification systems, DNA repair pathways, colicins, toxin-antitoxin systems, Tellurite resistance, and transcription factors. Only two of these 38 domains were also detected in *domain2Pfam* experiment, indicating again that most of these Pfam IDs are not individually detectable as associated with prophage density but, when annotated to GO, the variation patterns of these genomic elements are aggregated in the function level and eventually contribute

for the association to emerge. Additionally, a total of five (13.16%) of these domains are more represented in regions of viral origin than in host's genomes, and 28 of them (73.7%) are found in at least one bacteriophage genome, indicating that the pool of stress response domains is present in both host and viral genomes

Among the stress-response domains we also found some candidates for cellular responses to viral infections, such as the DNA repair pathways and restriction-modification mechanisms. It may be appealing to assume that an increase in genes fulfilling this biological role is the consequence of an adaptation process of host cells to the presence of integrated prophages.

However, domains belonging to these categories were observed both in regions of viral origin and host's genomes, providing additional evidence that anti-viral mechanisms carried out by bacteriophages may comprise an advantage for both hosts and integrated viruses by preventing competition through additional viral infection. We also found other stress response domains that may confer a fitness increase, such as virulence factors, warfare mechanisms and defense systems.

Associations with prophage density after removal of genes of viral origin

When searching for biological functions associated with prophage density in *E. coli* genomes, we know beforehand the location of all predicted prophages. Therefore, it is possible to remove all genes predicted as having viral origin (Supplementary File 1, section "Removal of genes of viral origin") and objectively evaluate the effect of this procedure on associated annotation terms. Such genomes lacking genes located within prophage genomes were annotated using InterProScan (Jones et al. 2014), ordered according to their original prophage densities before the removal of viral genes and evaluated using CALANGO with the same criteria for significance.

For the *domain2Pfam* annotation schema after excluding genes of viral origin, we found 86 Pfam IDs still associated with prophage density (Supplementary Table 2, sheet "domain2PfamCountLessPhages"), 57 of which in common with the ones found in the original *domain2Pfam* experiment including viral genes (Supplementary Table 2, sheet "domain2PfamCount").

The 125 distinct Pfam IDs manually curated as of viral origin and associated with prophage density were found to occur 23,310 times across the *E. coli* proteomes. The

removal of known genes of viral origin decreased the occurrence of these domains to 6,040, a reduction of 74%, therefore demonstrating we have been able to remove the majority of such domains. Furthermore, only two out of 125 domains were still found to be associated with prophage density in this experiment.

The first one is PF06316 (*Enterobacterial Ail/Lom protein*), a protein domain found virulence-related outer membrane protein family that is observed in bacteriophage genes, where it plays a role in lysogenic cycles (Pulkkinen e Miller 1991), and also contributes to a pathogenicity phenotype in gram-negative bacteria by allowing both resistance to complement activity and the ability to adhere and invade host cells (Cirillo et al. 1996). Furthermore, this domain has 289 copies in bacterial genomes before the removal of genes of viral origin, but only 14 copies (4.84%) remain after the removal of such genes. These observations suggest a scenario of bacteriophage-mediated horizontal gene transfer followed by prophage degeneration and the eventual maintenance of virulence factors in pathogenic lineages as a consequence of fitness increase. The second Pfam domain is PF07799 (*Protein of Unknown Function* (DUF1643)), a DUF found in several proteins in Archaea, Bacteria and bacteriophages that remains to be characterized (Jones et al. 2014) and was observed in 11 copies in both experiments. Together, these data indicate that the mechanism that associates the annotation terms previously classified through manual curation as having roles in the viral life cycle and found to be associated with prophage density is indeed the presence of genes of viral origin, rather than by other potential mechanisms.

As for the set of 56 protein domains expected to contribute to a pathogenicity phenotype in *E. coli*, 42 of them (75%) are still significantly associated with prophage density after the removal of genes of viral origin (Supplementary Table 2, sheets domain2PfamCount" and "domain2PfamCountLessPhages"). Additionally, 27 and 39 of such domains have exactly the same number of occurrences or differ by one, respectively, when comparing the output of the *domain2Pfam* experiments with and without viral genes (Supplementary Table 2, sheet "virulelence_factors"). In contrast with the protein domains annotated as playing roles in the viral life cycle, the removal of viral genes did not alter the occurrence of the majority of virulence factors, as most of them are located outside predicted prophages.

We performed the same *in silico* procedure in *domain2GO* annotation schema to

remove genes of viral origin and evaluate GO terms that remain associated with prophage density, again finding that the vast majority of annotation terms describing viral lifestyle functions not to be significantly associated with prophage densities once viral genes are removed (Supplementary Table 2, sheet "domain2GOCountLessPhages"). Most GO terms describing pathogenicity mechanisms, on the other hand, are also observed in *E. coli* genomes after removing the genes of viral origin, suggesting most of the domains annotated to these GO terms are located outside the regions of integrated prophages.

Protein domains independently expanded in taller angiosperms

It is worth noting that the domain with the domain with the smallest value of occurrence in our dataset and that is also associated with maximum height was observed 1108 times (PF11721, with 16 copies in *A. thaliana* and 62 copies in *E. grandis*, Supplementary Table 4, sheet "associated_domains"). Therefore, these domains comprise relatively large expansions in both absolute and relative terms, and therefore are not likely to be artifacts caused by the common pitfalls found in genome assembly and annotation, a particularly challenging field in plant genomics (Salzberg 2019).

The Malectin domain (PF11721) was initially characterized in the model organism *Xenopus laevis*, where it monitors protein glycosylation in the endoplasmic reticulum (Schallus et al. 2008). However, the crystal structure of this domain in plants revealed the absence of critical amino acids for the interaction with diglucosidues that are present in the animal enzymes, suggesting distinct functional roles for proteins containing this domain in these lineages (Xiao et al. 2019). We found this domain to have 15 and 62 copies *A. thaliana* and *E. grandis*, respectively (4.13X increase) (Fig. 3C, Supplementary Table 4, sheet "associated_domains").

In *A. thaliana*, enzymes containing this domain have been functionally characterized as cell wall sensors that regulate development, reproduction and resistance to various stresses (V. Kumar, Donev et al. 2020). Interestingly, this domain has been previously reported to be greatly expanded in land plants when compared with other eukaryotes (H. Yang et al. 2021), and also expanded in the genome of *Populus trichocarpa*, a model organism for tree plant biology, when compared with *A. thaliana* (V. Kumar, Hainaut et al. 2019). Genes containing this domain are also upregulated in the developing wood tissue of *P. trichocarpa* and *E. grandis* (V. Kumar, Hainaut et al. 2019; Pinard et al.

2015), and the expansion of malectin-containing genes in *P. trichocarpa* appears to be a key player in the development of wood tissue in this species (V. Kumar, Hainaut et al. 2019).

We have not used *P. trichocarpa* in our analysis, as this species has been subject of a whole-duplication event and therefore failed to fulfill our genome quality metrics (Tuskan et al. 2006). Therefore, the data automatically produced by CALANGO independently strengthens the hypothesis that malectin-containing genes play a role in wood tissue development and are independently expanded in taller plants. This is yet another showcase of how CALANGO can be used to find potential causal relationships that can be surveyed through downstream experiments.

ankyrin repeats (PF12796) mediate protein-protein interactions in a wide range of biological processes, and is one of the most common and phylogenetically diverse domain in public sequence databases, being observed in viruses, prokaryotes and eukaryotes, with the later comprising the vast majority of entries (Mosavi et al. 2004). We found 126 and 530 copies of PF12796 in *A. thaliana* and *E. grandis*, respectively (4.12x increase, Fig. 3C). In *A. thaliana* proteins containing this domain play several roles in early embryogenesis and organ development (*e.g.* ANK6 - *ankyrin repeat protein 6*, KEG - *KEEP ON GOING*, EMB506 - *embryo defective 506*), even though a considerable fraction of these genes lacks functional characterization and are, therefore, interesting targets for future functional characterization.

Two domains (PF00560 and PF13855) comprise leucine-rich repeats, whose are frequently involved in protein-protein interaction processes. In the flowering plants, these domains are commonly found in Leucine-Rich Repeats Receptor-Like Kinases (LRR-RLKs), which is the case for the majority of *A. thaliana* genes (Supplementary Table 4, sheet "Arabidopsis_genes"). LRR-RLKs is one of the largest and most complex gene family in this species, playing roles in developmental pathways and immunity, perception of environmental conditions and stress response (Dufayard et al. 2017). We found 179 and 582 copies of domain PF00560 in the *A. thaliana* and *E. grandis* non-redundant proteomes, respectively (3.25x increase). As for PF13855, these proteomes have respectively 439 and 1295 copies of it (2.95x increase).

In *A. thaliana*, several of the genes containing these domains lack functional characterization, which is certainly true for an even greater fraction of *E. grandis* genes. The

characterized genes in the thale cress that code for these domains are highly diverse in their functional aspects. We found this domain to occur in regulators of floral development (BAM1 - *BARELY ANY MERISTEM 1*, AT1G11130 - *STRUBBELIG*), modulators of controlled cell death (SERK5 - *somatic embryogenesis receptor-like kinase 5*), components of hormone signaling pathways (AT3G13380 - *BRI1-LIKE 3*), and agents of plant resistance to pathogens (AT4G26090 - *RESISTANT TO P. SYRINGAE* 2). The expansion of LRR-RLKs suggests that both developmental processes and immunity pathways are expanded in taller plants. As taller plants also have longer generation times, we hypothesize that the expansion of components of the immune system in these species may a consequence of the selective pressure caused by chronic pathogenic infections, a problem likely to be far more critical for long-living species that may take years before achieving fertility.

Estimation of ancestor state for maximum height in Angiosperms

We estimated the ancestral height state for all the internal nodes of the phylogeny based on the height values of the 54 extant species with high-quality genomes available (see Material and Methods, section "Estimation of ancestor states for heigh in Angiosperms"). Due to the high degree of height variation observed, most internal nodes were given average values, and, based on visual inspection, we found no evolutionary trend on height variation across the distinct angiosperm lineages (Fig. S-5B). On the contrary, we notice the increase and decrease in plant height to be spread across different clades. These results indicate that many independent events of increase and decrease in plant height have occurred in the evolutionary history of flowering plants, suggesting that distinct evolutionary strategies under diverse selective pressures might explain the height variation in of angiosperms (*e.g.*, higher longevity in taller species vs annual plants in shorter species) (Lanfear et al. 2013).

2.11.2 CALANGO package and dependencies

The CALANGO package is designed as an open-source tool for comparative genomics. CALANGO was developed as a CRAN-compliant R package (35) and makes full use of the existing R ecosystem to implement its internal routines using code that has been validated by other researchers and developers. Our tool is implemented as an open-source R

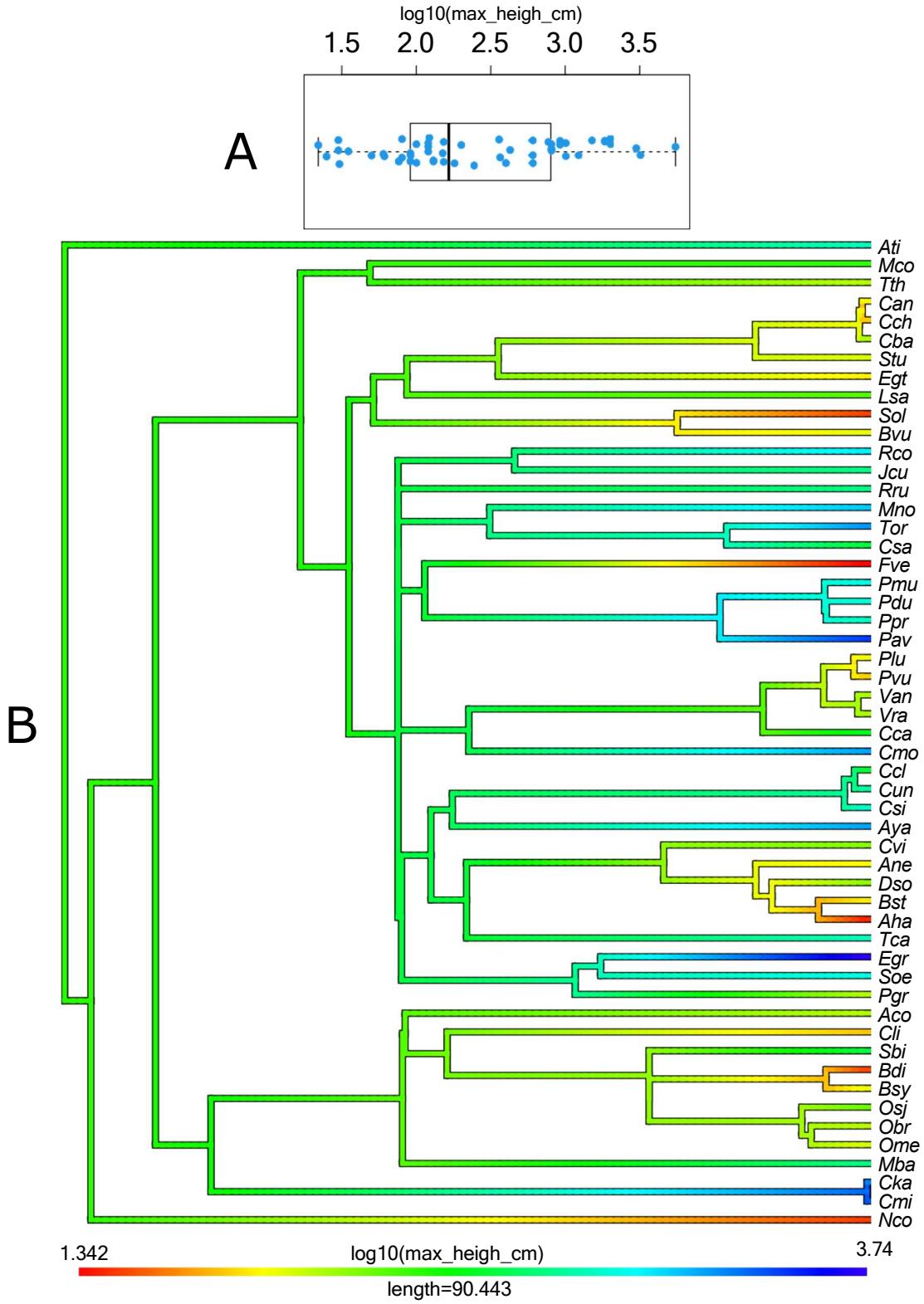


Figura S-5: Evolution of maximum height in Angiosperms. A) Phenotypic variation of maximum height in the 54 Angiosperm species used in this work. B) Ancestral state reconstruction for maximum height.

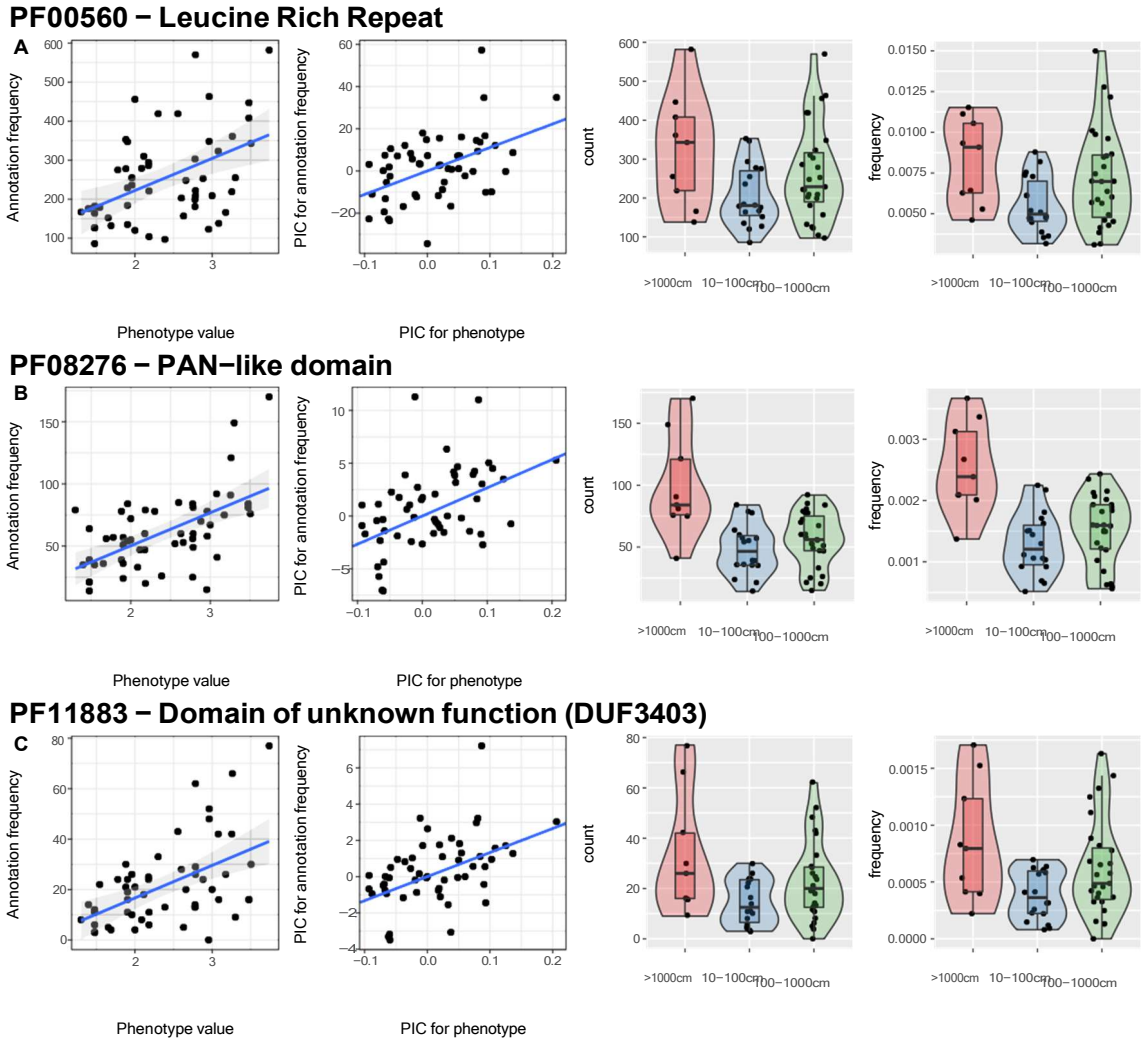


Figura S-6: Additional examples of Pfam terms associated with maximum height in Angiosperms. From left to right: linear models from raw count data, phylogeny-aware linear models; boxplot with raw counts; boxplot with normalized counts.

package distributed through the Comprehensive R Archive Network (CRAN). The package outputs results in the form of fully-functional dynamical HTML5 sites containing several summary statistics and other useful quantities, as well as interactive visual representations. CALANGO also outputs all results and reproducibility parameters as a list object, allowing easy integration with other bioinformatics pipelines (Fig. 1C).

CALANGO uses several R libraries to handle the different data types needed. The CALANGO analysis routines import functions from packages *ape* (Paradis e Schliep 2019) (to read nexus and newick phylogenetic trees and resolve multichotomies, and to calculate phylogeny-independent contrasts and the correlation structures arising from phylogenetic relationships); *taxize* (Chamberlain e Szöcs 2013) (to retrieve and process taxonomical hierarchies); *GO.db* (M 2019) and *AnnotationDbi* (Pagès et al. 2022) (to process GO annotation data); *KEGGREST* (Tenenbaum D 2022) (to process KEGG databases); and *nlme* (Pinheiro, Bates e R Core Team 2022) (to fit models using generalized least squares). CALANGO also imports functions from several packages to compose its visual output, namely: *dendextend* (Galili 2015), *rmarkdown* (Allaire et al. 2022) *heatmaply* (Galili et al. 2017), *heatmaply* (Sievert 2016), *plotly* (Wickham 2020), *DT* (Y. Xie, Cheng e Tan 2022), *htmltools* (Cheng et al. 2021) and *htmlwidgets* (Vaidyanathan et al. 2021). Other general-purpose packages used within CALANGO are pbmcapply (Kuang, Kong e Napolitano 2022) (for progress bars when using parallel processing); *assertthat* (Wickham 2019) (for input verification); *BiocManager* (Morgan 2022) (to retrieve and update dependencies from Bioconductor, namely *KEGGREST*, *GO.db* and *AnnotateDbi*); and *pkgdown* (Wickham, Hesselberth e Salmon 2022) (to automatically generate the project home page). Package updates on the CALANGO repository are automatically verified using Github Actions on the latest R versions for Windows and Mac OS, as well as for both the release and devel R versions on Ubuntu 20.04 LTS, to ensure code integrity. Future versions of CALANGO are planned to reduce the number of distinct dependencies so as to make the tool more resilient to changes in external package functionalities.

Tabela S-1: *Escherichia coli* genomic and phenotypic data. The full supplementary table 1 is available at https://bit.ly/CALANGO_tableS1.

Tabela S-2: Protein domains and Gene Ontology terms associated with prophage density in the *Escherichia coli* dataset. The full supplementary table 2 is available at https://bit.ly/CALANGO_tableS2.

Tabela S-3: Angiosperm genomic and phenotypic data. The full supplementary table 3 is available at https://bit.ly/CALANGO_tableS3.

Tabela S-4: *Arabidopsis thaliana* genes and protein domains positively associated with maximum height in Angiosperms. The full supplementary table 4 is available at https://bit.ly/CALANGO_tableS4.

2.11.3 Supplementary References

2.0, GenomeHubs (2022). *GoaT - Genomes on a Tree*. Last Accessed: August 19th 2022.

Adams, Dean C. e Michael L. Collyer (jul. de 2017). "Multivariate Phylogenetic Comparative: Evaluations, Comparisons, and Recommendations". Em: *Systematic Biology* 67.1, pp. 14–31. issn: 1063-5157. doi: 10.1093/sysbio/syx055.

Adl, Sina M. et al. (2012). "The Revised Classification of Eukaryotes". Em: *Journal of Eukaryotic Microbiology* 59.5, pp. 429–514. doi: https://doi.org/10.1111/j.1550-7408.2012.00644.x.

Allaire, JJ et al. (2022). *rmarkdown: Dynamic Documents for R*. R package version 2.14.

andILeitch, Michael Bennet (2005). "CHAPTER 2 - Genome Size Evolution in Plants". Em: *The Evolution of the Genome*. Ed. por T. Ryan Gregory. Burlington: Academic Press, pp. 89–162. isBn: 978-0-12-301463-4. doi: https://doi.org/10.1016/B978-012301463-4/50004-8.

Arndt, David et al. (mai. de 2016). "PHASTER: a better, faster version of the PHAST phage search tool". Em: *Nucleic Acids Research* 44.W1, W16–W21. issn: 0305-1048. doi: 10.1093/nar/gkw387.

Ball, Steven et al. (jan. de 2011). "The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis". Em: *Journal of Experimental Botany* 62.6, pp. 1775–1801. issn: 0022-0957. doi: 10.1093/jxb/erq411.

Bar-On, Yinon M., Rob Phillips e Ron Milo (2018). "The biomass distribution on Earth". Em: *Proceedings of the National Academy of Sciences* 115.25, pp. 6506–6511. doi: 10.1073/pnas.1711842115.

Barr, Jeremy J. et al. (2013). "Bacteriophage adhering to mucus provide a non-host-derived immunity". Em: *Proceedings of the National Academy of Sciences* 110.26, pp. 10771–10776. doi: 10.1073/pnas.1305923110.

Beaulieu, Jeremy M. et al. (2007). "Correlated evolution of genome size and seed mass". Em: *New Phytologist* 173.2, pp. 422–437. doi: https://doi.org/10.1111/j.1469-8137.2006.01919.x.

Bennett, Michael D. (1987). "VARIATION IN GENOMIC FORM IN PLANTS AND ITS ECOLOGICAL IMPLICATIONS". Em: *New Phytologist* 106.s1, pp. 177–200. doi: https://doi.org/10.1111/j.1469-8137.1987.tb04689.x.

Bentsink, Leónie et al. (2006). “Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis”. Em: *Proceedings of the National Academy of Sciences* 103.45, pp. 17042–17047. doi: 10.1073/pnas.0607877103.

Berardini, Tanya Z. et al. (2015). “The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome”. Em: *genesis* 53.8, pp. 474–485. doi: https://doi.org/10.1002/dvg.22877.

Bertozzi Silva, Juliano, Zachary Storms e Dominic Sauvageau (jan. de 2016). “Host receptors for bacteriophage adsorption”. Em: *FEMS Microbiology Letters* 363.4. fnw002. issn: 0378-1097. doi: 10.1093/femsle/fnw002.

Blázquez, Miguel A., David C. Nelson e Dolf Weijers (2020). “Evolution of Plant Hormone Response Pathways”. Em: *Annual Review of Plant Biology* 71.1. PMID: 32017604, pp. 327–353. doi: 10.1146/annurev-arplant-050718-100309.

Buchfink, Benjamin, Chao Xie e Daniel H. Huson (jan. de 2015). “Fast and sensitive protein alignment using DIAMOND”. Em: *Nature Methods* 12.1, pp. 59–60. issn: 1548-7105. doi: 10.1038/nmeth.3176.

Carrillo-Barral, Néstor, María del Carmen Rodríguez-Gacio e Angel Jesús Matilla (2020). “Delay of Germination-1 (DOG1): A Key to Understanding Seed Dormancy”. Em: *Plants* 9.4. issn: 2223-7747. doi: 10.3390/plants9040480.

Carta, Angelino et al. (2022). “Correlated evolution of seed mass and genome size varies among life forms in flowering plants”. Em: *Seed Science Research* 32.1, pp. 46–52. doi: 10.1017/S0960258522000071.

Cavalcanti, João Henrique F et al. (set. de 2018). “An L,L-diaminopimelate aminotransferase mutation leads to metabolic shifts and growth inhibition in Arabidopsis”. Em: *Journal of Experimental Botany* 69.22, pp. 5489–5506. issn: 0022-0957. doi: 10.1093/jxb/ery325.

Challis, Richard J. et al. (mai. de 2017). “GenomeHubs: simple containerized setup of a custom Ensembl database and web server for any species”. Em: *Database* 2017. bax039. issn: 1758-0463. doi: 10.1093/database/bax039.

Chamberlain, Scott A. e Eduard Szöcs (2013). “taxize: taxonomic search and retrieval in R”. Em: *F1000 Research* 2, p. 191. doi: 10.12688/f1000research.2-191.v1.

Chan, Patricia P. e Todd M. Lowe (dez. de 2015). “GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes”. Em: *Nucleic Acids Research* 44.D1, pp. D184–D189. issn: 0305-1048. doi: 10.1093/nar/gkv1309.

Chanderbali, Andre S et al. (mar. de 2016). “Evolving Ideas on the Origin and Evolution of Flowers: New Perspectives in the Genomic Era”. Em: *Genetics* 202.4, pp. 1255–1265. issn: 1943-2631. doi: 10.1534/genetics.115.182964.

Chaudhuri, Roy R. e Ian R. Henderson (2012). “The evolution of the Escherichia coli phylogeny”. Em: *Infection, Genetics and Evolution* 12.2, pp. 214–226. issn: 1567-1348. doi: https://doi.org/10.1016/j.meegid.2012.01.005.

Chen, Haixu et al. (nov. de 2021). “BRAD V3.0: an upgraded Brassicaceae database”. Em: *Nucleic Acids Research* 50.D1, pp. D1432–D1441. issn: 0305-1048. doi: 10.1093/nar/gkab1057.

Chen, Lu et al. (mar. de 2014). “Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity”. Em: *Molecular Biology and Evolution* 31.6, pp. 1402–1413. issn: 0737-4038. doi: 10.1093/molbev/msu083.

Cheng, Joe et al. (2021). *htmltools: Tools for HTML*. R package version 0.5.2.

Cirillo, D M et al. (1996). “Identification of a domain in Rck, a product of the Salmonella typhimurium virulence plasmid, required for both serum resistance and cell invasion”. Em: *Infection and Immunity* 64.6, pp. 2019–2023. doi: 10.1128/iai.64.6.2019-2023.1996.

Coghlan, Avril et al. (jan. de 2019). “Comparative genomics of the major parasitic worms”. Em: *Nature Genetics* 51.1, pp. 163–174. issn: 1546-1718. doi: 10.1038/s41588-018-0262-1.

Cornwell, Will e Shinichi Nakagawa (2017). “Phylogenetic comparative methods”. Em: *Current Biology* 27.9, R333–R336. issn: 0960-9822. doi: https://doi.org/10.1016/j.cub.2017.03.049.

