TOWARDS PREDICTION OF PHENOTYPE FROM GENOTYPE

APPROVED BY SUPERVISORY COMMITTEE

Nick V. Grishin, Ph. D.; Advisor

Zbyszek Otwinowski, Ph. D.; Committee Chair

Johann Deisenhofer, Ph. D.

Helen H. Hobbs, M. D., Ph. D.

DEDICATION

To those who care

TOWARDS PREDICTION OF PHENOTYPE FROM GENOTYPE

by

QIAN CONG

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

May, 2017

TOWARDS PREDICTION OF PHENOTYPE FROM GENOTYPE

Publication No. _____

Qian Cong, Ph. D.

The University of Texas Southwestern Medical Center at Dallas, 2017

Supervising Professor: Nick V. Grishin, Ph. D.

Predicting phenotype from genotype represents the epitome of biological questions. As a multiscale problem, it starts from predicting exons and culminates with modeling of whole organisms. Focusing on the molecular level, I studied the relationship between sequences and protein spatial structures and analyzed proteins with similar sequences but different structures. To aid the assessment of structure prediction, I developed a method to rank the predictions of proteins with new folds, a very challenging problem that was previously addressed by expert inspection. Then, I developed a set of computer programs and scripts to predict various structural and functional properties of proteins from their sequences and implemented them as a public web-server. I applied these methods to important agricultural (citrus disease) and

medical (*Ebolavirus*) problems. Moving on to organismal level predictions, I sequenced, annotated and analyzed complete genomes of butterflies and suggested hypotheses about genetic determinants of their behavior and other phenotypic traits. Taken together, these applications highlight the achievements possible today and challenges that lie ahead.

ACKNOWLEDGEMENTS

I am grateful to my mentor, Dr. Nick Grishin for teaching me how to do research, continuous help, encouragement and inspiration. I feel lucky to be able to participate in very diverse projects and learn a very broad spectrum of knowledge and skills.

I am thankful to my committee members, Drs. Hans Deisenhofer, Helen Hobbs and Zbyszek Otwinowski for discussions, advice and wisdom that taught me how to concentrate on what is important. I am indebted to Dominika Borek for her patience and teaching me how to do experiments and write papers.

Thanks to everyone in the Grishin lab, particularly Lisa Kinch, Jimin Pei, Jeremy Semeiks, Bong-Hyun Kim, Wenlin Li, Jinhui Shen, Jing Zhang and Ming Tang for their friendship, collaboration, help with many questions, and being a wonderful team of colleagues to turn to in any difficulty. Thanks to my collaborators, in particularly, William Israelsen and Siqi Liu, for brining me interesting scientific questions and sharing their experience. Thanks to everybody in UT Southwestern Medical Center, and in particular, people located in ND10, for maintaining such an open, friendly, and inspiring working environment.

Finally, I am grateful to my parents, Chengrui Cong and Yan Chen, and my best friend, Yunhan Wang for their constant care, support and the full trust in my decisions.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

1. **Cong Q**[*], Shen J[*], Li W, Borek D, Robbins RK, Otwinowski Z, Grishin NV. (2016) The first complete genome of a Metalmark butterfly. (submitted)
2. **Cong Q**[*], Shen J[*], Borek D, Robbins RK, Opler, PA, Otwinowski Z, Grishin NV. (2016) When COI barcodes deceive: complete genomes reveal introgression in hairstreaks. (submitted)
3. **Cong Q**, Grishin NV. (2016) Comparative analysis of Swallowtail transcriptomes reveals molecular determinants for speciation and adaptation. (submitted).
4. Shen J[*], **Cong Q**[*], Borek D, Otwinowski Z, Grishin NV. (2016) Complete genome of *Achalarus lyciades*, the first representative of the Eudaminae subfamily of Skippers. Current Genomics. (in press)
5. Shen J[*], **Cong Q**[*], Kinch LN, Borek D, Robbins RK, Otwinowski Z, Grishin NV. (2016) Complete genome of *Pieris rapae*, a resilient alien, a cabbage pest, and a source of anti-cancer proteins. *F1000Research***,** 5:2631 (doi: 10.12688/f1000research.9765.1)
6. Shen J, **Cong Q**, Grishin NV. (2016) The complete mitogenome of *Achalarus lyciades* (Lepidoptera: Hesperiidae). Mitochondrial DNA Part B: 1(1):581-583
7. **Cong Q**[*], Shen J[*], Borek D, Robbins RK, Otwinowski Z, Grishin NV. (2016) Complete genomes of Hairstreak butterflies, their speciation, and nucleo-mitochondrial incongruence. *Scientific Reports* 6:24863. doi: 10.1038/srep24863.
8. **Cong Q**[*], Shen J[*], Warren AD, Borek D, Otwinowski Z, Grishin NV. (2016) Speciation in Cloudless Sulphurs gleaned from complete genomes. *Genome Biology and Evolution.* evw045 doi: 10.1093/gbe/evw045, first published online: March 6, 2016.

9. **Cong Q**, Grishin NV. (2016) The complete mitochondrial genome of *Lerema accius* and its phylogenetic implications. *PeerJ.* 4: e1546.
10. **Cong Q**, Borek D, Otwinowski Z, Grishin NV. (2015) Skipper genome sheds light on unique phenotypic traits and phylogeny. *BMC Genomics*. 16: 639.
11. **Cong Q**, Pei J, Grishin NV. (2015) Predictive and comparative analysis of *Ebolavirus* proteins. *Cell Cycle* 14 (17): 2785-2797.
12. Shen J, **Cong Q**, Grishin NV. (2015) The complete mitochondrial genome of *Papilio glaucus* and its phylogenetic implications. *Meta gene* 5: 68-83.
13. **Cong Q**, Borek D, Otwinowski Z, Grishin NV. (2015) Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. *Cell reports* 10 (6): 910-919 (cover illustration).
14. Liu S, Cai X, Wu J, **Cong Q**, Chen X, Li T, Du F, Ren J, Wu YT, Grishin NV, Chen ZJ. (2015) Phosphorylation of innate immune adaptor proteins MAVS, STING, and TRIF induces IRF3 activation. *Science* 347 (6227): aaa2630.
15. Shiraiwa K, **Cong Q**, Grishin NV. (2014) A new *Heraclides* swallowtail (Lepidoptera, Papilionidae) from North America is recognized by the pattern on its neck. *Zookeys*. 468: 85-135.
16. **Cong Q**, Grishin NV. (2014) A new *Hermeuptychia* (Lepidoptera, Nymphalidae, Satyrinae) is sympatric and synchronic with *H. sosybius* in southeast US coastal plains, while another new *Hermeuptychia* species inhabits south Texas and northeast Mexico. *ZooKeys*. 379: 43-91.

17. Ji R, **Cong Q**, Li W, Grishin NV. (2013) M2SG: mapping human disease-related genetic variants to protein sequences and genomic loci. *Bioinformatics.* 29 (22): 2953-2954.

18. Li W, **Cong Q**, Kinch LN, Grishin NV. (2013) Seq2Ref: a web server to facilitate functional interpretation. *BMC bioinformatics*. 14: 30.

19. Li W, **Cong Q**, Pei J, Kinch LN, Grishin NV. (2012) The ABC transporters in *Candidatus* Liberibacter asiaticus. *Proteins.* 80 (11): 2614-2628.

20. **Cong Q**, Grishin NV. (2012) MESSA: Meta-Server for protein Sequence Analysis. *BMC Biology.* 10:82.

21. **Cong Q**, Kinch LN, Kim BH, Grishin NV. (2012) Predictive sequence analysis of the *Candidatus* Liberibacter asiaticus proteome**.** *PLoS One.* 7(7): e41071.

22. Kinch LN, Shi S, Cheng H, **Cong Q**, Pei J, Mariani V, Schwede T, Grishin NV. (2011) CASP9 target classification. *Proteins.* 79 Suppl 10: 21-36.

23. Kinch LN, Shi S, **Cong Q**, Cheng H, Liao Y, Grishin NV. (2011) Free Modeling structure prediction. *Proteins.* 79 Suppl 10: 59-73.

24. **Cong Q**, Kinch LN, Pei J, Shi, S, Grishin VN, Li W, Grishin NV. (2011) An automatic method for CASP9 free modeling structure prediction assessment. *Bioinformatics.* 27 (24): 3371-3378.

25. Kim BH, **Cong Q**, Grishin NV. (2010) HangOut: generating clean PSI-BLAST profiles for domains with long insertions. *Bioinformatics.* 26 (12): 1564-1565.

26. **Cong Q**, Kim BH, Kinch LN, Grishin NV. (2010) Structural Differences between Proteins with Similar Sequences, *Proceedings 2010 IEEE International Conference on Bioinformatics and Bioengineering.* bibe: 250-256.

* Authors contributed equally

CHAPTER ONE
General Introduction

Understanding how the phenotype of living organisms is encoded in the genotype is probably the most fundamental and important problem in biological sciences. Despite decades of research and many fundamental discoveries, the major breakthroughs on this front lie ahead. Prediction goes hand-in-hand with understanding. Developing predictive approaches to deduce various phenotypic features from gene sequences is essential both for understanding and practical applications, like suggesting hypotheses for experimental tests. Connections between genotype and phenotype can be studied at all scales. One of the simplest phenotypic features is spatial structure of proteins encoded in their sequences. Prediction of 3D structure and learning the connection between sequence and structure is the most basic puzzle at the molecular level. The next step is to predict functions of proteins and study how they interact with each other. Such predictions can be done on genomic level to investigate functional landscape of individual genes. Finally, at organismal level, one can think about predicting morphology of animals from their genomic sequences.

I studied the general problem of genotype-phenotype connection at all levels. I analyzed the data with existing software tools and developed new algorithms, applying them to specific problems of medical and agricultural importance. To address the problem at the level of whole organisms, I sequenced and analyzed complete genomes of butterflies and suggested hypotheses about their unique features and their functional implications. Here, I introduce

specific problems and summarize the results I have obtained, starting from the analysis of sequence-to-structure connection in proteins to genome-to-phenotypic traits connection in animals.

Similarity between protein sequences is usually predictive of similarity in structures. However, in some rare cases protein domains with significant sequence similarity adopt different structures. Here, we carry out a survey of protein domain pairs with high sequence similarity (measured by HHsearch probability) and low structural similarity (measured by Dali Z-score), aiming to identify the reasons for this discordance. Besides methodological problems with either sequences or structures of domains, we find and describe novel examples of homologs with structural changes.

Manual inspection has been applied to and is well-accepted for assessing CASP Free Modeling (FM) category predictions over the years. Such manual assessment requires expertise and significant time investment, yet has the problems of being subjective and unable to differentiate models of similar quality. It is beneficial to incorporate the ideas behind manual inspection to an automatic score system, which could provide subjective and reproducible assessment of structure models. Inspired by our experience in CASP9 FM category assessment, we developed an automatic superimposition independent method named Quality Control Score (QCS) for structure prediction assessment. QCS captures both global and local structural features, with emphasis on global topology. We applied this method to all FM targets from CASP9, and overall the results showed the best agreement with Manual Inspection Scores (MIS) among

automatic prediction assessment methods previously applied in CASPs, such as Global Distance Test Total Score (GDT_TS) and Contact Score (CS). As one of the important components to guide our assessment of CASP9 FM category predictions, this method correlates well with other scoring methods and yet is able to reveal good-quality models that are missed by GDT_TS.

Computational sequence analysis, i.e. prediction of local sequence properties, spatial structure and function from the sequence of a protein, offers an efficient way to obtain needed information about proteins under study. Since reliable sequence analysis is usually based on many computer programs to derive consensus and integrate evidence, meta severs have been developed to fit such needs. Most meta servers focus on one aspect of sequence analysis, while others incorporate more information, such as PredictProtein for local sequence feature predictions, SMART for domain architecture and sequence motifs annotation and Genesilico for secondary and spatial structure prediction. However, as predictions of local sequence properties, structure and function are usually intertwined, it is beneficial to address them together. We developed a MEta Server for Sequence Analysis (MESSA) to facilitate comprehensive protein sequence analysis. For an input protein sequence, the server incorporates a number of select tools to predict local sequence characteristics, detect homologous proteins, assign the query into related protein families and identify spatial structure templates. MESSA is designed for experimental biologists to gain structural and functional predictions about their protein of interest. We tested MESSA on the proteome of *Candidatus Liberibacter asiaticus*. Manual curation shows that the results provided by

MESSA could predict the 3D structure of around 75% of the residues and annotate the function of over 80% of the proteins in this entire proteome. MESSA is freely available for non-commercial use at http://prodata.swmed.edu/messa

Candidatus Liberibacter asiaticus (*Ca.* L. asiaticus) is a parasitic Gram-negative bacterium that is closely associated with citrus greening, a worldwide citrus disease. Given the difficulty in culturing the bacterium and thus in its experimental characterization, computational analyses of the whole *Ca.* L. asiaticus proteome can provide much needed insights into the mechanisms of the disease and guide the development of treatment strategies. In this study, we applied state-of-the-art sequence analysis tools to every *Ca.* L. asiaticus protein. The results are available as a public website at http://prodata.swmed.edu/liberibacter_asiaticus/. In particular, we manually curated the results to predict the structure and function of all *Ca.* L. asiaticus proteins. This extensive information should facilitate the study of *Ca.* L. asiaticus proteins aimed at understanding the biology of the bacterium and the mechanism of citrus greening. Pilot studies based on the information from this website have revealed several potential virulence factors.

Ebolavirus is the pathogen for Ebola Hemorrhagic Fever (EHF). This disease exhibits a high fatality rate and has recently reached a historically epidemic proportion in West Africa. Out of the five known *Ebolavirus* species, only *Reston ebolavirus* has lost human pathogenicity, while retaining the ability to cause EHF in long-tailed macaque. Significant efforts have been spent to determine the three-dimensional (3D) structures of *Ebolavirus* proteins, to study

their interaction with host proteins, and to identify the functional motifs in these viral proteins. Here, in light of these experimental results, we apply computational analysis to predict the 3D structures and functional sites for *Ebolavirus* protein domains with unknown structure, including a zinc-finger domain of VP30, the RNA-dependent RNA polymerase catalytic domain and a methyltransferase domain of protein L. In addition, we compare sequences of proteins that interact with *Ebolavirus* proteins from RESTV-resistant primates with those from RESTV-susceptible monkeys. The host proteins that interact with GP and VP35 show an elevated level of sequence divergence between the RESTV-resistant and RESTV-susceptible species, suggesting that they may be responsible for host specificity. Meanwhile, we detect variable positions in protein sequences that are likely associated with the loss of human pathogenicity in RESTV, map them onto the 3D structures and compare their positions to known functional sites. VP35 and VP30? are significantly enriched in these potential pathogenicity determinants and the clustering of such positions on the surfaces of VP35 and GP suggests possible uncharacterized interaction site with host proteins that contributes to the virulence of *Ebolavirus*.

Predicting phenotype from genotype represents the epitome of biological questions. Comparative genomics of appropriate model organisms holds the promise of making it possible. We sequenced, assembled, and comparatively analyzed a genome of the Eastern Tiger Swallowtail (*Papilio glaucus*), a showy butterfly with remarkable biological traits and challenging speciation puzzles. This highly heterozygous 376 Mb genome was obtained from a single male using a new cost-effective protocol. Comparison of its 15,000 genes with

available butterfly genomes suggests the molecular basis of phenotypic traits: e.g., a uniquely expanded family of isoprenoid synthesis enzymes could produce predator-repelling terpenes secreted by the swallowtail-specific caterpillar organ osmeterium. Only 4% of genes show divergence between *P. glaucus* and its sister species, *P. canadensis*, offering insights into phenotypic differences between them: e.g., species-specific mutations decidedly enriched in all 4 key circadian clock proteins may be responsible for conditional versus obligate pupal diapause distinguishing the two species. We deduce that *P. appalachiensis*, a species originated by hybridization of *P. glaucus* and *P canadensis*, inherited 80% of its genes from the latter, including the circadian clock components and thus obligate diapause. However, 6-phosphogluconate dehydrogenase, an enzyme linked to mimetic black female morph absent in *P. canadensis*, was among those inherited from *P. glaucus*. Finally, we propose several nuclear DNA barcodes, i.e., gene regions that can confidently identify closely related insect species, as a possible alternative to widely used mitochondrial DNA barcodes.

