

SUPPLEMENTARY TEXT

Conserved protein folds underpin the diversification of secreted proteins in a fungal pathogen

Thaís C. S. Dal’Sasso^{1,2*}, Eva H. Stukenbrock^{1,2*}

Affiliations:

¹Environmental Genomics, Christian-Albrechts University of Kiel, Am Botanischen Garten 9, 24118, Kiel, Germany

²Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306, Plön, Germany

***Corresponding authors:**

tdalsasso@bot.uni-kiel.de (TCSD)

estukenbrock@bot.uni-kiel.de (EHS)

SUPPLEMENTARY RESULTS

Structural annotation of *Z. passerinii* Zpa796 secreted proteins

We selected the 670 AlphaFold2 (AF2)-predicted structures of *Z. passerinii* Zpa796 for our downstream analyses. A total of 479 (71.49%) protein structures from Zpa796 were successfully annotated, that is, they shared a similar fold with entries in at least one of the six structural databases we analyzed (Fig. 2, Additional File 1: Table S4). The highest number of protein structures shared a similar fold with entries in the AFDB [1], with 461 and 423 annotated proteins with the AFDB (Proteome) and AFDB (Swiss-Prot) sections, respectively (Fig. 2A, Additional File 2: Figs. S4A and S8A). Zpa796 proteins also matched experimentally determined structures, with 389 proteins identified with a similar fold in PDB (Figs 2A, Additional File 2: S4A and S8A). Structural databases (CATH, ECOD, and SCOPe) [2–4] had fewer matches to Zpa796 proteins (340, 293, and 255, respectively) (Fig. 2A, Additional File 2: Figs. S4A and S8A). We found overlapping for structures annotated across the databases, with 247 structures from Zpa796 sharing a similar fold entry in all six structural databases (Fig. 2A). Smaller overlaps for combinations of databases were also observed. In total, 46 proteins structures of Zpa796 had a best hit in only one of the six structural databases, with 38 of these found exclusively in the AFDB (Proteome). No exclusive best hits were found in the ECOD and SCOPe databases.

We explored the distribution of annotated structures according to mature protein length (Fig. 2B-C, Additional File 2: Fig. S6). Proteins ranging from 101 to 600 amino acids were annotated more frequently than shorter proteins (up to 100 residues) or longer

ones (more than 600 residues) (Fig. 2B-C). The distribution of annotated proteins by length and structural confidence (pLDDT) followed similar trends across all databases, peaking at 401 amino acids and pLDDT = 94.26 (Fig. 2C-D, Additional File 2: Figs. S6 and S7).

Structure-based annotation of effector and antimicrobial proteins

Using EffectorP [5] we predicted 310 effectors in the secretome of *Z. passerinii* Zpa796 (Additional File 2: Fig. S4B, Additional File 1: Table S1). We predicted structures for 251 (80.98%) putative effectors (Additional File 2: Fig. S4B) using AF2 and annotated 143 (56.97%) of them. Zpa796 putative effectors shared a similar fold with entries in AFDB (135 in AFDB (Proteome), 115 in AFDB (Swiss-Prot)), followed by entries in CATH (99 similar entries), PDB (98), ECOD (71), and SCOPe (65) (Additional File 2: Fig. S4B). In total, 78 putative effectors were exclusively annotated by at least one of the four main databases (AFDB (Proteome and Swiss-Prot sections), CATH, PDB) (Additional File 2: Fig. S8B).

Using AF2-predicted structures as input and the machine learning pipeline AMAPEC [6], we predicted 341 antimicrobial proteins (AMPs) (Additional File 2: Fig. S4C) and annotated 273 (80.06%) of them. The number of predicted AMPs that shared a similar fold in the structural databases were: 264 (AFDB (Proteome)), 243 (AFDB (Swiss-Prot)), 222 (PDB), 197 (CATH), 170 (ECOD), and 147 (SCOPe) (Additional File 2: Fig. S4C). In total, 68 predicted AMPs were exclusively annotated by at least one of the four main databases (AFDBs (Proteome and Swiss-Prot), CATH, PDB) (Additional File 2: Fig. S8C).