Correa, Adrienne M. S. et al. (ago. de 2021). “Revisiting the rules of life for viruses of microorganisms”. Em: *Nature Reviews Microbiology* 19.8, pp. 501–513. issn: 1740-1534. doi: 10.1038/s41579-021-00530-x.

Cotter, Paul D., R. Paul Ross e Colin Hill (fev. de 2013). “Bacteriocins — a viable alternative to antibiotics?” Em: *Nature Reviews Microbiology* 11.2, pp. 95–105. issn: 1740-1534. doi: 10.1038/nrmicro2937.

Crooks, Gavin E. et al. (2004). “WebLogo: A Sequence Logo Generator”. Em: *Genome Research* 14.6, pp. 1188–1190. doi: 10.1101/gr.849004.

Cunningham, Fiona et al. (nov. de 2021). “Ensembl 2022”. Em: *Nucleic Acids Research* 50.D1, pp. D988–D995. issn: 0305-1048. doi: 10.1093/nar/gkab1049.

Dedrick, Rebekah M. et al. (jan. de 2017). “Prophage-mediated defence against viral attack and viral counter-defence”. Em: *Nature Microbiology* 2.3, p. 16251. issn: 2058-5276. doi: 10.1038/nmicrobiol.2016.251.

Dekkers, Bas J.W. et al. (2016). “The Arabidopsis DELAY OF GERMINATION 1 gene affects ABSCISIC ACID INSENSITIVE 5 (ABI5) expression and genetically interacts with ABI3 during Arabidopsis seed development”. Em: *The Plant Journal* 85.4, pp. 451–465. doi: https://doi.org/10.1111/tpj.13118.

Dobritsa, Anna A. e Daniel Coerper (nov. de 2012). “The Novel Plant Protein INAPERTURATE POLLEN1 Marks Distinct Cellular Domains and Controls Formation of Apertures in the Arabidopsis Pollen Exine ”. Em: *The Plant Cell* 24.11, pp. 4452–4464. issn: 1040-4651. doi: 10.1105/tpc.112.101220.

Dong, Qunfeng, Shannon D. Schlueter e Volker Brendel (jan. de 2004). “PlantGDB, plant genome database and analysis tools”. Em: *Nucleic Acids Research* 32.suppl_1, pp. D354–D359. issn: 0305-1048. doi: 10.1093/nar/gkh046.

Dubois, Emeline et al. (abr. de 2011). “Homologous Recombination Is Stimulated by a Decrease in dUTPase in Arabidopsis”. Em: *PLOS ONE* 6.4, pp. 1–8. doi: 10.1371/journal.pone.0018658.

Dufayard, Jean-François et al. (2017). “New Insights on Leucine-Rich Repeats Receptor-Like Kinase Orthologous Relationships in Angiosperms”. Em: *Frontiers in Plant Science* 8, p. 381. doi: 10.3389/fpls.2017.00381.

Dunn, Casey W. e Catriona Munro (2016). “Comparative genomics and the diversity of life”. Em: *Zoologica Scripta* 45.S1, pp. 5–13. doi: https://doi.org/10.1111/zsc.12211.

Durand, Eléonore et al. (2020). “Evolution of self-incompatibility in the Brassicaceae: Lessons from a textbook example of natural selection”. Em: *Evolutionary Applications* 13.6, pp. 1279–1297. doi: https://doi.org/10.1111/eva.12933.

Eddy, Sean R. (out. de 2011). “Accelerated Profile HMM Searches”. Em: *PLOS Computational Biology* 7.10, pp. 1–16. doi: 10.1371/journal.pcbi.1002195.

Edgar, Robert C. (mar. de 2004). “MUSCLE: multiple sequence alignment with high accuracy and high throughput”. Em: *Nucleic Acids Research* 32.5, pp. 1792–1797. issn: 0305-1048. doi: 10.1093/nar/gkh340.

Ehrbar, Kristin e Wolf-Dietrich Hardt (2005). “Bacteriophage-encoded type III effectors in Salmonella enterica subspecies 1 serovar Typhimurium”. Em: *Infection, Genetics and Evolution* 5.1, pp. 1–9. issn: 1567-1348. doi: https://doi.org/10.1016/j.meegid.2004.07.004.

Ekstrom, Alexander et al. (ago. de 2014). “PlantCAZyme: a database for plant carbohydrate-active enzymes”. Em: *Database* 2014. bau079. issn: 1758-0463. doi: 10.1093/database/bau079.

Falster, Daniel S. e Mark Westoby (2003). “Plant height and evolutionary games”. Em: *Trends in Ecology & Evolution* 18.7, pp. 337–343. issn: 0169-5347. doi: https://doi.org/10.1016/S0169-5347(03)00061-2.

Fedak, Halina et al. (2016). “Control of seed dormancy in Arabidopsis by a cis-acting noncoding antisense transcript”. Em: *Proceedings of the National Academy of Sciences* 113.48, E7846–E7855. doi: 10.1073/pnas.1608827113.

Felsenstein, Joseph (1985). “Phylogenies and the Comparative Method”. Em: *The American Naturalist* 125.1, pp. 1–15. issn: 00030147, 15375323.

Fernández, Lucía, Ana Rodríguez e Pilar García (mai. de 2018). “Phage or foe: an insight into the impact of viral predation on microbial communities”. Em: *The ISME Journal* 12.5, pp. 1171–1179. issn: 1751-7370. doi: 10.1038/s41396-018-0049-5.

Fischer, Steve et al. (2011). “Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups”. Em: *Current Protocols in Bioinformatics* 35.1, pp. 6.12.1–6.12.19. doi: https://doi.org/10.1002/0471250953.bi0612s35.

Fleischmann, Andreas et al. (out. de 2014). “Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new es-

timate of the minimum genome size in angiosperms". Em: *Annals of Botany* 114.8, pp. 1651–1663. issn: 0305-7364. doi: 10.1093/aob/mcu189.

Fu, Limin et al. (out. de 2012). "CD-HIT: accelerated for clustering the next-generation sequencing data". Em: *Bioinformatics* 28.23, pp. 3150–3152. issn: 1367-4803. doi: 10.1093/bioinformatics/bts565.

Galili, Tal (2015). "dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering". Em: *Bioinformatics*. doi: 10.1093/bioinformatics/btv428.

Galili, Tal et al. (2017). "heatmaply: an R package for creating interactive cluster heatmaps for online publishing". Em: *Bioinformatics*. doi: 10.1093/bioinformatics/btx657.

El-Gebali, Sara et al. (out. de 2018). "The Pfam protein families database in 2019". Em: *Nucleic Acids Research* 47.D1, pp. D427–D432. issn: 0305-1048. doi: 10.1093/nar/gky995.

Goffeau, A. et al. (1996). "Life with 6000 Genes". Em: *Science* 274.5287, pp. 546–567. doi: 10.1126/science.274.5287.546.

González-Morales, Sandra Isabel et al. (2016). "Regulatory network analysis reveals novel regulators of seed desiccation tolerance in Arabidopsis thaliana". Em: *Proceedings of the National Academy of Sciences* 113.35, E5232–E5241. doi: 10 .1073 / pnas . 1610985113.

Goodstein, David M. et al. (nov. de 2011). "Phytozome: a comparative platform for green plant genomics". Em: *Nucleic Acids Research* 40.D1, pp. D1178–D1186. issn: 0305-1048. doi: 10.1093/nar/gkr944.

Gordillo Altamirano, Fernando et al. (fev. de 2021). "Bacteriophage-resistant Acinetobacter baumannii are resensitized to antimicrobials". Em: *Nature Microbiology* 6.2, pp. 157–161. issn: 2058-5276. doi: 10.1038/s41564-020-00830-7.

Granzotto, Adriana e Guilherme Marcello Queiroga Cruz (2015). "Regulação de Elementos de Transposição: Mecanismos Epigenéticos de Silenciamento, Autorregulação e Ativação por Estresse". Em: *Elementos de transposição: diversidade, evolução, aplicações e impacto nos genomas dos seres vivos*. Ed. por Claudia Marcia Aparecida Carareto, Claudia Barros Monteiro-Vitorello e Marie-Anne Van Sluys. São José do Rio Preto: Editora FIOCRUZ, pp. 91–113. isBn: 978-85-7541-462-0. doi: https://doi.org/10.7476/9788575415672.

Greilhuber, Johann e I J. Leitch (2013). “Genome Size and the Phenotype”. Em: *Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes*. Ed. por Johann Greilhuber, Jaroslav Dolezel e Jonathan F. Wendel. Vienna: Springer Vienna, pp. 323–344. ISBN: 978-3-7091-1160-4. doi: 10.1007/978-3-7091-1160-4_20.

Groth, Philip et al. (set. de 2006). “PhenomicDB: a new cross-species genotype/phenotype resource”. Em: *Nucleic Acids Research* 35.suppl_1, pp. D696–D699. issn: 0305-1048. doi: 10.1093/nar/gkl662.

Harvey, Paul H, Mark D Pagel et al. (1991). *The comparative method in evolutionary biology*. Vol. 239. Oxford university press Oxford.

Haynes, Winston A., Aurelie Tomczak e Purvesh Khatri (jan. de 2018). “Gene annotation bias impedes biomedical research”. Em: *Scientific Reports* 8.1, p. 1362. issn: 2045-2322. doi: 10.1038/s41598-018-19333-x.

Heather, James M. e Benjamin Chain (2016). “The sequence of sequencers: The history of sequencing DNA”. Em: *Genomics* 107.1, pp. 1–8. issn: 0888-7543. doi: https://doi.org/10.1016/j.ygeno.2015.11.003.

Hidalgo, Oriane et al. (2017). “Is There an Upper Limit to Genome Size?” Em: *Trends in Plant Science* 22.7, pp. 567–573. issn: 1360-1385. doi: https://doi.org/10.1016/j.tplants.2017.04.005.

Hongo, Jorge Augusto et al. (2021). “CALANGO: an annotation-based, phylogeny-aware comparative genomics framework for exploring and interpreting complex genotypes and phenotypes”. Em: *bioRxiv*. doi: 10.1101/2021.08.25.457574.

Hung, Jui-Hung et al. (set. de 2011). “Gene set enrichment analysis: performance evaluation and usage guidelines”. Em: *Briefings in Bioinformatics* 13.3, pp. 281–291. issn: 1467-5463. doi: 10.1093/bib/bbr049.

Huo, Heqiang, Shouhui Wei e Kent J. Bradford (2016). “DELAY OF GERMINATION1 (DOG1) regulates both seed dormancy and flowering time through microRNA pathways”. Em: *Proceedings of the National Academy of Sciences* 113.15, E2199–E2206. doi: 10.1073/pnas.1600558113.

IHGSC et al. (fev. de 2001). “Initial sequencing and analysis of the human genome”. Em: *Nature* 409.6822, pp. 860–921. issn: 1476-4687. doi: 10.1038/35057062.

Initiative, The Arabidopsis Genome (dez. de 2000). “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana”. Em: *Nature* 408.6814, pp. 796–815. issn: 1476-4687. doi: 10.1038/35048692.

Jones, Philip et al. (jan. de 2014). “InterProScan 5: genome-scale protein function classification”. Em: *Bioinformatics* 30.9, pp. 1236–1240. issn: 1367-4803. doi: 10.1093/bioinformatics/btu031.

Kang, Ming et al. (2014). “Adaptive and nonadaptive genome size evolution in Karst endemic flora of China”. Em: *New Phytologist* 202.4, pp. 1371–1381. doi: https://doi.org/10.1111/nph.12726.

Kattge, Jens et al. (2020). “TRY plant trait database – enhanced coverage and open access”. Em: *Global Change Biology* 26.1, pp. 119–188. doi: https://doi.org/10.1111/gcb.14904.

Kawahara, Yoshihiro et al. (fev. de 2013). “Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data”. Em: *Rice* 6.1, p. 4. issn: 1939-8433. doi: 10.1186/1939-8433-6-4.

Kawashima, Tomokazu et al. (jul. de 2015). “Diversification of histone H2A variants during plant evolution”. Em: *Trends in Plant Science* 20.7, pp. 419–425. issn: 1360-1385. doi: 10.1016/j.tplants.2015.04.005.

Knight, Charles A., Nicole A. Molinari e Dmitri A. Petrov (jan. de 2005). “The Large Genome Constraint Hypothesis: Evolution, Ecology and Phenotype”. Em: *Annals of Botany* 95.1, pp. 177–190. issn: 0305-7364. doi: 10.1093/aob/mci011.

Koornneef, Maarten, Leónie Bentsink e Henk Hilhorst (2002). “Seed dormancy and germination”. Em: *Current Opinion in Plant Biology* 5.1, pp. 33–36. issn: 1369-5266. doi: https://doi.org/10.1016/S1369-5266(01)00219-9.

Kopriva, Stanislav e Andreas P M Weber (jan. de 2021). “Genetic encoding of complex traits”. Em: *Journal of Experimental Botany* 72.1, pp. 1–3. issn: 0022-0957. doi: 10.1093/jxb/eraa498.

Krishnakumar, Vivek et al. (nov. de 2014). “Araport: the Arabidopsis Information Portal”. Em: *Nucleic Acids Research* 43.D1, pp. D1003–D1009. issn: 0305-1048. doi: 10.1093/nar/gku1200.

Kuang, Kevin, Quyu Kong e Francesco Napolitano (2022). *pbmcapply: Tracking the Progress of Mc*pply with Progress Bar*. R package version 1.5.1.

Kumar, Sudhir, Glen Stecher et al. (mai. de 2018). “MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms”. Em: *Molecular Biology and Evolution* 35.6, pp. 1547–1549. issn: 0737-4038. doi: 10.1093/molbev/msy096.

Kumar, Sudhir, Michael Suleski et al. (ago. de 2022). “TimeTree 5: An Expanded Resource for Species Divergence Times”. Em: *Molecular Biology and Evolution* 39.8. msac174. issn: 1537-1719. doi: 10.1093/molbev/msac174.

Kumar, Vikash, Evgeniy N. Donev et al. (2020). “Genome-Wide Identification of Populus Malectin/Malectin-Like Domain-Containing Proteins and Expression Analyses Reveal Novel Candidates for Signaling and Regulation of Wood Development”. Em: *Frontiers in Plant Science* 11, p. 588846. doi: 10.3389/fpls.2020.588846.

Kumar, Vikash, Matthieu Hainaut et al. (2019). “Poplar carbohydrate-active enzymes: whole-genome annotation and functional analyses based on RNA expression data”. Em: *The Plant Journal* 99.4, pp. 589–609. doi: https://doi.org/10.1111/tpj.14417.

Lanfear, Robert et al. (mai. de 2013). “Taller plants have lower rates of molecular evolution”. Em: *Nature Communications* 4.1, p. 1879. issn: 2041-1723. doi: 10.1038/ncomms2836.

Lee, Byung Ha et al. (jul. de 2021). “A species-specific functional module controls formation of pollen apertures”. Em: *Nature Plants* 7.7, pp. 966–978. issn: 2055-0278. doi: 10.1038/s41477-021-00951-9.

Lee, Heewook et al. (2012). “Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing”. Em: *Proceedings of the National Academy of Sciences* 109.41, E2774–E2783. doi: 10.1073/pnas.1210309109.

Lei, Bingkun e Frédéric Berger (2020). “H2A Variants in Arabidopsis: Versatile Regulators of Genome Activity”. Em: *Plant Communications* 1.1, p. 100015. issn: 2590-3462. doi: https://doi.org/10.1016/j.xplc.2019.100015.

Leitch, A. R. e I. J. Leitch (2012). “Ecological and genetic factors linked to contrasting genome dynamics in seed plants”. Em: *New Phytologist* 194.3, pp. 629–646. doi: https://doi.org/10.1111/j.1469-8137.2012.04105.x.

Leitch, I. J., Mark W. Chase e Michael D. Bennett (dez. de 1998). “Phylogenetic Analysis of DNA C-values Provides Evidence for a Small Ancestral Genome Size in Flowering

Plants". Em: *Annals of Botany* 82.suppl_1, pp. 85–94. issn: 0305-7364. doi: 10.1006/anbo.1998.0783.

Leitch, I. J., D. E. Soltis et al. (jan. de 2005). "Evolution of DNA Amounts Across Land Plants (Embryophyta)". Em: *Annals of Botany* 95.1, pp. 207–217. issn: 0305-7364. doi: 10.1093/aob/mci014.

León, M e R Bastías (2015). "Virulence reduction in bacteriophage resistant bacteria." Em: *Frontiers in Microbiology* 343.6. doi: http://dx.doi.org/10.3389/fmicb.2015.00343.

Li, Fay-Wei et al. (jul. de 2018). "Fern genomes elucidate land plant evolution and cyanobacterial symbioses". Em: *Nature Plants* 4.7, pp. 460–472. issn: 2055-0278. doi: 10.1038/s41477-018-0188-8.

Li, Linzhou et al. (set. de 2020). "The genome of Prasinoderma coloniale unveils the existence of a third phylum within green plants". Em: *Nature Ecology &, Evolution* 4.9, pp. 1220–1231. issn: 2397-334X. doi: 10.1038/s41559-020-1221-7.

Li, Weizhong e Adam Godzik (mai. de 2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences". Em: *Bioinformatics* 22.13, pp. 1658–1659. issn: 1367-4803. doi: 10.1093/bioinformatics/btl158.

Liolios, Konstantinos et al. (nov. de 2009). "The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata". Em: *Nucleic Acids Research* 38.suppl_1, pp. D346–D354. issn: 0305-1048. doi: 10.1093/nar/gkp848.

Lisch, Damon (jan. de 2013). "How important are transposons for plant evolution?" Em: *Nature Reviews Genetics* 14.1, pp. 49–61. issn: 1471-0064. doi: 10.1038/nrg3374.

Liu, Jian-Zhong e Steven A. Whitham (2013). "Overexpression of a soybean nuclear localized type–III DnaJ domain-containing HSP40 reveals its roles in cell death and disease resistance". Em: *The Plant Journal* 74.1, pp. 110–121. doi: https:// doi.org/10.1111/tpj.12108.

M, Carlson (2019). *GO.db: A set of annotation maps describing the entire Gene Ontology*. R package version 3.8.2.

Ma, Xuelian et al. (set. de 2022). "PlantGSAD: a comprehensive gene set annotation database for plant species". Em: *Nucleic Acids Research* 50.D1, pp. D1456–D1467. issn: 0305-1048. doi: 10.1093/nar/gkab794.

ANEXO C - EVIDENCE ON THE ORIGIN OF DELAY OF GERMINATION1 GENE FAMILY IN AN ANCESTRAL OF LAND PLANTS WITHIN CHAROPHYTA

Mukherjee, Supratim et al. (out. de 2016). “Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements”. Em: *Nucleic Acids Research* 45.D1, pp. D446–D456. issn: 0305-1048. doi: 10.1093/nar/gkw992.

Nagy, László G et al. (jan. de 2020). “Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing”. Em: *Nucleic Acids Research* 48.5, pp. 2209–2219. issn: 0305-1048. doi: 10.1093/nar/gkz1241.

Nakabayashi, Kazumi et al. (jul. de 2012). “The Time Required for Dormancy Release in Arabidopsis Is Determined by DELAY OF GERMINATION1 Protein Levels in Freshly Harvested Seeds”. Em: *The Plant Cell* 24.7, pp. 2826–2838. issn: 1040-4651. doi: 10.1105/tpc.112.100214.

Nasrallah, June B. e Mikhail E. Nasrallah (mar. de 2014). “S-locus receptor kinase signalling”. Em: *Biochemical Society Transactions* 42.2, pp. 313–319. issn: 0300-5127. doi: 10.1042/BST20130222.

Niklas, Karl J. e Ulrich Kutschera (2010). “The evolution of the land plant life cycle”. Em: *New Phytologist* 185.1, pp. 27–41. doi: https://doi.org/10.1111/j.1469-8137.2009.03054.x.

Nishimura, Noriyuki et al. (jun. de 2018). “Control of seed dormancy and germination by DOG1-AHG1 PP2C phosphatase complex via binding to heme”. Em: *Nature Communications* 9.1, p. 2132. issn: 2041-1723. doi: 10.1038/s41467-018-04437-9.

Nishiyama, Eri et al. (2021). “Ancient and recent gene duplications as evolutionary drivers of the seed maturation regulators DELAY OF GERMINATION1 family genes”. Em: *New Phytologist* 230.3, pp. 889–901. doi: https://doi.org/10.1111/nph.17201.

Nishiyama, Takashi et al. (jan. de 2013). “The structure of the deacetylase domain of Escherichia coli PgaB, an enzyme required for biofilm formation: a circularly permuted member of the carbohydrate esterase 4 family”. Em: *Acta Crystallographica Section D* 69.1, pp. 44–51. doi: 10.1107/S0907444912042059.

O’Leary, Nuala A. et al. (nov. de 2015). “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. Em: *Nucleic Acids Research* 44.D1, pp. D733–D745. issn: 0305-1048. doi: 10.1093/nar/gkv1189.

Pagès, H et al. (2022). *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. R package version 1.58.0.

Pang, Shuai et al. (mai. de 2015). “GPU MrBayes V3.1: MrBayes on Graphics Processing Units for Protein Sequence Data”. Em: *Molecular Biology and Evolution* 32.9, pp. 2496–2497. issn: 0737-4038. doi: 10.1093/molbev/msv129.

Paradis, E. e K. Schliep (2019). “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R”. Em: *Bioinformatics* 35, pp. 526–528.

Park, Beom Seok e Jie-Oh Lee (dez. de 2013). “Recognition of lipopolysaccharide pattern by TLR4 complexes”. Em: *Experimental & Molecular Medicine* 45.12, e66–e66. issn: 2092-6413. doi: 10.1038/emm.2013.97.

Pasha, Asher et al. (jul. de 2020). “Araport Lives: An Updated Framework for Arabidopsis Bioinformatics”. Em: *The Plant Cell* 32.9, pp. 2683–2686. issn: 1040-4651. doi: 10.1105/tpc.20.00358.

Pawluk, April et al. (2014). “A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa”. Em: *mBio* 5.2, e00896–14. doi: 10.1128/mBio.00896-14.

Peiffer, Jason A et al. (abr. de 2014). “The Genetic Architecture Of Maize Height”. Em: *Genetics* 196.4, pp. 1337–1356. issn: 1943-2631. doi: 10.1534/genetics.113.159152.

Pellicer, Jaume, Michae F. Fay e I. J. Leitch (set. de 2010). “The largest eukaryotic genome of them all?” Em: *Botanical Journal of the Linnean Society* 164.1, pp. 10–15. issn: 0024-4074. doi: 10.1111/j.1095-8339.2010.01072.x.

Pellicer, Jaume, Oriane Hidalgo et al. (2018). “Genome Size Diversity and Its Impact on the Evolution of Land Plants”. Em: *Genes* 9.2. issn: 2073-4425. doi: 10.3390/genes9020088.

Pellicer, Jaume e I J. Leitch (2020). “The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies”. Em: *New Phytologist* 226.2, pp. 301–305. doi: https://doi.org/10.1111/nph.16261.

Petrov, Dmitri A. (jan. de 2001). “Evolution of genome size: new approaches to an old problem”. Em: *Trends in Genetics* 17.1, pp. 23–28. issn: 0168-9525. doi: 10.1016/S0168-9525(00)02157-0.

— (2002). “Mutational Equilibrium Model of Genome Size Evolution”. Em: *Theoretical Population Biology* 61.4, pp. 531–544. issn: 0040-5809. doi: https://doi.org/10.1006/tpbi.2002.1605.

Pinard, Desre et al. (mai. de 2015). “Comparative analysis of plant carbohydrate active enZymes and their role in xylogenesis”. Em: *BMC Genomics* 16.1, p. 402. issn: 1471-2164. doi: 10.1186/s12864-015-1571-8.

Pinheiro, José, Douglas Bates e R Core Team (2022). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-157.

Plazzi, Federico, Ronald R. Ferrucci e Marco Passamonti (abr. de 2010). “Phylogenetic representativeness: a new method for evaluating taxon sampling in evolutionary studies”. Em: *BMC Bioinformatics* 11.1, p. 209. issn: 1471-2105. doi: 10.1186/1471-2105-11-209.

Proost, Sebastian et al. (out. de 2014). “PLAZA 3.0: an access point for plant comparative genomics”. Em: *Nucleic Acids Research* 43.D1, pp. D974–D981. issn: 0305-1048. doi: 10.1093/nar/gku986.

Pulido, Pablo e Dario Leister (2018). “Novel DNAJ-related proteins in Arabidopsis thaliana”. Em: *The New Phytologist* 217.2, pp. 480–490. issn: 0028646X, 14698137.

Pulkkinen, W S e S I Miller (1991). “A Salmonella typhimurium virulence protein is similar to a Yersinia enterocolitica invasion protein and a bacteriophage lambda outer membrane protein”. Em: *Journal of Bacteriology* 173.1, pp. 86–93. doi: 10.1128/jb.173.1.86-93.1991.

Puttick, Mark N., James Clark e Philip C. J. Donoghue (2015). “Size is not everything: rates of genome size evolution, not C-value, correlate with speciation in angiosperms”. Em: *Proceedings of the Royal Society B: Biological Sciences* 282.1820, p. 20152289. doi: 10.1098/rspb.2015.2289.

Rambaut, Andrew et al. (abr. de 2018). “Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7”. Em: *Systematic Biology* 67.5, pp. 901–904. issn: 1063-5157. doi: 10.1093/sysbio/syy032.

Ramisetty, Bhaskar Chandra Mohan e Pavithra Anantharaman Sudhakari (2019). “Bacterial ‘Grounded’ Prophages: Hotspots for Genetic Renovation and Innovation”. Em: *Frontiers in Genetics* 10. issn: 1664-8021. doi: 10.3389/fgene.2019.00065.

Ren, Ren et al. (2018). “Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms”. Em: *Molecular Plant* 11.3. Genome Biology, pp. 414–428. issn: 1674-2052. doi: https://doi.org/10.1016/j.molp.2018.01.002.

Revell, Liam J. (2012). “phytools: an R package for phylogenetic comparative biology (and other things)”. Em: *Methods in Ecology and Evolution* 3.2, pp. 217–223. doi: https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Roff, Derek A. (1997). *Evolutionary Quantitative Genetics*. New York: Springer New York. isBn: 978-1-4615-4080-9. doi: https://doi.org/10.1007/978-1-4615-4080-9.

Sall, Khadidiatou et al. (2019). “DELAY OF GERMINATION 1-LIKE 4 acts as an inducer of seed reserve accumulation”. Em: *The Plant Journal* 100.1, pp. 7–19. doi: https://doi.org/10.1111/tpj.14485.

Salzberg, Steven L. (mai. de 2019). “Next-generation genome annotation: we still struggle to get it right”. Em: *Genome Biology* 20.1, p. 92. issn: 1474-760X. doi: 10.1186/s13059-019-1715-2.

Sandoval, Francisco J., Yi Zhang e Sanja Roje (nov. de 2008). “Flavin Nucleotide Metabolism in Plants: MONOFUNCTIONAL ENZYMES SYNTHESIZE FAD IN PLASTIDS *”. Em: *Journal of Biological Chemistry* 283.45, pp. 30890–30900. issn: 0021-9258. doi: 10.1074/jbc.M803416200.

Sanger, F. e A.R. Coulson (1975). “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. Em: *Journal of Molecular Biology* 94.3, pp. 441–448. issn: 0022-2836. doi: https://doi.org/10.1016/0022-2836(75)90213-2.

Sayers, Eric W et al. (out. de 2019). “GenBank”. Em: *Nucleic Acids Research* 48.D1, pp. D84–D86. issn: 0305-1048. doi: 10.1093/nar/gkz956.

Schäffer, Alejandro A. et al. (jul. de 2001). “Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements”. Em: *Nucleic Acids Research* 29.14, pp. 2994–3005. issn: 0305-1048. doi: 10.1093/nar/29.14.2994.

Schallus, Thomas et al. (2008). “Malectin: A Novel Carbohydrate-binding Protein of the Endoplasmic Reticulum and a Candidate Player in the Early Steps of Protein N-Glycosylation”. Em: *Molecular Biology of the Cell* 19.8. PMID: 18524852, pp. 3404–3414. doi: 10.1091/mbc.e08-04-0354.

Schneider, Rene e Staffan Persson (2015). “Another brick in the wall”. Em: *Science* 350.6257, pp. 156–157. doi: 10.1126/science.aad3200.

Schuster, Stephan C. (jan. de 2008). “Next-generation sequencing transforms today’s biology”. Em: *Nature Methods* 5.1, pp. 16–18. issn: 1548-7105. doi: 10.1038/nmeth1156.