For 200 years zoologists have relied on phenotypes to learn about the evolution of animals. A glance at the genotype, even through several gene markers, revolutionized our understanding of animal phylogeny. Recent advances in sequencing techniques allowed researchers to obtain complete genomes much easier, and opened unprecedented opportunities to study genetics and evolution. The genomic landscape of *Heliconius* butterflies challenged our view of speciation, and revealed inter-species hybridization as a powerful mechanism to shape adaptive evolution in butterflies. Comparison of complete genomes of closely related taxa is promising to shed light on speciation mechanisms and the link between genotype and phenotype. We assembled

a complete genome of the Cloudless Sulphur (*Phoebis sennae eubule*) from a single wild-caught specimen. This genome was used as reference to compare genomes of 6 individuals, 3 from the eastern populations (Oklahoma and North Texas), referred to as a subspecies *Phoebis sennae eubule*, and 3 from the southwestern populations (South Texas) known as a subspecies *Phoebis sennae marcellina*. While the two subspecies differ only subtly in phenotype and COI mitochondrial DNA barcodes, comparison of their complete genomes revealed consistent and significant differences, which are more prominent than those between tiger swallowtails *Pterourus canadensis* and *Pterourus glaucus*. The reasons for low (0.5%) mitochondrial divergence in *Phoebis* compared to its high (1.8%) nuclear divergence remain unclear. The two Sulphur taxa differed in histone methylation regulators, chromatin-associated proteins, circadian clock, and early development proteins. Phylogenetically, complete genomes place family Pieridae away from Papilionidae, which is consistent with previous analyses based on several gene markers. We sequenced and assembled the first genomes from the family Pieridae. Comparative analyses suggest that *Phoebis sennae marcellina* from the southwestern United States and Latin America and *Phoebis sennae eubule* from the southeastern United States, may both be considered species-level taxa, and revealed the mutation hotspots associated with the divergence between the two taxa. This work lays the foundation for Pieridae genomics and provides rich sequence datasets for comparative studies.

Hesperiidae (skippers) was traditionally viewed as a basal group of butterflies based on its moth-like morphology and darting flight habits with fast wing beats. However, DNA-based studies suggest that Papilionidae is the basal group. The moth-like features and the

controversial position of skippers in Lepidoptera phylogeny make them valuable models for comparative genomics. We obtained the 310 Mb draft genome of the Clouded Skipper (*Lerema accius*) from a wild-caught specimen using a cost-effective strategy that overcomes the high (1.5%) heterozygosity problem. Comparative analysis of *Lerema accius* and another highly heterozygous genome of *Papilio glaucus* reveals difference in patterns of SNP distribution, but similarity in functions of genes that are enriched in non-synonymous SNPs. Comparison of Lepidoptera genomes reveals possible molecular bases for unique traits of skippers: duplication of electron transport chain components could result in efficient energy supply for their rapid flight; a diversified family of predicted cellulases might allow them to feed on the cellulose-enriched grasses; expansion of pheromone-binding proteins and enzymes for pheromone synthesis implies a more efficient mate-recognition system, which compensates the lack of clear visual cues due to the similarity in wing colors and patterns of many species of skippers. Phylogenetic analysis of several Lepidoptera genomes suggests that the position of Hesperiidae remains uncertain and the tree topology varied depending on the evolutionary models. This is the first genome of the Hesperiidae family. Comparative analyses reveal potential genetic bases for the unique phenotypic traits of skippers. This work lays the foundation for future experimental studies of skippers and provides a rich dataset for comparative genomics and phylogenetic studies of Lepidoptera.

CHAPTER TWO
Structural Differences Between Proteins With Similar Sequences

INTRODUCTION

From the early days of protein structural biology, researchers have been surprised by the resistance of protein spatial structures to evolutionary changes [1]. This remarkable structural robustness combined with the limited number of available 3D structures has lead to a view that the abstract protein structure space is discrete, can be divided into a number of folds, and protein evolution mostly proceeds within the framework of the same fold [2]. Today, with the rapidly increasing number of protein structures, arguably, the majority of protein structural patterns have been experimentally determined and a new view of structural continuity of folding patterns is starting to emerge [3,4]. Many examples of proteins with statistically significant sequence similarity that display substantial structural differences have been documented [5,6]. Such phenomenon demonstrates the evolutionary bridges between structurally different proteins and profoundly influences our understanding of protein structure evolution. On one hand, the notion that protein structures are evolutionarily plastic and changeable has important applications in protein design. This idea opens new frontiers in engineering proteins that possess desired functional properties, such as potentially creating proteins with condition-dependent folds [7]. On the other hand, the existence of proteins with similar sequences but different structures hinders homology modeling methods, making our ability to detect such cases from sequence crucial. To study the mechanisms and paths of protein fold change in evolution, we undertook a comprehensive comparative analysis of

SCOP (Structural Classification of Proteins) [8] domains and found domain pairs with significant sequence similarity, but pronounced structural differences. The reasons for structural differences in sequence-similar pairs were analyzed. We found that many cases are caused by various technical problems with sequence or structure, but the remaining pairs reveal interesting evolutionary changes in structure or possible convergence in sequence.

MATERIALS AND METHODS

PDB-style files for SCOP (Structural Classification of Proteins) 1.71 [9] domains from 4 classes: all α, all β, α/β and α+β were obtained from ASTRAL [10] and filtered for 40 percent sequence identity, resulting in 7805 domains. For these domains, all-to-all sequence comparison by HHsearch (Version 1.5) [11] (measured by probability) and structure comparison by DaliLite (version 2.4.4) [12] (measured by Z-score) were performed with default parameters. Profiles for HHsearch were built with PSI-BLAST [13] (E-value threshold: 0.001; maximum iterations: 8; protein sequence database: NCBI non-redundant (nr)) using buildali.pl script generously provided by Johannes Soding. Domain pairs with high HHsearch probabilities but low Dali Z-scores satisfying the following conditions: (1) HHsearch probability > Dali Z-score / 30 + 0.9; (2) Dali Z-score > 0 were chosen for this study. All the resulting 1804 domain pairs fell into 120 SCOP superfamily pairs. The domain pair with the highest Dali Z-score was selected from each superfamily pair as a representative for the manual study. Briefly, sequence alignment was checked manually and sometimes confirmed by HHpred [14] and/or PSI-BLAST; structure and structure alignment were visualized in Pymol.

Besides, information from SCOP, ASTRAL, PDB (Protein Data Bank) [15] VAST (Vector Alignment Search Tool) [16], and literature were used.

RESULTS AND DISCUSSION

Relationship between sequence similarity and structural similarity in domain pairs

To measure sequence similarity, a well-established protein profile-profile comparison tool HHsearch was used. Based on sequence profiles and secondary structure predictions, the probability estimate given by HHsearch is a more sensitive indicator of remote homology than simple sequence identity [11,17]. To measure structural similarity, Dali Z-score was used, because Dali is one of the best performing methods for structure comparison.

Among 25,109,240 (7085 * 7086/2 + 7085 = 25109240) domain pairs compared, the majority exhibit dissimilarity in sequence (HHsearch probability lower than 0.20) and structure (Dali Z-score lower than 3). A distribution of scores for a random sample of domain pairs with HHsearch probability above 0.2 is shown in Fig. 1a. While the HHsearch probability is positively correlated with the Dali Z-score for probabilities above 0.6, correlation is not obvious for lower HHsearch probabilities. To focus on domain pairs with comparatively low structural similarity, the region with Dali Z-score below 10 is shown in Fig. 1b.

Points in the upper left corner in Fig. 1b represent those rare cases of similar sequences adopting different structures. Unexpectedly, the narrow region with HHsearch probability above 0.9 harbors higher density than the region immediately below, indicating that many domains with significant sequence similarity are structurally different. A triangle area (in Fig. 1b) delineates the 1804 domain pairs chosen for this analysis. The domain pairs with Dali Z-

score 0 were not considered since preliminary analysis indicated that many of these pairs were structurally similar, but DaliLite failed to produce good alignments. This study set was further narrowed down by superfamily to 120 representative pairs for the detailed analysis (see Materials and Methods). The majority of these 1804 domain pairs (62.1%) belong to a single superfamily, P-loop containing nucleoside triphosphate hydrolases. This superfamily is a well-characterized large group of domains that contain highly conserved NTP binding Walker A and Walker B motifs but adopt diverse structures. [18-21]

Detailed study of the representative domain pairs

For all representative domain pairs, the reasons for the discordance between the sequence similarity and structural dissimilarity were studied and classified into three categories (Summarized in Table I): (1) problems with the sequence or sequence alignment; (2) problems with structure or structure alignment; (3) events of interest for protein evolution and biology.

1. Problem with the sequence or sequence alignment

Three causes for sequence problems have been detected:

1.1 Similar secondary structure pattern: Occasionally, a pair of sequences is attributed a high HHsearch probability due to similar secondary structure patterns that biases amino acid usage. First, HHsearch explicitly uses secondary structure predictions in scoring [11]. Besides, the restriction on the amino acid frequency from secondary structure (amino acid propensity) may result in two sequences having higher probability to share similar amino acids due to amino acid bias [22-24] rather than homology. Such domain pairs have similar secondary structure

patterns but typically differ in topology and spatial alignment of these secondary structural elements, which suggests that they are not likely to be homologous. The aligned sequence segments, while being quite long (more than 50 residues) in most cases, do not show typical patches of aligned residues indicative of functional motifs.

1.2 Profile corruption: PSI-BLAST can incorporate a non-homologous sequence or sequence segment into the position-specific score matrix (PSSM). In subsequent iterations, these sequences will promote further inclusion of non-homologous regions, causing the profile to deteriorate [25]. Domain A will get an unduly high HHsearch probability with domain B if the profile of domain A is largely corrupted by homologous sequences of domain B. Cases we identified as profile corruption share several common features: (1) the structures of the two domains are highly dissimilar (usually belong to different folds), suggesting that they are not homologous; (2) domain A is adjacent in sequence to a homolog of domain B (in 8 out of the total 10 cases, the homolog of domain B is inserted into domain A); (3) the alignment encompasses the boundary between domain A and domain B's homolog. For instance, the structurally distinct GroEL intermediate domain (SCOP ID: d1kp8a3) is inserted into the middle of the GroEL equatorial domain (SCOP ID: d1kp8a1) [26], and thus both profiles are easily corrupted by each other.

1.3 Artifact in sequence: non-homologous pairs can get high HHsearch probabilities just because of the inclusion of expression vector sequence. For instance, in one case, both domains are flanked by "GSSGSSG" at the N-terminus and "SGPSSG" at the C-terminus. Such contaminant sequences are aligned in HHsearch, making the high HHsearch probability meaningless.

2. problem with structure or structure alignment

Four sources for structure problems have been identified:

2.1 low quality of structure: Some clearly homologous pairs are supported by pronounced sequence similarity and a structural resemblance apparent in manual observation. DaliLite, however, fails to generate high Z-scores due to two reasons: (1) the structure quality is low, e.g. low resolution X-ray or poorly defined NMR structure (in 24 cases out of 27, NMR structure is involved), as revealed by loose packing of secondary structure elements and distorted α-helices. (2) domain size is too small (50-100 residues) to substantiate a high Z-score. The two reasons usually coincide, so we did not discriminate them. An example of this category is shown in Fig. 2.[27,28]

2.2 Incompletely defined domain: rarely, the boundary of one domain is incorrectly defined in SCOP, resulting in an incomplete domain composed of only a few (3 or 4) secondary structural elements. As a result, homologous domains with significant sequence similarity produce low Dali Z-scores due to insufficient number of alignable residues. In all cases we found, it was possible to extend the boundary of these incomplete domains to cover the full range and to obtain strong structural similarity as measured by the Z-score.

2.3 Improper truncation: The N-terminal domain in a multidomain protein may be truncated as a result of false prediction of its start by gene finding algorithms. When cloned and structurally characterized, such proteins reveal distorted and partly disordered N-terminal domain structure due to the absence of some essential interactions. As a result, despite significant sequence similarity, such structures diverge from complete and well-structured

homologs. For example, the N-terminal segment of the DNA-binding protein Tfx [29] is composed of two short helices that do not appear compact (Fig. 3b). HHsearch matches this region to the C-terminal domain of Sigma factor sigma-28 [30], which is structured as an HTH-containing helical bundle (Fig. 3c). Inspection of BLAST [31] hits for DNA-binding protein Tfx sequence reveals a 100% identity match that contains additional 16 residues at the N-terminus. This longer variant is likely to represent the biological unit because the 16-residue segment matches the sequence of the N-terminus of HTH domain.

2.4 Artifact in structure: while in most cases X-ray structures represent proteins in physiological conditions, in some rare instances, probably due to experimental conditions, structural changes occur. We found a single example, namely Pleiotropic regulator of virulence genes, SarA[32] and Hypothetical protein AF2008[33]. Both domains are from the "Winged" helix DNA-binding domain superfamily, but the structure of SarA (Fig. 4b) is not similar to a typical winged helix domain and does not resemble its close homolog Staphylococcal accessory regulator A homolog, SarR[34,35] (Fig. 4d). In accord with our finding, the workers that determined the SarA structure hypothesized that without a carrier protein or sufficiently long DNA, SarA might contain anomalously folded region [36].

3. Interesting phenomena for protein evolution and biology

This category can be divided into 3 sub-categories:

3.1. Analogs adopting similar functional motif: The reason for significant sequence similarity can be either origin from a common ancestor (homologs), or convergent evolution (analogs) as a result of stringent functional requirements dictated by physics. Convincing examples of

analogy resulting in strong sequence similarity are hard to find[37,38]. They typically reveal domains with very different structures that bind metal ions or heme, as interaction with these cofactors requires a specific arrangement and chemical nature of the amino acids, causing sequence convergence. In contrast, if physical restriction on the sequence is not strict, which means there are several ways to carry out the function, it is most likely that similarity is the result of evolutionary descent. Three examples of possible sequence analogy were found and one of them is shown in Fig. 5. Both proteins are multiheme cytochromes[39,40]. Despite a short common heme-binding motif, distinct structures around the binding site and different topology argue for their convergence (analogy).

3.2. Domain swap: Well-known exchanges of equivalent structural segments between domains exist in some oligomeric structures[41,42]. Many homologous families contain examples of both swapped and unswapped domains, and the Dali Z-score for a match between the swapped and the unswapped domain is frequently low. We place domain swaps in a separate category instead of among other homologs with structural differences, since they are extensively studied and result from interactions between domains rather than evolutionary changes within a domain.

3.3. Homologs with different structures: This category includes the most interesting examples found in this work. Detectable sequence similarity in functionally important regions is reflected in local structural similarity and indicates homology. However, global structures of such homologs can be quite different due to deletions, insertions and significant structural rearrangements [5,43]. For instance, the P-loop (Walker A) motif was found by HHsearch in proteins from three SCOP folds: P-loop containing nucleoside triphosphate hydrolases, PEP

carboxykinase-like, and MurCD N-terminal domain. Because of significant structural differences between these domains, they are placed in different SCOP folds, but statistically supported sequence similarity suggests homology. Most of the 39 cases belong to this category (listed in table II) have been noticed before, and several interesting and novel examples are described here.

4. Novel examples of homologs with different structures

4.1. Ataxin-1 AXH and DnaB intein domains

Ataxin-1 AXH domain [44] and DnaB intein domain[45] represent the SCOP superfamilies of "AXH domain" and "Hedgehog/intein domain" (Fig. 6). They are homologs and share significant sequence similarity detected by HHsearch (probability 99.8%) and PSI-BLAST. The ataxin-1 AXH domain sequence finds an intein domain (PDB ID: 1VDE) in the non-redundant database (nr) with a significant E-value (6e-6) at iteration 4. However, the structural similarity of this pair is low (Dali Z-score: 2.6). Taken together, these data may point to a fold change resulting from several evolutionary events: extension, insertion, duplication, domain swapping and circular permutation. The common region, i.e. evolutionary core inherited from their common ancestor, is composed of 6 β-strands and it is colored in rainbow in Figs. 6b and c. In the AXH domain, the evolutionary core harbors an insertion of an α-helix between β5 and β6, and is extended with an α-helix and β-strands at the N-terminus. In the intein, the core is duplicated (Fig. 6c) and forms two subdomains: β1-β6 and β1'-β6'. These two subdomains swap a β-strand (β1 and β1') with each other. The order of strands in the intein sequence is: β3'-β4'-β5'-β6'-β1-β2-β3-β4-β5-β6-β1'-β2', while the AXH domain has the order: β1-β2-β3-

β4-β5-β6. A circular permutation positioned β1'-β2' at the C-terminus in the DnaB intein after duplication.