Sequence-based annotation of the secretome

In addition to the structure-based annotation, we annotated the 743 predicted secreted proteins using amino acid sequences to identify conserved domains (Additional File 2: Fig. S4A, Additional File 1: Table S1). The sequence-based databases that annotated the highest number of secreted proteins were InterPro (386 annotated proteins, 51.95% of total secreted proteins) and eggNOG (378 proteins, 50.87%) (Additional File 2: Fig. S4A). When compared to the four main structural databases used (AFDBs – both Proteome and Swiss-Prot sections, CATH, PDB), only 38 proteins (5.11% of total secreted proteins) were annotated exclusively by InterPro and eggNOG (Additional File 2: Fig. S8A). Among the 310 putative effectors, InterPro and eggNOG annotated 59 (19.03%) and 69 (22.26%) proteins, respectively (Additional File 2: Fig. S4B). Only 10 (3.23%) effectors were annotated exclusively by one of these two sequence-based databases and not by any of the four main structural databases (Additional File 2: Fig. S8B). Among the 341 predicted AMPs, InterPro and eggNOG annotated 208 (61%) and 209 (61.30%) proteins, respectively (Additional File 2: Fig. S4C). Fifteen (4.40%) predicted AMPs were exclusively annotated by one of these two sequence-based databases when compared to structural annotations (Additional File 2: Fig. S8C). In total, 226 (30.42%) secreted proteins, 157 (50.65%) putative effectors, and 53 (15.54%) predicted AMPs were not annotated by any of the two sequence-based databases (InterPro, eggNOG) and the four structural-based databases (AFDB (Proteome), AFDB (Swiss-Prot), CATH, and PDB) (Additional File 2: Fig. S8).

Network statistics and robustness of structural clusters

To characterize the topology of the structural similarity network, we computed general network statistics and centrality measures across all nodes. Overall, network statistics were consistent with the fragmented nature of the network. The average degree was 4.813, indicating that each protein was structurally similar to roughly five others on average. The network density was 0.013, supporting a sparse network with few connections among all possible protein pairs. For centrality measures, the average degree centrality was 0.013, reflecting the proportion of nodes each protein is directly connected to. The average betweenness centrality was close to 0.000, indicating that most proteins do not act as bridges. The average closeness centrality was 0.016, supporting long paths between proteins within the network. Due to the high fragmentation of the network, we did not include statistics based solely on the largest connected component (i.e. average clustering coefficient, shortest path length, or network diameter) as these metrics would not capture the global architecture of the network.

To assess the robustness of the structural clusters and enrichment results for the Zpa796 secretome, we rebuilt structural similarity networks using increasing mean AF2 pLDDT thresholds from >50 to >95 and repeated the functional enrichment analyses (Additional File 2: Fig. S9). These rebuilt networks reproduced the original graph, whereas the number of proteins and cluster sizes decreased as the confidence threshold became more stringent (Additional File 2: Fig. S9A). Notably, effector-enriched clusters were stable under moderate-to-high confidence filters (pLDDT > 60 to pLDDT > 85), with five of the original effector-enriched groups remaining significant: G.03, G.04, G.06, G.12, and G.14 (Additional File 2: Fig. S9B). Cluster G.21 was significant at the least stringent thresholds (pLDDT > 50 and pLDDT > 55). Under stringent pLDDT > 90 filter, only G.06

and G.14 remained significantly enriched, consistent with the strong reduction in network size and statistical power under this cutoff (Additional File 2: Fig. S9). No structural component was significantly enriched in predicted AMPs under any pLDDT threshold. Overall, this sensitivity analysis shows that the main effector-enrichment patterns are retained across moderate-to-high confidence filters, supporting the robustness of the structural clustering results.

Functional landscape of structurally similar clusters

Although not significantly enriched after p-value correction ($p\text{-value}_{\text{adj}} > 0.05$), 10 small structural clusters (composed of 2–4 proteins) contained only proteins predicted to be effectors (Fig. 3). Some of these clusters were annotated into known protein folds or related to characterized biological functions, including: G.31 (OB-Fold; nucleotide-binding and antifungal activity), G.34 (Translation initiation factors EIF1), G.36 (Hydrophobins; structural proteins), G.41 (Pectate-lyases C), G.44 (Glucanases), and G.50 (Barwin-domain, endoglucanases) (Additional File 1: Table S4). Moreover, five small clusters (G.29, G.31, G.44, G.50, and G.58) were composed exclusively of putative antimicrobial effectors (Fig. 3).