SHAPIRO, S. S. e M. B. WILK (dez. de 1965). “An analysis of variance test for normality (complete samples)”. Em: *Biometrika* 52.3-4, pp. 591–611. doi: 10.1093/biomet/52.3-4.591.

Sievert, Carson (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. isBn: 978-3-319-24277-4.

Silveira, Cynthia B. e Forest L. Rohwer (jul. de 2016). “Piggyback-the-Winner in host-associated microbial communities”. Em: *npj Biofilms and Microbiomes* 2.1, p. 16010. issn: 2055-5008. doi: 10.1038/npjbiofilms.2016.10.

Simmons, Emilia L. et al. (2020). “Biofilm Structure Promotes Coexistence of Phage-Resistant and Phage-Susceptible Bacteria”. Em: *mSystems* 5.3, e00877–19. doi: 10.1128/mSystems.00877-19.

Sørensen, Iben et al. (2011). “The charophycean green algae provide insights into the early origins of plant cell walls”. Em: *The Plant Journal* 68.2, pp. 201–211. doi: https://doi.org/10.1111/j.1365-313X.2011.04686.x.

Steyert, Susan R. e James B. Kaper (2012). “Contribution of Urease to Colonization by Shiga Toxin-Producing Escherichia coli”. Em: *Infection and Immunity* 80.8, pp. 2589–2600. doi: 10.1128/IAI.00210-12.

Subburaj, Saminathan et al. (jun. de 2016). “Phylogenetic Analysis, Lineage-Specific Expansion and Functional Divergence of seed dormancy 4-Like Genes in Plants”. Em: *PLOS ONE* 11.6, pp. 1–24. doi: 10.1371/journal.pone.0153717.

Tello-Ruiz, Marcela K et al. (nov. de 2020). “Gramene 2021: harnessing the power of comparative genomics and pathways for plant research”. Em: *Nucleic Acids Research* 49.D1, pp. D1452–D1463. issn: 0305-1048. doi: 10.1093/nar/gkaa979.

Tenenbaum D, Maintainer B (2022). *KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)*. R package version 1.36.3.

Tomaž, Špela, Kristina Gruden e Anna Coll (2022). “TGA transcription factors—Structural characteristics as basis for functional variability”. Em: *Frontiers in Plant Science* 13. issn: 1664-462X. doi: 10.3389/fpls.2022.935819.

Tong, Chao et al. (jan. de 2020). “Comparative Genomics Identifies Putative Signatures of Sociality in Spiders”. Em: *Genome Biology and Evolution* 12.3, pp. 122–133. issn: 1759-6653. doi: 10.1093/gbe/evaa007.

Touchon, Marie et al. (jan. de 2009). “Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths”. Em: *PLOS Genetics* 5.1, pp. 1–25. doi: 10.1371/journal.pgen.1000344.

Tuskan, G. A. et al. (2006). “The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray)”. Em: *Science* 313.5793, pp. 1596–1604. doi: 10.1126/science.1128691.

Ung, Huoi, Wolfgang Moeder e Keiko Yoshioka (set. de 2014). “Arabidopsis Triphosphate Tunnel Metalloenzyme2 Is a Negative Regulator of the Salicylic Acid-Mediated Feedback Amplification Loop for Defense Responses ”. Em: *Plant Physiology* 166.2, pp. 1009–1021. issn: 0032-0889. doi: 10.1104/pp.114.248757.

Vaidya, Gaurav, David J. Lohman e Rudolf Meier (2011). “SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information”. Em: *Cladistics* 27.2, pp. 171–180. doi: https://doi.org/10.1111/j.1096-0031.2010.00329.x.

Vaidyanathan, Ramnath et al. (2021). *htmlwidgets: HTML Widgets for R*. R package version 1.5.4.

Vandecraen, Joachim et al. (2017). “The impact of insertion sequences on bacterial genome plasticity and adaptability”. Em: *Critical Reviews in Microbiology* 43.6. PMID: 28407717, pp. 709–730. doi: 10.1080/1040841X.2017.1303661.

Veselý, Pavel, Petr Bureš e Petr Šmarda (ago. de 2013). “Nutrient reserves may allow for genome size increase: evidence from comparison of geophytes and their sister non-geophytic relatives”. Em: *Annals of Botany* 112.6, pp. 1193–1200. issn: 0305-7364. doi: 10.1093/aob/mct185.

Vinogradov, Alexander E (2003). “Selfish DNA is maladaptive: evidence from the plant Red List”. Em: *Trends in Genetics* 19.11, pp. 609–614. issn: 0168-9525. doi: https://doi.org/10.1016/j.tig.2003.09.010.

Vitti, Joseph J., Sharon R. Grossman e Pardis C. Sabeti (2013). “Detecting Natural Selection in Genomic Data”. Em: *Annual Review of Genetics* 47.1. PMID: 24274750, pp. 97–120. doi: 10.1146/annurev-genet-111212-133526.

Vogel, Christine e Cyrus Chothia (mai. de 2006). “Protein Family Expansions and Biological Complexity”. Em: *PLOS Computational Biology* 2.5, pp. 1–13. doi: 10.1371/journal.pcbi.0020048.

Wang, B et al. (2019). “[The China National GeneBank owned by all, completed by all and shared by all]”. Em: *Yi Chuan* 20.41, pp. 761–772. doi: 10.16288/j.yczz..

Wang, Dandan et al. (2021). “Which factors contribute most to genome size variation within angiosperms?” Em: *Ecology and Evolution* 11.6, pp. 2660–2668. doi: https://doi.org/10.1002/ece3.7222.

Wang, Xiaoxue et al. (dez. de 2010). “Cryptic prophages help bacteria cope with adverse environments”. Em: *Nature Communications* 1.1, p. 147. issn: 2041-1723. doi: 10.1038/ncomms1146.

Waterhouse, Robert M et al. (dez. de 2017). “BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics”. Em: *Molecular Biology and Evolution* 35.3, pp. 543–548. issn: 0737-4038. doi: 10.1093/molbev/msx319.

Wendel, Jonathan F. et al. (mai. de 2002). “Feast and famine in plant genomes”. Em: *Genetica* 115.1, pp. 37–47. issn: 1573-6857. doi: 10.1023/A:1016020030189.

Wickham, Hadley (2019). *assertthat: Easy Pre and Post Assertions*. R package version 0.2.1.

— (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman & Hall. ISBN: 9781138331457.

Wickham, Hadley, Jay Hesselberth e Maëlle Salmon (2022). *pkgdown: Make Static HTML Documentation for a Package*. R package version 2.0.3.

Willi, Yvone e Ary A. Hoffman (2009). “Demographic factors and genetic variation influence population persistence under environmental change”. Em: *Journal of Evolutionary Biology* 22.1, pp. 124–133. doi: https://doi.org/10.1111/j.1420-9101.2008.01631.x.

Wolf, Andrea J. e David M. Underhill (abr. de 2018). “Peptidoglycan recognition by the innate immune system”. Em: *Nature Reviews Immunology* 18.4, pp. 243–254. issn: 1474-1741. doi: 10.1038/nri.2017.136.

Wolf, Jason B. (2002). “The geometry of phenotypic evolution in developmental hyperspace”. Em: *Proceedings of the National Academy of Sciences* 99.25, pp. 15849–15851. doi: 10.1073/pnas.012686699.

Xiao, Yu et al. (mai. de 2019). “Mechanisms of RALF peptide perception by a heterotypic receptor complex”. Em: *Nature* 572.7768, pp. 270–274. issn: 1476-4687. doi: 10.1038/s41586-019-1409-7.

Xie, Yihui, Joe Cheng e Xianying Tan (2022). *DT: A Wrapper of the JavaScript Library 'DataTables'*. R package version 0.23.

Xue, Han et al. (out. de 2021). “qPTMplants: an integrative database of quantitative post-translational modifications in plants”. Em: *Nucleic Acids Research* 50.D1, pp. D1491–D1499. issn: 0305-1048. doi: 10.1093/nar/gkab945.

Yang, He et al. (2021). “Malectin/Malectin-like domain-containing proteins: A repertoire of cell surface molecules with broad functional potential”. Em: *The Cell Surface* 7, p. 100056. issn: 2468-2330. doi: https://doi.org/10.1016/j.tcsw.2021.100056.

Yang, Xiaohan et al. (set. de 2019). “Comparative genomics can provide new insights into the evolutionary mechanisms and gene function in CAM plants”. Em: *Journal of Experimental Botany* 70.22, pp. 6539–6547. issn: 0022-0957. doi: 10.1093/jxb/erz408.

Yelagandula, Ramesh et al. (jul. de 2014). “The Histone Variant H2A.W Defines Heterochromatin and Promotes Chromatin Condensation in Arabidopsis”. Em: *Cell* 158.1, pp. 98–109. issn: 0092-8674. doi: 10.1016/j.cell.2014.06.006.

Zhang, Jian et al. (fev. de 2020). “The hornwort genome and early land plant evolution”. Em: *Nature Plants* 6.2, pp. 107–118. issn: 2055-0278. doi: 10.1038/s41477-019-0588-4.

Zu, Pengjuan e Florian P. Schiestl (2017). “The effects of becoming taller: direct and pleiotropic effects of artificial selection on plant height in Brassica rapa”. Em: *The Plant Journal* 89.5, pp. 1009–1019. doi: https://doi.org/10.1111/tpj.13440.

Zwickl, Derrick J. e David M. Hillis (jul. de 2002). “Increased Taxon Sampling Greatly Reduces Phylogenetic Error”. Em: *Systematic Biology* 51.4, pp. 588–598. issn: 1063-5157. doi: 10.1080/10635150290102339.

Capítulo 3

Evidence on the origin of *Delay of Germination1* gene family in an ancestral of land plants within Charophyta

Evidence on the origin of *Delay of Germination1* gene family in an ancestral of land plants within Charophyta

Alison Pelri Albuquerque Menezes[1] João Victor dos Anjos Almeida[2]
Luiz-Eduardo Del-Bem[3] Francisco Pereira Lobo[1,*]

[1] *Departamento de Genética, Ecologia e Evolução, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.*

[2] *Universidade Estadual Paulista Júlio de Mesquita Filho, Jaboticabal, São Paulo*

[3] *Departamento de Botânica, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.*

**To whom correspondence should be addressed. Tel: +55 31 34093072; Fax: +55 31 34092567; Email: franciscolobo@ufmg.br, franciscolobo@gmail.com*

The Delay of Germination1 (DOG1) protein domain is a key component in the regulation of several important physiological processes, such as germination and flowering time. This domain is found in the DOG1 gene family (DGF) and TGACG motif-binding transcription factor proteins (TGA). DGF is a small and diverse gene family, related mainly to seed germination and flowering - however, it has been found in seedless tracheophytes and bryophytes. TGA was reported in land plants and in the charophyte *Klebsormidium nitens*. To investigate the evolution of the DOG1 protein domain family we searched for the DOG1 protein domain in the predicted proteomes of green algae species. Using methods based on the Hidden Markov Model we detected novel DGF proteins in charophytes and novel TGA proteins in charophyte and chlorophyte species, revealing an older origin and divergence between DGFs and TGAs than previously thought. Although we have searched for the DOG1 protein domain in four Chlorophyta classes, we only found it in Trebouxiophyceae, suggesting that this could have been an event of horizontal gene transfer.

KEYWORDS: DOG1, phylogenetics, evolution, delay of germination, seed dor-

mancy.

3.1 INTRODUCTION

The evolutionary success of seed plants depends on the correct timing of seed germination. Dormancy is essential for seeds to survive adverse environmental circumstances until conditions are appropriate for germination (Koornneef, Bentsink e Hilhorst 2002). Thus, seed dormancy is one of the most important mechanisms in the life cycle of a seed plant. Consequently, the mechanisms involved in seed dormancy and germination are tightly regulated and under strong selection (Fedak et al. 2016). The DELAY OF GERMINATION1 (DOG1) family genes (DGFs) are master transcriptional regulators of seed dormancy (Bentsink et al. 2006; Nakabayashi et al. 2012; Dekkers et al. 2016; Nishimura et al. 2018). The whole process through which DGFs regulate seed germination is not yet fully characterized, with previous studies indicating that it involves a complex network of abscisic acid (ABA) signaling, miRNA-regulated pathways, and antisense non-coding transcripts (asDOG1) (Fedak et al. 2016; Huo, Wei e Bradford 2016; Carrillo-Barral, Rodríguez-Gacio e Matilla 2020). It has been demonstrated they are also involved in the regulation of seed reserve accumulation and maturation, tolerance to desiccation, and flowering (Dekkers et al. 2016; González-Morales et al. 2016; Huo, Wei e Bradford 2016; Sall et al. 2019).

All DGF proteins share a conserved DOG1 domain, which is also found in the TGACG motif-binding transcription factor proteins (TGAs), one of the most well-characterized transcription factors of plants and a key component of biotic and abiotic stress responses, developmental processes, and circadian rhythm (Tomaž, Gruden e Coll 2022). A major difference between canonical DGF and TGA proteins is the presence of a bZIP (IPR004827) domain that can be found only in the N-terminal region of TGAs (Sall et al. 2019). Previous works on the DOG1 domain evolution could not define whether the bZIP domain was lost in DGFs or gained in TGAs (Sall et al. 2019). They suggest that DGFs diverged from the TGAs, and then gave rise to four different clusters within DGF: DOG1 (encompassing DOG1 and DOG-Like 1-3), DOGL4, DOGL5, and DOGL6 (Sall et al. 2019; E. Nishiyama et al. 2021). The diversity of DGF lineages and functions indicate that the gene family diversified in the most recent common ancestor (MRCA) of angi-

osperms as a consequence of gene duplication events (Carrillo-Barral, Rodríguez-Gacio e Matilla 2020; E. Nishiyama et al. 2021).

Although the function of DGFs has only been studied in angiosperms so far, they have been found in all major clades of land plants and in the charophyte *K. nitens* (Subburaj et al. 2016; E. Nishiyama et al. 2021). More recently, a third set comprising two paralog genes found in angiosperms that also code for the DOG1 domain (INNAPERTURATE POLLEN1 and 2 - INP1/INP2) has been characterized as a molecular network needed for pollen development during fertilization (B. H. Lee et al. 2021). The phylogenetic origin of INPs when compared with DGFs and TGAs, however, remains to be determined.

In this study, we surveyed 171 high-quality, non-redundant Archaeplastida proteomes, including species from key lineages, for the understanding of the origin and diversification of the DOG1 domain. For that, we used a search strategy based on *de novo* proteome annotation using either *InterProScan* or a Hidden-Markov Model (HMM) built from DOG1 sequences from early-branching Archaeplastida species. We report the presence of DGF-like proteins in charophytes, indicating a more ancient origin for this gene family than previously reported (E. Nishiyama et al. 2021), predating the evolution of land plants. We also observed two TGAs in Trebouxiophyceae, a group of Chlorophyta, in a scenario compatible with horizontal gene transfer (HGT) from seed plants happening twice, independently.

3.2 MATERIAL AND METHODS

3.2.1 Data Collection and Functional Annotation

We gathered 171 high-quality non-redundant predicted proteomes from 5 genomic databases representing all major extant lineages of Archaeplastida lineages (Supplementary Table S1; Methods S1). We searched for sequences containing DOG1 domains using *InterProScan* 5 (Jones et al. 2014) to perform a *de novo* annotation of all predicted proteomes. Then, we extracted information on the sequences containing the InterPro accession "IPR025422", which represents the DOG1 domain through the signatures PF14144 - Seed dormancy control, from the Pfam database and PS51806 - DOG1 domain profile, From the ProSiteProfiles database. At this point, we also collected information about the presence of putative bZIP domains located in the N-terminal region and selected the longest

bZIP domain signature as the canonical representation of this domain (see Supplementary Table S2 for all bZIP signatures considered in this study). The putative DOG1 domains found in charophytes and chlorophytes at this point were used as queries in BLAST searches to survey sequence databases for the presence of this domain in other Archaeplastida species not included in our initial dataset.

3.2.2 Detecting DOG1 domain sequences in Archaeplastida

We used 12 sequences annotated with IPR025422 from 8 green algae species (6 charophytes - *Chara braunii*, *Coleochaete irregularis*, *Cylindrocystis cushleckae*, *K. nitens*, *Mesotaenium caldariorum*, and *Staurodesmus omearii* - and 2 chlorophytes - *Coccomyxa subellipsoidea* C-169 and *Coccomyxa* sp.) to construct a Hidden Markov Model (HMM) using *HMMER* software (Eddy 2011) and eventually recover additional DOG1 domain sequences not found by the InterPro search.

3.2.3 Phylogenetic reconstruction

We extracted the DOG1 domain sequences using the coordinates as predicted by either InterProScan or HMMER and used this data to perform a phylogenetic analysis of the DOG1 domain with two different datasets: 1) all the *Arabidopsis thaliana* proteins; 2) all *A. thaliana*, plus all the green algae proteins *Chara braunii*, *C. subellipsoidea* C-169, *Coleochaete irregularis*, *Cylindrocystis cushleckae*, *K. nitens*, *Mesotaenium caldariorum*, *Staurodesmus omearii*, and *Trebouxia* sp. A1-2; 3) sequences from green algae species and the following embryophytes: *A. thaliana*, *Amborella trichopoda*, *Brachypodium distachyon*, *Marchantia polymorpha*, *Oryza sativa*, *Selaginella moellendorffii*, *Theobroma cacao*, *Asparagus officinalis*, *Ceratodon purpureus*, *Pinus sylvestris*, *Nymphaea colorata*, *Nymphaea thermarum*, *Anthoceros angustus*, *Salvinia cucullata*, *Thuja plicata*, and *Welwitschia mirabilis*.

We aligned the domain sequences using the *Muscle* v3.8.1551 (Edgar 2004) on default parameters. A maximum likelihood phylogeny was inferred in *IQ-TREE* v2.0.3 (Minh et al. 2020) using ultrafast bootstrap set to run 1000 bootstraps and the remaining parameters as default. The software automatically detects the best-fitting evolutionary model for the data according to BIC. For the tree including only *A. thaliana* sequences and the one with *A. thaliana* and green algae sequences, the best fitting model was the LG+G4

model. For the tree including green algae and embryophyte species, the chosen model was the JTT+F+R6 model.

3.2.4 Analysis of conservation patterns

We created a sequence logo representation of the DOG1 domain sequences from the 12 genes used to build our HMM using WebLogo with default parameters (Crooks et al. 2004). We also created a representation of the domain architecture of all DOG1-containing sequences from *A. thaliana* and all the green algae sequences. Then, we combined this domain architecture with a phylogeny of the same sequences.

3.2.5 Protein classification and nomenclature

We classified the proteins we surveyed in DGFs, TGAs, and INPs based on domain architecture, phylogenetic data, and literature. Proteins containing the DOG1 and bZIP domains were classified as TGAs, while DGF proteins have the DOG1 domain, but lack the bZIP domain. Based on the phylogenetic distribution of these proteins and the previous classification using the *A. thaliana* proteins as a model (E. Nishiyama et al. 2021), the DGF superfamily was further divided into 6 families: DOG1, DOGL1-5, and DOGL proteins. Although AtINP1 and AtINP2 present a domain architecture of a DGF they were previously classified as INP (Dobritsa e Coerper 2012) and we kept the nomenclature, however in our data analysis we treated them as members of the DGF superfamily.

3.3 RESULTS

Our search for proteins containing the DOG1 domain found 2010 occurrences of this conserved region in plant species across all major clades of Viridiplantae, but no homologous proteins were found in Rhodophyta or Glaucophyta (Supplementary Table S3). We used the presence of a bZIP at the N-terminal region as the domain architecture proxy to distinguish between TGAs and DGFs/INPs. As we demonstrate below, and in contrast to previous literature, our results indicate that DGFs most likely emerged in the last common ancestor of Charophyta and Embrophyta. We also found evidence of the presence of TGA genes in Trebouxiophyceae (Chlorophyta).

The search using InterproScan found 11 different proteins containing the DOG1 domain (6 DGFs and 7 TGAs) in 6 species of charophytes, belonging to 4 different classes: Charophyceae (*Chara braunii*), Coleochaetophycaee (*Coleochaete irregularis*), Klebsormidsiophyceae (*K. nitens*), and Zygnemophyceae (*Cylindrocystis cushleckae*, *Mesotaenium caldarium*, and *Staurodesmus omearii*). *InterProScan* 5 also found a TGA sequence in the Trebouxiophyceae Coccomyxa subellipsoidea. Our search for similar sequences in the NCBI database using *PSI-BLAST* (Schäffer et al. 2001) returned a protein (BDA44423.1) of another Coccomyxa species containing the DOG1 domain.

We used all the 12 green algae protein domains found by *InterProScan* 5 and our BLAST search to create an HMM and look for ancient homologs of DOG1-containing proteins in the same Viridiplantae species. We observed a high degree of conservation between the DOG1 domains in green algae species (Fig. S-1). The vast majority of proteins with the DOG1 domain were detected by both InterProScan and HMMER annotation (1982 proteins, 98,6%), therefore demonstrating a high agreement between both search strategies. Our *HMMER* (Eddy 2011) search returned the same proteins for Charophyta and Chlorophyta species and one extra TGA protein for a Trebouxiophyceae *Trebouxia* sp.

We reconstructed a phylogeny of the 20 *A. thaliana* DOG1 domain homologs, including AtINP1, AtINP2, and 10 AtTGA protein domains, to further characterize their evolutionary history. The DOG1 domains from *A. thaliana* DGF genes clustered according to their designated families, recovering the phylogenetic relationships previously reported (E. Nishiyama et al. 2021). The TGA protein domains also clustered together, forming a monophyletic sister group of DGF protein domains. The AtINP1 and AtINP2 DOG1 domains were not included in the previous phylogenetic analysis; in our tree, the AtINPs form a polytomy low support with the branches of TGA and DGF in our tree the AtINPs form a polytomy with low support (Fig. 1A, Fig. S-2).

We further evaluated the phylogenetic relationships of DOG1 domains by constructing a tree including all 11 green algae sequences together with the ones found in 16 embryophytes, including the species used by E. Nishiyama et al. (2021) and others representing major groups of Archaeplastida (Fig. 1B). Again, we observed a clear separation of the DOG1 domains from the DGF and TGA gene families, with a few domains from DGF-like proteins found in the TGA cluster. The AtINP1 and AtINP2 protein domains

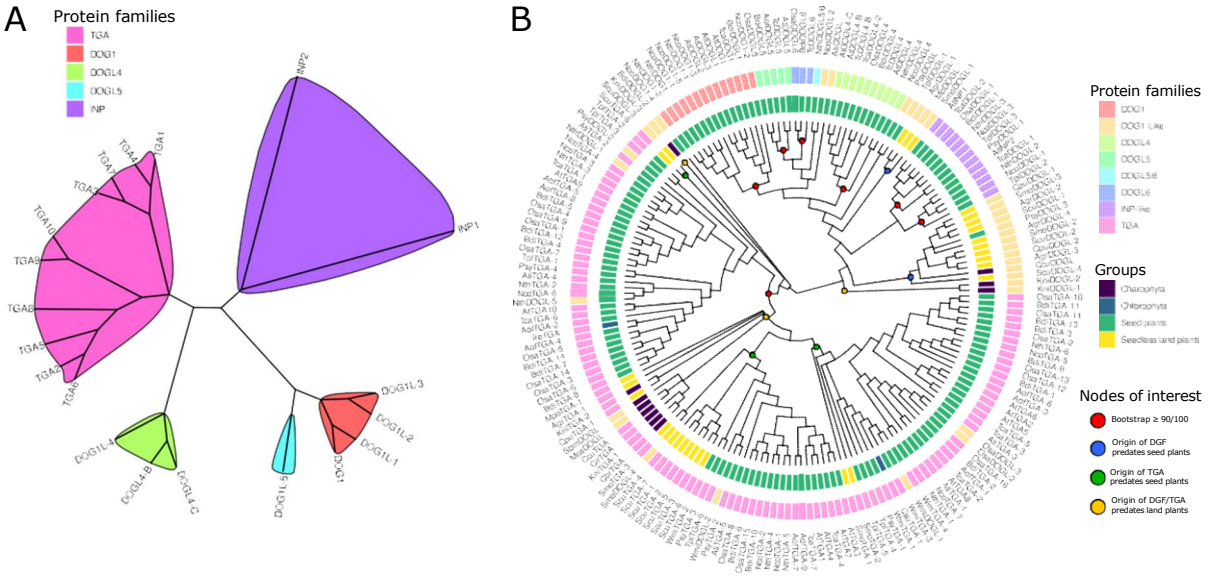


Figura 1: Phylogenetic reconstruction of DOG1 domains. A) phylogeny of DOG1 protein domains from all protein sequences recovered from *Arabidopsis thaliana*; B) DOG1 phylogeny reconstructed from from 189 DOG1 protein domains from 33 species across all major clades of Viridiplantae.

are grouped with other seed plant species in a monophyletic group separated into two different clades corresponding to the two A. thaliana genes. From now on, we refer to these genes as the INP-Like cluster.

The subgroups of the DGF gene family previously reported by E. Nishiyama et al. (2021), together with their clustering patterns, were also found in our analysis. We found the DGF group formed by DOG1, DOGL4, DOGL5, DOGL5/6, and DOGL6, as well as the larger group of DGFs including the INP-like to be sister groups of sequence clusters of seedless land plants, indicating an origin of these groups predating the origin of Spermatophyta, also in accordance with E. Nishiyama et al. (2021) (Fig. 1B, blue and green circles for DGFs and TGAs, respectively).

We found three clusters of DOG1 domains from charophytes that strongly suggest the occurrence of the DGF and TGA gene families in species from this group (Fig 1B., orange circles). The first cluster comprises two DOG1 sequences from *K. nitens* that are a sister group of all other DGFs from land plants. These DGF-like sequences all lack the bZIP domain that is characteristic of the TGA gene family. The second cluster is an early branch from a polytomy containing a subgroup of DOG1 domains classified as TGAs. Even though not fully resolved, this topology, together with domain architecture data,

is compatible with the presence of TGAs in Charophytes. The third group is a single sequence from *K. nitens*, and is again a polytomy between a DOG1 domain from the charophyte *K. nitens* and the remaining TGAs from the land plants. In this phylogeny, the two protein domains from Trebouxiophyceae species cluster separately with other seed plant protein domains.

3.4 DISCUSSION

In this work, we employed a *de novo* annotation of 171 high-quality Archaeplastida proteomes to investigate the phylogenetic origin of the DOG1 domain, a key component of several transcriptional programs of plants found in the DGF and TGA gene families. Previous analysis of the evolution of the DOG1 domain reported the expansion of the DGF in the angiosperms (E. Nishiyama et al. 2021). They also described a complex pattern of gene gains and losses in individual angiosperm species and provided a robust classification of TGA and DGFs families taking into account both phylogenetic and domain architecture information. These authors also described the presence of TGA genes in the Charophyte *K. nitens*. More recently, two *A. thaliana* proteins with the DOG1 domain (INP1 and INP2) have been functionally characterized as major players in pollen development (B. H. Lee et al. 2021). Even though they lack bZIP domains, their phylogenetic relationship with other DGFs and TGAs has never been addressed. Our *de novo* annotation strategy managed to recover all 20 known genes from *A. thaliana* annotated as either DGFs, TGAs, or INPs. For comparison, the work by E. Nishiyama et al. (2021), using BLAST-based searches, recovered 15 DOG1-containing proteins in *A. thaliana*, and missed the two INP sequences.

Although AtINP1 and AtINP2 proteins have been previously named after their mutant phenotype, according to their domain architecture they are DGFs, as both AtINP proteins contain a DOG1 domain and lack the N-terminal extension containing a bZIP domain typical of TGA proteins (Dobritsa e Coerper 2012; E. Nishiyama et al. 2021). Our phylogenetic analyses have grouped AtINP1 and AtINP2 in a clade with protein domains from other seed plants, suggesting this clade as a DGF lineage. Previous evolutionary and functional studies also suggested homology between AtINP2 and TcaDOGL-2unusual (XP_007023253) and BdiDOGL-1unsual (PNT69181), which were also found as INPs in

our study (B. H. Lee et al. 2021; Mazuecos-Aguilera et al. 2021). Therefore, we provide further evidence that INPs are likely to be an embryophyte-specific group of DGFs with a possible neofunctionalization.