Homology between the two domains provides functional insights. The AXH domain is essential for RNA binding [46,47], and the colored β-strands (Fig. 6b) are likely to be responsible for such interaction as evidenced by: 1) the N-terminal β-strands and α-helix are buried at the dimer interface and are not accessible to RNA; 2) the β-strand packing resembles an OB-fold, a motif presents in various oligo-nucleotide binding proteins or nucleases[44,48]. The function of the intein domain is to excise itself and rejoin the remaining segments of the host protein[49]. DnaB intein carries a homing endonuclease domain inserted between β6 and β1'[45]. In the available DnaB intein structure, the homing endonuclease domain is not included. The full DnaB intein (the intein together with homing endonuclease) can recognize specific sites in DNA, cleave the DNA and trigger double-strand break homologous recombination[50]. Being a homolog of the nucleic-acid binding AXH domain, the intein domain might also assist the endonuclease in DNA binding.

4.2. *C-terminal subdomain of CPS large subunit ATP-binding domain and Acetyl-CoA carboxylase BC-C subdomain*

The carbamoyl phosphate synthetase (CPS) large subunit ATP-binding domain[51] and the acetyl-CoA carboxylase, biotin Carboxylase C-terminal (BC-C) subdomain[52] share significant sequence similarity supported by HHsearch (probability 98.9%). The biotin carboxylase middle domain (BC-M)[52], an adjacent subdomain to the BC-C, belongs to the same SCOP family as the ATP-binding domain. The combined structure of BC-C and BC-M subdomains (Fig. 7d) is remarkably similar to the ATP-binding domain (Fig. 7b). Such

evidence suggests that the ATP-binding domain contains two subdomains, which are homologs of BC-C subdomain and BC-M subdomain, respectively.

Notably, the aligned segments (shown in the same color in Fig. 7b and 7d) seem to differ in handedness: the red α-helix, the green and yellow β-strands form a left-handed unit in the ATP-binding domain, while the unit is right-handed in the BC-C subdomain. Since handedness is unlikely to change in evolution, we hypothesize that a deletion of the green and blue β-strands in the ATP-binding domain resulted in a partially incorrect alignment by HHsearch. The green β-strand in the ATP-binding domain should be aligned to the blue β-strand in the BC-C subdomain, which keeps the handedness the same. To test this hypothesis, the C-terminal half (residue 491-527) of the aligned sequence is taken from the BC-C subdomain as a query for HHpred (the green and blue β-strands excluded). HHpred still finds the ATP-binding domain with a probability of 93.9%, indicating that the N-terminal half is not necessary to detect this similarity. Therefore the HHsearch sequence alignment in Fig. 7a was incorrectly extended towards the N-terminus because of the deletion. Compared to homologs in the same SCOP superfamily, the BC-C subdomain undergoes several deletions, making it more likely that in the ATP-binding domain, yet another deletion greatly changed the fold.

4.3. *Phthalate dioxygenase reductase C-terminal domain and Dihydroorotate dehydrogenase B, PyrK subunit*

The phthalate dioxygenase reductase (PDR) C-terminal domain [53] and the dihydroorotate dehydrogenase B, PyrK subunit [54] represent the superfamily pair of "2Fe-2S ferredoxin-like" and "Ferredoxin reducatase-like and C-terminal NADP-linked domain". They share significant sequence similarity (HHsearch probability 96.4%). On the first glance, stringent requirements

on sequence and local structure within the binding site (boxed in Fig. 8a) imposed by interactions with the 2Fe-2S cluster might suggest analogy (refer to the example of analogs in Fig. 5). However, structural evidence strongly argues for their homology: besides a similar loop that binds to the 2Fe-2S cluster, the β-strands (the red and orange β-strands in Fig. 8b and 8c) adjacent to the loop are also remarkably similar. Without apparent physical restriction on these β-strands that leads to evolutionary convergence, common ancestry is the likely explanation for this similarity.

The structural change (Dali Z-score 1.7) from the PDR C-terminal domain to the PyrK subunit might stem from a deletion of the segment in the black frame in Fig. 8b. This deletion causes a reorientation of secondary structural elements in the PyrK subunit and results in the unusual packing of β-strands shown in the black frame in Fig. 8c. Since the blue β-strand and green α-helix in the PDR C-terminal domain were possibly lost in evolution to generate the PyrK subunit, they should not be aligned to any PyrK sequence segment. Therefore, we excluded this sequence and selected only the C-terminal half (residue 266-295) of the initially aligned PDR C-terminal domain sequence (Fig. 8a) as a query for the online server of HHsearch, HHpred. The PyrK subunit was found with a probability of 97.8%, implying that the blue β-strand and the green α-helix do not significantly contribute to the sequence alignment and that they might be absent in PyrK subunit due to deletion.

conclusions

CONCLUSIONS

On the one hand, the dataset of structurally different proteins with strong sequence similarity is plagued with various technical problems, which encompass over half of representative domain pairs and make the examination a tedious task. These problems arise at all stages, from experiment (genetic construct, structure determination) to data processing (generating PDB file and SCOP domain) and data analysis (profile, alignment, structure superposition). On the other hand, careful investigation reveals interesting examples of homologs with distinct structures and advances our understanding of protein evolution. We see that insertions, extensions, and duplications decorate and expand the evolutionary core; deletions reduce the core, sometimes beyond recognition, potentially resulting in reorientation of structural elements. Topology and mutual arrangement of secondary structures may change due to circular permutation or domain swapping. Finally, combination of several such events makes for the largest structural differences between homologs.

TABLE I. SUMMARY OF SYSTEMATICAL STUDY OF REPRESENTATIVES

Category	*representatives*	*domain pairs*
similar secondary structure pattern	21	40
profile corruption	10	54
artifact in sequence	2	2
Total of problem with sequence	34	111
low quality of structure	27	143
incomplete domain	4	22
improper truncation	2	3
artifact in structure	1	7
Total of problem with structure	34	175
homolog with different structures	39	1441
domain swap	6	42
analog adopt the same functional motif	3	31
Total of case with biological meaning	48	1514
complicated case	5	19
Total of all	120	1804

* For each category listed in the first column, the number of superfamily representatives in the category is listed in the second column. The number of domain pairs in each category is deduced from the number of domain pairs that each superfamily representative stands for.

TABLE II. SUMMARY OF HOMOLOGOUS SUPERFAMILY PAIRS

SCOP superfamily 2	*SCOP superfamily 1*	*Case*
P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases	1121
EF-hand	EF-hand	106
P-loop containing nucleoside triphosphate	PEP carboxykinase-like hydrolases	72
Immunoglobulin	Immunoglobulin	35
NAD(P)-binding Rossmann-fold domains	FAD/NAD(P)-binding domain	17
Cytochrome c	Cytochrome c	16
lambda repressor-like DNA-binding domains	Homeodomain-like	9
Prokaryotic type KH domain (KH-domain type II)	Eukaryotic type KH-domain (KH-domain type I)	7
Hedgehog/intein (Hint) domain	AXH domain	5
ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine	Sporulation response regulatory protein Spo0B	5
Ferredoxin reductase-like, C-terminal NADP-linked domain	2Fe-2S ferredoxin-like	4
S-adenosyl-L-methionine-dependent methyltransferases	S-adenosyl-L-methionine-dependent methyltransferases	4
Chaperone J-domain	Chaperone J-domain	3
"Winged helix" DNA-binding domain	Homeodomain-like	3
5' to 3' exonuclease, C-terminal subdomain	RuvA domain 2-like	3
NAD(P)-binding Rossmann-fold domains	Nucleotide-binding domain	3
lambda repressor-like DNA-binding domains	"Winged helix" DNA-binding domain	2
Pyrimidine nucleoside phosphorylase C-terminal domain	Single hybrid motif	2
Glutathione synthetase ATP-binding domain-like	Rudiment single hybrid motif	2
SET domain	SET domain	2
Rhodanese/Cell cycle control phosphatase	(Phosphotyrosine protein) phosphatases II	2
lambda repressor-like DNA-binding domains	lambda repressor-like DNA-binding domains	1
Homeodomain-like	ARID-like	1
Putative DNA-binding domain	"Winged helix" DNA-binding domain	1
RuvA domain 2-like	Rad51 N-terminal domain-like	1
RuvA domain 2-like	DNA-glycosylase	1
5' to 3' exonuclease, C-terminal subdomain	Rad51 N-terminal domain-like	1
C-terminal (heme d1) domain of cytochrome cd1-nitrite reductase	GyrA/ParC C-terminal domain-like	1
GyrA/ParC C-terminal domain-like	DPP6 N-terminal domain-like	1
GyrA/ParC C- terminal domain-like	WD40 repeat-like	1
GyrA/ParC C-terminal domain-like	RCC1/BLIP-II	1
GyrA/ParC C-terminal domain-like	Tricorn protease domain 2	1
Single hybrid motif	Single hybrid motif	1
Rudiment single hybrid motif	Rudiment single hybrid motif	1
NAD(P)-binding Rossmann-fold domains	NAD(P)-binding Rossmann-fold domains	1
Tryptophan synthase beta subunit-like PLP-dependent enzymes	NAD(P)-binding Rossmann-fold domains	1
PreATP-grasp domain	FAD/NAD(P)-binding domain	1
P-loop containing nucleoside triphosphate hydrolases	MurCD N-terminal domain	1
Nucleotide-binding domain	FAD/NAD(P)-binding domain	1

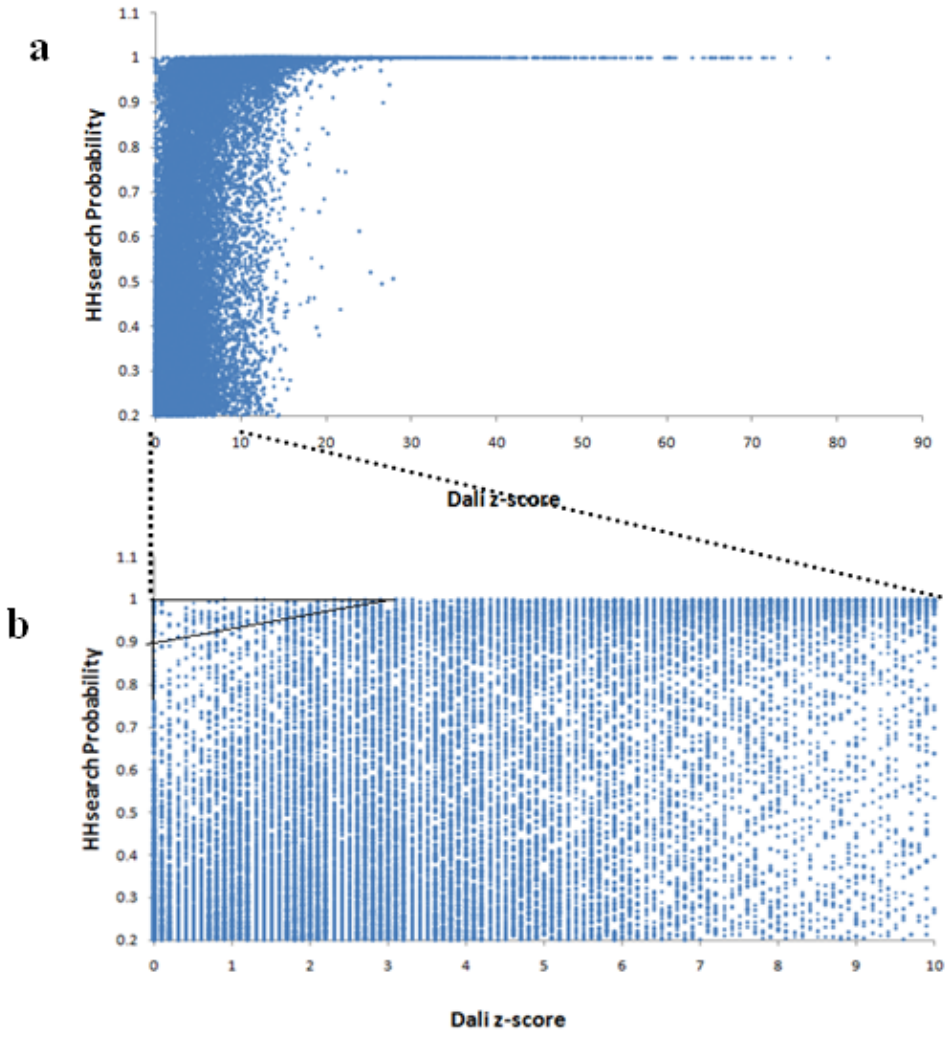


Figure 1. Distribution of HHsearch probability and Dali Z-score of a randomly selected sample: **a.** A sample of 28430 domain pairs with HHsearch probability higher than 0.2. **b.** Enlarged version of a subregion (Dali Z-score 0-10) in a. The triangle area in black frame is selected for study.

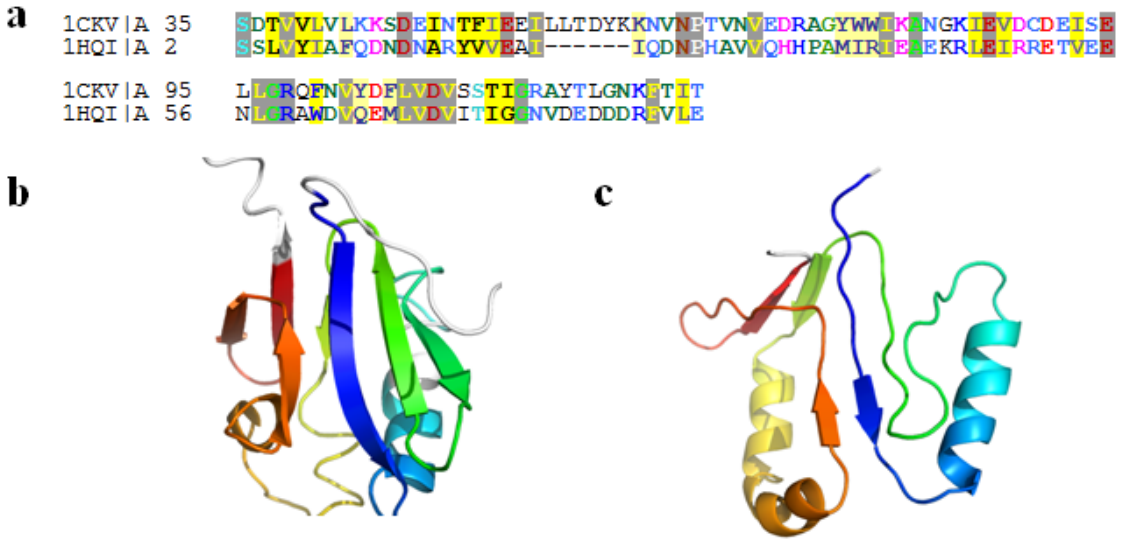


Figure 2. Low quality of structure: **a.** sequence alignment of Soluble methane monooxygenase regulatory protein B (SCOP ID: d1ckva_) and Phenol hydroxylase P2 protein (SCOP ID: d1hqia_) by HHsearch. The sequence alignment is colored according to similarity between residues using Chroma: briefly, identical residues are on grey background; aligned hydrophobic residues are highlighted yellow; if aligned residues share similarities in their charge or size, they are shown in color other than black. **b.** structure of Soluble methane monooxygenase regulatory protein B determined by NMR; **c.** structure of Phenol hydroxylase P2 protein determined by NMR. (the segment shown in sequence alignment above is colored in rainbow in both b. and c.)