Physicochemical properties of proteins across the network

To contextualize the lack of AMP enrichment in the network, we compiled six physicochemical properties and compared these properties across the structurally similar clusters (Additional File 2: Fig. S11, Additional File 1: Tables S9 and S11). We limited comparisons to the 46 structural clusters with at least three proteins. The six properties

showed significant differences across structural clusters according to Kruskal-Wallis tests ($p\text{-value} < 0.001$) (Additional File 1: Table S9). Post hoc pairwise comparisons using Dunn's test showed several significant differences ($p\text{-value}_{\text{adj}} < 0.05$) between clusters for five properties (Additional File 1: Table S11). There were 495 significant differences between pairs of structural clusters for mature protein length, 123 for net charge, 81 for hydrophobic moment, 75 for mean hydrophobicity, and one for surface hydrophobicity ($p\text{-value}_{\text{adj}} < 0.05$) (Additional File 1: Table S11). Overall, we identified significant differences between effector-enriched and effector-mainly clusters when compared to clusters composed of putative non-effector proteins (predicted as AMPs only and other secreted proteins) (Additional File 1: Table S11). Despite a significant Kruskal-Wallis test for charge density (Additional File 1: Tables S9), post hoc pairwise comparisons did not show any significant pairwise comparison between clusters after p -value correction ($p\text{-value}_{\text{adj}} > 0.05$) (Additional File 1: Table S11). This is likely because only the most pronounced cluster-specific differences were identified as statistically significant due to multiple-testing correction for 1035 comparisons using the Benjamini and Hochberg (BH) method.

Protein evolution among homologs of *Z. passerinii* and other *Zymoseptoria* spp.

We used the OrthoFinder pipeline [7] to identify homologous relationships between proteins from *Z. passerinii* Zpa796 and other *Zymoseptoria* species. Among the 364 proteins from the structural similarity network of isolate Zpa796, five proteins — jg9754.t1 (G.03), jg9578.t1 (G.04), jg2591.t1 (G.07), jg2596.t1 (G.37), and jg5893.t1 (G.37) — had no homologs in any other *Zymoseptoria* species or in *Z. passerinii* Zpa63 (Additional File 1: Table S12), suggesting they may be isolate-specific. The remaining 359 Zpa796

proteins shared homologs with at least one other species (Additional File 1: Table S12). Some, such as the predicted AMP jg2656.t1 (G.10), had a single ortholog in all compared individuals, including the reference isolate *Z. tritici* IPO323 [8–10] (Additional File 1: Table S12). Others had variable number of homologs in the other species, such as jg9741.t1 (G.14). Homologs of this putative antimicrobial effector varied from one (in *Z. passerinii* Zpa63, *Z. pseudotritici* Zp13, and *Z. tritici* Zt469) [11,12] to five copies in *Z. ardabilae* Za17 [11] (Fig. 6, Additional File 1: Table S12).

The structural alignments of *Z. passerinii* Zpa796 proteins (jg3663.t1 and jg9741.t1, both putative antimicrobial effectors) and their effector homologs in *Z. tritici* IPO323 showed overall structural conservation, but we identified local variations primarily in loop regions (Fig. 7). The superposition of jg3663.t1 (G.12, KP4-like) and ZtIPO323_026410.1 structures showed a conserved core fold (TM-score = 0.87, RMSD = 2.08), with major differences in loop extensions (small disruptions reflecting indels in the diagonal pattern of the contact map, Fig. 7) and terminal regions. Similarly, there was overall structural conservation between jg9741.t1 (G.04, alpha-beta plait/ferredoxin-like) and its IPO323 homologs ZtIPO323_104450.1 and ZtIPO323_014270.1 (TM-scores ranging from 0.80 to 0.82 and RMSD values between 1.46 and 1.55) (Fig. 7). The core helices and beta-strands were conserved, whereas we identified structural differences primarily in the loop regions, resulting from indels. The most prominent structural deviations were in the loop regions of the alignment between jg9741.t1 and ZtIPO323_014270.1, with the *Z. tritici* IPO323 protein having more extended and less constrained β 1- α 1 and β 2- β 3 loops than its homolog in *Z. passerinii* Zpa796.

SUPPLEMENTARY METHODS

Sequence-based protein functional annotation

Protein domains were annotated using InterProScan v5.54-87.0 [13] with the applications: SMART-7.1, SUPERFAMILY-1.75, CDD-3.18, TIGRFAM-15.0, Pfam v34.0, and Gene3D-4.3.0. EuKaryotic Orthologous Groups (KOG), Gene Ontology, and KEGG terms were assigned to the proteome using eggNOG-mapper v2 [14] against the eggNOG 5.0 database [15]. Carbohydrate-active enzymes (CAZymes) were predicted using the dbCAN2 HMM profile database v7.0 [16] and *hmmScan* from HMMER v3.1b2 [17]. Lipases were predicted using HMM profiles from the Lipase Engineering Database v3.0 [18] and *hmmScan*. Significant matches for CAZymes and lipases were parsed using as cutoff E-value = $1e-4$ and coverage = 0.35. Proteases were predicted using the fungal proteome as queries and the MEROPS database v12.1 [19] with BLASTp run locally (thresholds set as E-value = $1e-4$, minimum identity = 0.40, coverage = 0.80). The secondary metabolite biosynthesis gene clusters in the genome were predicted with antiSMASH v7.1.0 [20] for fungi using the following features: KnownClusterBlast, ClusterBlast, SubClusterBlast, MIBiG Cluster, ActiveSiteFinder, RREFinder, ClusterPfam, Pfam-based GO Term Annotation, and TIGRFam.