DGF is widely distributed in land plants and especially diverse in angiosperms (E. Nishiyama et al. 2021). The work by E. Nishiyama et al. 2021) found a single TGA gene in the charophyte *K. nitens*, even though these authors suggest that the DGF gene family may also predate the origin of land plants. We found additional evidence of the presence of TGA genes in the charophytes *C. braunii*, *C. irregularis*, *S. omearii*, *C. cushleckae*, and *M. caldariorum*. More importantly, the phylogenetic location of the three clusters of proteins from charophytes harboring the DOG1 domain, together with their domain architecture, provides compelling evidence that the DGF also predates the emergence of land plants.

After searching for DOG1-containing proteins in 4 different classes of chlorophytes (namely Chlorophyceae, Chloropicophyceae, Mamiellophyceae, and Trebouxiophyceae) our analysis found evidence of TGA genes only in Trebouxiophyceae species. Based on our search and the separation of these two sequences in the phylogeny, we hypothesize that this protein was horizontally transferred to this clade. Our results indicate that this HGT might have happened at least twice independently in Trebouxiophyceae. Clearly, further studies including new genomic data from these groups are needed to determine the most parsimonious scenario to explain the presence of TGA genes in Trebouxiophyceae.

Our findings indicate that DGFs diverged from TGAs before than previously thought, in the charophytes. Also, we reported TGA sequences in chlorophyte species, most likely transferred horizontally from another species. Further genetic investigations are needed to screen the presence of DOG1 homologs in other Trebouxiophyceae species and determine the origin and age of this transfer. DGF proteins have been increasingly studied in flowering plants. Along the course of evolution, the DGF proteins were likely co-opted for the regulation of flowering and germination, as their origin is older than the origin of seeds and flowers in Viridiplantae. Although these proteins must have been involved in different processes originally, identifying the presence of DGF proteins in the green algae allows us to further investigate what their original functions could be and what networks they were part of.

3.5 ACKNOWLEDGEMENT

We would like to thank Dr. Anderson Vieira Chaves and Dr. Tetsu Sakamoto for the insightful discussions during the elaboration of this research.

3.6 FUNDING

We are grateful to the Graduate Program in Genetics, the Graduate Program in Bioinformatics, the "Centro de Processamento de Alto Desempenho/ICB"(CEPAD/SAGARANA cluster), and the Vice Dean for Research from Universidade Federal de Minas Gerais, Brazil, for the financial and computational support for this research. This work was partially funded by CAPES/Brazil (Grant 001).

3.7 CONFLICT OF INTEREST

All authors declare no conflict of interest for this publication.

3.8 REFERENCES

2.0, GenomeHubs (2022). *GoaT - Genomes on a Tree*. Last Accessed: August 19th 2022.

Adams, Dean C. e Michael L. Collyer (jul. de 2017). “Multivariate Phylogenetic Comparative: Evaluations, Comparisons, and Recommendations”. Em: *Systematic Biology* 67.1, pp. 14–31. issn: 1063-5157. doi: 10.1093/sysbio/syx055.

Adl, Sina M. et al. (2012). “The Revised Classification of Eukaryotes”. Em: *Journal of Eukaryotic Microbiology* 59.5, pp. 429–514. doi: https://doi.org/10.1111/j.1550-7408.2012.00644.x.

Allaire, JJ et al. (2022). *rmarkdown: Dynamic Documents for R*. R package version 2.14.

andILeitch, Michael Bennet (2005). “CHAPTER 2 - Genome Size Evolution in Plants”. Em: *The Evolution of the Genome*. Ed. por T. Ryan Gregory. Burlington: Academic Press, pp. 89–162. isBn: 978-0-12-301463-4. doi: https://doi.org/10.1016/B978-012301463-4/50004-8.

Arndt, David et al. (mai. de 2016). “PHASTER: a better, faster version of the PHAST phage search tool”. Em: *Nucleic Acids Research* 44.W1, W16–W21. issn: 0305-1048. doi: 10.1093/nar/gkw387.

Ball, Steven et al. (jan. de 2011). “The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis”. Em: *Journal of Experimental Botany* 62.6, pp. 1775–1801. issn: 0022-0957. doi: 10.1093/jxb/erq411.

Bar-On, Yinon M., Rob Phillips e Ron Milo (2018). “The biomass distribution on Earth”. Em: *Proceedings of the National Academy of Sciences* 115.25, pp. 6506–6511. doi: 10.1073/pnas.1711842115.

Barr, Jeremy J. et al. (2013). “Bacteriophage adhering to mucus provide a non-host-derived immunity”. Em: *Proceedings of the National Academy of Sciences* 110.26, pp. 10771–10776. doi: 10.1073/pnas.1305923110.

Beaulieu, Jeremy M. et al. (2007). “Correlated evolution of genome size and seed mass”. Em: *New Phytologist* 173.2, pp. 422–437. doi: https://doi.org/10.1111/j.1469-8137.2006.01919.x.

Bennett, Michael D. (1987). “VARIATION IN GENOMIC FORM IN PLANTS AND ITS ECOLOGICAL IMPLICATIONS”. Em: *New Phytologist* 106.s1, pp. 177–200. doi: https://doi.org/10.1111/j.1469-8137.1987.tb04689.x.

Bentsink, Leónie et al. (2006). “Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis”. Em: *Proceedings of the National Academy of Sciences* 103.45, pp. 17042–17047. doi: 10.1073/pnas.0607877103.

Berardini, Tanya Z. et al. (2015). “The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome”. Em: *genesis* 53.8, pp. 474–485. doi: https://doi.org/10.1002/dvg.22877.

Bertozzi Silva, Juliano, Zachary Storms e Dominic Sauvageau (jan. de 2016). “Host receptors for bacteriophage adsorption”. Em: *FEMS Microbiology Letters* 363.4. fnw002. issn: 0378-1097. doi: 10.1093/femsle/fnw002.

Blázquez, Miguel A., David C. Nelson e Dolf Weijers (2020). “Evolution of Plant Hormone Response Pathways”. Em: *Annual Review of Plant Biology* 71.1. PMID: 32017604, pp. 327–353. doi: 10.1146/annurev-arplant-050718-100309.

Buchfink, Benjamin, Chao Xie e Daniel H. Huson (jan. de 2015). “Fast and sensitive protein alignment using DIAMOND”. Em: *Nature Methods* 12.1, pp. 59–60. issn: 1548-7105. doi: 10.1038/nmeth.3176.

Carrillo-Barral, Néstor, María del Carmen Rodríguez-Gacio e Angel Jesús Matilla (2020). “Delay of Germination-1 (DOG1): A Key to Understanding Seed Dormancy”. Em: *Plants* 9.4. issn: 2223-7747. doi: 10.3390/plants9040480.

Carta, Angelino et al. (2022). “Correlated evolution of seed mass and genome size varies among life forms in flowering plants”. Em: *Seed Science Research* 32.1, pp. 46–52. doi: 10.1017/S0960258522000071.

Cavalcanti, João Henrique F et al. (set. de 2018). “An L,L-diaminopimelate aminotransferase mutation leads to metabolic shifts and growth inhibition in Arabidopsis”. Em: *Journal of Experimental Botany* 69.22, pp. 5489–5506. issn: 0022-0957. doi: 10.1093/jxb/ery325.

Challis, Richard J. et al. (mai. de 2017). “GenomeHubs: simple containerized setup of a custom Ensembl database and web server for any species”. Em: *Database* 2017. bax039. issn: 1758-0463. doi: 10.1093/database/bax039.

Chamberlain, Scott A. e Eduard Szöcs (2013). “taxize: taxonomic search and retrieval in R”. Em: *F1000 Research* 2, p. 191. doi: 10.12688/f1000research.2-191.v1.

Chan, Patricia P. e Todd M. Lowe (dez. de 2015). “GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes”. Em: *Nucleic Acids Research* 44.D1, pp. D184–D189. issn: 0305-1048. doi: 10.1093/nar/gkv1309.

Chanderbali, Andre S et al. (mar. de 2016). “Evolving Ideas on the Origin and Evolution of Flowers: New Perspectives in the Genomic Era”. Em: *Genetics* 202.4, pp. 1255–1265. issn: 1943-2631. doi: 10.1534/genetics.115.182964.

Chaudhuri, Roy R. e Ian R. Henderson (2012). “The evolution of the Escherichia coli phylogeny”. Em: *Infection, Genetics and Evolution* 12.2, pp. 214–226. issn: 1567-1348. doi: https://doi.org/10.1016/j.meegid.2012.01.005.

Chen, Haixu et al. (nov. de 2021). “BRAD V3.0: an upgraded Brassicaceae database”. Em: *Nucleic Acids Research* 50.D1, pp. D1432–D1441. issn: 0305-1048. doi: 10.1093/nar/gkab1057.

Chen, Lu et al. (mar. de 2014). “Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity”. Em:

Molecular Biology and Evolution 31.6, pp. 1402–1413. issn: 0737-4038. doi: 10.1093/molbev/msu083.

Cheng, Joe et al. (2021). *htmltools: Tools for HTML*. R package version 0.5.2.

Cirillo, D M et al. (1996). "Identification of a domain in Rck, a product of the Salmonella typhimurium virulence plasmid, required for both serum resistance and cell invasion". Em: *Infection and Immunity* 64.6, pp. 2019–2023. doi: 10.1128/iai.64.6.2019-2023.1996.

Coghlan, Avril et al. (jan. de 2019). "Comparative genomics of the major parasitic worms". Em: *Nature Genetics* 51.1, pp. 163–174. issn: 1546-1718. doi: 10.1038/s41588-018-0262-1.

Cornwell, Will e Shinichi Nakagawa (2017). "Phylogenetic comparative methods". Em: *Current Biology* 27.9, R333–R336. issn: 0960-9822. doi: https://doi.org/10.1016/j.cub.2017.03.049.

Correa, Adrienne M. S. et al. (ago. de 2021). "Revisiting the rules of life for viruses of microorganisms". Em: *Nature Reviews Microbiology* 19.8, pp. 501–513. issn: 1740-1534. doi: 10.1038/s41579-021-00530-x.

Cotter, Paul D., R. Paul Ross e Colin Hill (fev. de 2013). "Bacteriocins — a viable alternative to antibiotics?" Em: *Nature Reviews Microbiology* 11.2, pp. 95–105. issn: 1740-1534. doi: 10.1038/nrmicro2937.

Crooks, Gavin E. et al. (2004). "WebLogo: A Sequence Logo Generator". Em: *Genome Research* 14.6, pp. 1188–1190. doi: 10.1101/gr.849004.

Cunningham, Fiona et al. (nov. de 2021). "Ensembl 2022". Em: *Nucleic Acids Research* 50.D1, pp. D988–D995. issn: 0305-1048. doi: 10.1093/nar/gkab1049.

Dedrick, Rebekah M. et al. (jan. de 2017). "Prophage-mediated defence against viral attack and viral counter-defence". Em: *Nature Microbiology* 2.3, p. 16251. issn: 2058-5276. doi: 10.1038/nmicrobiol.2016.251.

Dekkers, Bas J.W. et al. (2016). "The Arabidopsis DELAY OF GERMINATION 1 gene affects ABSCISIC ACID INSENSITIVE 5 (ABI5) expression and genetically interacts with ABI3 during Arabidopsis seed development". Em: *The Plant Journal* 85.4, pp. 451–465. doi: https://doi.org/10.1111/tpj.13118.

Dobritsa, Anna A. e Daniel Coerper (nov. de 2012). "The Novel Plant Protein INAPERTURATE POLLEN1 Marks Distinct Cellular Domains and Controls Formation of

Apertures in the Arabidopsis Pollen Exine ". Em: *The Plant Cell* 24.11, pp. 4452–4464. issn: 1040-4651. doi: 10.1105/tpc.112.101220.

Dong, Qunfeng, Shannon D. Schlueter e Volker Brendel (jan. de 2004). "PlantGDB, plant genome database and analysis tools". Em: *Nucleic Acids Research* 32.suppl_1, pp. D354–D359. issn: 0305-1048. doi: 10.1093/nar/gkh046.

Dubois, Emeline et al. (abr. de 2011). "Homologous Recombination Is Stimulated by a Decrease in dUTPase in Arabidopsis". Em: *PLOS ONE* 6.4, pp. 1–8. doi: 10.1371/journal.pone.0018658.

Dufayard, Jean-François et al. (2017). "New Insights on Leucine-Rich Repeats Receptor-Like Kinase Orthologous Relationships in Angiosperms". Em: *Frontiers in Plant Science* 8, p. 381. doi: 10.3389/fpls.2017.00381.

Dunn, Casey W. e Catriona Munro (2016). "Comparative genomics and the diversity of life". Em: *Zoologica Scripta* 45.S1, pp. 5–13. doi: https://doi.org/10.1111/zsc.12211.

Durand, Eléonore et al. (2020). "Evolution of self-incompatibility in the Brassicaceae: Lessons from a textbook example of natural selection". Em: *Evolutionary Applications* 13.6, pp. 1279–1297. doi: https://doi.org/10.1111/eva.12933.

Eddy, Sean R. (out. de 2011). "Accelerated Profile HMM Searches". Em: *PLOS Computational Biology* 7.10, pp. 1–16. doi: 10.1371/journal.pcbi.1002195.

Edgar, Robert C. (mar. de 2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". Em: *Nucleic Acids Research* 32.5, pp. 1792–1797. issn: 0305-1048. doi: 10.1093/nar/gkh340.

Ehrbar, Kristin e Wolf-Dietrich Hardt (2005). "Bacteriophage-encoded type III effectors in Salmonella enterica subspecies 1 serovar Typhimurium". Em: *Infection, Genetics and Evolution* 5.1, pp. 1–9. issn: 1567-1348. doi: https://doi.org/10.1016/j.meegid.2004.07.004.

Ekstrom, Alexander et al. (ago. de 2014). "PlantCAZyme: a database for plant carbohydrate-active enzymes". Em: *Database* 2014. bau079. issn: 1758-0463. doi: 10.1093/database/bau079.

Falster, Daniel S. e Mark Westoby (2003). "Plant height and evolutionary games". Em: *Trends in Ecology & Evolution* 18.7, pp. 337–343. issn: 0169-5347. doi: https://doi.org/10.1016/S0169-5347(03)00061-2.

Fedak, Halina et al. (2016). “Control of seed dormancy in Arabidopsis by a cis-acting noncoding antisense transcript”. Em: *Proceedings of the National Academy of Sciences* 113.48, E7846–E7855. doi: 10.1073/pnas.1608827113.

Felsenstein, Joseph (1985). “Phylogenies and the Comparative Method”. Em: *The American Naturalist* 125.1, pp. 1–15. issn: 00030147, 15375323.

Fernández, Lucía, Ana Rodríguez e Pilar García (mai. de 2018). “Phage or foe: an insight into the impact of viral predation on microbial communities”. Em: *The ISME Journal* 12.5, pp. 1171–1179. issn: 1751-7370. doi: 10.1038/s41396-018-0049-5.

Fischer, Steve et al. (2011). “Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups”. Em: *Current Protocols in Bioinformatics* 35.1, pp. 6.12.1–6.12.19. doi: https://doi.org/10.1002/0471250953.bi0612s35.

Fleischmann, Andreas et al. (out. de 2014). “Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms”. Em: *Annals of Botany* 114.8, pp. 1651–1663. issn: 0305-7364. doi: 10.1093/aob/mcu189.

Fu, Limin et al. (out. de 2012). “CD-HIT: accelerated for clustering the next-generation sequencing data”. Em: *Bioinformatics* 28.23, pp. 3150–3152. issn: 1367-4803. doi: 10.1093/bioinformatics/bts565.

Galili, Tal (2015). “dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering”. Em: *Bioinformatics*. doi: 10.1093/bioinformatics/btv428.

Galili, Tal et al. (2017). “heatmaply: an R package for creating interactive cluster heatmaps for online publishing”. Em: *Bioinformatics*. doi: 10.1093/bioinformatics/btx657.

El-Gebali, Sara et al. (out. de 2018). “The Pfam protein families database in 2019”. Em: *Nucleic Acids Research* 47.D1, pp. D427–D432. issn: 0305-1048. doi: 10.1093/nar/gky995.

Goffeau, A. et al. (1996). “Life with 6000 Genes”. Em: *Science* 274.5287, pp. 546–567. doi: 10.1126/science.274.5287.546.

González-Morales, Sandra Isabel et al. (2016). “Regulatory network analysis reveals novel regulators of seed desiccation tolerance in Arabidopsis thaliana”. Em: *Proceedings*

of the National Academy of Sciences 113.35, E5232–E5241. doi: 10 .1073 / pnas . 1610985113.

Goodstein, David M. et al. (nov. de 2011). "Phytozome: a comparative platform for green plant genomics". Em: *Nucleic Acids Research* 40.D1, pp. D1178–D1186. issn: 0305-1048. doi: 10.1093/nar/gkr944.

Gordillo Altamirano, Fernando et al. (fev. de 2021). "Bacteriophage-resistant Acinetobacter baumannii are resensitized to antimicrobials". Em: *Nature Microbiology* 6.2, pp. 157–161. issn: 2058-5276. doi: 10.1038/s41564-020-00830-7.

Granzotto, Adriana e Guilherme Marcello Queiroga Cruz (2015). "Regulação de Elementos de Transposição: Mecanismos Epigenéticos de Silenciamento, Autorregulação e Ativação por Estresse". Em: *Elementos de transposição: diversidade, evolução, aplicações e impacto nos genomas dos seres vivos*. Ed. por Claudia Marcia Aparecida Carareto, Claudia Barros Monteiro-Vitorello e Marie-Anne Van Sluys. São José do Rio Preto: Editora FIOCRUZ, pp. 91–113. isBn: 978-85-7541-462-0. doi: https://doi.org/10.7476/9788575415672.

Greilhuber, Johann e I J. Leitch (2013). "Genome Size and the Phenotype". Em: *Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes*. Ed. por Johann Greilhuber, Jaroslav Dolezel e Jonathan F. Wendel. Vienna: Springer Vienna, pp. 323–344. isBn: 978-3-7091-1160-4. doi: 10.1007/978-3-7091-1160-4_20.

Groth, Philip et al. (set. de 2006). "PhenomicDB: a new cross-species genotype/phenotype resource". Em: *Nucleic Acids Research* 35.suppl_1, pp. D696–D699. issn: 0305-1048. doi: 10.1093/nar/gkl662.

Harvey, Paul H, Mark D Pagel et al. (1991). *The comparative method in evolutionary biology*. Vol. 239. Oxford university press Oxford.

Haynes, Winston A., Aurelie Tomczak e Purvesh Khatri (jan. de 2018). "Gene annotation bias impedes biomedical research". Em: *Scientific Reports* 8.1, p. 1362. issn: 2045-2322. doi: 10.1038/s41598-018-19333-x.

Heather, James M. e Benjamin Chain (2016). "The sequence of sequencers: The history of sequencing DNA". Em: *Genomics* 107.1, pp. 1–8. issn: 0888-7543. doi: https://doi.org/10.1016/j.ygeno.2015.11.003.

Hidalgo, Oriane et al. (2017). “Is There an Upper Limit to Genome Size?” Em: *Trends in Plant Science* 22.7, pp. 567–573. issn: 1360-1385. doi: https://doi.org/10.1016/j.tplants.2017.04.005.

Hongo, Jorge Augusto et al. (2021). “CALANGO: an annotation-based, phylogeny-aware comparative genomics framework for exploring and interpreting complex genotypes and phenotypes”. Em: *bioRxiv*. doi: 10.1101/2021.08.25.457574.

Hung, Jui-Hung et al. (set. de 2011). “Gene set enrichment analysis: performance evaluation and usage guidelines”. Em: *Briefings in Bioinformatics* 13.3, pp. 281–291. issn: 1467-5463. doi: 10.1093/bib/bbr049.

Huo, Heqiang, Shouhui Wei e Kent J. Bradford (2016). “DELAY OF GERMINATION1 (DOG1) regulates both seed dormancy and flowering time through microRNA pathways”. Em: *Proceedings of the National Academy of Sciences* 113.15, E2199–E2206. doi: 10.1073/pnas.1600558113.

IHGSC et al. (fev. de 2001). “Initial sequencing and analysis of the human genome”. Em: *Nature* 409.6822, pp. 860–921. issn: 1476-4687. doi: 10.1038/35057062.

Initiative, The Arabidopsis Genome (dez. de 2000). “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana”. Em: *Nature* 408.6814, pp. 796–815. issn: 1476-4687. doi: 10.1038/35048692.

Jones, Philip et al. (jan. de 2014). “InterProScan 5: genome-scale protein function classification”. Em: *Bioinformatics* 30.9, pp. 1236–1240. issn: 1367-4803. doi: 10.1093/bioinformatics/btu031.

Kang, Ming et al. (2014). “Adaptive and nonadaptive genome size evolution in Karst endemic flora of China”. Em: *New Phytologist* 202.4, pp. 1371–1381. doi: https://doi.org/10.1111/nph.12726.

Kattge, Jens et al. (2020). “TRY plant trait database – enhanced coverage and open access”. Em: *Global Change Biology* 26.1, pp. 119–188. doi: https://doi.org/10.1111/gcb.14904.

Kawahara, Yoshihiro et al. (fev. de 2013). “Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data”. Em: *Rice* 6.1, p. 4. issn: 1939-8433. doi: 10.1186/1939-8433-6-4.

Kawashima, Tomokazu et al. (jul. de 2015). “Diversification of histone H2A variants during plant evolution”. Em: *Trends in Plant Science* 20.7, pp. 419–425. issn: 1360-1385. doi: 10.1016/j.tplants.2015.04.005.

Knight, Charles A., Nicole A. Molinari e Dmitri A. Petrov (jan. de 2005). “The Large Genome Constraint Hypothesis: Evolution, Ecology and Phenotype”. Em: *Annals of Botany* 95.1, pp. 177–190. issn: 0305-7364. doi: 10.1093/aob/mci011.

Koornneef, Maarten, Leónie Bentsink e Henk Hilhorst (2002). “Seed dormancy and germination”. Em: *Current Opinion in Plant Biology* 5.1, pp. 33–36. issn: 1369-5266. doi: https://doi.org/10.1016/S1369-5266(01)00219-9.

Kopriva, Stanislav e Andreas P M Weber (jan. de 2021). “Genetic encoding of complex traits”. Em: *Journal of Experimental Botany* 72.1, pp. 1–3. issn: 0022-0957. doi: 10.1093/jxb/eraa498.

Krishnakumar, Vivek et al. (nov. de 2014). “Araport: the Arabidopsis Information Portal”. Em: *Nucleic Acids Research* 43.D1, pp. D1003–D1009. issn: 0305-1048. doi: 10.1093/nar/gku1200.

Kuang, Kevin, Quyu Kong e Francesco Napolitano (2022). *pbmcapply: Tracking the Progress of Mc*pply with Progress Bar*. R package version 1.5.1.

Kumar, Sudhir, Glen Stecher et al. (mai. de 2018). “MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms”. Em: *Molecular Biology and Evolution* 35.6, pp. 1547–1549. issn: 0737-4038. doi: 10.1093/molbev/msy096.

Kumar, Sudhir, Michael Suleski et al. (ago. de 2022). “TimeTree 5: An Expanded Resource for Species Divergence Times”. Em: *Molecular Biology and Evolution* 39.8. msac174. issn: 1537-1719. doi: 10.1093/molbev/msac174.

Kumar, Vikash, Evgeniy N. Donev et al. (2020). “Genome-Wide Identification of Populus Malectin/Malectin-Like Domain-Containing Proteins and Expression Analyses Reveal Novel Candidates for Signaling and Regulation of Wood Development”. Em: *Frontiers in Plant Science* 11, p. 588846. doi: 10.3389/fpls.2020.588846.

Kumar, Vikash, Matthieu Hainaut et al. (2019). “Poplar carbohydrate-active enzymes: whole-genome annotation and functional analyses based on RNA expression data”. Em: *The Plant Journal* 99.4, pp. 589–609. doi: https://doi.org/10.1111/tpj.14417.

Lanfear, Robert et al. (mai. de 2013). “Taller plants have lower rates of molecular evolution”. Em: *Nature Communications* 4.1, p. 1879. issn: 2041-1723. doi: 10.1038/ncomms2836.

Lee, Byung Ha et al. (jul. de 2021). “A species-specific functional module controls formation of pollen apertures”. Em: *Nature Plants* 7.7, pp. 966–978. issn: 2055-0278. doi: 10.1038/s41477-021-00951-9.

Lee, Heewook et al. (2012). “Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing”. Em: *Proceedings of the National Academy of Sciences* 109.41, E2774–E2783. doi: 10.1073/pnas.1210309109.

Lei, Bingkun e Frédéric Berger (2020). “H2A Variants in Arabidopsis: Versatile Regulators of Genome Activity”. Em: *Plant Communications* 1.1, p. 100015. issn: 2590-3462. doi: https://doi.org/10.1016/j.xplc.2019.100015.

Leitch, A. R. e I. J. Leitch (2012). “Ecological and genetic factors linked to contrasting genome dynamics in seed plants”. Em: *New Phytologist* 194.3, pp. 629–646. doi: https://doi.org/10.1111/j.1469-8137.2012.04105.x.

Leitch, I. J., Mark W. Chase e Michael D. Bennett (dez. de 1998). “Phylogenetic Analysis of DNA C-values Provides Evidence for a Small Ancestral Genome Size in Flowering Plants”. Em: *Annals of Botany* 82.suppl_1, pp. 85–94. issn: 0305-7364. doi: 10.1006/anbo.1998.0783.

Leitch, I. J., D. E. Soltis et al. (jan. de 2005). “Evolution of DNA Amounts Across Land Plants (Embryophyta)”. Em: *Annals of Botany* 95.1, pp. 207–217. issn: 0305-7364. doi: 10.1093/aob/mci014.

León, M e R Bastías (2015). “Virulence reduction in bacteriophage resistant bacteria.” Em: *Frontiers in Microbiology* 343.6. doi: http://dx.doi.org/10.3389/fmicb.2015.00343.

Li, Fay-Wei et al. (jul. de 2018). “Fern genomes elucidate land plant evolution and cyanobacterial symbioses”. Em: *Nature Plants* 4.7, pp. 460–472. issn: 2055-0278. doi: 10.1038/s41477-018-0188-8.

Li, Linzhou et al. (set. de 2020). “The genome of Prasinoderma coloniale unveils the existence of a third phylum within green plants”. Em: *Nature Ecology &, Evolution* 4.9, pp. 1220–1231. issn: 2397-334X. doi: 10.1038/s41559-020-1221-7.

Li, Weizhong e Adam Godzik (mai. de 2006). “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”. Em: *Bioinformatics* 22.13, pp. 1658–1659. issn: 1367-4803. doi: 10.1093/bioinformatics/btl158.

Liolios, Konstantinos et al. (nov. de 2009). “The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata”. Em: *Nucleic Acids Research* 38.suppl_1, pp. D346–D354. issn: 0305-1048. doi: 10.1093/nar/gkp848.

Lisch, Damon (jan. de 2013). “How important are transposons for plant evolution?” Em: *Nature Reviews Genetics* 14.1, pp. 49–61. issn: 1471-0064. doi: 10.1038/nrg3374.

Liu, Jian-Zhong e Steven A. Whitham (2013). “Overexpression of a soybean nuclear localized type–III DnaJ domain-containing HSP40 reveals its roles in cell death and disease resistance”. Em: *The Plant Journal* 74.1, pp. 110–121. doi: https:// doi.org/10.1111/tpj.12108.

M, Carlson (2019). *GO.db: A set of annotation maps describing the entire Gene Ontology*. R package version 3.8.2.

Ma, Xuelian et al. (set. de 2022). “PlantGSAD: a comprehensive gene set annotation database for plant species”. Em: *Nucleic Acids Research* 50.D1, pp. D1456–D1467. issn: 0305-1048. doi: 10.1093/nar/gkab794.

Manni, Mosè et al. (jul. de 2021). “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes”. Em: *Molecular Biology and Evolution* 38.10, pp. 4647–4654. issn: 1537-1719. doi: 10.1093/molbev/msab199.

Marks, Rose A. et al. (dez. de 2021). “Representation and participation across 20 years of plant genome sequencing”. Em: *Nature Plants* 7.12, pp. 1571–1578. issn: 2055-0278. doi: 10.1038/s41477-021-01031-8.

Mashau, Aluoneswi C. et al. (2021). “Plant height and lifespan predict range size in southern African grasses”. Em: *Journal of Biogeography* 48.12, pp. 3047–3059. doi: https://doi.org/10.1111/jbi.14261.

Maslov, Sergei e Kim Sneppen (jan. de 2017). “Population cycles and species diversity in dynamic Kill-the-Winner model of microbial ecosystems”. Em: *Scientific Reports* 7.1, p. 39642. issn: 2045-2322. doi: 10.1038/srep39642.