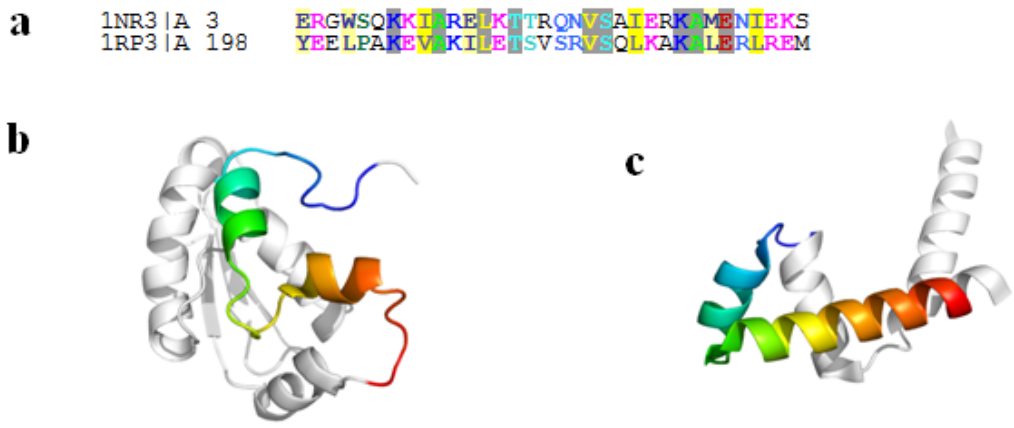


Figure 3. Improper truncation: **a.** sequence alignment (colored according to similarity between residues) of DNA-binding protein Tfx (SCOP ID: d1nr3a_) to C-terminal domain of Sigma factor sigma-28 (SCOP ID: d1rp3a2) by HHsearch. **b.** structure of DNA-binding protein Tfx **c.** structure of C-terminal domain of Sigma factor sigma-28 (the segment shown in sequence alignment above is colored in rainbow in both b. and c.)

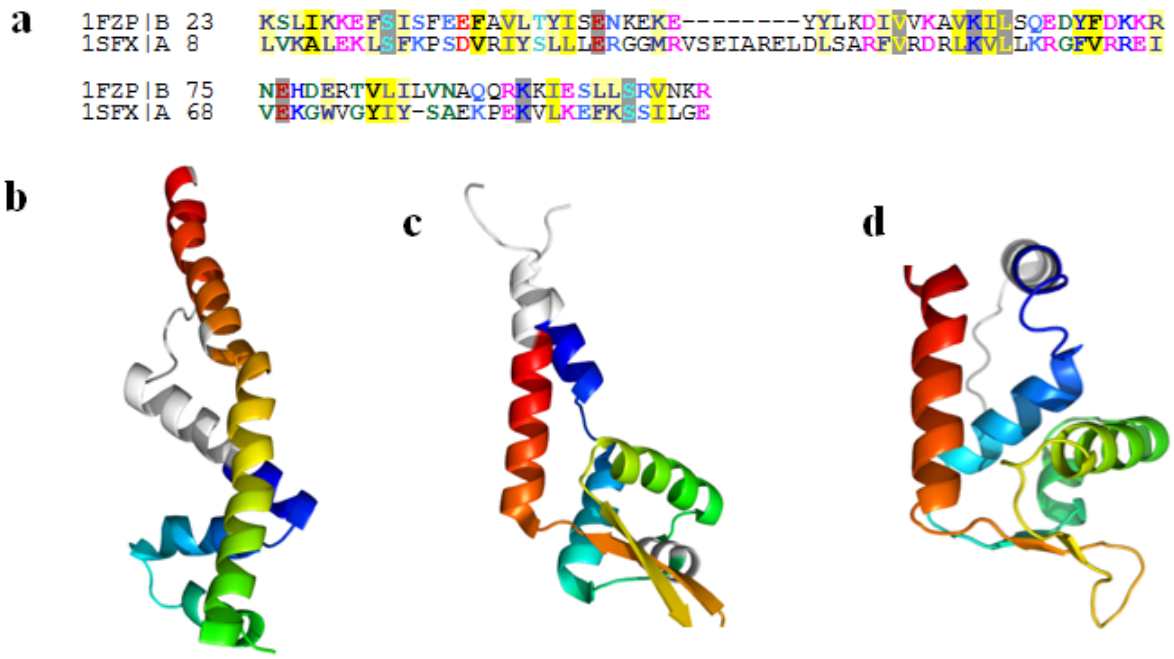


Figure 4. Artifact in structure: **a.** sequence alignment (colored according to similarity between residues) of Pleiotropic regulator of virulence genes, SarA (SCOP ID: d1fzpb_) and Hypothetical protein AF2008 (SCOP ID: d1sfxa_) by HHsearch. **b.** structure of Pleiotropic regulator of virulence genes **c.** structure of Hypothetical protein AF2008. (The segment shown in sequence alignment above is colored in rainbow in b. and c.) **d.** structure of Staphylococcal accessory regulator A homolog (SarR, SCOP ID: d1hsja1, close homolog of SarA) and the structure is colored in rainbow according to its sequence alignment to SarA.

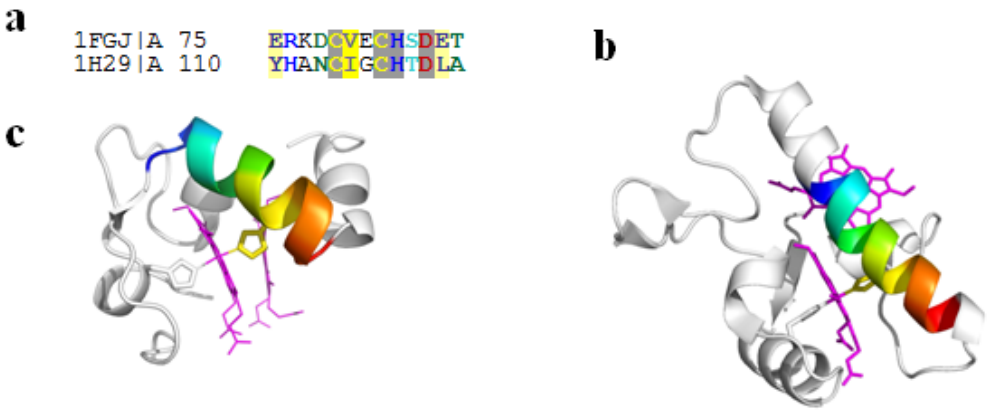


Figure 5. Analogs adopting similar functional motif: **a.** sequence alignment (colored according to similarity between residues) of Hydroxylamine oxidoreductase (SCOP ID: d1fgja_) to 16-heme cytochrome c HmcA (SCOP ID: d1h29a_) by HHsearch. **b.** structure of a segment (residue 65-99, 144-166) of Hydroxylamine oxidoreductase. **c.** structure of a segment (residue 60-140) of 16-heme cytochrome c HmcA. (the segment shown in sequence alignment above is colored in rainbow and hemes are colored in purple in both b. and c.)

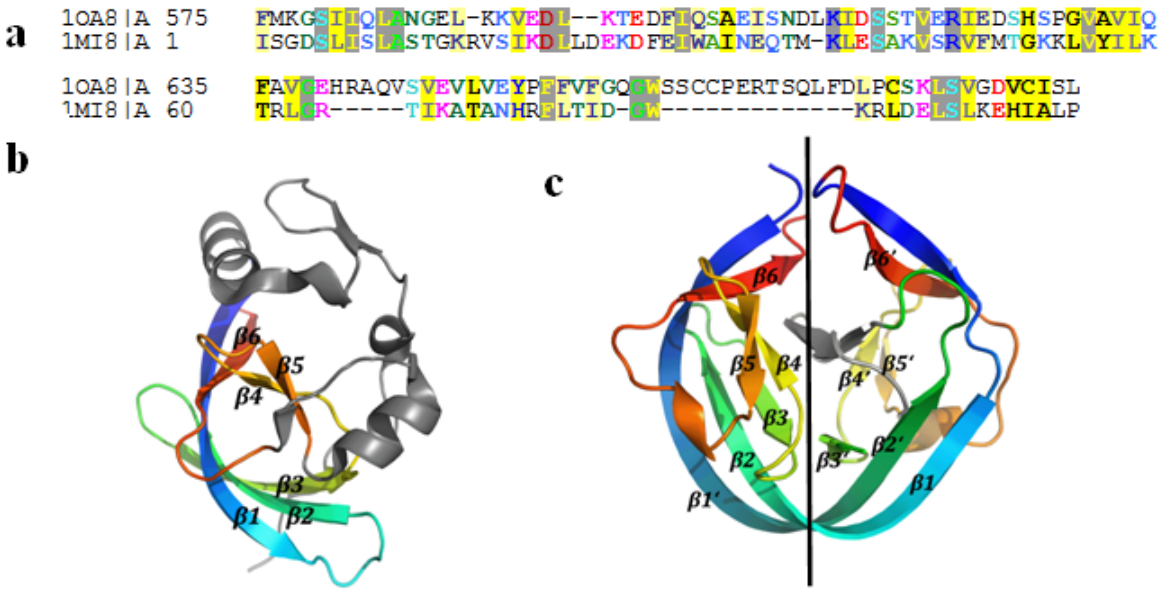


Figure 6. Ataxin-1 AXH domain and DnaB intein domain: **a.** sequence alignment (colored according to similarity between residues) of Ataxin-1 AXH domain (SCOP ID: d1oa8a_) and DnaB intein domain (SCOP ID: d1mi8a_) by HHsearch. **b.** structure of Ataxin-1 AXH domain; the 6 β-strands from the evolutionary core are colored in rainbow and labeled with β1-β6 from N- to C-terminus. **c.** structure of DnaB intein domain; the evolutionary core is colored and labeled with β1-β6 and β1'-β6' according to the structure of AXH domain in b. so that the corresponding secondary structure elements between two structures have the same color and label.

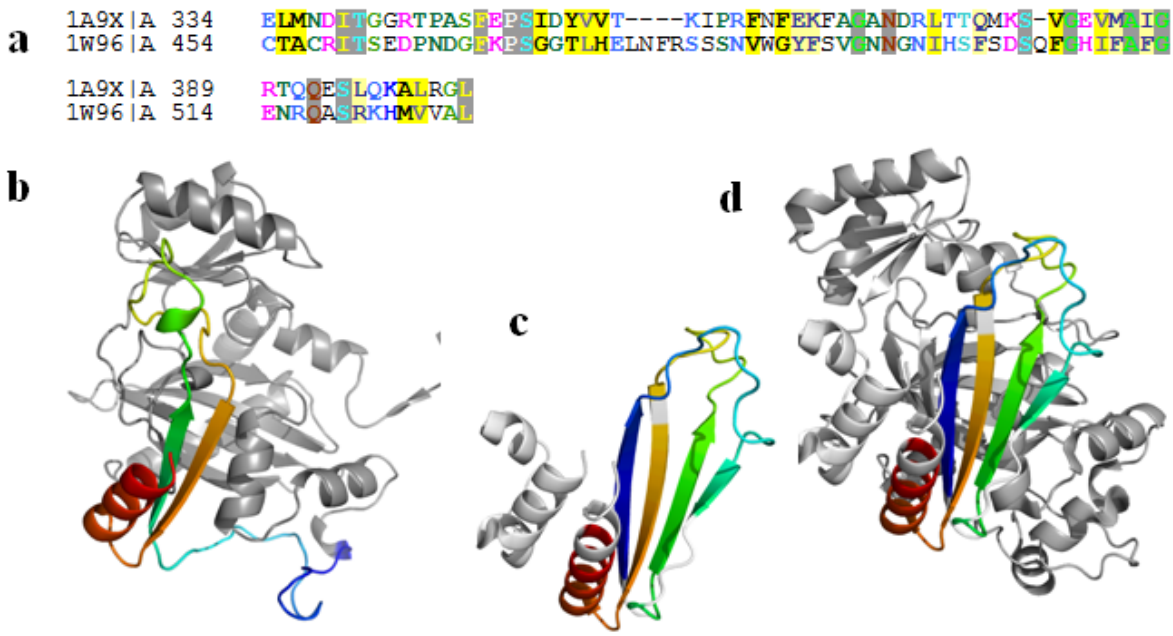


Figure 7. CPS large subunit ATP-binding domain and Acetyl-CoA carboxylase BC-C subdomain: **a.** sequence alignment (colored according to similarity between residues) of CPS large subunit ATP-binding domain (SCOP ID: d1a9xa5) and Acetyl-CoA carboxylase BC-C subdomain (SCOP ID: d1w96a1) by HHsearch. **b.** structure of CPS large subunit ATP-binding domain. **c.** structure of Acetyl-CoA carboxylase BC-C subdomain. (the segment shown in sequence alignment above is colored in rainbow in both b. and c.) **d.** combined structure of Acetyl-CoA carboxylase BC-C subdomain and BC-M subdomain (SCOP ID: d1w96a3)

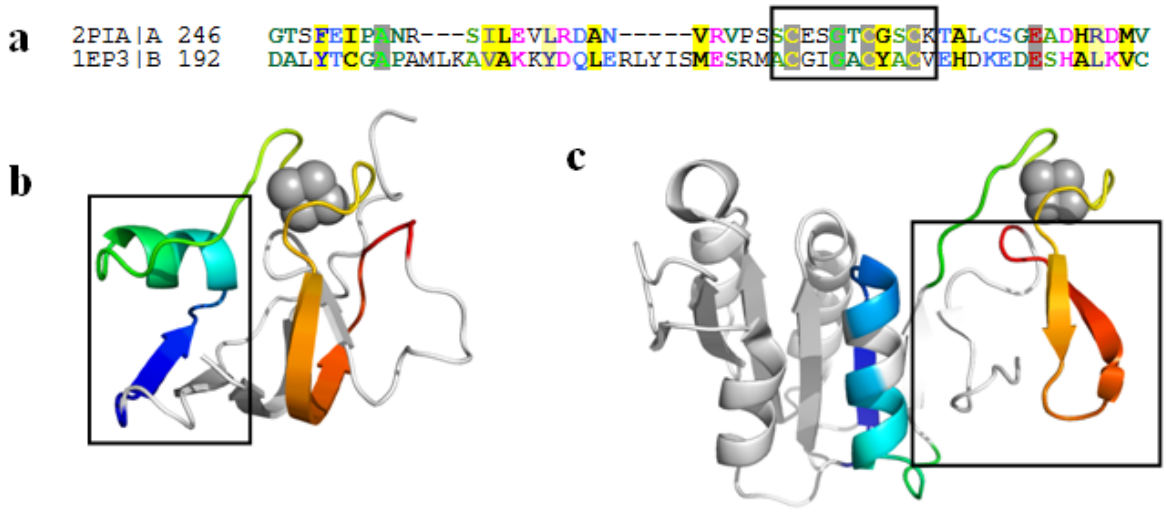


Figure 8. PDR C-terminal domain and Dihydroorotate dehydrogenase B, PyrK subunit: **a.** sequence alignment (colored according to similarity between residues) of PDR C-terminal domain (SCOP ID: d2piaa3) and Dihydroorotate dehydrogenase B, PyrK subunit (SCOP ID: d1ep3b2) by HHsearch. **b.** structure of PDR C-terminal domain and 2Fe-2s cluster is shown as gray balls. **c.** structure of PyrK subunit. (the segment shown in sequence alignment above is colored in rainbow in both b. and c.)

REFERENCES

[1] C. Chothia and A.M. Lesk, "The relation between the divergence of sequence and structure in proteins", EMBO J. 5(4): 823–826, April 1986.

[2] A.N. Lupas, K.K. Koretke, Evolution of Protein Folds. In *Computational Structural Biology: Methods and Applications*. Edited by M. Pitsch, T. Schwede, N.J. Hackensack, World Scientific; 2008:131-152.

[3] J. Skolnick, A.K. Arakaki, S.Y.Lee, M. Brylinski, "The continuity of protein structure space is an intrinsic property of proteins", Proc Natl Acad Sci U S A. 106(37):15690-15695, Sep. 2009.

[4] Y. Zhang, I.A. Hubner, A.K. Arakaki, E. Shakhnovich, J. Skolnick, "On the origin and highly likely completeness of single-domain protein structures", Proc Natl Acad Sci U S A. 103(8): 2605–2610, Feb. 2006.

[5] N.V. Grishin, "Fold change in evolution of protein structures", J Struct Biol. 134(2-3):167-185, May-Jun. 2001.