Fungal secretome modelling by AlphaFold2 and ESMFold

We compared structural prediction methods using the entire secretome of *Z. passerinii* Zpa796 and by sequence ranges based on the mature protein lengths. Proteins

were grouped into 100-amino-acid length groups (e.g., 1–100, 101–200, up to proteins with more than 1,000 residues). To compare the quality of structural predictions by AF2 and ESMFold for secreted proteins, we examined the correlation between the two methods by comparing pLDDT values. We assessed data normality with the Shapiro-Wilk test ($p\text{-value} < 0.05$), followed by Spearman's correlation test. Distributions of pLDDT values for the entire predicted secretome and for each length category were compared using the Wilcoxon rank-sum test and effect sizes were calculated using Cliff's delta. Additionally, we used TM-align v20220412 [21] with default parameters to compare structural similarity between prediction methods by assessing the TM-score and RMSD values obtained from pairwise structural alignments of proteins with predicted structures by both methods. Differences in structural similarity across protein length categories were evaluated for each similarity index using the Kruskal-Wallis test followed by Dunn's test with Benjamini and Hochberg ($p\text{-value}_{\text{adj}} < 0.05$). To measure how similar the AF2 and ESMFold predictions were to crystallized fungal protein structures, we compared TM-scores and RMSD values between structures predicted by both methods and crystalized structures retrieved from the Protein Data Bank (PDB). Statistical comparisons were performed using the Kruskal-Wallis and Dunn's tests in R v4.3.3 [22].

Phylogenetic and selection analyses

To investigate the phylogenetic relationships of proteins from selected structural clusters, we aligned amino acid sequences for each cluster separately using the L-INS-i algorithm implemented in MAFFT v7.525 [23]. Alignments were trimmed with trimAl v1.5.rev0 [24] using the parameters -gt 0.9 -cons 60 -w 3. We determined the best-fit

evolutionary model for each dataset using ModelFinder, as implemented in IQ-TREE v2.3.6 [25], based on the Bayesian Information Criterion (BIC). The best-fit models were WAG+G4 for cluster G.04 and WAG+I+G4 for cluster G.12.

To test for selection at the codon level, we analyzed the *Z. passerinii* Zpa796 proteins within selected structural clusters. For each cluster, the corresponding CDS sequences were combined with the mature protein alignments to generate codon-aware nucleotide alignments using MAFFT L-INS-i. These alignments were trimmed using the trimAl and used to infer maximum-likelihood trees in IQ-TREE using the parameters described above. Selection analyses were conducted in HyPhy v2.5.64 [26] using the methods MEME and FEL. For each method, we retained sites with significant support under the default parameters.

Physicochemical properties calculated by AMAPEC

Among the six physicochemical properties we compared for the secreted proteins of *Z. passerinii* Zpa796, four were already calculated as part of AMAPEC v1.0b pipeline [6]: net charge, mean hydrophobicity, and hydrophobic moment, and surface hydrophobicity. Net charge, mean hydrophobicity, and hydrophobic moment were computed from mature protein sequences using the *Peptides* package [27] in R v4.3.1, as implemented in AMAPEC with default parameters. Net charge was calculated at pH = 7 and using 'pKscale = "EMBOSS"'. Mean hydrophobicity was calculated using the Eisenberg scale and the arithmetic mean per-residue values across the mature protein sequence. Hydrophobic moment (amphipathicity) was quantified with a protein rotational angle of 100 and a sliding window of 11 residues; for each protein, the maximum value

across all sliding windows was retained to capture the strongest local helical amphipathicity. Surface hydrophobicity was computed from solvent-accessible surface areas via Bio.PDB.DSSP module from Biopython v1.83 [28].