Mazuecos-Aguilera, Ismael et al. (2021). “The Role of INAPERTURATE POLLEN1 as a Pollen Aperture Factor Is Conserved in the Basal Eudicot Eschscholzia californica (Papaveraceae)”. Em: *Frontiers in Plant Science* 12. issn: 1664-462X. doi: 10.3389/fpls.2021.701286.

Minelli, Alessandro (2018). “Introducing Plant Evo-Devo”. Em: *Plant Evolutionary Developmental Biology: The Evolvability of the Phenotype*. Cambridge University Press, pp. 1–29. doi: 10.1017/9781139542364.002.

Minh, Bui Quang et al. (fev. de 2020). “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era”. Em: *Molecular Biology and Evolution* 37.5, pp. 1530–1534. issn: 0737-4038. doi: 10.1093/molbev/msaa015.

Moles, Angela T. et al. (2009). “Global patterns in plant height”. Em: *Journal of Ecology* 97.5, pp. 923–932. doi: https://doi.org/10.1111/j.1365-2745.2009.01526.x.

Morgan, Martin (2022). *BiocManager: Access the Bioconductor Project Package Repository*. R package version 1.30.18.

Mosavi, Leila K. et al. (2004). “The ankyrin repeat as molecular architecture for protein recognition”. Em: *Protein Science* 13.6, pp. 1435–1448. doi: https://doi.org/10.1110/ps.03554604.

Mukherjee, Supratim et al. (out. de 2016). “Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements”. Em: *Nucleic Acids Research* 45.D1, pp. D446–D456. issn: 0305-1048. doi: 10.1093/nar/gkw992.

Nagy, László G et al. (jan. de 2020). “Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing”. Em: *Nucleic Acids Research* 48.5, pp. 2209–2219. issn: 0305-1048. doi: 10.1093/nar/gkz1241.

Nakabayashi, Kazumi et al. (jul. de 2012). “The Time Required for Dormancy Release in Arabidopsis Is Determined by DELAY OF GERMINATION1 Protein Levels in Freshly Harvested Seeds”. Em: *The Plant Cell* 24.7, pp. 2826–2838. issn: 1040-4651. doi: 10.1105/tpc.112.100214.

Nasrallah, June B. e Mikhail E. Nasrallah (mar. de 2014). “S-locus receptor kinase signalling”. Em: *Biochemical Society Transactions* 42.2, pp. 313–319. issn: 0300-5127. doi: 10.1042/BST20130222.

Niklas, Karl J. e Ulrich Kutschera (2010). “The evolution of the land plant life cycle”. Em: *New Phytologist* 185.1, pp. 27–41. doi: https://doi.org/10.1111/j.1469-8137.2009.03054.x.

Nishimura, Noriyuki et al. (jun. de 2018). “Control of seed dormancy and germination by DOG1-AHG1 PP2C phosphatase complex via binding to heme”. Em: *Nature Communications* 9.1, p. 2132. issn: 2041-1723. doi: 10.1038/s41467-018-04437-9.

Nishiyama, Eri et al. (2021). “Ancient and recent gene duplications as evolutionary drivers of the seed maturation regulators DELAY OF GERMINATION1 family genes”. Em: *New Phytologist* 230.3, pp. 889–901. doi: https://doi.org/10.1111/nph.17201.

Nishiyama, Takashi et al. (jan. de 2013). “The structure of the deacetylase domain of Escherichia coli PgaB, an enzyme required for biofilm formation: a circularly permuted member of the carbohydrate esterase 4 family”. Em: *Acta Crystallographica Section D* 69.1, pp. 44–51. doi: 10.1107/S0907444912042059.

O’Leary, Nuala A. et al. (nov. de 2015). “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. Em: *Nucleic Acids Research* 44.D1, pp. D733–D745. issn: 0305-1048. doi: 10.1093/nar/gkv1189.

Pagès, H et al. (2022). *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. R package version 1.58.0.

Pang, Shuai et al. (mai. de 2015). “GPU MrBayes V3.1: MrBayes on Graphics Processing Units for Protein Sequence Data”. Em: *Molecular Biology and Evolution* 32.9, pp. 2496–2497. issn: 0737-4038. doi: 10.1093/molbev/msv129.

Paradis, E. e K. Schliep (2019). “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R”. Em: *Bioinformatics* 35, pp. 526–528.

Park, Beom Seok e Jie-Oh Lee (dez. de 2013). “Recognition of lipopolysaccharide pattern by TLR4 complexes”. Em: *Experimental & Molecular Medicine* 45.12, e66–e66. issn: 2092-6413. doi: 10.1038/emm.2013.97.

Pasha, Asher et al. (jul. de 2020). “Araport Lives: An Updated Framework for Arabidopsis Bioinformatics”. Em: *The Plant Cell* 32.9, pp. 2683–2686. issn: 1040-4651. doi: 10.1105/tpc.20.00358.

Pawluk, April et al. (2014). “A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa”. Em: *mBio* 5.2, e00896–14. doi: 10.1128/mBio.00896-14.

Peiffer, Jason A et al. (abr. de 2014). “The Genetic Architecture Of Maize Height”. Em: *Genetics* 196.4, pp. 1337–1356. issn: 1943-2631. doi: 10.1534/genetics.113.159152.

Pellicer, Jaume, Michae F. Fay e I. J. Leitch (set. de 2010). “The largest eukaryotic genome of them all?” Em: *Botanical Journal of the Linnean Society* 164.1, pp. 10–15. issn: 0024-4074. doi: 10.1111/j.1095-8339.2010.01072.x.

Pellicer, Jaume, Oriane Hidalgo et al. (2018). “Genome Size Diversity and Its Impact on the Evolution of Land Plants”. Em: *Genes* 9.2. issn: 2073-4425. doi: 10.3390/genes9020088.

Pellicer, Jaume e I J. Leitch (2020). “The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies”. Em: *New Phytologist* 226.2, pp. 301–305. doi: https://doi.org/10.1111/nph.16261.

Petrov, Dmitri A. (jan. de 2001). “Evolution of genome size: new approaches to an old problem”. Em: *Trends in Genetics* 17.1, pp. 23–28. issn: 0168-9525. doi: 10.1016/S0168-9525(00)02157-0.

— (2002). “Mutational Equilibrium Model of Genome Size Evolution”. Em: *Theoretical Population Biology* 61.4, pp. 531–544. issn: 0040-5809. doi: https://doi.org/10.1006/tpbi.2002.1605.

Pinard, Desre et al. (mai. de 2015). “Comparative analysis of plant carbohydrate active enZymes and their role in xylogenesis”. Em: *BMC Genomics* 16.1, p. 402. issn: 1471-2164. doi: 10.1186/s12864-015-1571-8.

Pinheiro, José, Douglas Bates e R Core Team (2022). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-157.

Plazzi, Federico, Ronald R. Ferrucci e Marco Passamonti (abr. de 2010). “Phylogenetic representativeness: a new method for evaluating taxon sampling in evolutionary studies”. Em: *BMC Bioinformatics* 11.1, p. 209. issn: 1471-2105. doi: 10.1186/1471-2105-11-209.

Proost, Sebastian et al. (out. de 2014). “PLAZA 3.0: an access point for plant comparative genomics”. Em: *Nucleic Acids Research* 43.D1, pp. D974–D981. issn: 0305-1048. doi: 10.1093/nar/gku986.

Pulido, Pablo e Dario Leister (2018). “Novel DNAJ-related proteins in Arabidopsis thaliana”. Em: *The New Phytologist* 217.2, pp. 480–490. issn: 0028646X, 14698137.

Pulkkinen, W S e S I Miller (1991). “A Salmonella typhimurium virulence protein is similar to a Yersinia enterocolitica invasion protein and a bacteriophage lambda outer membrane protein”. Em: *Journal of Bacteriology* 173.1, pp. 86–93. doi: 10.1128/jb.173.1.86-93.1991.

Puttick, Mark N., James Clark e Philip C. J. Donoghue (2015). “Size is not everything: rates of genome size evolution, not C-value, correlate with speciation in angiosperms”. Em: *Proceedings of the Royal Society B: Biological Sciences* 282.1820, p. 20152289. doi: 10.1098/rspb.2015.2289.

Rambaut, Andrew et al. (abr. de 2018). “Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7”. Em: *Systematic Biology* 67.5, pp. 901–904. issn: 1063-5157. doi: 10.1093/sysbio/syy032.

Ramisetty, Bhaskar Chandra Mohan e Pavithra Anantharaman Sudhakari (2019). “Bacterial ‘Grounded’ Prophages: Hotspots for Genetic Renovation and Innovation”. Em: *Frontiers in Genetics* 10. issn: 1664-8021. doi: 10.3389/fgene.2019.00065.

Ren, Ren et al. (2018). “Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms”. Em: *Molecular Plant* 11.3. Genome Biology, pp. 414–428. issn: 1674-2052. doi: https://doi.org/10.1016/j.molp.2018.01.002.

Revell, Liam J. (2012). “phytools: an R package for phylogenetic comparative biology (and other things)”. Em: *Methods in Ecology and Evolution* 3.2, pp. 217–223. doi: https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Roff, Derek A. (1997). *Evolutionary Quantitative Genetics*. New York: Springer New York. isbn: 978-1-4615-4080-9. doi: https://doi.org/10.1007/978-1-4615-4080-9.

Sall, Khadidiatou et al. (2019). “DELAY OF GERMINATION 1-LIKE 4 acts as an inducer of seed reserve accumulation”. Em: *The Plant Journal* 100.1, pp. 7–19. doi: https://doi.org/10.1111/tpj.14485.

Salzberg, Steven L. (mai. de 2019). “Next-generation genome annotation: we still struggle to get it right”. Em: *Genome Biology* 20.1, p. 92. issn: 1474-760X. doi: 10.1186/s13059-019-1715-2.

Sandoval, Francisco J., Yi Zhang e Sanja Roje (nov. de 2008). “Flavin Nucleotide Metabolism in Plants: MONOFUNCTIONAL ENZYMES SYNTHESIZE FAD IN PLASTIDS

*". Em: *Journal of Biological Chemistry* 283.45, pp. 30890–30900. issn: 0021-9258. doi: 10.1074/jbc.M803416200.

Sanger, F. e A.R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". Em: *Journal of Molecular Biology* 94.3, pp. 441–448. issn: 0022-2836. doi: https://doi.org/10.1016/0022-2836(75)90213-2.

Sayers, Eric W et al. (out. de 2019). "GenBank". Em: *Nucleic Acids Research* 48.D1, pp. D84–D86. issn: 0305-1048. doi: 10.1093/nar/gkz956.

Schäffer, Alejandro A. et al. (jul. de 2001). "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements". Em: *Nucleic Acids Research* 29.14, pp. 2994–3005. issn: 0305-1048. doi: 10.1093/nar/29.14.2994.

Schallus, Thomas et al. (2008). "Malectin: A Novel Carbohydrate-binding Protein of the Endoplasmic Reticulum and a Candidate Player in the Early Steps of Protein N-Glycosylation". Em: *Molecular Biology of the Cell* 19.8. PMID: 18524852, pp. 3404–3414. doi: 10.1091/mbc.e08-04-0354.

Schneider, Rene e Staffan Persson (2015). "Another brick in the wall". Em: *Science* 350.6257, pp. 156–157. doi: 10.1126/science.aad3200.

Schuster, Stephan C. (jan. de 2008). "Next-generation sequencing transforms today's biology". Em: *Nature Methods* 5.1, pp. 16–18. issn: 1548-7105. doi: 10.1038/nmeth1156.

SHAPIRO, S. S. e M. B. WILK (dez. de 1965). "An analysis of variance test for normality (complete samples)". Em: *Biometrika* 52.3-4, pp. 591–611. doi: 10.1093/biomet/52.3-4.591.

Sievert, Carson (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. isbn: 978-3-319-24277-4.

Silveira, Cynthia B. e Forest L. Rohwer (jul. de 2016). "Piggyback-the-Winner in host-associated microbial communities". Em: *npj Biofilms and Microbiomes* 2.1, p. 16010. issn: 2055-5008. doi: 10.1038/npjbiofilms.2016.10.

Simmons, Emilia L. et al. (2020). "Biofilm Structure Promotes Coexistence of Phage-Resistant and Phage-Susceptible Bacteria". Em: *mSystems* 5.3, e00877–19. doi: 10.1128/mSystems.00877-19.

Sørensen, Iben et al. (2011). “The charophycean green algae provide insights into the early origins of plant cell walls”. Em: *The Plant Journal* 68.2, pp. 201–211. doi: https://doi.org/10.1111/j.1365-313X.2011.04686.x.

Steyert, Susan R. e James B. Kaper (2012). “Contribution of Urease to Colonization by Shiga Toxin-Producing Escherichia coli”. Em: *Infection and Immunity* 80.8, pp. 2589–2600. doi: 10.1128/IAI.00210-12.

Subburaj, Saminathan et al. (jun. de 2016). “Phylogenetic Analysis, Lineage-Specific Expansion and Functional Divergence of seed dormancy 4-Like Genes in Plants”. Em: *PLOS ONE* 11.6, pp. 1–24. doi: 10.1371/journal.pone.0153717.

Tello-Ruiz, Marcela K et al. (nov. de 2020). “Gramene 2021: harnessing the power of comparative genomics and pathways for plant research”. Em: *Nucleic Acids Research* 49.D1, pp. D1452–D1463. issn: 0305-1048. doi: 10.1093/nar/gkaa979.

Tenenbaum D, Maintainer B (2022). *KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)*. R package version 1.36.3.

Tomaž, Špela, Kristina Gruden e Anna Coll (2022). “TGA transcription factors—Structural characteristics as basis for functional variability”. Em: *Frontiers in Plant Science* 13. issn: 1664-462X. doi: 10.3389/fpls.2022.935819.

Tong, Chao et al. (jan. de 2020). “Comparative Genomics Identifies Putative Signatures of Sociality in Spiders”. Em: *Genome Biology and Evolution* 12.3, pp. 122–133. issn: 1759-6653. doi: 10.1093/gbe/evaa007.

Touchon, Marie et al. (jan. de 2009). “Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths”. Em: *PLOS Genetics* 5.1, pp. 1–25. doi: 10.1371/journal.pgen.1000344.

Tuskan, G. A. et al. (2006). “The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray)”. Em: *Science* 313.5793, pp. 1596–1604. doi: 10.1126/science.1128691.

Ung, Huoi, Wolfgang Moeder e Keiko Yoshioka (set. de 2014). “Arabidopsis Triphosphate Tunnel Metalloenzyme2 Is a Negative Regulator of the Salicylic Acid-Mediated Feedback Amplification Loop for Defense Responses ”. Em: *Plant Physiology* 166.2, pp. 1009–1021. issn: 0032-0889. doi: 10.1104/pp.114.248757.

Vaidya, Gaurav, David J. Lohman e Rudolf Meier (2011). “SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon

information". Em: *Cladistics* 27.2, pp. 171–180. doi: https://doi.org/10.1111/j.1096-0031.2010.00329.x.

Vaidyanathan, Ramnath et al. (2021). *htmlwidgets: HTML Widgets for R*. R package version 1.5.4.

Vandecraen, Joachim et al. (2017). "The impact of insertion sequences on bacterial genome plasticity and adaptability". Em: *Critical Reviews in Microbiology* 43.6. PMID: 28407717, pp. 709–730. doi: 10.1080/1040841X.2017.1303661.

Veselý, Pavel, Petr Bureš e Petr Šmarda (ago. de 2013). "Nutrient reserves may allow for genome size increase: evidence from comparison of geophytes and their sister non-geophytic relatives". Em: *Annals of Botany* 112.6, pp. 1193–1200. issn: 0305-7364. doi: 10.1093/aob/mct185.

Vinogradov, Alexander E (2003). "Selfish DNA is maladaptive: evidence from the plant Red List". Em: *Trends in Genetics* 19.11, pp. 609–614. issn: 0168-9525. doi: https://doi.org/10.1016/j.tig.2003.09.010.

Vitti, Joseph J., Sharon R. Grossman e Pardis C. Sabeti (2013). "Detecting Natural Selection in Genomic Data". Em: *Annual Review of Genetics* 47.1. PMID: 24274750, pp. 97–120. doi: 10.1146/annurev-genet-111212-133526.

Vogel, Christine e Cyrus Chothia (mai. de 2006). "Protein Family Expansions and Biological Complexity". Em: *PLOS Computational Biology* 2.5, pp. 1–13. doi: 10.1371/journal.pcbi.0020048.

Wang, B et al. (2019). "[The China National GeneBank owned by all, completed by all and shared by all]". Em: *Yi Chuan* 20.41, pp. 761–772. doi: 10.16288/j.yczz..

Wang, Dandan et al. (2021). "Which factors contribute most to genome size variation within angiosperms?" Em: *Ecology and Evolution* 11.6, pp. 2660–2668. doi: https://doi.org/10.1002/ece3.7222.

Wang, Xiaoxue et al. (dez. de 2010). "Cryptic prophages help bacteria cope with adverse environments". Em: *Nature Communications* 1.1, p. 147. issn: 2041-1723. doi: 10.1038/ncomms1146.

Waterhouse, Robert M et al. (dez. de 2017). "BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics". Em: *Molecular Biology and Evolution* 35.3, pp. 543–548. issn: 0737-4038. doi: 10.1093/molbev/msx319.

Wendel, Jonathan F. et al. (mai. de 2002). “Feast and famine in plant genomes”. Em: *Genetica* 115.1, pp. 37–47. issn: 1573-6857. doi: 10.1023/A:1016020030189.

Wickham, Hadley (2019). *assertthat: Easy Pre and Post Assertions*. R package version 0.2.1.

— (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman & Hall. isbn: 9781138331457.

Wickham, Hadley, Jay Hesselberth e Maëlle Salmon (2022). *pkgdown: Make Static HTML Documentation for a Package*. R package version 2.0.3.

Willi, Yvone e Ary A. Hoffman (2009). “Demographic factors and genetic variation influence population persistence under environmental change”. Em: *Journal of Evolutionary Biology* 22.1, pp. 124–133. doi: https : / / doi . org / 10 . 1111 / j . 1420 - 9101.2008.01631.x.

Wolf, Andrea J. e David M. Underhill (abr. de 2018). “Peptidoglycan recognition by the innate immune system”. Em: *Nature Reviews Immunology* 18.4, pp. 243–254. issn: 1474-1741. doi: 10.1038/nri.2017.136.

Wolf, Jason B. (2002). “The geometry of phenotypic evolution in developmental hyperspace”. Em: *Proceedings of the National Academy of Sciences* 99.25, pp. 15849–15851. doi: 10.1073/pnas.012686699.

Xiao, Yu et al. (mai. de 2019). “Mechanisms of RALF peptide perception by a heterotypic receptor complex”. Em: *Nature* 572.7768, pp. 270–274. issn: 1476-4687. doi: 10.1038/s41586-019-1409-7.

Xie, Yihui, Joe Cheng e Xianying Tan (2022). *DT: A Wrapper of the JavaScript Library ’DataTables’*. R package version 0.23.

Xue, Han et al. (out. de 2021). “qPTMplants: an integrative database of quantitative post-translational modifications in plants”. Em: *Nucleic Acids Research* 50.D1, pp. D1491–D1499. issn: 0305-1048. doi: 10.1093/nar/gkab945.

Yang, He et al. (2021). “Malectin/Malectin-like domain-containing proteins: A repertoire of cell surface molecules with broad functional potential”. Em: *The Cell Surface* 7, p. 100056. issn: 2468-2330. doi: https://doi.org/10.1016/j.tcsw.2021.100056.

Yang, Xiaohan et al. (set. de 2019). “Comparative genomics can provide new insights into the evolutionary mechanisms and gene function in CAM plants”. Em: *Journal*

of Experimental Botany 70.22, pp. 6539–6547. issn: 0022-0957. doi: 10.1093/jxb/erz408.

Yelagandula, Ramesh et al. (jul. de 2014). "The Histone Variant H2A.W Defines Heterochromatin and Promotes Chromatin Condensation in Arabidopsis". Em: *Cell* 158.1, pp. 98–109. issn: 0092-8674. doi: 10.1016/j.cell.2014.06.006.

Zhang, Jian et al. (fev. de 2020). "The hornwort genome and early land plant evolution". Em: *Nature Plants* 6.2, pp. 107–118. issn: 2055-0278. doi: 10.1038/s41477-019-0588-4.

Zu, Pengjuan e Florian P. Schiestl (2017). "The effects of becoming taller: direct and pleiotropic effects of artificial selection on plant height in Brassica rapa". Em: *The Plant Journal* 89.5, pp. 1009–1019. doi: https://doi.org/10.1111/tpj.13440.

Zwickl, Derrick J. e David M. Hillis (jul. de 2002). "Increased Taxon Sampling Greatly Reduces Phylogenetic Error". Em: *Systematic Biology* 51.4, pp. 588–598. issn: 1063-5157. doi: 10.1080/10635150290102339.

3.9 SUPPLEMENTARY INFORMATION

3.9.1 Supplementary methods

To study the evolution of the DOG1 protein family in plants we analyzed a rich dataset of DOG1-containing protein sequences, our method is represented in the form of a flowchart in Figure 1. To do so, we first built a comprehensive and phylogenetic diverse genomic dataset for 171 species of Archaeplastida. To do so we screened 5 different databases, National Center for Biotechnology Information (NCBI) databases - RefSeq (O'Leary et al. 2015) and GenBank (Sayers et al. 2019), Phytozome (Goodstein et al. 2011), Gymno PLAZA (Proost et al. 2014), FernBase (F.-W. Li et al. 2018), and CNGB (B. Wang et al. 2019). We downloaded the assembled genomic data for all Embryophyta species in the NCBI databases. Then, we complemented our dataset with predicted proteomes belonging to relevant and underrepresented lineages with data from other public databases. To reduce the redundancy generated by the presence of different isoforms of the same genes and avoid bias towards model organisms, we applied two different protocols to our dataset. The first protocol was used on NCBI data, this in-house pipeline keeps only the longest protein sequence of each locus based on the "locus_tag" or "gene_id"

information. However, proteomes from other databases do not contain these fields in their sequence headers and could not be submitted to our pipeline. To reduce redundancy for those proteomes we used the CD-HIT software (W. Li e Godzik 2006; Fu et al. 2012) with the threshold set to 1. We evaluated the assembling quality of each proteome, using the BUSCO (Manni et al. 2021) software to assess their gene completeness based on Eukaryota_odb10. We kept in our dataset proteomes with completeness higher than 70% and rates lower than 20% for duplicated and fragmented genes, based on Arabidopsis thaliana results (Supplementary Table S1). This way, we have generated a non-redundant dataset of high-quality proteomes from genomic data. We searched for sequences containing DOG1 domain using textitInterProScan 5 (Jones et al. 2014) to perform a *de novo* annotation for all the 171 non-redundant predicted proteomes in our dataset (Supplementary Table S1). Then, we extracted information on the sequences containing the InterPro ID "IPR025422", associated with DOG1 domain family annotations.

3.9.2 Supplementary materials

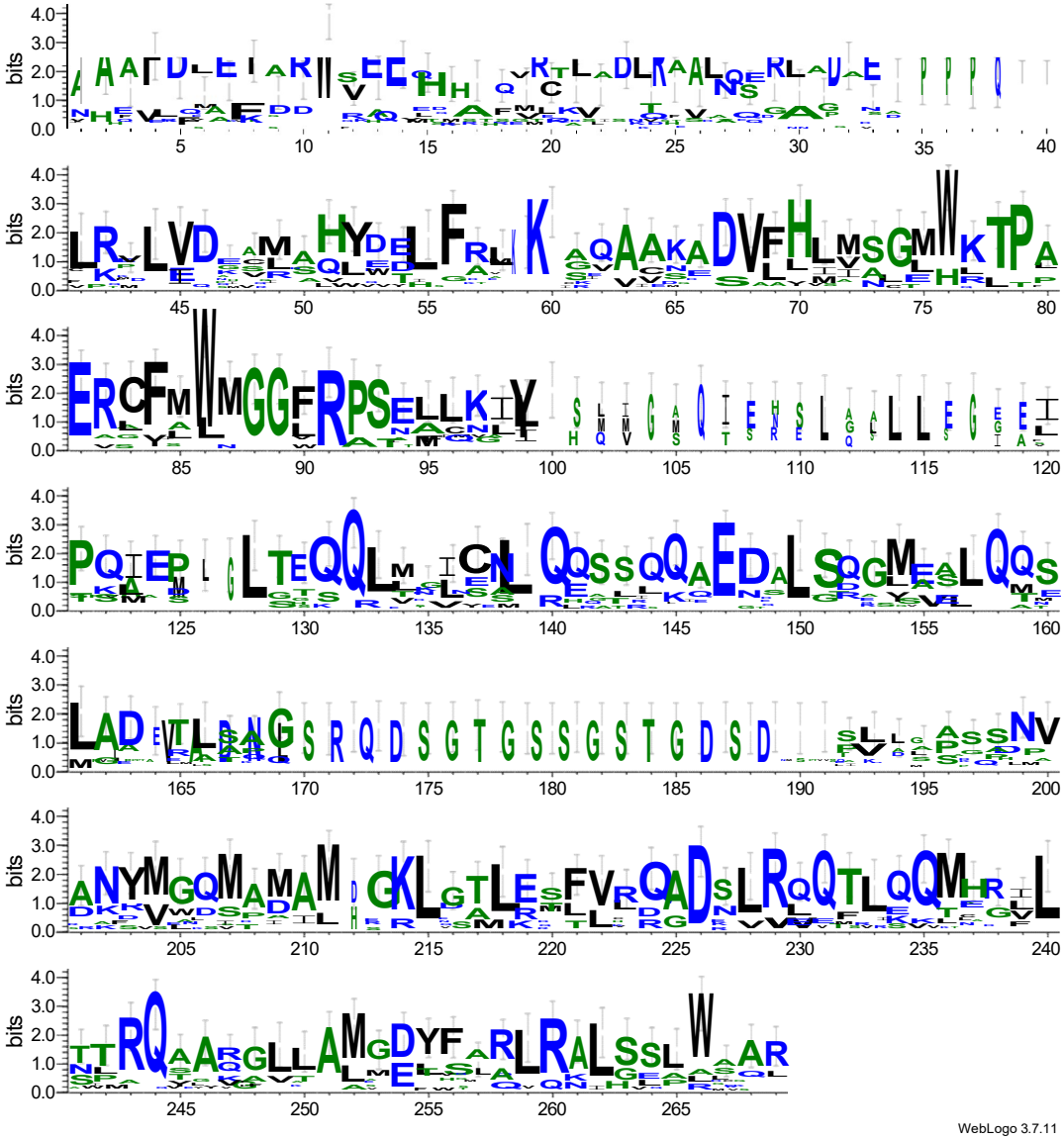


Figura S-1: logo sequence comparison of 12 DOG1 domains from the following green algae species: *Chara braunii*, *C. subellipsoidea* C-169, *Coleochaete irregularis*, *Cylindrocystis cushleckae*, *K. nitens*, *Mesotaenium caldariorum*, *Staurodesmus omearii*, and *Trebouxia* sp. A1-2; 3).

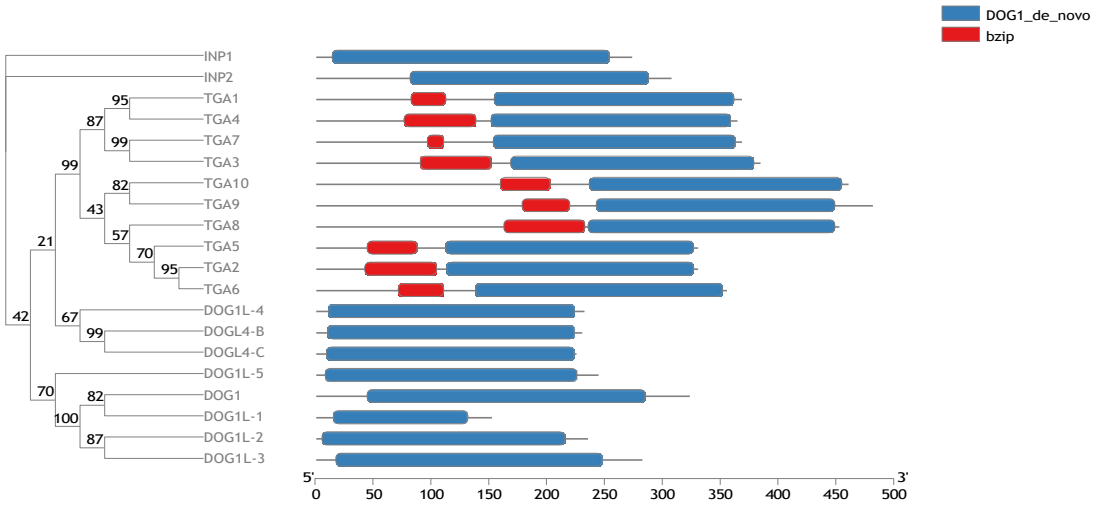


Figura S-2: Domain architecture representation of the 20 DOG1-containing proteins from *Arabidopsis thaliana*. Red bars represent bZIP domains from TGA superfamily, while blue bars represent DOG1 domain.