[6] A. Andreeva, A.G. Murzin, "Evolution of protein fold in the presence of functional constraints", Curr Opin Struct Biol. 16(3):399-408, Jun. 2006.

[7] X.I. Ambroggio, B. Kuhlman, "Design of protein conformational switches", Curr Opin Struct Biol. 16(4):525-530, Aug. 2006

[8] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures", J Mol Biol. 247(4):536-540, Apr. 1995

[9] A. Andreeva, D. Howorth, J.M. Chandonia, S.E. Brenner, J.T. Hubbard, C. Chothia, A.G. Murzin, “Data growth and its impact on the SCOP database: new developments”, Nucleic Acids Res. 36:D419-425, Jan. 2008

[10] J.M. Chandonia, G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, S.E. Brenner, “The ASTRAL Compendium in 2004”, Nucleic Acids Res. 32:D189-192, Jan. 2004.

[11] J. Söding, “Protein homology detection by HMM-HMM comparison”, Bioinformatics. 21(7):951-960. Apr. 2005.

[12] L. Holm, J. Park, “DaliLite workbench for protein structure comparison”, Bioinformatics. 16(6):566-7,Jun. 2000.

[13] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”, Nucleic Acids Res. 25(17):3389-3402, Sep. 1997.

[14] J. Söding, A. Biegert, A.N. Lupas, “The HHpred interactive server for protein homology detection and structure prediction”, Nucleic Acids Res. 33: W244-248, Jul. 2005.

[15] N. Deshpande, K.J. Addess, W.F. Bluhm, J.C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, R.K. Green, J.L. Flippen-Anderson, J. Westbrook, H.M. Berman, P.E. Bourne, “The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema”, Nucleic Acids Res. 33:D233-7, Jan. 2005.

[16] J.F. Gibrat, T. Madej, S.H. Bryant, “Surprising similarities in structure comparison”, Curr Opin Struct Biol. 6(3):377-385, Jun. 1996.

[17] G.P.Raghava, G.J.Barton, “Quantification of the variation in percentage identity for protein sequence alignments”, BMC Bioinformatics, 19(7):415, Sep., 2006

[18] D.D.Leipe, Y.I.Wolf, E.V.Koonin, L.Aravind,“Classification and evolution of P-loop GTPases and related ATPases”,J Mol Biol.317(1):41-72, Mar. 2002.

[19] L.M. Iyer, D.D. Leipe, E.V. Koonin, L. Aravind, “Evolutionary history and higher order classification of AAA+ ATPases”, J Struct Biol. 146(1-2):11-31, Apr-May. 2004.

[20] D.D. Leipe, E.V. Koonin, L. Aravind, “Evolution and classification of P-loop kinases and related proteins”, J Mol Biol. 333(4):781-815, Oct. 2003.

[21] S. Cheek, H. Zhang, N.V. Grishin, “Sequence and structure classification of kinases”, J Mol Biol. 320(4):855-881, Jul. 2002.

[22] A. Street, S. Mayo, “Intrinsic beta-sheet propensities result from van der waals interactions between side chains and the local backbone”, Proc Natl Acad Sci USA. 96:9074–9076, Aug. 1999.

[23] T.P. Creamer, G.D. Rose, “Alpha-helix-forming propensities in peptides and proteins” Proteins 19:85–97, Jun. 1994.

[24] G. Bellesia, A.I. Jewett, J.E. Shea, “Sequence periodicity and secondary structure propensity in model proteins”, Protein Sci. 19(1):141-154, Jan. 2010.

[25] M.M. Lee, M.K. Chan, R. Bundschun, “Simple is beautiful: a straightforward approach to improve the delineation of true and false positives in PSI-BLAST searches”, Bioinformatics. 24(11):1339-1343, Jun. 2008.

[26] J. Wang, D.C. Boisvert, "Structural basis for GroEL-assisted protein folding from the crystal structure of (GroEL-KMgATP)14 at 2.0A resolution", J Mol Biol. 327(4):843-855, Apr. 2003.

[27] H. Qian, U. Edlund, J. Powlowski, V. Shingler, I. Sethson, "Solution structure of phenol hydroxylase protein component P2 determined by NMR spectroscopy", Biochemistry. 36(3):495-504, Jan. 1997.

[28] K.J. Walters, G.T. Gassner, S.J. Lippard, G. Wagner, "Structure of the soluble methane monooxygenase regulatory protein B", Proc Natl Acad Sci U S A. 96(14):7877-7882, Jul. 1999.

[29] Y. Shen, G. Liu, R. Bhaskaran, A. Yee, C. Arrowsmith, T. Szyperski, "Solution structure of the protein MTH0916: the northeast structural genomics consortium target TT212", to be published.

[30] M.K. Sorenson, S.S. Ray, S.A. Darst, "Crystal structure of the flagellar sigma/anti-sigma complex sigma(28)/FlgM reveals an intact sigma factor in an inactive conformation", Mol Cell. 14(1):127-138, Apr. 2004.

[31] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, "Basic local alignment search tool", J Mol Biol. 215(3):403-410, Oct. 1990.

[32] M.A. Schumacher, B.K. Hurlburt, R.G. Brennan, "Crystal structures of SarA, a pleiotropic regulator of virulence genes in S. aureus", Nature. 409(6817):215-219, Jan. 2001.

[33] J. Osipiuk, T. Skarina, A. Savchenko, A. Edwards, M. Cymborowski, W. Minor, A. Joachimiak, "X-ray crystal structure of putative HTH transcription regulator from Archaeoglobus fulgidus", to be published.

[34] Y. Liu, A. Manna, R. Li, W.E. Martin, R.C. Murphy, A.L. Cheung, G. Zhang, "Crystal structure of the SarR protein from Staphylococcus aureus", Proc Natl Acad Sci U S A. 98(12):6877-6882, Jun. 2001.

[35] M. Gao, J. Skolnick, "DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions", Nucleic Acids Res. 36(12):3978-3992, Jul.,2008.

[36] M.A. Schumacher, B.K. Hurlburt, R.G. Brennan, "Crystal structures of SarA, a pleiotropic regulator of virulence genes in S. aureus", Nature. 414(6859):85, Nov. 2001.

[37] R.F. Doolittle, "Similar amino acid sequences: chance or common ancestry?", Science. 214(4517):149-159, Oct. 1981.

[38] S.S. Krishna, R.I. Sadreyev, N.V. Grishin, "A tale of two ferredoxins: sequence similarity and structural differences", BMC Struct Biol. 6:8, Apr. 2006.

[39] N. Igarashi, H. Moriyama, T. Fujiwara, Y. Fukumori, N. Tanaka, "The 2.8 A structure of hydroxylamine oxidoreductase from a nitrifying chemoautotrophic bacterium, Nitrosomonas europaea", Nat Struct Biol. 4(4):276-284, Apr. 1997.

[40] P.M. Matias, A.V. Coelho, F.M. Valente, D. Plácido, J. LeGall, A.V. Xavier, I.A. Pereira, M.A. Carrondo, "Sulfate respiration in Desulfovibrio vulgaris Hildenborough. Structure of the 16-heme cytochrome c HmcA AT 2.5-A resolution and a view of its role in transmembrane electron transfer", J Biol Chem. 277(49):47907-16, Dec. 2002.

[41] A.M. Gronenborn, "Protein acrobatics in pairs--dimerization via domain swapping", Curr Opin Struct Biol. 19(1):39-49, Feb. 2009.

[42] Y. Liu, D. Eisenberg, "3D domain swapping: as domains continue to swap", Protein Sci. 11(6):1285-1299, Jun. 2002

[43] A. Andreeva, A.G. Murzin, “Evolution of protein fold in the presence of functional constraints”, Curr Opin Struct Biol. 16(3):399-408, Jun. 2006.

[44] Y.W. Chen, M.D. Allen, D.B. Veprintsev, J. Löwe, M. Bycroft, “The structure of the AXH domain of spinocerebellar ataxin-1”, J Biol Chem. 279(5):3758-3765, 2004 Jan. 2004.

[45] Y. Ding, M.Q. Xu, I. Ghosh, X. Chen, S. Ferrandon, G. Lesage, Z. Rao, “Crystal structure of a mini-intein reveals a conserved catalytic module involved in side chain cyclization of asparagine during protein splicing”, J Biol Chem. 278(40):39133-39142, Oct. 2003.

[46] S. Yue, H.G. Serra, H.Y. Zoghbi, H.T. Orr, “The spinocerebellar ataxia type 1 protein, ataxin-1, has RNA-binding activity that is inversely affected by the length of its polyglutamine tract”, Hum Mol Genet. 10:25–30, Jan. 2001.

[47] A. Matilla-Dueñas, R. Goold, P. Giunti, “Clinical, genetic, molecular, and pathophysiological insights into spinocerebellar ataxia type 1”, Cerebellum.7(2):106-114, 2008.

[48] D.L. Theobald, R.M. Mitton-Fry, D.S. Wuttke, “Nucleic acid recognition by OB-fold proteins”, Annu Rev Biophys Biomol Struct. 32:115-133, 2003.

[49] L. Saleh, F.B. Perler, “Protein splicing in cis and in trans”, Chem Rec. 6(4):183-193, 2006.

[50] B.L. Stoddard, “Homing endonuclease structure and function”, Q Rev Biophys. 38(1):49-95, Feb. 2005.

[51] J.B. Thoden, S.G. Miran, J.C. Phillips, A.J. Howard, F.M. Raushel, H.M. Holden, “Carbamoyl phosphate synthetase: caught in the act of glutamine hydrolysis” Biochemistry. 37(25):8825-8831, Jun. 1998.

[52] Y. Shen, S.L. Volrath, S.C. Weatherly, T.D. Elich, L. Tong, “A mechanism for the potent inhibition of eukaryotic acetyl-coenzyme A carboxylase by soraphen A, a macrocyclic polyketide natural product”, Mol Cell. 16(6):881-891, Dec. 2004.

[53] C.C. Correll, C.J. Batie, D.P. Ballou, M.L. Ludwig, “Phthalate dioxygenase reductase: a modular structure for electron transfer from pyridine nucleotides to [2Fe-2S]”, Science. 258(5088):1604-1610, Dec. 1992.

[54] P. Rowland, S. Nørager, K.F. Jensen, S. Larsen, “Structure of dihydroorotate dehydrogenase B: electron transfer between two flavin groups bridged by an iron-sulphur cluster”, Structure. 8(12):1227-1238, Dec. 2000.

CHAPTER THREE
An automatic method for CASP9 free modeling structure prediction assessment

INTRODUCTION

Critical Assessment of protein Structure Prediction (CASP), is an experiment running for 16 years that has been absolutely critical for evaluating progress (or lack of thereof) in prediction, spotting and encouraging most successful methods and stimulating discussions in the field of structure prediction (Kryshtafovych *et al.,* 2005; Moult, 2006; Moult *et al*., 2009). For each biannual CASP prediction period, organizers collect sequences with 3D structures in the works and release them to predictors; predictors deliver structure models and assessors critically evaluate the quality of predictions after the experimental structures have been determined. By separating the process of prediction and assessment, CASP provides an objective basis for comprehensive evaluation of models (Moult *et al*., 1995).

Based on the availability of structural templates and the prediction difficulty, targets in CASP are currently divided into two categories: Template-Based Modeling (TBM) and Free Modeling (FM) (Kinch *et al*., 2011a). Without an easily detectable template, targets in the FM category are the most challenging and predicted models are usually of low quality. FM category models are traditionally evaluated by manual inspection (Ben-David *et al*., 2009; Jauch *et al*., 2007; Tai *et al*., 2005) because well-established structure comparison measures, such as RMSD or even GDT-TS may miss promising models (Jauch *et al*., 2007). For instance, GDT-like scores may emphasize on small but precisely modeled substructure (such as a long

α-helix) rather than decent general fold and topology. However, model evaluation by human experts is subjective and time-consuming, and it is impossible to carefully examine all the models within the time frame of a CASP experiment. A practical compromise (Ben-David *et al*., 2009; Jauch *et al*., 2007; Tai *et al*., 2005; Aloy *et al*., 2003) is to limit manual inspection to the top models selected by a scoring system (e.g. GDT_TS). However, this initial selection biases final results. To avoid the bias, recent CASP assessors utilized additional scores (e.g. Cα-Cα contacts or distances) to select candidates for visual inspection. Combination of different methods lowers the probability of missing reasonable models and improves the evaluation of structure prediction.

As the assessors of the CASP9 FM category, we introduced a novel automatic structure prediction assessment method named Quality Control Score (QCS). We suggest that the score is particularly useful to compare poor predictions. QCS reflects our manual evaluation experience and aims to capture global features of models defined by mutual arrangement of secondary structure elements (SSEs). Inter-residue contact component is included in QCS as to quantify the accuracy of modeling atomic details. Overall, QCS is in agreement with manual inspection and correlates well with GDT-TS. However, QCS can reveal models with better global topology that are missed by GDT-TS. QCS is not only suitable to select candidates for manual inspection in the CASP assessment, but can be used as an independent and subjective method to assess the quality of structure prediction with emphasis on the global topology. Moreover, QCS can be expanded as a fold comparison tool and applied to remote homology inference and protein fold classification.

METHODS

CASP9 targets and models were downloaded from the prediction center web site (http://predictioncenter.org/). Representative evaluation units (T0531, T0534 domains 1 and 2, T0537, T0550 domains 1 and 2, T0561, T0578, T0581, T0604 domain 1, T0608 domain 1, T0618, T0621, T0624) from the CASP9 FM category (Kinch *et al*., 2011a) were assessed by manual inspection during CASP9 season. Briefly, for each target, a set of criteria (points) was developed based on the target structural features, including the size and orientation of SSEs, key contacts between SSEs, and any additional unusual structural features such as kink in the helix. Models were visually compared to the targets to evaluate whether the model agrees with the target on these criteria (points) without superposition. Expert Manual Inspection Scores (MISs) were recorded as a percentage of the maximum points assigned to each target (Kinch *et al*., 2011b).

Building on the experience in manual assessment, QCS (details described in *RESULTS AND DISCUSSION*) focuses on global features of models on the basis of SSEs (the SSE length, the relative position, angle and key interactions between SSE pairs and the handedness of the structure). To discriminate the local structure details between models, all inter-residue contacts were assessed as well.

All the evaluation units from CASP9 FM category (a total of 29 protein domains) were assessed by QCS, GDT-TS (Zemla *et al*., 1999b, 2001, 2003), CS, TenS (a consensus-based method used in CASP5 and CASP9), TM-align, Mammoth and SOV. To test QCS on easier targets, the 21 single domain TBM targets were assessed by both QCS and GDT-TS.

The performance of QCS was first examined on the subset of FM models that were assigned nonzero MISs. The agreement between QCS and MIS was investigated and compared with other automatic methods by the general correlation and the overlap in top models. Comparison between QCS and other similarity scores was then carried out on all CASP9 FM targets and TBM representatives by investigating correlation and visually comparing the top models selected by various methods. Finally, we tested QCS on the Template Free Modeling category targets from CASP7 and CASP8 and compared the results to those obtained by previous assessors.

RESULTS AND DISCUSSION

Components of QCS

QCS calculation uses only Cα atoms and it relies heavily on SSEs that defines protein's architecture and topology. We used PALSSE (Majumdar *et al*., 2005), a sensitive secondary structure assignment program to define SSEs from the target 3D coordinates (Fig. 1A) and propagated these SSE definitions to models (Fig. 1C) by residue numbers. Thus, the target and the model were simplified to a set of SSE vectors (Fig. 1B and 1D). Several features were compared between them, and scores were assigned for each feature.