REFERENCES

1. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 2022; 50:D439–D444. <https://doi.org/10.1093/nar/gkab1061>
2. Knudsen M, Wiuf C. The CATH database. *Hum Genomics.* 2010;4: 207–212. <https://doi.org/10.1186/1479-7364-4-3-207>
3. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, et al. ECOD: An Evolutionary Classification of Protein Domains. *PLoS Comput Biol.* 2014; 10:e1003926. <https://doi.org/10.1371/journal.pcbi.1003926>
4. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins - Extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* 2014; 42:D304–D309. <https://doi.org/10.1093/nar/gkt1240>
5. Sperschneider J, Dodds PN. EffectorP 3.0: Prediction of apoplastic and cytoplasmic effectors in fungi and oomycetes. *Mol Plant Microbe Interact.* 2022; 35:146–156. <https://doi.org/10.1094/MPMI-08-21-0201-R>
6. Mesny F, Wolf V, López-Moral A, Kraege A, Punt W, Liu S, et al. Plant-associated fungi co-opt ancient antimicrobials for host manipulation. *Sci Adv.* 2026;12:1406. <https://doi.org/10.1126/sciadv.aec1406>
7. Emms DM, Kelly S. OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 2015; 16:1–14. <https://doi.org/10.1186/s13059-015-0721-2>
8. Goodwin SB, M'Barek S Ben, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, et al. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet.* 2011; 7(6): e1002070. <https://doi.org/10.1371/journal.pgen.1002070>
9. Grandaubert J, Bhattacharyya A, Stukenbrock EH. RNA-seq-Based gene annotation and comparative genomics of four fungal grass pathogens in the genus *Zymoseptoria* identify novel orphan genes and species-specific invasions of transposable elements. *G3: Genes, Genomes, Genetics.* 2015;5:1323–33. <https://doi.org/10.1534/g3.115.017731>
10. Lapalu N, Lamothe L, Petit Y, Genissel A, Delude C, Feurtey A, et al. Improved gene annotation of the fungal wheat pathogen *Zymoseptoria tritici* based on combined Iso-Seq and RNA-Seq evidence. *Mol Plant Microbe Interact.* 2025;38:834–47; <https://doi.org/https://doi.org/10.1094/MPMI-07-25-0077-TA>

11. Feurtey A, Lorrain C, Croll D, Eschenbrenner C, Freitag M, Habig M, et al. Genome compartmentalization predates species divergence in the plant pathogen genus *Zymoseptoria*. BMC Genomics. 2020; 21:588. <https://doi.org/10.1186/s12864-020-06871-w>
12. Fagundes WC, Hansen R, Rojas Barrera IC, Caliebe F, Feurtey A, Haueisen J, et al. Host specialization defines the emergence of new fungal plant pathogen populations. bioRxiv. 2024; <https://doi.org/10.1101/2024.09.30.615799>
13. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: Genome-scale protein function classification. Bioinformatics. 2014;30:1236–40. <https://doi.org/10.1093/bioinformatics/btu031>
14. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021;38:5825–9. <https://doi.org/10.1093/molbev/msab293>
15. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:D309–14. <https://doi.org/10.1093/nar/gky1085>
16. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. DbCAN2: A meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46:W95–101. <https://doi.org/10.1093/nar/gky418>
17. Finn RD, Clements J, Eddy SR. HMMER: Profile hidden Markov models for biological sequence analysis. <http://hmmer.org>. Accessed 11 Aug 2025
18. Fischer M, Pleiss J. The Lipase Engineering Database: A navigation and analysis tool for protein families. Nucleic Acids Res. 2003;31:319–21. <https://doi.org/10.1093/nar/gkg015>
19. Rawlings ND, Barrett AJ, Bateman A. MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2012;40:343–50. <https://doi.org/10.1093/nar/gkr987>
20. Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, et al. antiSMASH 7.0: New and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Res. 2023;51. <https://doi.org/10.1093/nar/gkad344>
21. Zhang Y, Skolnick J. TM-align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33:2302–9. <https://doi.org/10.1093/nar/gki524>
22. R Core Team. R: A language and environment for statistical computing. <https://www.R-project.org/>. Accessed 11 Aug 2025
23. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013; 30:772–780. <https://doi.org/10.1093/molbev/mst010>
24. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25:1972–1973. <https://doi.org/10.1093/bioinformatics/btp348>
25. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32:268–74. <https://doi.org/10.1093/molbev/msu300>

26. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5 - A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol.* 2020; 37:295–9. <https://doi.org/10.1093/molbev/msz197>
27. Osorio D, Rondón-Villarreal P, Rodrigo Torres. Peptides: A package for data mining of antimicrobial peptides. *R J.* 2015;7:4–14. <https://doi.org/10.32614/RJ-2015-001>
28. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 2009;25:1422–3. <https://doi.org/10.1093/bioinformatics/btp163>