Tabela S1: List of species we surveyed and their respective database of origin, major clade, family, species. species code, BUSCO (Manni et al. 2021) results, and the number of DOG1 domain copies.

Database	Major clade	Species	Code	BUSCO quality assessment						# DOG1
				C	S	D	F	M	n	
phytozome	Angiosperm	*Amaranthus hypochondriacus*	Ahy	83.9	72.9	11	9.8	6.3	255	18
NCBI	Angiosperm	*Amborella trichopoda*	Ati	98	92.9	5.1	0.8	1.2	255	6
phytozome	Angiosperm	*Spirodela polyrhiza*	Spo	90.6	88.6	2	4.7	4.7	255	11
NCBI	Angiosperm	*Asparagus officinalis*	Aof	89.4	69.4	20	4.3	6.3	255	10
NCBI	Angiosperm	*Lactuca saligna*	Lsa	93.3	79.2	14.1	2	4.7	255	24
NCBI	Angiosperm	*Lactuca sativa*	Lst	99.6	81.6	18	0	0.4	255	27
CNGB	Angiosperm	*Podophyllum peltatum*	Ppe	74.5	68.6	5.9	12.5	13	255	8
phytozome	Angiosperm	*Arabidopsis halleri*	Aha	90.2	76.9	13.3	3.5	6.3	255	17
NCBI	Angiosperm	*Arabidopsis thaliana*	Ath	99.6	81.6	18	0	0.4	255	20
NCBI	Angiosperm	*Arabis nemorensis*	Ane	98.8	85.1	13.7	0	1.2	255	20
phytozome	Angiosperm	*Boechera stricta*	Bst	96	82.7	13.3	3.1	0.9	255	17
phytozome	Angiosperm	*Descurainia sophioides*	Dso	99.2	85.5	13.7	0.4	0.4	255	18
NCBI	Angiosperm	*Ananas comosus*	Aco	96.9	85.5	11.4	0.4	2.7	255	20
NCBI	Angiosperm	*Cannabis sativa*	Csa	95.7	86.7	9	1.2	3.1	255	11
NCBI	Angiosperm	*Parasponia andersonii*	Pan	94.5	91	3.5	4.7	0.8	255	16
NCBI	Angiosperm	*Trema orientale*	Tor	94.5	89.4	5.1	4.7	0.8	255	15

Table S1 continued from previous page

Database	Major clade	Species	Code	BUSCO quality assessment						# DOG1
				C	S	D	F	M	n	
NCBI	Angiosperm	*Carica papaya*	Cpa	79.7	77.3	2.4	12.9	7.4	255	15
NCBI	Angiosperm	*Cephalotus follicularis*	Cfo	98	89.4	8.6	2	0	255	12
NCBI	Angiosperm	*Beta vulgaris subsp. vulgaris*	Bvu	99.6	91.8	7.8	0	0.4	255	16
NCBI	Angiosperm	*Spinacia oleracea*	Sol	100	90.6	9.4	0	0	255	17
CNGB	Angiosperm	*Ascarina rubricaulis*	Arb	83.1	80.4	2.7	9.4	7.5	255	6
NCBI	Angiosperm	*Kingdonia uniflora*	Kun	83.2	76.1	7.1	13.7	3.1	255	20
phytozome	Angiosperm	*Cleome violacea*	Cvi	99.2	93.7	5.5	0.8	0	255	16
NCBI	Angiosperm	*Cuscuta australis*	Cau	95.3	90.6	4.7	2	2.7	255	9
NCBI	Angiosperm	*Cucumis melo*	Cme	96.5	90.6	5.9	2.4	1.1	255	15
NCBI	Angiosperm	*Cucumis melo var. makuwa*	Cmm	80	75.3	4.7	7.8	12.2	255	9
NCBI	Angiosperm	*Cucumis sativus*	Cst	98.4	92.5	5.9	1.2	0.4	255	15
NCBI	Angiosperm	*Momordica charantia*	Mch	98.1	91	7.1	1.2	0.7	255	15
NCBI	Angiosperm	*Carex littledalei*	Cli	92.9	80.4	12.5	1.2	5.9	255	17
CNGB	Angiosperm	*Hibbertia grossulariifolia*	Hgr	91.3	88.2	3.1	4.3	4.4	255	7
phytozome	Angiosperm	*Dioscorea alata*	Dal	98.5	92.2	6.3	1.6	0.1	255	15
NCBI	Angiosperm	*Rhododendron williamsianum*	Rwi	79.6	72.9	6.7	12.2	8.2	255	15
NCBI	Angiosperm	*Jatropha curcas*	Jcu	100	92.9	7.1	0	0	255	14
NCBI	Angiosperm	*Ricinus communis*	Rco	95.3	91	4.3	2.7	2	255	15

ANEXO D - VIGILÂNCIA GENÔMICA EM SAÚDE PÚBLICA

Col@b

Vigilância em Saúde

Interfaces entre a Saúde Pública e a Pesquisa Científica

Jean Ezequiel Limongi | Organizador

Vigilância em Saúde

Interfaces entre a Saúde Pública e a Pesquisa Científica

Atribuição - Não Comercial - Sem Derivações 4.0 Internacional

Direitos reservados à Editora Colab. É permitido download do arquivo (PDF) da obra, bem como seu compartilhamento, desde que sejam atribuídos os devidos créditos aos autores.

Não é permitida a edição/alteração de conteúdo, nem sua utilização para fins comerciais.

A responsabilidade pelos direitos autorais do conteúdo (textos, imagens e ilustrações) de cada capítulo é exclusivamente dos autores.

Autores:

Vários autores

Conselho Editorial e Responsabilidade Técnica

A Colab possui Conselho Editorial para orientação e revisão das obras, mas garante, ética e respeitosamente, a identidade e o direito autoral do material submetido à editora.

Conheça nossos Conselheiros Editorias em https://editoracolab.com/sobre-n%C3%B3s

Dados Internacionais de Catalogação na Publicação (CIP)

Vários autores.

Vigilância em Saúde [livro eletrônico]: Interfaces entre a Saúde Pública e a Pesquisa Científica

Jean Ezequiel Limongi | **Organizador**

Uberlândia, MG : Editora Colab, 2021.
3,0 MB; PDF
Bibliografia
ISBN: 978-65-86920-18-5
doi: http://dx.doi.org/10.51781/9786586920185

1. Saúde - Vigilância. 2. Pesquisa. 3. Ciência. 4. Pública. 5. Investigação

Índices para catálogo sistemático: Vigilância em Saúde

614 – Saúde Pública

APRESENTAÇÃO

O termo vigilância nos remete ao verbo vigiar, do latim vigilare, que pode ser entendido como estar atento, ter cautela, precaução, diligência, zelo, entre outros. No campo da saúde, a vigilância percorreu um longo caminho para alcançar o conceito atual de Vigilância em Saúde (VS).

A história da VS se faz desde as primeiras civilizações. Nos registros das epidemias, por exemplo, já eram coletadas informações sobre desfechos de saúde, fatores de risco e intervenções. No século XX, a concepção de que vigilância é informação para ação ganhou força, e fez da VS atribuição essencial da Saúde Pública, sendo sua gestão responsabilidade exclusiva do Estado.

No Brasil, a VS remonta ao início do século XVIII, seguindo o modelo português, com ações primordiais daquilo que hoje chamamos de vigilância sanitária, além do controle de epidemias e ações relacionadas ao saneamento. A partir do século XX, vários marcos históricos emblemáticos levaram à configuração e às práticas de VS atuais. Dentre estes, a organização de serviços federais de controle de doenças endêmicas (década de 1940), a criação do Sistema Nacional de Vigilância Epidemiológica (1975), da Secretaria Nacional de Vigilância Sanitária (1976), da Secretaria de Vigilância em Saúde (década de 1990), da Agência Nacional de Vigilância Sanitária (ANVISA) (1999), a implantação da área técnica de Vigilância em Saúde Ambiental (1999), a organização de forma descentralizada e regionalizada da Vigilância em Saúde do Trabalhador (2002) e, recentemente, a instituição da Política Nacional de Vigilância em Saúde (2018).

Durante o período de concepção dessa obra, em um contexto catastrófico da pandemia de COVID-19, a VS reforça o seu papel de destaque na saúde pública e expõe a importância da imunização, da notificação e investigação de doenças e agravos, da regulação, intervenção e atuação em condicionantes e determinantes da saúde e apresenta, para a sociedade, vertentes importantes da VS, como a vigilância genômica e a territorialização das ações de saúde.

Na presente obra, os seis capítulos iniciais apresentam revisões teóricas sobre vigilância em saúde ambiental, vigilância em saúde do trabalhador, vigilância alimentar e nutricional, vigilância genômica e a importância da medicina veterinária em ações de VS na perspectiva da Saúde Única em um mundo constantemente ameaçado por doenças zoonóticas. Estes capítulos preenchem importante lacuna existente nos livros-texto sobre estes temas e constituem material didático importante para estudantes de graduação, pós-graduação e profissionais de saúde.

Nos oito capítulos seguintes, são apresentadas pesquisas na área de promoção e vigilância de doenças não transmissíveis, vigilância alimentar e nutricional, integração da VS com a Atenção Básica, vigilância de doenças zoonóticas entre profissionais que manuseiam animais e mortalidade relacionada à COVID-19. Estas pesquisas possuem aplicabilidade direta na saúde pública, demonstrando a importância das universidades e institutos de pesquisa em promover as respostas necessárias para a construção do conhecimento no campo da saúde.

A VS é definida como um processo contínuo e sistemático de coleta, consolidação, análise de dados e disseminação de informações sobre eventos relacionados à saúde. Deste modo, é imperativo a sua integração com a pesquisa científica promovendo assim a superação da dicotomia entre os serviços e a academia.

Jean Ezequiel Limongi | **Organizador**

Como citar este trabalho:

LIMONGI, J.E. **Vigilância em Saúde: Interfaces entre a Saúde Pública e a Pesquisa Científica**. 1Ed. Uberlândia: Editora Colab, 2021. 238. p. http://dx.doi.org/10.51781/9786586920185

Sumário

Vigilância Genômica em Saúde Pública

Thiago Mendonça dos Santos
Doutorando em Bioinformática, Universidade Federal de Minas Gerais
mendonca_thiago22@hotmail.com

Alison Pelri Albuquerque Menezes
Doutorando em Genética, Universidade Federal de Minas Gerais
alisonpam@gmail.com

Wenderson Felipe Costa Rodrigues
Mestrando em Bioinformática, Universidade Federal de Minas Gerais
rod.wenderson@gmail.com

Luiz Eduardo Del Bem
Doutor em Genética e Biologia Molecular , Universidade Federal de Minas Gerais
delbem@ufmg.br

RESUMO: As práticas em vigilância em saúde vêm se tornando cada vez mais amplas e inclusivas. Ferramentas genômicas baseadas em sequenciamento de DNA como o Whole Genome Sequencing e a Metagenômica podem ser usadas na rápida detecção, monitoramento e estudo de potenciais patógenos que ameacem a saúde humana, animal ou ambiental. A determinação da sequência nucleotídica acoplada às análises computacionais de dado organismo de interesse, cultivado ou retirado diretamente de amostra ambiental ou clínica, fornece informações que ajudam a clarificar duas indagações críticas: quais os organismos presentes na amostra e o que eles fazem. Essas técnicas já são empregadas por diversos órgãos ou centros de pesquisa ao redor do mundo a fim de mensurar a diversidade de espécies, prevenir o surgimento de novos patógenos e acompanhar a evolução e transmissão de patógenos durante surtos em determinadas populações. Descreveremos a utilização dessas técnicas em áreas de interesse crítico à saúde humana e economia global, como zoonoses, vigilância de águas e esgotos, arboviroses e desenvolvimento de resistência antimicrobiana. Embora se trate de um campo relativamente recente e ainda em desenvolvimento de metodologias e protocolos, a vigilância genômica é uma prática cada vez mais adotada, tornando-se uma importante atualização ao sistema de saúde pública.

Palavras-chave: Vigilância; WGS; metagenômica.

Como citar este trabalho:
SANTOS, T.M.; MENEZES, A.P.A.; RODRIGUES, W.F.C.; DEL BEM, L.E. Vigilância Genômica em Saúde Pública. In: LIMONGI, J.E. **Vigilância em Saúde: Interfaces entre a Saúde Pública e a Pesquisa Científica**. 1Ed. Uberlândia: Editora Colab, 2021. p. 32-50.
doi: http://dx.doi.org/10.51781/97865869201853250

INTRODUÇÃO

As práticas em saúde pública e vigilância tornam-se cada vez mais amplas, inclusivas e focadas na prevenção (ROSEN, 1980; ACHESON, 1988; SOUZA, 2014; OMS, 2020). Com o surgimento e avanço das técnicas de análise genômica, tornou-se possível sua aplicação na prevenção e promoção de saúde em sentido amplo. Ferramentas genômicas podem ser usadas na detecção, monitoramento e estudo de potenciais patógenos que ameacem a saúde humana, animal ou ambiental direta ou indiretamente (DJORDJEVIC et al., 2020).

Histórico

Na Europa, com a consolidação dos estados nacionais modernos, a população passou a ser considerada um bem valioso, e como tal, deveria ser protegida e ampliada. Devido a importância de se manter uma população saudável, a própria saúde passa a ser considerada item de interesse governamental (ROSEN, 1980). Atualmente, a saúde pública é definida pela Organização Mundial da Saúde (OMS) como "a arte e ciência de prevenir doenças, prolongar a vida e promover a saúde através de esforços organizados da sociedade" (ACHESON, 1988; OMS, 2020). A saúde pública engloba todas as práticas que visem garantir saúde e bem-estar da população, incluindo a vigilância em saúde pública.

O campo de Vigilância em Saúde surgiu com o foco em doenças infecciosas e ao longo do tempo expandiu cada vez mais sua área de atuação (FIOCRUZ, 2021). Atualmente, a vigilância em saúde pública envolve a coleta, análise, interpretação e disseminação sistemática e contínua de dados de saúde para o planejamento, implementação e avaliação de ações para saúde pública (OMS, 2021). As práticas de vigilância em saúde pública têm como objetivo não só a atenção e promoção à saúde pública, mas principalmente a prevenção de doenças. Com a ampliação do que se entende por saúde pública e vigilância, bem como o avanço das tecnologias de sequenciamento, montagem e análise de genomas, tornou-se possível usar a genômica para detecção de surtos precocemente, rastreio de variantes patogênicas em uma determinada população e prevenção e combate de doenças infectocontagiosas, mesmo sem a presença de sintomas (PEACOCK et al., 2018). Esse sistema denomina-se vigilância genômica e, combinado com dados ambientais e epidemiológicos, tem sido uma ferramenta importante na vigilância em saúde.

Saúde Única (One Health)

É possível notar uma clara tendência da saúde pública em se tornar cada vez mais abrangente e inclusiva. Seguindo essa tendência, a abordagem da "Saúde Única" foi criada com base na Medicina Comparativa, uma abordagem popular no pré-modernismo que usava estudos realizados em animais para extrapolação em humanos (RYU et al., 2017). A Saúde Única é uma abordagem transdisciplinar e transversal que envolve esforços para garantir a promoção da saúde pública, não apenas do ponto de vista de saúde humana, mas também ambiental e animal, afinal, esses setores encontram-se conectados, e eventos que acontecem em uma área podem ter consequências em todas (OMS, 2017). As preocupações da Saúde Única incluem qualquer ameaça comum à saúde humana, animal e ambiental, sendo as principais a emergência de zoonoses, o surgimento de patógenos resistentes a drogas antimicrobianas e a segurança alimentar (CDC, 2018). Visto que a abordagem de Saúde Única trata a saúde de forma ampla e única, a vigilância genômica se enquadra como uma de suas possíveis práticas. A vigilância de potenciais ameaças e o planejamento precoce

de prevenção ou combate a elas são imprescindíveis para a manutenção da saúde pública em qualquer esfera.

Genômica

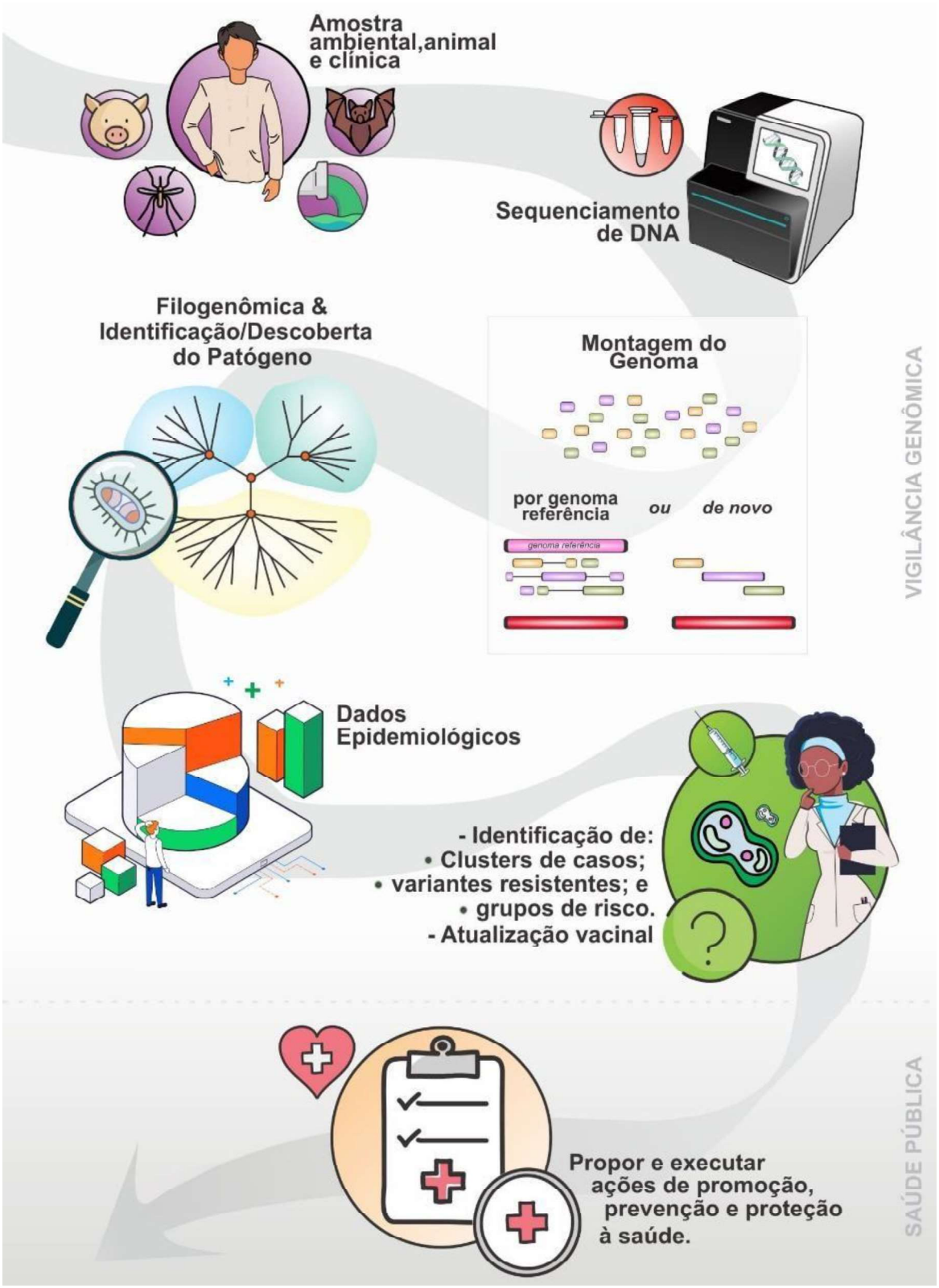
A parceria entre vigilância em saúde e genômica só foi possível graças ao desenvolvimento de tecnologias de sequenciamento de DNA. Desde a década de 1970, as técnicas de sequenciamento de trechos de DNA têm sido aplicadas em estudos genéticos. Em 1977, Sanger et al. (1977) desenvolveram um método de sequenciamento baseado em eletroforese capilar que permitiu que fossem sequenciados fragmentos completos de genes ou genomas. O surgimento de tecnologias baseadas no método de Sanger (1977) deu grande impulso aos estudos genômicos, inclusive o sequenciamento do genoma humano (INTERNATIONAL HUMAN GENOME CONSORTIUM, 2003). Porém, nas últimas décadas, o uso de sequenciamento de DNA aumentou exponencialmente graças ao surgimento de tecnologias mais rápidas e menos dispendiosas por base sequenciada. Para atender a essa demanda, surgiram os métodos de sequenciamento massivo de DNA ou Next Generation Sequencing (NGS). O primeiro método a surgir foi a técnica de sequenciamento por síntese, na qual, durante a síntese de novas moléculas de DNA, a partir do material fornecido em uma amostra, cada nova base adicionada à nova fita sintetizada libera um, sinal permitindo sua identificação e, à medida que a síntese ocorre, é possível determinar a sequência nucleotídica do material sequenciado (SCHUSTER, 2008).

Atualmente existem diversas plataformas de NGS com técnicas diferentes, mas todas elas mantêm o princípio básico que foi o grande avanço desses novos métodos em relação ao Sanger (1977), a utilização de sequenciamento por síntese a partir de clonagem in vitro de moléculas aderidas a um suporte sólido, permitindo sequenciar um número maior de moléculas ao mesmo tempo, reduzindo tempo e custos com sequenciamento, tornando a obtenção de dados genômicos muito mais acessível (CARVALHO; SILVA 2009; GRADA; WEINBRECHT 2013).

A quantidade e a qualidade da informação produzida pelas plataformas de NGS e o baixo custo relativo para o uso dessa metodologia em grandes populações é o principal atrativo da técnica. Em contrapartida, o alto custo de implementação dessas plataformas, os recursos computacionais e os conhecimentos específicos em bioinformática necessários para a manipulação desses dados são o principal limitante para sua popularização. Justamente pela baixa popularização do método, apesar de sua comprovada eficiência, os protocolos de coleta e tratamento dos dados não estão padronizados. Por se tratar de uma prática de saúde pública, é interessante que haja um esforço no sentido de criar protocolos padronizados de como tratar esse tipo de dado, de modo que a comparação entre dados seja mais fácil e otimizada.

No geral, o fluxo de trabalho da Vigilância Genômica (Figura 1), inicia com a amostra a ser analisada.

Figura 1. Fluxo de trabalho da vigilância genômica.



Fonte: Autores (2021). O fluxo descreve como amostras das mais variadas origens podem ser sequenciadas e genomas de patógenos podem ser parcial ou completamente montados, permitindo utilizar de metodologias de reconstrução filogenética para identificar ou até descobrir novos patógenos com base em dados de sequência. A união destes com dados epidemiológicos pode orientar decisões diversas em saúde pública.

A depender do objetivo, a amostra pode proceder da vigilância contínua a partir de amostras ambientais como esgoto, amostras biológicas de animais selvagens ou domésticos, artrópodes hematófagos, amostras clínicas humanas, dentre outras. O DNA dessas amostras deve ser sequenciado e computacionalmente montado com auxílio de genomas de referência ou de novo (termo em Latim que designa algo que começa a partir do zero) e, a partir das sequências montadas, relações filogenômicas são estabelecidas no intuito de identificar os patógenos ou inferir relações entre diferentes variantes de um mesmo tipo. Ao final, essas informações são associadas à vigilância epidemiológica e geram resultados indiretos que podem ser utilizados na promoção de políticas públicas em saúde. Esses passos serão tratados com mais detalhes nos próximos tópicos.

A aplicação da vigilância genômica rotineira tem como principais resultados a detecção, monitoramento e estudo de patógenos potenciais e já conhecidos que ameacem a saúde humana, animal ou ambiental direta ou indiretamente (DJORDJEVIC et al., 2020). Entretanto, algumas consequências indiretas advindas da coleção de dados genômicos obtidos também estão atreladas a esse processo. Uma delas é a identificação de sequências codificantes virais alvo para atualização vacinal. Dependendo do patógeno alvo de uma vacina, sua taxa de substituição de aminoácidos em proteínas pode exigir uma constante atualização do material imunizante da vacina, seja proteína ou mRNA. Estas sequências podem também ser alvo de modelagem computacional das proteínas virais, permitindo inferir propriedades imunogênicas e de interação com proteínas humanas em geral, e proteínas receptoras virais em particular. Outros resultados indiretos derivados da vigilância incluem a identificação de cepas bacterianas resistentes à antibióticos e identificação de clusters de casos causados por cada variante. Ainda, análises conjuntas com dados epidemiológicos podem ajudar na identificação de grupos de risco associados à determinada variante.

MÉTODOS GENÔMICOS E ANÁLISES EMPREGADAS NA VIGILÂNCIA GENÔMICA

Whole Genome Sequencing (WGS)

O advento do NGS abriu caminho para a aplicação de métodos de sequenciamento de DNA em vigilância genômica como o Whole Genome Sequencing (WGS) ou Sequenciamento de Genoma Completo, em português. WGS é o processo de determinação de toda a sequência de DNA do genoma de um organismo, em um único experimento, utilizando-se de tecnologias de NGS. Uma grande vantagem dessa técnica é a geração de genomas completos, com maior resolução, e mais precisos do que por meio de outras técnicas, que não miram o genoma completo (GILCHRIST et al., 2015). A obtenção do material genético a ser utilizado em WGS pode ser derivado de qualquer célula ou microrganismo previamente isolado em cultura. No campo

da vigilância, o material utilizado visa identificar bactérias, vírus, fungos e outros patógenos, podendo ser estes previamente conhecidos ou não. Há vários exemplos do uso de WGS para a vigilância genômica e seu uso em rotinas públicas de vigilância sanitária (FORD et al., 2018). WGS pode ser utilizado durante um surto de alguma doença desconhecida, fornecendo informações sobre o patógeno causador (como seus mecanismos de virulência), revelando o caminho da transmissão da doença dentro de uma população e identificando a fonte e origem do surto através do compartilhamento de informações entre pesquisadores (GILCHRIST et al., 2015).

WGS gera uma grande quantidade de dados brutos a serem analisados, o que requer análises de bioinformática complexas a fim de extrair informações relevantes para a vigilância genômica. Dado que o DNA é fragmentado em segmentos pequenos (tipicamente < 500 nt) antes do sequenciamento, há a necessidade de montar as sequências (reads) resultantes para obtenção do genoma completo, como um quebra-cabeça, o que torna esse passo extremamente complexo. Uma primeira abordagem na montagem do genoma sequenciado por WGS é a comparação, através de alinhamento, das reads obtidas com sequências referência, ou seja, sequências de genomas extremamente bem caracterizadas depositadas em banco de dados de genomas, como o National Center for Biotechnology Information (NCBI). É um método mais rápido e computacionalmente simples, que permite uma anotação mais precisa dos genes para casos de genomas recém sequenciados (NARZISI; MISHRA, 2011). Essa técnica foi utilizada na identificação do SARS-CoV-2 durante o início do surto de COVID-19 (ZHU et al., 2019). Entretanto, há um risco da montagem ou anotação não ocorrer de forma correta caso o genoma sequenciado e o de referência sejam muito divergentes entre si ou apresentem quebras de sintenia ou variações estruturais, perdendo, assim, informação importante acerca do organismo em análise (GILCHRIST et al., 2015). A segunda abordagem é realizada através do sequenciamento de novo que permite a montagem do genoma de organismos desconhecidos a partir de reads curtos não se utilizando de genomas de referência, sendo, por isso, mais desafiadora computacionalmente. Apesar disso, novas tecnologias desenvolvidas como o sequenciador Oxford Nanopore, que permite o sequenciamento de reads de DNA mais longas, facilitam a montagem do genoma de novo por WGS (KOREN; PHILLIPPY, 2015).