3.2.1 The length of SSE vectors

As we propagated the SSE definition from a target to models, we expect the length of a certain SSE in the model to agree with that in the target if the secondary structures of residues are modeled correctly. The SSE lengths in the model ($L_i(M)$, M indicates SSEs or measurements in the model) and in the target ($L_i(T)$, T represents SSEs or measurement in the target) were

used to calculate a length score ($s_{Length}(i)$, Eq. 1) for SSE *i*. The average length score over all SSEs weighted by number of residues in each SSE (Eq. 2) was applied to assess the secondary structure quality.

$$s_{Length}(i) = \exp\{-\ln 2[L_i(M) - L_i(T)/0.25 \times L_i(T)]^2\} \quad (1)$$

$$S_L = \sum_i w_i \times s_{Length}(i) \Big/ \sum_i w_i \quad (2)$$

3.2.2 The global position of SSEs

The position of SSEs was evaluated by their pairwise distances and the distances were measured in two ways. In the first SSE position measurement ($S1_P$), each SSE was divided into three equal segments and reduced to three points by averaging Cα coordinates. Position scores were assigned by comparing the distances between all these points (*i* and *j*, except points within one SSE) in the target ($D_{i,j}(T)$) and in the model ($D_{i,j}(M)$) (Eq. 3). In this measurement only models with correct alignment would be favored, as SSEs were defined only in the target and the definitions were propagated by residue numbers to the model. This meaningful dependence on correct alignment might over-penalize models based on correct template but erroneous alignment. To balance this effect, we introduced the second SSE position measurement ($S2_P$) that is less sensitive to shifts in alignment. We compared closest Cα distances between SSEs *i* and *j* in the model ($D_{i,j}(M)$) and in the target ($D_{i,j.}(T)$) to assess their relative positions ($s2_{Position}(i,j)$, Eq. 5) Combining these two scoring functions resulted in a balance between rewarding reasonable structure traces (templates) and high quality of alignment.

$$s1_{Position}(i,j) = \exp\{-\ln 2[D_{i,j}(M) - D_{i,j}(T)/0.5 \times D_{i,j}(T)]^2\}$$

(3)

$$S1_P = \sum_{i,j} w_{i,j} \times s1_{Position}(i,j) \Big/ \sum_{i,j} w_{i,j} \quad (4)$$

$$s2_{Position}(i,j) = \exp\{-\ln 2[D_{i,j}(M) - D_{i,j}(T)/0.5 \times D_{i,j}(T)]^2\} \quad (5)$$

$$S2_P = \sum_{i,j} w_{i,j} \times s2_{Position}(i,j) \Big/ \sum_{i,j} w_{i,j} \quad (6)$$

$$S_P = (S1_P + S2_P)/2 \quad (7)$$

3.2.3 The angle between SSE vectors

To assess angle between SSEs *i* and *j*, we transformed the 3D coordinates of the model so that one SSE vector (*i(M)*) is aligned in direction to the corresponding vector (*i(T)*) in the target and the centers of other two SSE vectors (*j(M)* and *j(T)*) are superimposed. After the transformation, the angle ($A_{i,j}(M,T)$) between *j(M)* and *j(T)* (illustrated in Fig. 2A) represents the discrepancy in angle. An angle score $s_{Angle}(i,j)$ was thus assigned as shown in Eq. 8. The average of angle scores over all SSE pairs, weighted by the residue numbers of the pair of SSEs (*Ni* and *Nj*) and the distance between central part of the two SSEs (*Di,j*) (Eq. 9 and Eq. 10) was taken to evaluate the accuracy of the packing angles between SSEs.

$$s_{Angle}(i,j) = \exp\{-\ln 2[A_{i,j}(M,T)/0.7]^2\} \quad (8)$$

$$w_{i,j} = N_i N_j / D_{i,j}$$

(9)

$$S_A = \sum_{i,j} w_{i,j} \times s_{Angle}(i,j) \Big/ \sum_{i,j} w_{i,j} \quad (10)$$

3.2.4 Handedness

When more than two SSEs are considered, handedness is the key to distinguish correct topology. Handedness defines the position of a third SSE (*k*) in relative to the plane specified by two reference SSEs (*i* and *j*). Fig. 2B explains our quantification of handedness. Handedness can be clearly defined when *k(M)* and *k(T)* are not very close to the reference plane. Moreover, when the reference SSEs are far from each other, reversal of handedness should not be penalized as much as when the reference vectors are directly interacting. Based on these considerations, we designed the handedness score as in Eq. 11, where the penalty negatively correlates with the distance ($D_{i,j}(T)$) between *i(T)* and *j(T)* and positively correlates with the shorter distance between *k(T)* or *k(M)* and the reference plane.

$$s_{Hand}(i,j,k) = 1 - 2\min(D_{k,P}(M), D_{k,P}(T)) \big/ D_{i,j}(T) \quad (11)$$

$$S_H = \sum_{i,j,k} w_{i,j,k} \times s_{Hand}(i,j,k) \Big/ \sum_{i,j,k} w_{i,j,k} \quad (12)$$

3.2.5 The interaction between SSE vectors

Interactions between SSEs *i* and *j* were represented by the closest pair of residues with distance below 8.5 Å as a cutoff. Interacting residue pairs defined in the target (or certain model) were propagated to the model (or the target) by residue number. The distances between these

interacting residue pairs could be different in the model from those in the target, resulting from either missing correct interaction or forming incorrect contacts. By comparing the Cα distances of the interacting residues in the target ($D_{i,j}(T)$) and in the model ($D_{i,j}(M)$), we assigned interaction scores ($s_{Interaction}(i, j)$) for each pre-defined interactions (Eq. 13 and Eq. 14). The average of these scores, weighted by the product of the residue numbers of the SSEs was the final interaction score (Eq. 15).

$$ts_{Inter}(i,j) = \exp\{-\ln 2[D_{i,j}(M) - D_{i,j}(T)/D_{i,j}(T)]^2\} \quad (13)$$

$$ms_{Inter}(i,j) = \exp\{-\ln 2[D_{i,j}(M) - D_{i,j}(T)/D_{i,j}(M)]^2\} \quad (14)$$

$$S_I = [\sum_{i,j} ts_{Inter}(i,j) \times w_{t,i,j} + \sum_{i,j} ms_{Inter}(i,j) \times w_{m,i,j}] \Big/ (\sum_{i,j} w_{m,i,j} + \sum_{i,j} w_{t,i,j}) \quad (15)$$

3.2.6 The contact score

Scores based on inter-residue contacts or distances were another commonly used method by previous assessors (Ben-David *et al*., 2009, Jauch *et al*., 2007). We incorporated a Contact Score (Shi *et al*., 2009) into QCS to quantify the atomic details of the models. In concept, it is similar to our interaction scores for SSEs, except that it evaluates all C-alpha contacts in the target. Contact score ($s_{Contact}(i)$) was calculated as in Eq. 16 and 17, where $D_i(M)$ is the distance in model and $D_i(T)$ is the distance in target, N is the total number of defined contacts.

$$s_{Contact}(i,j) = \exp\{-\ln 2[D_{i,j}(M) - D_{i,j}(T)/0.2)]^2\} \quad (16)$$

$$S_C = \sum_i s_{Contact}(i) \Big/ N \quad (17)$$

3.2.7 The QCS is the weighted sum of the six components

The QCS was defined as a weighted sum of all 6 scores discussed above. The weight of each component could be adjusted to accentuate certain aspect of the models. In this work, by default, all the components were weighted equally. To adjust the scale of QCS, we performed a transformation per Eq. 19. The parameter a, specific for each target, was obtained from random models. Ten random models were generated by circularly permutating the target structure to abolish the correspondence between the sequence and the 3D coordinates. For CASP9 FM targets and TBM representatives, these random models acquired average QCSs from 28 to 45. By hyperbolic transformation and adjusting the value of a, we rescaled the average random QCS for each target to 20. As a result, the scores from different targets are comparable to each other. The transformed scores correspond to the final QCS.

$$QCS = \frac{100}{\sum_{i=1}^{6} w_i}(w_1 S_P + w_2 S_L + w_3 S_H + w_4 S_A + w_5 S_I + w_6 S_C) \quad (18)$$

$$QCS_{rescaled} = QCS_{original}\,(a-1) \Big/ (a - QCS_{original}) \quad (19)$$

Agreement between QCS and manual assessment

The traditional and well-accepted way to assess CASP template free structure prediction is manual inspection by experts. To test the performance of QCS, we first compared QCS with the MIS on CASP9 FM models that obtained a non-zero MIS (zero MIS means either the global topology of the model is completely wrong or redundant models). Only models that correctly

predicted at least part of the structure core would attain a non-zero MIS, and thus these models were of relatively good quality. On these models, QCS correlates well with MIS (shown in Fig. 3) with Pearson correlation coefficient of 0.86.

QCS harbors the highest correlation coefficients with MIS among all the structure comparison methods we tested, including GDT_TS, CS, TM-align (Zhang *et al.,* 2005) and other traditional methods for structure comparison (Ortiz *et al*., 2002; Zemla *et al*., 1999a) (see Table 1). It is within our expectation as several QCS criteria were derived from the experience of manual inspection and both QCS and MIS emphasize on the global features of the models. Notably, GDT_TS and contact score show satisfactory correlation with manual judgment as well, which is consistent with previous experiences from CASP assessment (Ben-David *et al*., 2009; Jauch *et al*., 2007).

Three out of the four correlation coefficients listed in Table 1 (the Pearson's correlation coefficient (r), the Spearman's rank correlation coefficient (ρ) and the Kendall tau rank correlation coefficient (ιc) for all pairs of models) estimate the agreement in both ranking models of one target and comparing the relative prediction quality among different targets. From both aspects, QCS agrees with MIS the best. The fourth coefficient (Kendall tau rank correlation coefficient (ιi) computed by comparing only models that are from the same targets), however, is the most indicative for the ability of ranking models for a particular target. QCS and MIS obtain a ιi of 0.49, suggesting for 75% of all cases, QCS and MIS agree in their judgments.

Other similarity scores acquire even lower ιi. Moderate agreement between MIS and automatic scores likely results from 3 reasons: (1) MIS works differently from all automatic

methods by design. On one hand, different from QCS (similar to GDT_TS and other superimposition dependent methods), MIS positively scores only SSEs appearing in a correct local mutual arrangement (e.g. a helical hairpin). On the other hand, different from GDT_TS (similar to QCS), MIS assigns scores on the basis of the whole SSEs, considering their packing and interactions. (2) Low quality of FM predictions and the similarity among models made it impossible to clearly discern a "better" model in many cases. The ranking was thus highly sensitive to the differences in the criteria implemented by different method. This effect was exaggerated as only the ranking of relatively good models were examined and the fact that many of these models were generated by refining or selecting the predictions from several well-performing servers. If we considered all models by including the zero MISs, MIS correlated with automatic scores much better and QCS displayed the highest ιi of 0.67. (3) MIS contains minor errors and the scores are sometimes inconsistent, as the time devoted to each model is quite short, limited by the time frame in the CASP season.

The correlation coefficients between QCS components and MIS are shown in Table 1. The contact score alone (S_C) displays the best correlation with MIS. Although other components, taken separately, show lower correlation; taken together (S_5 in Table 1) they correlate with MIS even better than contact score (shown in table 1). Similarly, none of the other components dominates the performance of QCS. Each individual component assesses a specific aspect of the model, and their combination evaluates comprehensive features required for a good model and lowers the possibility of assigning a favorable score to a poor model due to a random match to the target.

In addition to combining all the component scores with equal weights, we optimized the weights on correlation coefficients with MIS. (QCS_r, QCS_ρ, QCS_{tc}, and QCS_{ti} in Table 1 stand for the optimized result on r, ρ, ιc and ιi respectively). Optimization can only boost the correlation slightly. This is firstly due to the absence of high agreement between any similarity score and MIS as discussed above. Moreover, as models of higher quality are usually favored by all the QCS components, change of weights does not lead to substantial change in QCS ranking (Kendall tau rank correlation between QCS_r, QCS_ρ, QCS_{tc}, QCS_{ti} and QCS are all above 0.82).

The correlation between QCS and other methods

We compared QCS with other assessment methods used in CASP9 and CASP8, including GDT_TS, CS, TenS, TenS components for CASP9 (Kinch *et al*., 2011b) and GDT_TS, Mammoth, Q scores for CASP8. QCS shows higher correlation with GDT_TS, Qcomb, TenS and CS (Kendall tau rank correlation coefficient above 0.65). These 4 scores are proved useful in previous CASPs (Kinch *et al*., 2003; Ben-David *et al*., 2009), and similarly to QCS, they balance between local and global features.

We compared QCS and GDT_TS on CASP9 TBM representatives, and the overall Kendall tau correlation coefficient is about 0.75. The general trend is that as the target becomes easier for predictors and thus the overall performance of all groups gets better, the correlation increases. This close correlation with GDT_TS for TBM targets indicates that the QCS method can also be applied to TBM model assessment. For the TBM category, even though most of

models get the global features correct, S_I and S_C in QCS still can reveal the difference in model quality.

Ability of revealing best models

An essential task of CASP assessment is to identify the best models. To focus on the ability of identifying best models, we studied the overlap between top models selected by automatic methods and by MIS. The top 5 models (including ties) were taken for comparison, and QCS top models overlap the most with MIS (43% overlap overall, shown in Table 2). Likewise, QCS ranks top models by MIS the highest, while GDT_TS and CS ranks them slightly lower than QCS did.

This moderate overlap is likely due to similar reasons as discussed in section 3.3. For T0534d1 and T0534d2, as all the models failed to predict the topology correctly, clearly best models do not exist. In contrast, for T0537 and T0550d1, many models were based on the same correct template and only precise measurement could differentiate the model quality. In both cases, the top models selected by MIS are questionably ideal. There are also a few cases where MIS top models are worse than top models detected by other methods after careful manual inspection. Without special attention to selecting the best few models, the models with highest MIS might result from subjective judgment without careful study in the limited time frame of CASP season.

In the development stage of QCS, we devoted special attention to ensuring the top 10 models correspond to or are comparable with the best models by careful manual inspection.

Top 10 models selected by QCS and top 5 models according to MIS and other methods are at http://prodata.swmed.edu/congqian/casp_sum.html.

We designed QCS on the basis of our experience in CASP9 assessment. The criteria for assessing structure prediction we implemented could be different from the standard of others. In the CASP8 experiment, all the best FM models selected by the assessors corresponded to the ones with highest GDT_TS (Ben-David *et al*., 2009). This perfect overlap might either indicate their great emphasis on the model's ability in superimposing to the target or reflect the bias placed by GDT_TS on the assessors: their manual assessors were likely to be aware of GDT_TS rankings and only a small portion of models ranked high by GDT_TS were manually inspected, which in some cases represent only a group of similar models (Ben-David *et al*., 2009).

In contrast, QCS agrees with CASP7 assessors' manual inspection results better than GDT_TS and the contact-based score (named CMO) designed by CASP7 assessors. Even though GDT_TS and CMO top 25 models were used as candidates, the best models selected after 3 rounds of careful manual inspection are ranked higher by QCS than by either GDT_TS or CMO. Out of the 45 best models for 18 targets, 25 are in the top 5 ranks by QCS, while 15 of them overlap with GDT_TS top 5 models and only 6 are among the CMO's top 5 models. Moreover, for most targets, the average QCS ranks of the best models are higher or about the same as GDT_TS and CMO ranks. This good agreement between QCS selection and CASP7 assessors' manual inspection results independently supports the value of QCS in revealing the best models.

For 3 targets (T0296, T0309, T0314), QCS ranked the best models lower than GDT_TS and CMO did. However, for T0296 and T0314, no predictions modeled the topology of the structure correctly (Jauch *et al*., 2007) and the best models selected by previous assessors are not clearly better than QCS top models. Only for T0309, the best manually selected models seemed to be of better quality than QCS picks. This target is a domain-swapped octamer. Manually selected models placed the strands that involve in oligomerization correctly, somewhat neglecting other parts of the molecule, while QCS preferred models that packed the rest of the molecule correctly. Manual inspectors paid more attention on the oligomerization strands since they form the core of the octamer. However, as the oligomerization strands are loosely packed in the monomer, QCS, by design laid less emphasis on them. Such a priority defined by the specific features of certain target is the unique advantage of manual inspection, and it signifies the importance of manual assessment.

QCS reveal models of superior global topology

Best models selected by QCS were compared with best models suggested by GDT_TS. In most cases the best models selected by both scores agree with each other (shown in Table 3). For some cases, QCS did reveal models with good features that were missed by GDT_TS. Three such examples are shown in Fig. 5-7.