Metagenômica

O Sequenciamento de genomas de espécies isoladas vem agregando muitas informações novas para a ciência, entretanto, essa abordagem molecular possui algumas limitações. A primeira limitação é a perda da informação ambiental na qual o organismo estava inserido, dado que se deve isolá-lo, não levando em consideração sua interação com outros microrganismos. A segunda limitação é a necessidade de cultivo do microrganismo previamente ao sequenciamento, o que demanda mais tempo, além disso, apenas uma

pequena fração de microrganismos é cultivável.

O desenvolvimento da metagenômica, ciência que estuda o metagenoma, permitiu a superação dessas limitações. O metagenoma é definido como o conjunto de todas as informações genômicas extraídas diretamente do ambiente em estudo, sem a necessidade de cultivo prévio. Devido a esse fato, a metagenômica é capaz de evidenciar a relação entre os microrganismos sequenciados, levando em consideração suas comunidades e habitats. Consegue também revelar os microrganismos que normalmente não seriam observados em técnicas que necessitam de cultivo prévio (GARRIDO-CARDENAS; MANZANO-AGUGLIARO, 2017). Todo (ou quase todo) o DNA da amostra será extraído e sequenciado, o que inclui toda e qualquer espécie presente. As amostras utilizadas podem ser extraídas de diversos ambientes, como solos, rios, mar e até mesmo amostras do ar (ALTEIO et al., 2020; BEHZAD; GOJOBORI; MINETA, 2015; COUTINHO et al., 2017; PINTO et al., 2020). Encontra-se crescente também o ramo da metagenômica clínica, que consiste na análise, a partir de amostras do paciente, de material genético de parasita e hospedeiro, com consequente correlação clínica na saúde do paciente (CHIU; MILLER, 2019).

O primeiro passo na realização da metagenômica é a obtenção das amostras, as quais devem ser representativas do ambiente em estudo. Nessa etapa, também são recolhidos dados adicionais, chamados de metadados, que descrevem as amostras recolhidas, como local, cor, temperatura, pH, etc. (KUNIN et al., 2008). Após a extração de DNA das amostras, o sequenciamento de DNA poderá dar-se por duas maneiras: sequenciamento de amplicons (produtos de PCR) de marcadores de DNA (por exemplo, rDNA ou ITS) ou shotgun. A técnica de shotgun envolve a fragmentação do DNA extraído e consequentemente sequenciamento por completo dos fragmentos. Essa característica adiciona uma dificuldade a mais na montagem dos genomas sequenciados, já que não se sabe a diversidade de espécies presentes na amostra (WOOLEY et al., 2010). Já a técnica de sequenciamento de amplicons, como rDNA, envolve o sequenciamento somente de regiões definidas como 16S rDNA, frequentemente usados para procariotos e 18S, para eucariotos, e permite estimativa de diversidade de espécies na amostra. Apesar de ser mais caro, o método de shotgun possui uma resolução maior do que marcadores genômicos isolados, ou seja, pode fornecer informações taxonômicas mais precisas (nível de espécies), informações funcionais como os genes presentes, e consegue identificar organismos que não possuem genes tipicamente usados como alvo de amplificação como rDNAs, como vírus (WOOLEY et al., 2010). O processamento de dados montados de metagenomas gera agrupamentos de indivíduos altamente relacionados chamados de OTUs (Operational Taxonomic Unit). As OTUs são agrupadas por algoritmos de acordo com a similaridade na sequência de DNA e potencialmente representam, em diferentes níveis taxonômicos, as espécies ou populações que foram sequenciadas (KUNIN et al., 2008).

Especialmente em se tratando de viroma (metagenoma de vírus), algumas dificuldades associadas

ao processo experimental devem ser mencionadas. Em primeiro lugar, como a proporção de DNA viral é tipicamente menor do que a do hospedeiro ou de outros microrganismos, o preparo da amostra pode exigir alguns passos a mais como amplificação do DNA viral por PCR, filtragem do material genético por concentração ou tamanho molecular através de centrifugação ou, ainda, alguns ajustes durante a análise de bioinformática, como a eliminação de sequências de organismos abundantes já conhecidos. Sequenciamento de bibliotecas de RNAs pequenos também podem ser utilizada para montagem de genomas virais de RNA. Em segundo lugar, a grande quantidade de dados gerados requer extensivo trabalho computacional especializado e esbarra na falta de dados completos de viromas ambientais (ROSE et al., 2016).

Filogenômica

A quantidade de novos dados genômicos gerados permitiu o aprimoramento da filogenômica, a ciência que estuda in silico a reconstrução da história evolutiva das espécies através de comparações entre os genomas completos, uma derivação da filogenética. A reconstrução filogenômica baseia-se no uso de alinhamento múltiplo de genomas ou de grandes conjuntos concatenados de sequências proteicas, seguido pela inferência de árvores filogenômicas com auxílio de algoritmos implementados em softwares como MEGA ou PhyML que agrupam os genomas por similaridade (CHAN; RAGAN, 2013). Essa ciência é ferramenta essencial para vigilância genômica. Através dela, pode-se inferir relações evolutivas entre variantes de um patógeno responsável por um surto, inferir a provável origem desse patógeno, clados de variantes similares e identificação de linhagens emergentes.

A filogenômica pode ser utilizada a fim de acompanhar a evolução molecular de determinado patógeno, o que é extremamente útil no caso de vírus e bactérias, correlacionando o acúmulo de mutações e a evolução da virulência. Essa abordagem envolve o mapeamento de mutações em árvores filogenômicas de patógenos amostrados durante, ou entre surtos de doenças infecciosas em animais reservatórios e em novos hospedeiros (GEOGHEGAN; HOLMES, 2018). A identificação da fixação de mutações ocorrendo em paralelo dentro de árvores filogenômicas torna possível inferir pressões seletivas agindo no universo de mutações. Como exemplo, o surgimento repetido da mesma mutação seguido pela transmissão entre espécies do vírus da influenza aviária para humanos sugere que essa mutação afeta diretamente o leque de hospedeiros do vírus (TAUBENBERGER et al., 2005).

ÁREAS DE APLICAÇÃO DA VIGILÂNCIA GENÔMICA

Resistência antimicrobiana

Uma crescente preocupação mundial é o surgimento cada vez mais frequente de bactérias resistentes a antibióticos. Esse fenômeno leva a dezenas de milhares de mortes por ano mundialmente e é previsto um custo econômico indireto de US$ 1–3,4 trilhões de dólares no mundo em termos de morbidade, incapacidade, mortes prematuras e trabalho reduzido até 2030, se o aumento de bactérias resistentes a antibióticos não for contido (O'NEILL, 2016; WORLD BANK, 2017).

Dados de resistência antimicrobiana são essenciais a fim de informar o governo para elaboração de políticas públicas e adoção de medidas preventivas. As análises de rotina, baseadas em dados fenotípicos, para detecção de bactérias resistentes a antibióticos envolvem técnicas laboratoriais de difusão e diluição de antibióticos em discos de papel em culturas de bactérias em placas de ágar. Entretanto, as limitações dessa prática incluem dificuldades nas interpretações da Concentração Inibitória Mínima (CIM), manutenção da temperatura de cultivo, pH e condições atmosféricas e concentrações de íons no meio de cultura (WORLD HEALTH ORGANIZATION (WHO), 2020). O método de difusão pode não ser adequado para alguns antibióticos, como a colistina, por ser difícil a cultura de bactérias de crescimento lento e fastidioso. Além disso há a dificuldade de cultivo de bactérias anaeróbias ou de espécies raras, para as quais meios de cultura convencionais podem não ser efetivos (KHAN et al., 2019).

A abordagem com WGS, além de não sofrer com as desvantagens citadas para o método fenotípico, permite a identificação do conjunto de genes que conferem resistência aos microrganismos com base em comparações de sequência com bancos de dados de genes de resistência já anotados como o Comprehensive Antibiotic Resistance Database - Resistance Gene Identifier (CARD-RGI) (MCARTHUR et al., 2013). Essa prática foi corroborada com experimentos de validação em que verificou-se que há uma concordância de mais de 96% entre a presença de genes de resistência e a CIM de diversos antibióticos testados em enterobactérias (HENDRIKSEN et al., 2019). Amostras clínicas de pacientes, amostragem em ambientes hospitalares, na indústria de alimentos e agropecuária, podem ser utilizadas na vigilância genômica para resistência antimicrobiana. Ademais, esse método permite a detecção de microrganismos resistentes à multidrogas. O que antes era feito por testes a partir de três drogas, agora pode ser feito analisando vários conjuntos de genes ao mesmo tempo, revelando diferentes padrões de resistência a multidrogas e indicando a possibilidade de transferência gênica (BALLOUX et al., 2018; WORLD HEALTH ORGANIZATION (WHO), 2020).

Diversos órgãos públicos vêm adotando vigilância genômica como prática, ao menos complementar, aos testes microbiológicos em alguns países. Nos Estados Unidos, o sistema nacional de monitoramento de resistência antimicrobiana (National Antimicrobial Resistance Monitoring System) foca no acompanhamento de resistência microbiana em bactérias entéricas como a Salmonella. No Brasil, o Programa de Vigilância e Monitoramento da Resistência aos Antimicrobianos no Âmbito da Agropecuária (PAN-BR AGRO) faz uso de

WGS de forma complementar no monitoramento de resistência antimicrobiana no setor agropecuário, tendo possíveis consequências à saúde humana, dentro de uma abordagem de Saúde Única.

Doenças transmitidas pela água

A vigilância genômica de águas e esgotos constitui-se um caso interessante da aplicação da metagenômica, objetivando-se a prevenção e auxílio no combate a doenças transmitidas por água contaminada com patógenos. Estudos com a adoção desse tipo de vigilância acontecem em empresas de saneamento, indústrias da agropecuária, alimentícia e de aquacultura, que visam a busca de microrganismos em tendência de crescimento e com potencial patogênico (NIEUWENHUIJSE; KOOPMANS, 2017). Diversos patógenos podem se transmitir pela água, como a Entamoeba histolytica, causadora da amebíase, Vibrio cholerae, causador da cólera, Rotavírus, que causa gastroenterite, os vírus das hepatites A e E, dentre muitos outros. Entretanto, a metagenômica na vigilância de águas é mais frequentemente utilizada para estudos de viroma, por ser a técnica mais adequada a este fim.

Muitos patógenos de transmissão fecal-oral são encontrados nesse tipo de ambiente devido a presença de resíduos humanos, podendo ser monitorados simultaneamente através de metagenômica, em uma única amostragem. Esse tipo de monitoramento permite o acompanhamento dos níveis de contaminação através da quantidade de material genético observada de determinado patógeno e a determinação e rastreamento dos locais onde a infecção é presente, constituindo-se uma fonte de dados epidemiológicos importantes para análise de novos vírus, vírus emergentes e outros patógenos conhecidos que possam ser excretados pela população local (FERNANDEZ-CASSI et al., 2018). Como exemplo, na Suécia, norovírus (causadores de diarreias), puderam ser relacionados a pacientes hospitalizados diagnosticados com a infecção viral nos arredores da coleta amostral. Interessantemente, foi observado um pico na abundância de material genético de norovírus semanas antes de um surto em hospitais nesse país (HELLMÉR et al., 2014).

Indo além dos vírus predominantemente fecais-orais, esgotos podem ser a base do monitoramento de outros patógenos como o SARS-CoV-2 que, apesar de ser um vírus respiratório, apresenta manifestações gastrointestinais nos infectados. Diversos órgãos públicos e grupos de pesquisas no Brasil e em outros países utilizaram águas de esgotos a fim de monitorar os níveis e abrangência de infecções por SARS-CoV-2 durante a pandemia de COVID-19 (FONGARO et al., 2020; MEDEMA et al., 2020; PECCIA et al., 2020).

Doenças zoonóticas

O surgimento dos vírus HIV, SARS-CoV, MERS, H1N1, e mais recentemente de SARS-CoV-2 tiveram pesadas consequências humanas e econômicas ao redor do mundo. A emergência desses novos patógenos

pode parecer imprevisível, entretanto há um padrão de surgimento nesses casos, todos tem origem zoonótica confirmada ou provável. Mais de 60% de doenças infecciosas novas identificadas no mundo desde 1940 tem origem zoonótica (JONES et al., 2008). Animais (selvagens ou domésticos) em lugares específicos de interfaces de contato com humanos, como fazendas, fronteiras ambientais e mercados de animais representam alvos importantes para a vigilância genômica. Para mais, essas novas infecções zoonóticas são majoritariamente dirigidas pela influência humana por meio de mudanças ecológicas, comportamentais ou socioeconômicas, portanto, possuem tendência de piora num cenário de mudanças climáticas e subdesenvolvimento social (MORSE et al., 2012).

Como o reservatório natural de alguns vírus são os animais, a vigilância genômica em zoonoses deve focar em animais conhecidos por serem carreadores de determinados vírus, principalmente os que já estiveram envolvidos em surtos anteriores, como porcos, aves, morcegos, roedores e macacos, principalmente em áreas tropicais e de alta biodiversidade. A vigilância genômica rotineira de amostras biológicas desses animais pode ser uma prática adotada para antever o surgimento de novas infecções ou realizar o controle de doenças conhecidas. Múltiplos estudos foram realizados com intuito de avaliar a capacidade de monitoramento em áreas sensíveis para surgimento de surtos zoonóticos. Macacos do gênero Alouatta tiveram amostras sequenciadas através de WGS em uma abordagem espaço-temporal a fim de monitorar a distribuição do vírus da febre amarela (YFV) em seu ciclo silvestre, durante epidemia ocorrida entre 2016 e 2018 no estado de São Paulo, no Brasil. Esse estudo permitiu a montagem de vários genomas de YFV, identificou as fases epidêmicas e sugeriu um lugar de onde se originou a epidemia (HILL et al., 2020).

Outro exemplo é a infecção pelo rotavírus, a principal causa de gastroenterite humana. É comum a morte por esse vírus em países em desenvolvimento na Ásia e na África (TATE et al., 2012). Além de infectar humanos, o rotavírus também infecta outros animais, como o porco, e pode ser considerada uma doença zoonótica. Devido a esse fato, várias tentativas foram feitas a fim de acompanhar a reprodução desse vírus em espécies não humanas, dado que fora observada transmissão direta de variantes de rotavírus de animais para humanos (GHOSH; KOBAYASHI, 2014). No Vietnã, amostras de fezes de porcos de áreas periurbanas foram utilizadas em uma abordagem WGS para detecção de variantes de rotavírus e, através de análises filogenéticas, descobriu-se que cocirculavam múltiplas variantes do rotavírus, muitas das quais não eram cobertas pela vacina disponível (PHAN et al., 2016).

Arboviroses

Uma crescente preocupação de saúde pública, especialmente em países tropicais como o Brasil, são as arboviroses, nome dado às doenças virais transmitidas por meio de artrópodes hematófagos, como

mosquitos e carrapatos. A maioria dessas doenças, com exceção da Febre amarela e da Dengue, não possuem vacina disponível e, por décadas, o controle principal se baseou no uso de inseticidas, instalação de barreiras físicas (telas e mosquiteiros) e controle de lugares de reprodução. As principais arboviroses são a Dengue, Zika, Chikungunya, doença do Oeste do Nilo e a Febre Amarela, responsáveis por milhares de mortes anualmente ao redor do mundo. O desenvolvimento de resistência aos inseticidas, surgimento de novas doenças e o aumento da área de propagação desses artrópodes devido às mudanças climáticas justificam a vigilância genômica desses invertebrados e suas populações de vírus (LONDONO-RENTERIA; TROUPIN, 2016).

A primeira abordagem para o monitoramento de arboviroses envolve o acompanhamento das infecções em populações sentinelas humanas, ou seja, um grupo de indivíduos aleatórios de determinada população que servirá de amostragem para a população como um todo. Geralmente utiliza-se sangue, urina ou saliva, somadas a metadados como sexo, idade e sintomas. Essas amostras são utilizadas em WGS ou metagenômica para busca do material genético dos arbovírus. No Brasil, esse método foi aplicado para reconstrução genômica em escala de tempo a fim de avaliar a distribuição do vírus da Zika nas Américas, confirmando que o surto se originou no Brasil e identificando que a transmissão desse vírus provavelmente permaneceu indetectada por mais de um ano antes do primeiro caso reportado (FARIA et al., 2017). Também a partir de amostras clínicas de pacientes, o vírus da Chikungunya foi sequenciado com a finalidade de esclarecer a dinâmica de transmissão no Rio de Janeiro, Brasil, revelando que a linhagem circulante foi introduzida no estado ao menos 5 meses antes do primeiro caso (XAVIER et al., 2019). Esses estudos também ilustram uma limitação na vigilância genômica de determinada doença antes dela ser reportada, demonstrando a necessidade de uma vigilância genômica constante.

A segunda abordagem envolve a captura em campo dos artrópodes vetores, acompanhado de análise genômica viral por meio de metagenômica ou do isolamento viral em cultura de células, seguido de WGS. Um estudo australiano foi capaz de identificar um vírus da família Reoviridae por meio de metagenômica, descrito anteriormente somente na China, presente em um mosquito do gênero Aedes capturado em campo, além da identificação de diversos outros vírus nunca descritos anteriormente (COFFEY et al., 2019). Mosquitos capturados são utilizados também em uma metodologia chamada xenovigilância, que consiste na vigilância indireta de vírus que infectam humanos (mas que não se reproduzem em mosquitos), e que foram ingeridos por esses animais, refletindo a ideia de mosquitos como "seringas biológicas voadoras". Estudos conseguem aferir a distribuição, prevalência e reprodução de determinados vírus em uma população a partir da xenovigilância, sendo capazes de reportar, por exemplo, a presença de herpesvirus, vírus da varíola e papilomavirus. Vírus epiteliais são especialmente encontrados por xenovigilância em mosquitos, já que os vírus são transferidos da pele humana para o intestino do animal durante sua alimentação (BRINKMANN et al., 2016).

PERSPECTIVAS

O rápido avanço das tecnologias de sequenciamento de DNA somado à urgente necessidade de prever o surgimento de novas doenças com potencial epidêmico ou pandêmico geram um ambiente favorável à criação e desenvolvimento de novas metodologias, tecnologias e protocolos para utilização na vigilância genômica. Além disso, atualmente, grande parte dos esforços realizados com fins de monitoramento genômico partem das instituições de pesquisa e ensino, porém, é esperado que governos locais ou centrais adotem as metodologias de vigilância genômica como práticas rotineiras.

O desenvolvimento de sequenciadores portáteis, como o Oxford Nanopore, abre caminho para a criação de tecnologias de captura de amostra automatizadas, que podem ser instaladas em locais de interesse como saguões de aeroportos, centros de grandes cidades e hospitais a fim de realizar o monitoramento genômico de forma remota, conectadas à internet das coisas, disponibilizando informação genômica em bancos de dados públicos em tempo real, acrescido de vários metadados, complementando a informação. Tecnologias portáteis e conectadas à smartphones permitiriam também a realização de testes in loco durante potenciais surtos em lugares remotos a partir de centros de testagem provisórios montados para esse fim, como tendas laboratoriais ou hospitais de campanha. Esse tipo de tecnologia poderia ser aplicada também no monitoramento da vida selvagem e do meio ambiente, em uma abordagem Saúde Única, visando a identificação de possíveis vazamentos zoonóticos, com a possibilidade de testagem local e sem a necessidade de transporte de equipamentos pesados para o ambiente natural (GARDY; LOMAN, 2018).

Durante a epidemia de Zika em 2015 no Brasil, pesquisadores utilizaram o HealthMap (sistema de monitoramento eletrônico que reúne informações de surtos ao redor do mundo) conjuntamente com o Google Trends para estimação da reprodução do vírus. Verificaram através de estimativas filogenômicas que os intervalos calculados por esta abordagem eram semelhantes aos obtidos com dados genômicos, demonstrando que outros tipos de dados podem ser conjuntamente utilizados no cálculo de parâmetros epidemiológicos (MAJUMDER et al., 2016). Isso foi feito durante a epidemia de Chikungunya em 2018, quando o número de buscas sobre a doença no Google foi positivamente correlacionado com os casos confirmados laboratorialmente dessa doença em Roraima, Brasil (NAVECA et al., 2019). Essa é uma abordagem recente, que vem se fundindo em conjunto com a vigilância genômica, e que pode se tornar bastante útil no futuro: a epidemiologia digital. Ela consiste no emprego de tecnologias digitais como celulares e computadores na forma de aplicativos, buscadores online e redes sociais, associados a dados genômicos, objetivando o monitoramento de doenças. Outras informações, como alunos faltantes em escolas ou faculdades, faltas no trabalho, compras na farmácia, usos de aplicativos de auxílio médico, crescente uso de determinadas palavras

em mídias sociais como Twitter ou Instagram que indicassem algum potencial surto poderiam direcionar o acionamento de testagem por metagenômica em determinada população (GARDY; LOMAN, 2018). A implementação de tais tecnologias dependerá não só de seu desenvolvimento técnico, mas também da percepção de importância pela iniciativa privada e poder público.

QUESTÕES ÉTICAS

Uma consequência e ao mesmo tempo um requisito de estudos de vigilância genômica é a construção de bancos de dados de informação genômica de uma grande quantidade de organismos ou indivíduos. Entretanto, esse acúmulo de informações também gera preocupação quanto ao seu uso, principalmente de dados humanos. A Declaração de Helsinque diz que "toda precaução deve ser tomada para proteger a privacidade dos sujeitos da pesquisa e a confidencialidade de suas informações pessoais". Nesse ponto é importante, então, chamar atenção para a preocupação com a proteção de dados identificáveis dos indivíduos envolvidos em estudos de vigilância genômica.

O desenvolvimento tecnológico tornou possível construir bases de dados forenses que combinam informação genética de indivíduos e outros identificadores biométricos - impressão digital e reconhecimento facial, por exemplo (WALLACE et al., 2014). Nas últimas décadas, tem crescido um importante debate entre sociedade civil, pesquisadores e organizações políticas de diferentes países sobre o crescimento dessas bases de dados e a aplicação de conceitos da vigilância genômica em humanos (WALLACE et al., 2014). Tais debates têm influenciado a criação de protocolos e legislações que regulamentem e guiem a criação e utilização de bases forenses de dados genéticos (MOREAU, 2019; FORENSIC GENETICS POLICY INITIATIVE, 2017). Isso gera questões sobre a segurança e confiabilidade dessas bases de dados, as implicações sociais para os indivíduos cujos dados estão armazenados e as bases éticas que permeiam a construção e manutenção de bases de dados como essas (WALLACE et al., 2014). Elas também podem ser usadas de forma abusiva por governos para reforçar políticas discriminatórias e autoritárias (MOREAU, 2019). Como exemplo, o governo chinês está criando uma base de dados genéticos nacional, que em 2019 já continha DNA (coletados compulsoriamente) de cerca de 17 milhões de pessoas (DIRKS; LEIBOLD, 2020), envolvendo indivíduos pertencentes a minorias étnicas, como a população tibetana e a população muçulmana de Xinjiang (MOREAU, 2019).

Já a área de diagnósticos baseados em sequenciamento, como os originados da metagenômica clínica, ainda se encontra na transição da pesquisa científica e a aplicação clínica. Por se tratar do cuidado às vidas, sejam elas humanas ou animais, é importante a precisão do diagnóstico. A diferenciação entre patógenos e comensais, a sensitividade e especificidade do teste, a reprodutibilidade, o custo da tecnologia genômica e

a correlação entre a presença do patógeno e a doença são os critérios adotados para afirmação de um teste genômico. O aumento astronômico do uso de NGS em surtos, como o que ocorreu na pandemia de COVID-19, e seus desdobramentos deverão gerar questões sobre a ética do monitoramento e divulgação de dados genéticos. Por isso deve-se antecipar as discussões sobre as possíveis consequências da utilização de abordagens não validadas, em teste ou recém implementadas.

Os aspectos sociais e civilizatórios também devem ser considerados durante a introdução de uma nova tecnologia, não deixando-se levar somente pelo entusiasmo científico. A vigilância genômico-digital, por exemplo, carrega pesadas implicações éticas em procedimentos como o uso público de informações pessoais de redes sociais e outras informações como localização pessoal e monitoramento de deslocamento. Identificar alguém como "super carreador" de determinado patógeno com base nessas informações pode ter tanto consequências positivas para o manejo de surtos em termos de isolamento como consequências negativas, como o estigma social potencial para esse indivíduo (GARDY; LOMAN, 2018).

As ciências genômicas têm evoluído rapidamente e, infelizmente, o debate público não tem acompanhado esse passo. Certamente há muito o que se aproveitar dessa área de pesquisa para benefício da sociedade. Entretanto é importante que os pesquisadores envolvidos, a população e os governos interessados debatam aberta e claramente esse tópico a fim de que essa tecnologia seja usada de forma responsável. Esse debate já começou e, mesmo que incipiente, tem influenciado a tomada de decisões em alguns países. Todavia, é interessante ao bem comum que essas discussões sejam elevadas a nível internacional, com a possível criação de agências reguladoras que construam protocolos e diretrizes a serem seguidas pelas instituições, a fim de ter o melhor uso das tecnologias de sequenciamento e suas bases de dados.

AGRADECIMENTOS

Este trabalho foi apoiado pelo Instituto Serrapilheira (projeto Serra-1812-26691).

REFERÊNCIAS

ACHESON, Donald. **Public Health in England**. Department of Health, London. 1988

ALTEIO, L. V. et al. Complementary Metagenomic Approaches Improve Reconstruction of Microbial Diversity in a Forest Soil. **mSystems**, v. 5, n. 2, 2020.

BALLOUX, F. et al. From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. **Trends in Microbiology**, v. 26, n. 12, p. 1035–1048, 2018. Disponível em: <https://doi.org/10.1016/j.tim.2018.08.004>.

BEHZAD, H.; GOJOBORI, T.; MINETA, K. Challenges and opportunities of airborne metagenomics. **Genome Biology and Evolution**, v. 7, n. 5, p. 1216–1226, 2015.
BRINKMANN, Annika; NITSCHE, Andreas; KOHL, Claudia. Viral metagenomics on blood-feeding arthropods as a tool for human disease surveillance. **International journal of molecular sciences**, v. 17, n. 10, p. 1743, 2016.

CARVALHO, Mayra C. C. G.; SILVA, Danielle C. G. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciência Rural**, v.40, n.3, p.735-744, 2010.

CHAN, C. X.; RAGAN, M. A. Next-generation phylogenomics. **Biology Direct**, v. 8, n. 1, p. 1–6, 2013.

CHIU, C. Y.; MILLER, S. A. Clinical metagenomics. **Nature Reviews Genetics**, v. 20, n. 6, p. 341–355, 2019. Disponível em: <http://dx.doi.org/10.1038/s41576-019-0113-7>.

COFFEY, Lark L. et al. Enhanced arbovirus surveillance with deep sequencing: Identification of novel rhabdoviruses and bunyaviruses in Australian mosquitoes. **Virology**, v. 448, p. 146-158, 2014.

COUTINHO, F. H. et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. **Nature Communications**. [S.l: s.n.]. , 2017

DIRKS, Emile; LEIBOLD, James. Genomic surveillance: Inside China's DNA dragnet. Australian Strategic Policy Institute. Disponivel em: <https://www.aspi.org.au/report/genomic-surveillance>. Acesso em: 28 de mai. de 2021.

DJORDJEVIC, Steven P. et al. Genomic Surveillance for One Health Antimicrobial Resistance: Understanding Human, Animal, and Environmental Reservoirs and Transmission. In: MANAIA, C., Donner E., Vaz-Moreira I., Hong P. (org.) **Antibiotic Resistance in the Environment. The Handbook of Environmental Chemistry**, Cidade: Springer, Cham. 2020. p. 71-100. Disponível em: <https://link.springer.com/chapter/10.1007/698_2020_626>. Acesso em: 18 de jun. de 2021

FARIA, Nuno R. et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. **Nature**, v. 546, n. 7658, p. 406-410, 2017.

FERNANDEZ-CASSI, X. et al. Metagenomics for the study of viruses in urban sewage as a tool for public health surveillance. **Science of the Total Environment**, v. 618, p. 870–880, 2018.

FONGARO, G. et al. SARS-CoV-2 in human sewage in Santa Catalina, Brazil, November 2019. **Science of The Total Environment**, n. January, 2020.

FORD, L. et al. Incorporating Whole-Genome Sequencing into Public Health Surveillance: Lessons from Prospective Sequencing of Salmonella Typhimurium in Australia. **Foodborne Pathogens and Disease**, v. 15, n. 3, p. 161–167, 2018.

FORENSIC GENETICS POLICY INITIATIVE. **Establishing Best Practice for Forensic DNA Databases**. Forensic Genetics Policy Initiative. Disponível em: <http://dnapolicyinitiative.org/report>. Acesso em: 19 de jun. de 2021.