The first example is target T0561 (Fig. 5A). QCS selects model TS295_2 (Fig. 5B) as the best model with a score of 62.4 and scores 46.9 for the other model TS324_5 (Fig. 5C), while GDT_TS favors model TS324_5 (GDT_TS: 39.4) over TS295_2 (GDT_TS: 31.5). GDT_TS favors model TS324_5 as its 3 helices at the N-terminus (colored in blue, green and yellow in

Fig. 5) can be precisely superimposed to the corresponding helices in the target. However, in terms of global topology, the two helices at the C-terminus of model TS324_5 (Fig. 5G) are packed in opposite orientations compared to the target (Fig. 5E). The incorrect packing of these helices in the model TS324_5 diminishes the quality of this model. On the contrary, the global topology of model TS295_2 (Fig. 5F) agrees exactly with that of the target. Moreover, out of the 3 key interactions (Fig. 5I, colored in magenta) defined in target, TS295_2 (Fig. 5J) predicts all of them correctly while in TS324_5 (Fig. 5K) only one of them is correct. Apparently, by paying attention to the global features, QCS has revealed models with superior global topology and interactions, which should be favored after closer inspection.

The model (TS096_4, Fig. 5D) selected by MIS also adopts correct topology (Fig. 5H). QCS ranks this model at 18 with a score of 58.0, after the cluster of models that assemble TS295_2. A superior feature of this model is that the N- and C- termini are placed close to each other as they are in the target. Nevertheless, the helices in this structure are over-predicted and thus the loop regions are inadequate to allow correct packing angles between the helices. Moreover, close study of the interactions shows that they are poorly predicted in this model. Such features downgraded the quality of this model and made it worse than TS295_2 by careful manual comparison.

The second example is the target T0618 (Fig. 6A). For this target QCS ranks TS386_4 (Fig. 6B) as the best model (QCS: 61.3, GDT_TS: 39.6), which is visually identical to the best model according to MIS. And GDT_TS selects TS380_4 (Fig. 6C) as the first model (QCS: 53.3, GDT_TS: 41.9). TS380_4 is worse in topology as the green helix in a completely wrong orientation, leading to opposite handedness between the green, yellow and orange (or red)

helices. Moreover, different from the real structure, the C- and N-terminal helices in TS380_4 are almost perpendicular and the shape of the whole protein is a poor representation of the reality. Comparatively, the best model selected by QCS almost correctly predicted the topology and the global shape of the model, promising an undoubtedly better model by manual inspection.

The third example is target T0621 (Fig. 7A). QCS favors model TS065_5 (shown in Fig. 7B, QCS: 65.3, GDT_TS: 32.2) over TS002_5 (shown in Fig. 7C, QCS: 53.2, GDT_TS: 34.0). The core of this target is a jelly roll β-sandwich and the model favored by QCS positions all the β-strands in the β-sandwich correctly. However, the model favored by GDT_TS failed to pack the N-terminal and C-terminal strands, even though it may have better superimposition with target because of better details in the shape of the β-sandwich. Similarly to QCS, MIS ranked model TS065_1 (QCS rank it as 2nd) as the best, which is very similar to TS065_5 in global topology, with differences mainly in the inserted helices and hairpin colored in cyan in Fig. 6A.

These three examples illustrate general properties of QCS. Compared to the well-established GDT_TS, this new method emphasizes more on the global topology, thus it can overcome the problem caused by domination of local features that is frequently revealed in GDT_TS. QCS defines all the SSEs and contacts in target and propagates these definitions to the model. Shift in the alignment will lead to incorrect definition of SSEs in the model and result in unfavorable QCS. As most structure prediction methods more or less take advantage of a template structure or template structure fragment, the correct alignment between the

template sequence and the target sequence during structure prediction procedure will be highly favored by QCS.

The best models selected by QCS and GDT_TS could be very different but both have certain advantages, one example is the target T0578 (Fig. 8A). This target has two hard to predict features: one is the unusual crossover between the green and yellow strands, which is predicted by none of the models. Another is the packing of the three helices. Model TS428_2 is among the few models that packed these helices almost correctly and that explains the high MIS it obtained. However, this model only predicted half of the β-sheet in the target, and failed to adopt an elongated shape as the target. On the contrary, QCS's top model, TS065_3 correctly predicted the shape of the protein and the major part of the β-sheet while it failed to model the topology of the helices. QCS favors such model likely because we designed QCS to emphasize on strands by weighing the residues of a strand twice as much as residues in a helix.

In such cases, top models selected by different methods reveal different positive features. By combining them, we can generate a better pool of candidates for best models and provide better assessment of structure predictions and facilitating development of methods.

CONCLUSIONS

We developed an automatic method for structure prediction assessment that inspired the manual assessment traditionally carried out by CASP assessors. Not dominated by local features of the prediction, QCS emphasizes on the global topology. QCS is a good complement for superimposition based scores as GDT_TS and can be used for CASP in the future and generally for automatic structure prediction assessment. Moreover, QCS can be upgraded into

a tool for general structure alignment and comparison. With emphasis on global structure, QCS or the ideas presented could be useful for remote homology detection and structure classification of proteins.

Table 1. Correlation coefficient between automatic scores and MIS

Score name	Weights of QCS components						r	ρ	ιc	ιi
	S_L	S_P	S_A	S_I	S_H	S_C				
QCS	1/6	1/6	1/6	1/6	1/6	1/6	0.86	0.87	0.69	0.49
S_L	0	1	0	0	0	0	0.70	0.69	0.49	0.37
S_P	1	0	0	0	0	0	0.67	0.68	0.50	0.32
S_A	0	0	1	0	0	0	0.69	0.70	0.53	0.34
S_I	0	0	0	1	0	0	0.67	0.68	0.49	0.35
S_H	0	0	0	0	1	0	0.55	0.51	0.37	0.23
S_C(CS)	0	0	0	0	0	1	0.73	0.74	0.53	0.40
S_5	1/5	1/5	1/5	1/5	1/5	0	0.85	0.86	0.68	0.48
QCS_r	0.07	0.17	0.03	0.07	0.23	0.43	0.88	0.89	0.71	0.51
QCS_ρ	0.10	0.17	0.07	0.03	0.23	0.40	0.88	0.89	0.72	0.51
$QCS_{\iota c}$	0.10	0.17	0.07	0.03	0.23	0.40	0.88	0.89	0.72	0.51
QCS ιi	0.07	0.20	0.07	0.03	0.20	0.43	0.88	0.89	0.71	0.51
GDT_TS	–	–	–	–	–	–	0.70	0.67	0.50	0.42
TenS	–	–	–	–	–	–	–	–	–	0.44
Tm-align	–	–	–	–	–	–	0.55	0.53	0.40	0.34
Mammoth	–	–	–	–	–	–	0.52	0.54	0.38	0.34
SOV	–	–	–	–	–	–	0.46	0.48	0.34	0.26

Table 2. Top model overlaps between MIS and automatic scores

Target	T531	T578	T581	T604d1	T608d1	T621	T624	
CS	0.60	0.31	0.83	0.40	0.33	0.80	0.50	
QCS	0.60	0.65	0.83	0.40	0.67	0.80	0.75	
GDT_TS	0.20	0.54	0.83	0.40	0.67	0.20	0.75	
TM	0.00	0.31	0.83	0.60	0.33	0.60	0.75	
Mammoth	0.00	0.69	0.67	0.40	0.67	0.60	0.75	
SOV	0.20	0.27	0.50	0.00	0.33	0.60	0.00	
TenS	0.17	0.67	0.00	0.60	0.00	0.40	0.67	
Target	T534d1	T534d2	T537	T550d1	T550d2	T561	T618	Overall
CS	0.00	0.00	0.00	0.33	0.21	0.00	0.00	0.31
QCS	0.00	0.25	0.40	0.33	0.14	0.00	0.17	0.43
GDT_TS	0.20	0.25	0.13	0.17	0.19	0.00	0.17	0.34
TM	0.40	0.00	0.00	0.00	0.29	0.00	0.17	0.31
Mammoth	0.40	0.25	0.00	0.00	0.29	0.00	0.07	0.34
SOV	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.15
TenS	0.75	0.00	0.38	0.00	0.43	0.00	0.25	0.31

Table 3. Comparison of best models selected by GDT_TS and QCS

Target	T0531		TS534d1		T0534d2		T0537		T0550d1		T0550d2		T0561	
Method	QCS	GDT_TS	QCS	GDT_TS	QCS	GDT_TS	QCS	GDT_TS	QCS	GDT_TS	QCS	GDT_TS	QCS	GDT_TS
First model	TS399_5	TS399_5	TS114_4	TS297_4	TS172_4	TS110_4	TS065_3	TS065_5	TS065_2	TS065_2	TS104_3	TS104_3	TS295_2	TS324_5
MIS	55.8	55.8	40.9	40.9	n/a	46.4	52.4	52.4	88.9	88.9	81.2	81.2	54	68
Careful inspection	Equal	Equal	Equal	Equal	Equal	Equal	Equal	Equal	Equal	Equal	Equal	Equal	Better	Worse
Target	T0578		T0581		T0604d1		T0608d1		T0618		T0621		T0624	
Method	QCS	GDT_TS	QCS	GDT_TS	QCS	GDT_TS	QCS	GDT_TS	QCS	GDT_TS	QCS	GDT_TS	QCS	GDT_TS
Top model	TS065_3	TS428_2	TS065_2	TS170_1	TS096_1	TS096_1	TS147_1	TS147_1	TS386_4	TS380_4	TS065_5	TS002_5	TS172_1	TS172_1
MIS	56.3	62.5	90.6	81.3	96.3	96.3	n/a	59.3	n/a	50.1	50	48.3	81.3	81.3
Careful inspection	Equal	Equal	Better	Worse	Equal	Equal	Equal	Equal	Better	Worse	Better	Worse	Equal	Equal

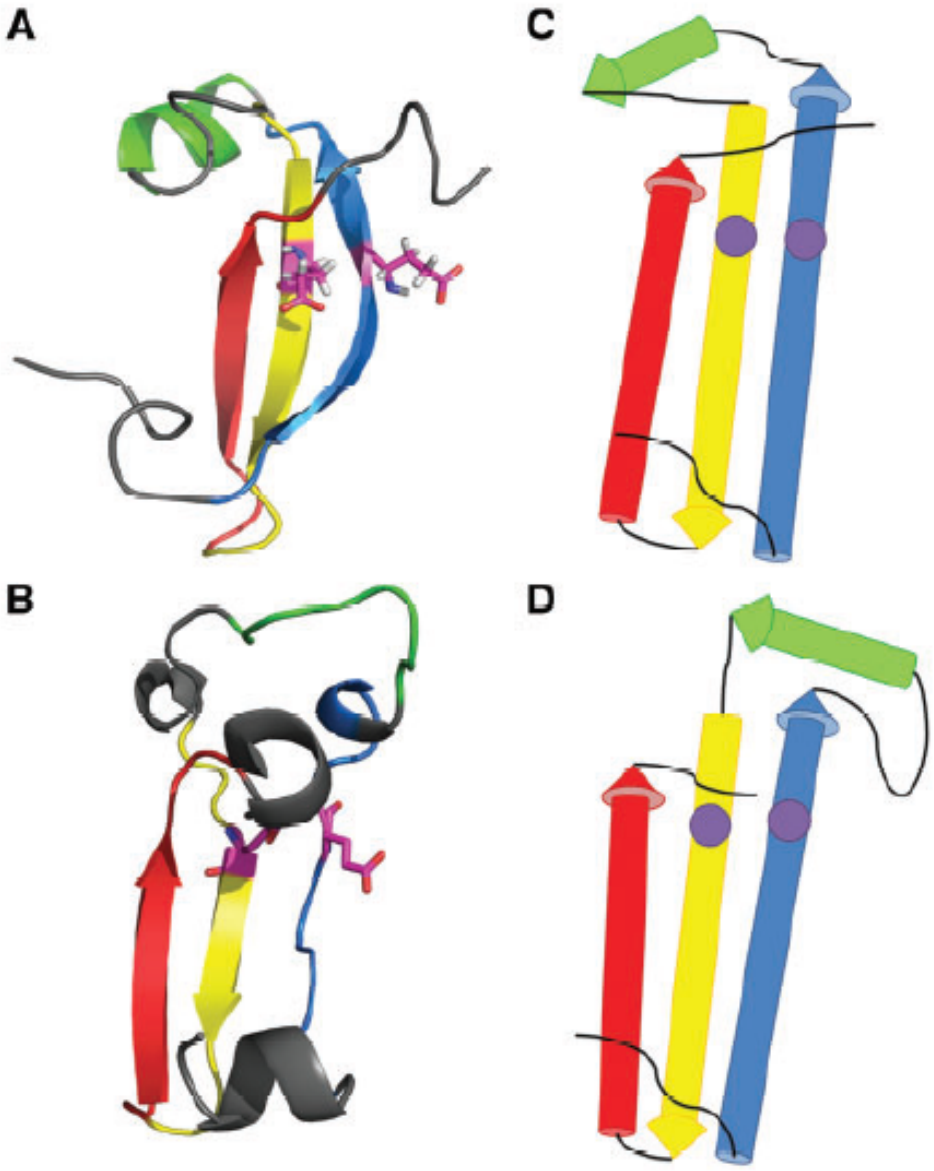


Fig. 1. Simplification of the target and models. (**A**) Target T0531: SSEs are colored in rainbow and one pair of residues where two SSEs interact with each other (defined as interaction) are highlighted in magenta. (**B**) Simplified T0531: the SSEs are represented by vectors and the interactions are represented by pairs of points illustrated by the purple dots. (**C**) A model for T0531 (TS399_4) colored in rainbow according to the target SSE definition with the same interaction defined in the target highlighted in magenta. (**D**) Simplified model TS399_4.

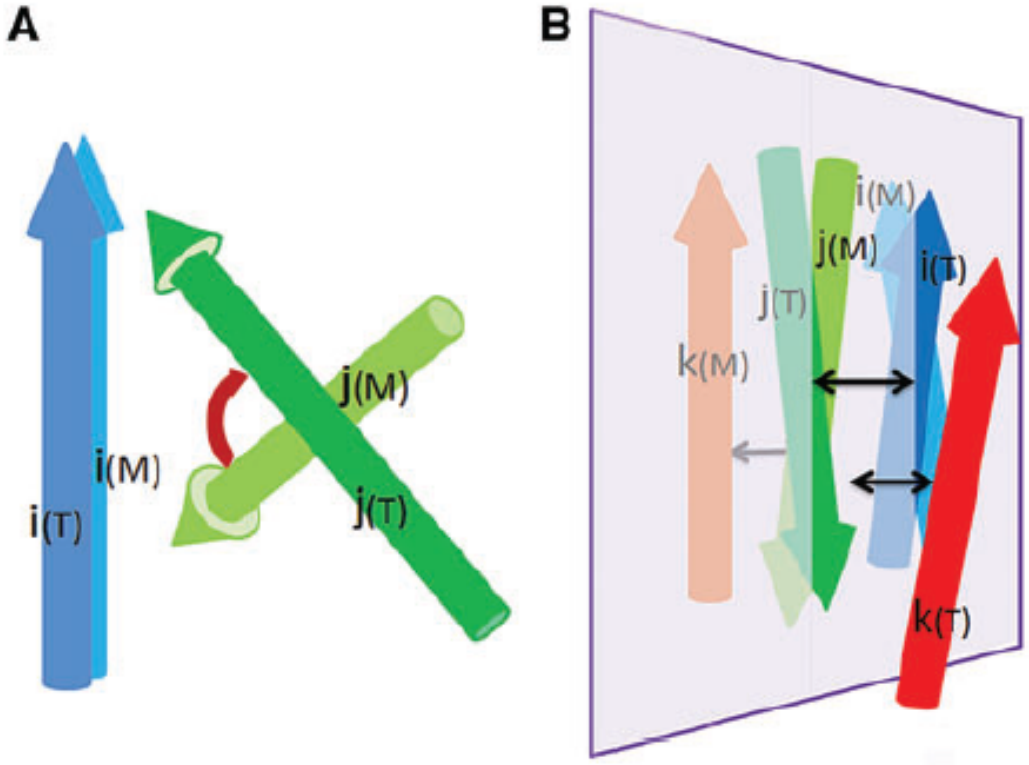


Fig. 2. Illustration of SSE angle and handedness measurement. (**A**) The dark blue and dark green vectors represent a pair of SSEs in the target. The blue and green vectors represent the corresponding SSE pair in the model. The red arrow indicates the angle discrepancy between the target and the model. (**B**) The 3 SSE vectors ($i(T)$, $j(T)$ and $k(T)$) from the target are colored in dark blue, dark green and red, and the corresponding SSEs ($i(M)$, $j(M)$ and $k(M)$) in the model are in blue, green and orange. In both the target and the model, we define the reference plane (colored in light purple) as the one that passes through the centers of i, j and parallel to the general orientation of i and j ($i+j$ when the angle between them is $<90°$ and $i-j$ when their angle is $>90°$). The cross product of i's projection on the reference plane and the vector connecting the centers of i and j represent the norm of this plane. After superimposing the reference planes in the target and in the model, the third vectors, $k(T)$ and $k(M)$ are on opposite sides of the plane, indicating an error in handedness. In such case, certain penalty would be introduced as in Equation (11). The black arrows show the distances ($D_{k,P}(M)$, $D_{k,P}(T)$ and $D_{i,j}(T)$) that are measured for handedness score.