GARDY, Jennifer L.; LOMAN, Nicholas J. Towards a genomics-informed, real-time, global pathogen surveillance system. **Nature Reviews Genetics**, v. 19, n. 1, p. 9, 2018.
GARRIDO-CARDENAS, J. A.; MANZANO-AGUGLIARO, F. The metagenomics worldwide research. **Current Genetics**, v. 63, n. 5, p. 819–829, 2017.

GEOGHEGAN, J. L.; HOLMES, E. C. The phylogenomics of evolving virus virulence. **Nature Reviews Genetics**, v. 19, n. 12, p. 756–769, 2018. Disponível em: <http://dx.doi.org/10.1038/s41576-018-0055-5>.

GHOSH, Souvik; KOBAYASHI, Nobumichi. Exotic rotaviruses in animals and rotaviruses in exotic animals. **Virusdisease**, v. 25, n. 2, p. 158-172, 2014.

GILCHRIST, C. A. et al. Whole-genome sequencing in outbreak analysis. **Clinical Microbiology Reviews**, v. 28, n. 3, p. 541–563, 2015.

GRADA, Ayman; WEINBRECHT, Kate. Next-Generation Sequencing: Methodology and Application. **Journal of Investigative Dermatology**, v.133, n.8, e.11, 2013.

HELLMÉR, Maria et al. Detection of pathogenic viruses in sewage provided early warnings of hepatitis A virus and norovirus outbreaks. **Applied and environmental microbiology**, v. 80, n. 21, p. 6771-6781, 2014.

HENDRIKSEN, R. S. et al. Using Genomics to Track Global Antimicrobial Resistance. **Frontiers in Public Health**, v. 7, 2019.

HILL, Sarah C. et al. Genomic surveillance of yellow fever virus epizootic in São Paulo, Brazil, 2016–2018. **PLoS pathogens**, v. 16, n. 8, p. e1008699, 2020.

INTERNATIONAL HUMAN CONSORTIUM. Initial sequencing and analysis of the human genome. **Nature**. v.409, p.860-921. 2001.

JONES, Kate E. et al. Global trends in emerging infectious diseases. **Nature**, v. 451, n. 7181, p. 990-993, 2008.
KHAN, Zeeshan A.; SIDDIQUI, Mohd F.; PARK, Seungkyung. Current and emerging methods of antibiotic susceptibility testing. **Diagnostics**, v. 9, n. 2, p. 49, 2019.

KOREN, S.; PHILLIPPY, A. M. One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. **Current Opinion in Microbiology**, v. 23, p. 110–120, 2015. Disponível em: <http://dx.doi.org/10.1016/j.mib.2014.11.014>.

KUNIN, V. et al. A Bioinformatician's Guide to Metagenomics. **Microbiology and Molecular Biology Reviews**, v. 72, n. 4, p. 557–578, 2008.

LONDONO-RENTERIA, Berlin; TROUPIN, Andrea; COLPITTS, Tonya M. Arbovirosis and potential transmission blocking vaccines. **Parasites & vectors**, v. 9, n. 1, p. 1-11, 2016.

MAJUMDER, Maimuna S. et al. Utilizing nontraditional data sources for near real-time estimation of transmission dynamics during the 2015-2016 Colombian Zika virus disease outbreak. **JMIR public health and surveillance**, v. 2, n. 1, p. e30, 2016.

MCARTHUR, Andrew G. et al. The comprehensive antibiotic resistance database. **Antimicrobial agents and chemotherapy**, v. 57, n. 7, p. 3348-3357, 2013.

MEDEMA, G. et al. Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in the Netherlands. **Environmental Science and Technology Letters**, v. 7, n. 7, p. 511–516, 2020.

MOREAU, Yves. Crack down on genomic surveillance. **Nature.** v.576, p.36-38. 2019.

MORSE, Stephen S. et al. Prediction and prevention of the next pandemic zoonosis. **The Lancet**, v. 380, n. 9857, p. 1956-1965, 2012.

NARZISI, G.; MISHRA, B. Comparing De Novo genome assembly: The long and short of it. **PLoS ONE**, v. 6, n. 4, 2011.

NAVECA, Felipe Gomes et al. Genomic, epidemiological and digital surveillance of Chikungunya virus in the Brazilian Amazon. **PLoS neglected tropical diseases**, v. 13, n. 3, p. e0007065, 2019.

NIEUWENHUIJSE, David F.; KOOPMANS, Marion PG. Metagenomic sequencing for surveillance of food-and waterborne viral diseases. **Frontiers in Microbiology**, v. 8, p. 230, 2017.

ONE HEALTH. World Health Organization, 2017. Disponível em: <https://www.who.int/news-room/q-a-detail/one-health>. Acesso em: 28 de mai. de 2021

ONE HEALTH: ONE HEALTH BASICS. CDC | Centers for Disease Control and Prevention, 2018. Disponível em: <https://www.cdc.gov/onehealth/basics/index.html>. Acesso em: 15 de jun. de 2021.

O'NEILL, Jim et al. Review on antimicrobial resistance: tackling drug-resistant infections globally: final report and recommendations. 2016.

PEACOCK, Sarah J. et al. Changing the paradigm for hospital outbreak detection by leading with genomic surveillance of nosocomial pathogens. **Microbiology**. v.164, n.10, p.1213–1219. 2018

PECCIA, J. et al. SARS-CoV-2 RNA concentrations in primary municipal sewage sludge as a leading indicator of COVID-19 outbreak dynamics. **medRxiv**, v. 1, n. 203, 2020.

PHAN, My VT et al. Unbiased whole-genome deep sequencing of human and porcine stool samples reveals circulation of multiple groups of rotaviruses and a putative zoonotic infection. **Virus evolution**, v. 2, n. 2, 2016.

PINTO, O. H. B. et al. Genome-resolved metagenomics analysis provides insights into the ecological role of Thaumarchaeota in the Amazon River and its plume. **BMC Microbiology**, v. 20, n. 1, p. 1–11, 2020.
Public health services. WHO | World Health Organization, 2012. Disponível em: <https://www.euro.who.int/en/health-topics/Health-systems/public-health-services/public-health-services>. Acesso em: 03 de jun. de 2021.

PUBLIC HEALTH SURVEILLANCE. WHO | World Health Organization, 2021. Disponível em <http://www.emro.who.int/health-topics/public-health-surveillance/index.html>. Acesso em: 28 de mai. de 2021.

ROSE, R. et al. Challenges in the analysis of viral metagenomes. **Virus Evolution**, v. 2, n. 2, p. 1–11, 2016.

ROSEN, George. **Da Polícia Médica à Medicina Social**. Rio de Janeiro: Graal; 1980.

RYU, Sukhyun et al. One Health Perspectives on Emerging Public Health Threats. **Journal of Preventive Medicine & Public Health**. v.50, n.6, p.411-414. 2017.

SANGER, Frederick et al. DNA sequencing with chain terminating inhibitors. **Proceedings of the National Academy of Sciences**, v.74, n.12, p.5463-5467, 1977.

SCHUSTER, Stephan C. Next-generation sequencing transforms today's biology. **Nature Methods**. v.5, p.16–18. 2008.

SOUZA, Luis P. E. F. Saúde Pública ou Saúde Coletiva? **Revista espaço para a saúde.** v. 15, n. 4, p. 01-21, 2014

TATE, Jacqueline E. et al. 2008 estimate of worldwide rotavirus-associated mortality in children younger than 5 years before the introduction of universal rotavirus vaccination programmes: a systematic review and meta-analysis. **The Lancet infectious diseases**, v. 12, n. 2, p. 136-141, 2012.

TAUBENBERGER, Jeffery K. et al. Characterization of the 1918 influenza virus polymerase genes. **Nature**, v. 437, n. 7060, p. 889-893, 2005.

VIGILÂNCIA EM SAÚDE - **SUS: O Que E? PenseSUS** | Fiocruz. Disponível em: <https://pensesus.fiocruz.br/vigilancia-em-saude>. Acesso em: 03 de jun. de 2021.

WALLACE, Helen M. et al. Forensic DNA databases–Ethical and legal standards: A global review. **Egyptian Journal of Forensic Sciences**, v.4, n.3, p.57-63. 2014.

WOOLEY, John C.; GODZIK, Adam; FRIEDBERG, Iddo. A primer on metagenomics. **PLoS Comput Biol**, v. 6, n. 2, p. e1000667, 2010.

WORLD BANK. Drug-resistant infections: a threat to our economic future. World Bank, 2017.

WORLD HEALTH ORGANIZATION (WHO). GLASS whole-genome sequencing for surveillance of antimicrobial resistance: Global Antimicrobial Resistance and Use Surveillance System (GLASS). [S.l: s.n.], 2020. Disponível em: <https://www.who.int/publications/i/item/9789240011007>.

XAVIER, Joilson et al. Circulation of chikungunya virus East/Central/South African lineage in Rio de Janeiro, Brazil. **PloS one**, v. 14, n. 6, p. e0217871, 2019.

ZHU, Na et al. A novel coronavirus from patients with pneumonia in China, 2019. **New England journal of medicine**, 2020.

ANEXO E - GENÔMICA COMPARATIVA

score de alinhamento entre si. Sequências dentro do mesmo genoma e com alto *score* são consideradas parálogas, enquanto aquelas em genomas distintos são potenciais ortólogas.

De maneira geral, a ordem dos genes nos genomas de grupos de organismos eucarióticos proximamente relacionados é bastante conservada. Sintenia refere-se aos genes (ou regiões genômicas) presentes no mesmo cromossomos de duas (ou mais) espécies diferentes; colinearidade diz respeito à conservação da ordem desses genes [60]. Dessa forma, a sintenia e colinearidade de alguns blocos gênicos também podem oferecer informação relevante para a determinação de grupos de homólogos. As abordagens mais bem sucedidas consistem no uso de informações de diferentes fontes (filogenia, heurística e sintenia) para a obtenção de agrupamentos de genes homólogos mais robustos [57]. A partir dos grupos de genes homólogos, uma série de análises de genômica comparativa estão disponíveis.

3.3 - Genômica comparativa de genes, módulos e vias

A genômica comparativa surge a partir de estudos genômicos que se dedicam a realizar análises mais aprofundadas das características de um ou mais genomas de interesse através da sua comparação com outros genomas. A genômica comparativa gera conhecimento biologicamente relevante utilizando informações sobre os padrões de conservação e variação entre os diferentes elementos genômicos, compartilhados ou não entre os genomas em análise. Uma vez que os organismos vivos, do ponto de vista molecular, organizam-se de maneira hierárquica, essas análises podem ser feito em diferentes níveis, podendo-se realizar a busca por padrões de conservação e variação em genes, módulos e vias bioquímicas, dentro da mesma espécie ou entre espécies distintas.

Análises intra-específicas permitem a detecção de características fenotípicas complexas associadas à variações genômicas dentro de uma mesma espécie, uma estratégia frequentemente utilizada para a identificação de genes potencialmente causadores e de vias celulares relevantes para doenças em humanos e para características de interesse agropecuário nas espécies domesticadas [61]. Adicionalmente, análises dos padrões de variação entre os genomas de organismos de uma mesma espécie permitem a elucidação de sua história evolutiva [62].

A detecção de grupos de genes homólogos compartilhados entre espécies distintas permite que um amplo leque de análises pós-genômicas seja utilizado para, a partir dos padrões de conservação e variação dos elementos, realizar análises como inferências sobre a história evolutiva de genes ou espécies, buscas por padrões evolutivos associados a fenótipos ou por genes envolvidos em processos adaptativos [63]. Os grupos de genes ortólogos 1-1, definidos como grupos específicos de genes homólogos que possuem uma única cópia em todos os genomas analisados, são comumente utilizados para a determinação a história evolutiva de espécies distintas a partir de seu conteúdo genômico total compartilhado [64]. Outro uso comum de elementos homólogos consiste na busca por grupos de elementos genômicos cujos padrões de expansão e contração em diferentes espécies estão associados à algum fenótipo de interesse, como o parasitismo [65] ou a complexidade biológica [66], o que permite identificar possíveis agentes moleculares que contribuem para a emergência desses fenótipos.

3.4 - Pan-genoma

Os estudos de genômica comparativa detectaram que, enquanto nos eucariotos a sintenia e o conteúdo gênico compartilhado mantem-se relativamente conservados em organismos próximos filogeneticamente (e.g. humanos e chimpanzés), a ordem dos genes e a sua presença ou ausência varia consideravelmente em organismos procarióticos, mesmo em uma mesma espécie [67]. Tettelin e colaboradores (2005), estudando diferentes linhagens patogênicas da espécie *Streptococcus agalactiae*, cunharam o termo "pan-genoma", utilizado para representar o repertório genético completo de um dado clado (e.g espécie, gênero ou família) [68]. O pan-genoma é composto pelos genes presentes em todas as linhagens de um clado (genoma central, usualmente responsável por aspectos básicos da biologia do táxon e pelos principais fenótipos do mesmo), os genes ausentes em uma ou mais linhagens (genoma acessório, o qual contribui para a diversidade do táxon, podendo incluir vias bioquímicas complementares e funções que, embora não sejam essenciais para o crescimento, podem conferir vantagem adaptativo, tais como colonização de novos nichos, resistência à antibióticos ou a colonização de novos hospedeiros) e genes presente em apenas uma linhagem (genes linhagem-específicos, usualmente são mais pobremente caracterizados e

parecem ser advindos preferencialmente de enventos de transdução mediada por bacteriófagos) [69].

A maioria dos estudos de pan-genômica se dedicam ao estudo de microrganismos, sobretudo bactérias, devido à facilidade de sequenciamento e montagem do genomas desses organismos. Recentemente, estudos pan- genômicos têm contemplado organismos com eucarióticos. Estudos do tipo com plantas têm demonstrado alta variação de genes presentes/ausentes em genomas de plantas cultivadas e associados esses genes a características de interesse agrícola [70].

3.5 - Busca computacional por genes adaptativos

O estudo molecular da ação da seleção natural em regiões codificadoras homólogas evidencia um claro viés na frequência de mutações não-sinônimas quando comparada à frequência de mutações sinônimas [71]. De maneira geral, alinhamentos múltiplos de códons de um dado grupo de genes homólogos possuem a vasta maioria das colunas do alinhamento sem variação no aminoácido codificado; já as mutações para códons sinônimos ocorrem em frequências consideravelmente maiores. Assim, a vasta maioria das regiões codificadoras possuem uma taxa de substituições sinônimas (dS) consideravelmente maior que a taxa de substituições não-sinônimas (dN). Este fenômeno ocorre porque mutações não-sinônimas usualmente reduzem a eficiência funcional da proteína codificada em comparação ao alelo não mutante fixado anteriormente. Assim, mutações não-sinônimas geralmente diminuem a aptidão evolutiva dos organismos que as possuem, e estes alelos menos funcionais são rapidamente removidos das populações através de seleção negativa ou purificadora [72]. Consequentemente, os valores de dS são consideravelmente maiores que os de dN.

Entretanto, algumas poucas posições em alguns poucos genes e em algumas poucas linhagens podem apresentar valores de dN significativamente maiores do que os valores de dS, indicando a tendência evolutiva da fixação de novos alelos em detrimento aos antigos, ou a tendência de variação de determinadas posições das sequências ao invés da sua conservação. Este fenômeno é denominado seleção positiva ou Darwiniana, sendo observado em códons e genes que codificam proteínas nas quais ocorre pressão seletiva para a

variação ao invés da conservação do aminoácido na posição em análise quando comparada ao restante das posições, ou em uma dada sequência quando comparada ao restante das sequências [73]. Dentre os fenômenos biológicos que comumente possuem grupos de genes homólogos evoluindo sob pressão seletiva positiva destacam-se genes envolvidos na ocupação de nichos ecológicos pelas linhagens em análise. No caso de metazoários, genes evoluíndo sob evidência de seleção positiva estão usualmente envolvidos nos fenômenos de percepção sensorial, reprodução, imunidade e na relação parasita-hospedeiro [72, 74-78].

Computacionalmente, a busca por seleção positiva necessita de duas estruturas de dados biológicos: um alinhamento múltiplo de códons, onde cada coluna corresponde a uma posição homóloga nos genes em questão, e uma árvore filogenética descrevendo as relações entre as sequências em análise, onde cada linhagem terminal corresponde a uma sequência homóloga às demais. Dentro desta estrutura de dados, é possível observar posições onde há conservação do aminoácido codificado por um dado códon homólogo entre as sequências e posições onde há variação do aminoácido codificado pelo códon homólogo.

A busca por seleção positiva Darwiniana pode ocorrer em dois grandes modos de análise: a busca por seleção em colunas do alinhamento, chamada de *site-model*. O outro modo, chamado de *branch-model*, consiste na busca por seleção positiva em uma determinada linhagem evolutiva da árvore quando comparada às demais. Assim, análises do tipo *branch-model* permitem interrogar, dentro de uma filogenia, quais são os genes que possuem elevados valores de dN em uma linhagem específica da filogenia. Um terceiro modo, denominado branch- site-model, busca por seleção positiva em algumas posições (colunas) do alinhamento que ocorram em algumas linhagens (linhas) do mesmo [79]. Diversas estruturas de dados são necessárias para a detecção adequada de grupos de homólogos com seleção positiva, tais como alinhamento múltiplo de códons, construção de árvores filogenéticas e detecção de potenciais fontes de erro. Consequentemente, alguns programas encontram-se disponíveis para realizar todas as etapas necessárias para a detecção de seleção positiva em escala genômica [80, 81].

Referências Bibliográficas

1. Levy SE, Myers RM: **Advancements in Next-Generation Sequencing.** *Annu Rev Genomics Hum Genet* 2016, **17:**95-115.
2. Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al: **Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication.** *Cell* 2012, **149:**912-922.
3. Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T, et al: **Early human dispersals within the Americas.** *Science* 2018, **362**.
4. Aronson SJ, Rehm HL: **Building the foundation for genomics in precision medicine.** *Nature* 2015, **526:**336-342.
5. Hobert O: **The impact of whole genome sequencing on model system genetics: get ready for the ride.** *Genetics* 2010, **184:**317-319.
6. Flores R, Hernandez C, Martinez de Alba AE, Daros JA, Di Serio F: **Viroids and viroid-host interactions.** *Annu Rev Phytopathol* 2005, **43:**117-139.
7. Koonin EV, Dolja VV, Krupovic M: **Origins and evolution of viruses of eukaryotes: The ultimate modularity.** *Virology* 2015, **479-480:**2-25.
8. Martinez-Cano DJ, Reyes-Prieto M, Martinez-Romero E, Partida-Martinez LP, Latorre A, Moya A, Delaye L: **Evolution of small prokaryotic genomes.** *Front Microbiol* 2014, **5:**742.
9. Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ: **Genome Size Diversity and Its Impact on the Evolution of Land Plants.** *Genes (Basel)* 2018, **9**.
10. Evans BJ, Upham NS, Golding GB, Ojeda RA, Ojeda AA: **Evolution of the Largest Mammalian Genome.** *Genome Biol Evol* 2017, **9:**1711-1724.
11. Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers A, Van den Berghe A, et al: **Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene.** *Nature* 1976, **260:**500-507.
12. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: **Nucleotide sequence of bacteriophage phi X174 DNA.** *Nature* 1977, **265:**687-695.
13. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al.: **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.** *Science* 1995, **269:**496-512.
14. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al: **Life with 6000 genes.** *Science* 1996, **274:**546, 563-547.
15. Consortium CeS: **Genome sequence of the nematode C. elegans: a platform for investigating biology.** *Science* 1998, **282:**2012-2018.
16. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al: **The genome sequence of Drosophila melanogaster.** *Science* 2000, **287:**2185-2195.
17. Arabidopsis Genome I: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408:**796-815.
18. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
19. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
20. Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al: **Initial**

sequencing and comparative analysis of the mouse genome. *Nature* 2002, **420:**520-562.
21. Koepfli KP, Paten B, Genome KCoS, O'Brien SJ: **The Genome 10K Project: a way forward.** *Annu Rev Anim Biosci* 2015, **3:**57-111.
22. Matasci N, Hung LH, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M, et al: **Data access for the 1,000 Plants (1KP) project.** *Gigascience* 2014, **3:**17.
23. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J: **NCBI prokaryotic genome annotation pipeline.** *Nucleic Acids Res* 2016, **44:**6614-6624.
24. Richardson EJ, Watson M: **The automatic annotation of bacterial genomes.** *Brief Bioinform* 2013, **14:**1-12.
25. **FastQC: a quality control tool for high throughput sequence data** [http://www.bioinformatics.babraham.ac.uk/projects/fastqc]
26. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30:**2114-2120.
27. Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Vinnere Pettersson O, Amselem J, Bouri L, Bocs S, Klopp C, et al: **Ten steps to get started in Genome Assembly and Annotation.** *F1000Res* 2018, **7**.
28. Sohn JI, Nam JW: **The present and future of de novo whole-genome assembly.** *Brief Bioinform* 2018, **19:**23-40.
29. Aldrup-Macdonald ME, Sullivan BA: **The past, present, and future of human centromere genomics.** *Genes (Basel)* 2014, **5:**33-50.
30. Altemose N, Miga KH, Maggioni M, Willard HF: **Genomic characterization of large heterochromatic gaps in the human genome assembly.** *PLoS Comput Biol* 2014, **10:**e1003628.
31. Rice ES, Green RE: **New Approaches for Genome Assembly and Scaffolding.** *Annu Rev Anim Biosci* 2019, **7:**17-40.
32. Kremer FS, McBride AJA, Pinto LS: **Approaches for in silico finishing of microbial genome sequences.** *Genet Mol Biol* 2017, **40:**553-576.
33. **The Animal Genome Size Database**
34. Yandell M, Ence D: **A beginner's guide to eukaryotic genome annotation.** *Nat Rev Genet* 2012, **13:**329-342.
35. Ayling M, Clark MD, Leggett RM: **New approaches for metagenome assembly with short reads.** *Brief Bioinform* 2019.
36. Koonin EV: **Orthologs, paralogs, and evolutionary genomics.** *Annu Rev Genet* 2005, **39:**309-338.
37. Hillier LW, Miller RD, Baird SE, Chinwalla A, Fulton LA, Koboldt DC, Waterston RH: **Comparison of C. elegans and C. briggsae genome sequences reveals extensive conservation of chromosome organization and synteny.** *PLoS Biol* 2007, **5:**e167.
38. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 2015, **31:**3210-3212.
39. Abbas Q, Raza SM, Biyabani AA, Jaffar MA: **A Review of Computational Methods for Finding Non-Coding RNA Genes.** *Genes (Basel)* 2016, **7**.
40. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics* 2011, **12:**491.
41. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30:**1236-1240.
42. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite.** *Nucleic Acids Res* 2008, **36:**3420-3435.

43. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A: **Protein function annotation by homology-based inference.** *Genome Biol* 2009, **10:**207.
44. Gaudet P, Livstone MS, Lewis SE, Thomas PD: **Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium.** *Brief Bioinform* 2011, **12:**449-462.
45. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, et al: **Clinical whole-exome sequencing for the diagnosis of mendelian disorders.** *N Engl J Med* 2013, **369:**1502-1511.
46. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S, et al: **Genetic diagnosis by whole exome capture and massively parallel DNA sequencing.** *Proc Natl Acad Sci U S A* 2009, **106:**19096-19101.
47. Bloss CS, Zeeland AA, Topol SE, Darst BF, Boeldt DL, Erikson GA, Bethel KJ, Bjork RL, Friedman JR, Hwynn N, et al: **A genome sequencing program for novel undiagnosed diseases.** *Genet Med* 2015, **17:**995-1001.
48. Schwarze K, Buchanan J, Taylor JC, Wordsworth S: **Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature.** *Genet Med* 2018, **20:**1122-1130.
49. Biesecker LG, Shianna KV, Mullikin JC: **Exome sequencing: the expert view.** *Genome Biol* 2011, **12:**128.
50. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, et al: **Exome sequencing identifies the cause of a mendelian disorder.** *Nat Genet* 2010, **42:**30-35.
51. Mundade R, Ozer HG, Wei H, Prabhu L, Lu T: **Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond.** *Cell Cycle* 2014, **13:**2847-2852.
52. Furey TS: **ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions.** *Nat Rev Genet* 2012, **13:**840-852.
53. Reinert K, Langmead B, Weese D, Evers DJ: **Alignment of Next-Generation Sequencing Reads.** *Annu Rev Genomics Hum Genet* 2015, **16:**133-151.
54. Xu C: **A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data.** *Comput Struct Biotechnol J* 2018, **16:**15-24.
55. Steinhauser S, Kurzawa N, Eils R, Herrmann C: **A comprehensive comparison of tools for differential ChIP-seq analysis.** *Brief Bioinform* 2016, **17:**953-966.
56. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147:**195-197.
57. Kristensen DM, Wolf YI, Mushegian AR, Koonin EV: **Computational methods for Gene Orthology inference.** *Brief Bioinform* 2011, **12:**379-391.
58. Zhou X, Shen XX, Hittinger CT, Rokas A: **Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets.** *Mol Biol Evol* 2018, **35:**486-503.
59. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci U S A* 1999, **96:**2896-2901.
60. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH: **Synteny and collinearity in plant genomes.** *Science* 2008, **320:**486-488.
61. Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D: **Benefits and limitations of genome-wide association studies.** *Nat Rev Genet* 2019, **20:**467-484.
62. Parker HG, Dreger DL, Rimbault M, Davis BW, Mullen AB, Carpintero-Ramirez G, Ostrander EA: **Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development.** *Cell Rep* 2017, **19:**697-708.
63. Koonin EV: **Darwinian evolution in the light of genomics.** *Nucleic Acids Res* 2009, **37:**1011-1034.

64. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al: **Whole-genome analyses resolve early branches in the tree of life of modern birds.** *Science* 2014, **346:**1320-1331.
65. International Helminth Genomes C: **Comparative genomics of the major parasitic worms.** *Nat Genet* 2018.
66. Vogel C, Chothia C: **Protein family expansions and biological complexity.** *PLoS Comput Biol* 2006, **2:**e48.
67. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, et al: **Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths.** *PLoS Genet* 2009, **5:**e1000344.
68. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al: **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome".** *Proc Natl Acad Sci U S A* 2005, **102:**13950-13955.
69. Vernikos G, Medini D, Riley DR, Tettelin H: **Ten years of pan-genome analyses.** *Curr Opin Microbiol* 2015, **23:**148-154.
70. Tao Y, Zhao X, Mace E, Henry R, Jordan D: **Exploring and Exploiting Pan-genomics for Crop Improvement.** *Mol Plant* 2019, **12:**156-169.
71. Yang Z, Bielawski JP: **Statistical methods for detecting molecular adaptation.** *Trends Ecol Evol* 2000, **15:**496-503.
72. Aguileta G, Refregier G, Yockteng R, Fournier E, Giraud T: **Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists.** *Infect Genet Evol* 2009, **9:**656-670.
73. Moretti S, Laurenczy B, Gharib WH, Castella B, Kuzniar A, Schabauer H, Studer RA, Valle M, Salamin N, Stockinger H, Robinson-Rechavi M: **Selectome update: quality control and computational improvements to a database of positive selection.** *Nucleic Acids Res* 2014, **42:**D917-921.
74. Yang Z, Wong WS, Nielsen R: **Bayes empirical bayes inference of amino acid sites under positive selection.** *Mol Biol Evol* 2005, **22:**1107-1118.
75. Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L: **Patterns of Positive Selection in Seven Ant Genomes.** *Mol Biol Evol* 2014.
76. Soyer Y, Orsi RH, Rodriguez-Rivera LD, Sun Q, Wiedmann M: **Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected Salmonella serotypes.** *BMC Evol Biol* 2009, **9:**264.
77. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A: **Patterns of positive selection in six Mammalian genomes.** *PLoS Genet* 2008, **4:**e1000144.
78. Aguileta G, Lengelle J, Marthey S, Chiapello H, Rodolphe F, Gendrault A, Yockteng R, Vercken E, Devier B, Fontaine MC, et al: **Finding candidate genes under positive selection in Non-model species: examples of genes involved in host specialization in pathogens.** *Mol Ecol* 2010, **19:**292-306.
79. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24:**1586-1591.
80. Sahm A, Bens M, Platzer M, Szafranski K: **PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes.** *Nucleic Acids Res* 2017, **45:**e100.
81. Hongo JA, de Castro GM, Cintra LC, Zerlotini A, Lobo FP: **POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes.** *BMC Genomics* 2015, **16:**567.