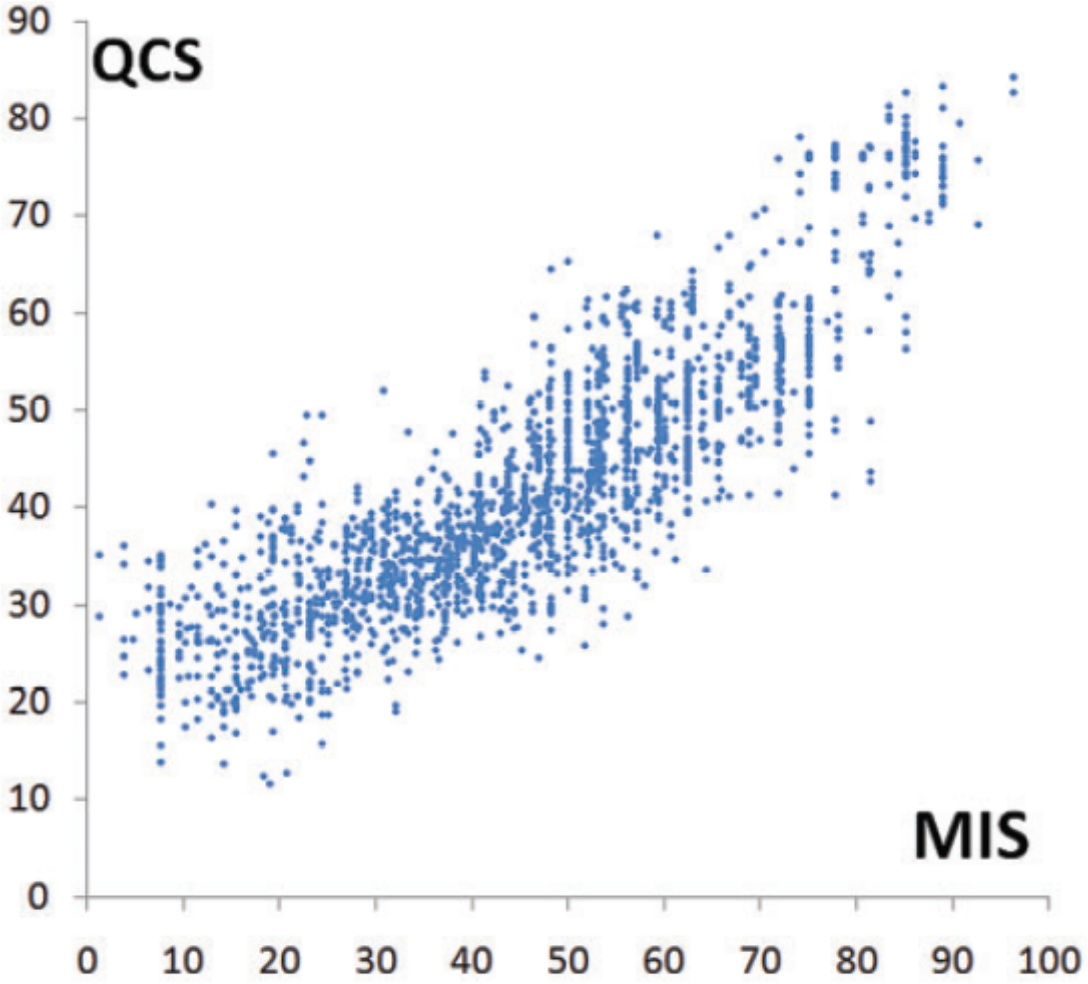


Fig. 3. The correlation between QCS and MIS on a set of CASP9 FM models, which was evaluated by manual inspection.

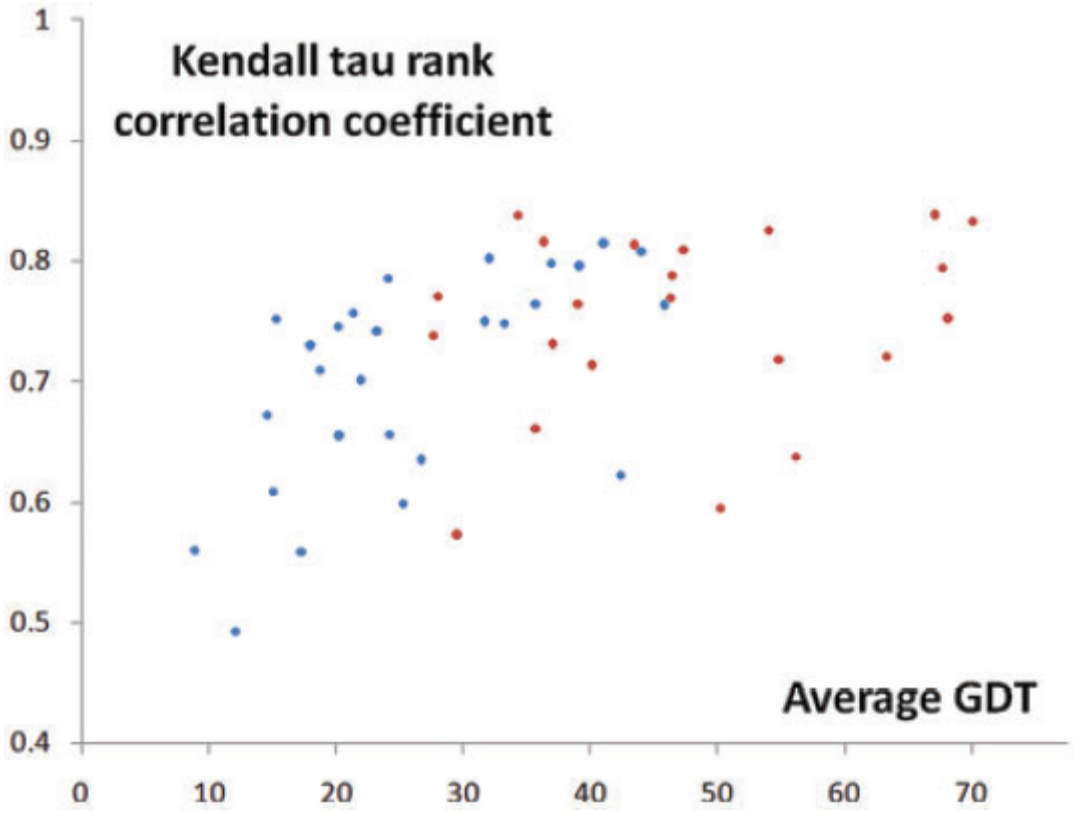


Fig. 4. The Kendall tau rank correlation coefficient between QCS and GDT_TS on CASP9 FM targets (represented by blue dots) and TBM representatives (represented by red dots).

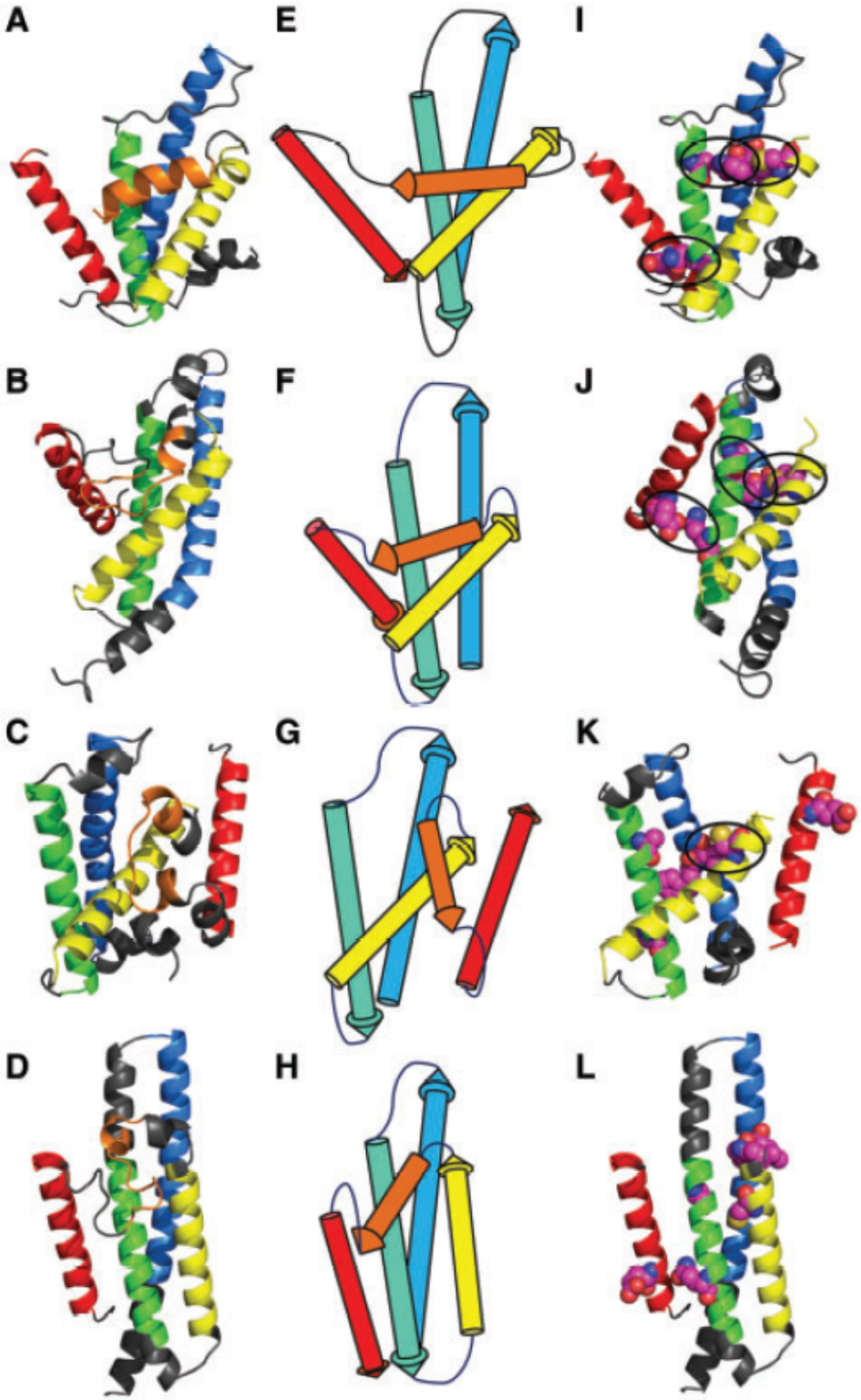

A
E
I
B
F
J
C
G
K
D
H
L

Fig. 5. Example (Target 561) of QCS revealing models with good global topology and correct interactions. The first panel: the target or model structures; the second panel: the topology diagrams; the last panel: the structures with interactions colored in magenta. (**A**), (**E**), (**I**) target T0561; (**B**), (**F**), (**J**) the best model selected by QCS; (**C**), (**G**), (**K**) the best model selected by GDT_TS; (**D**), (**H**), (**L**) the best model selected by MIS.

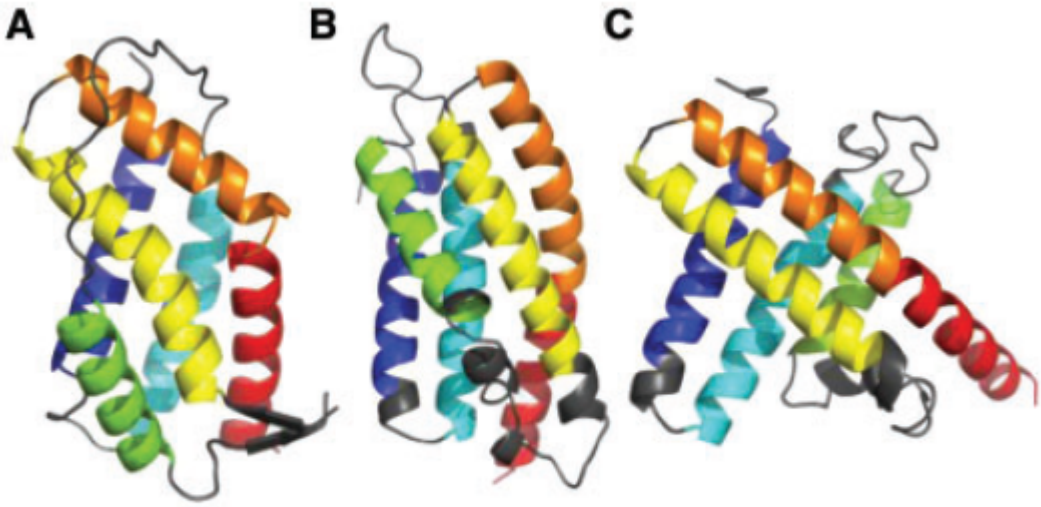


Fig. 6. Example (Target 618) of QCS revealing models with correct topology and global shape. (**A**) Structure of the target T0618 colored in rainbow; (**B**) the best model selected by QCS; (**C**) the best model selected by GDT_TS.

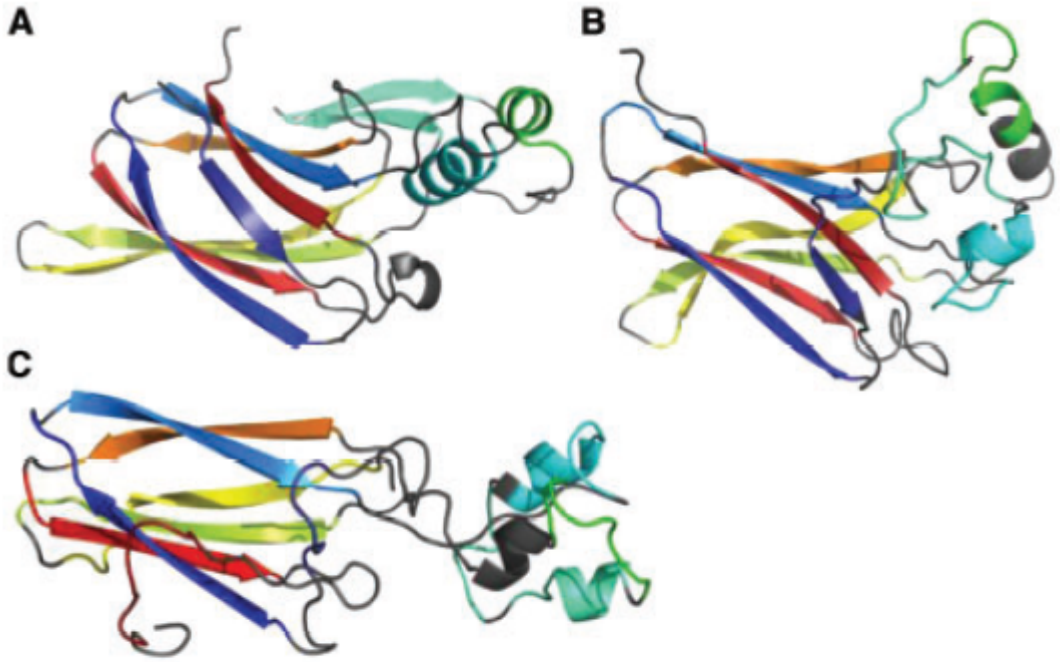


Fig. 7. Example (Target 621) of QCS revealing models with superior global features. (**A**) The structure of target T0621 colored in rainbow; (**B**) the best model selected by QCS; (**C**) the best model selected by GDT_TS.

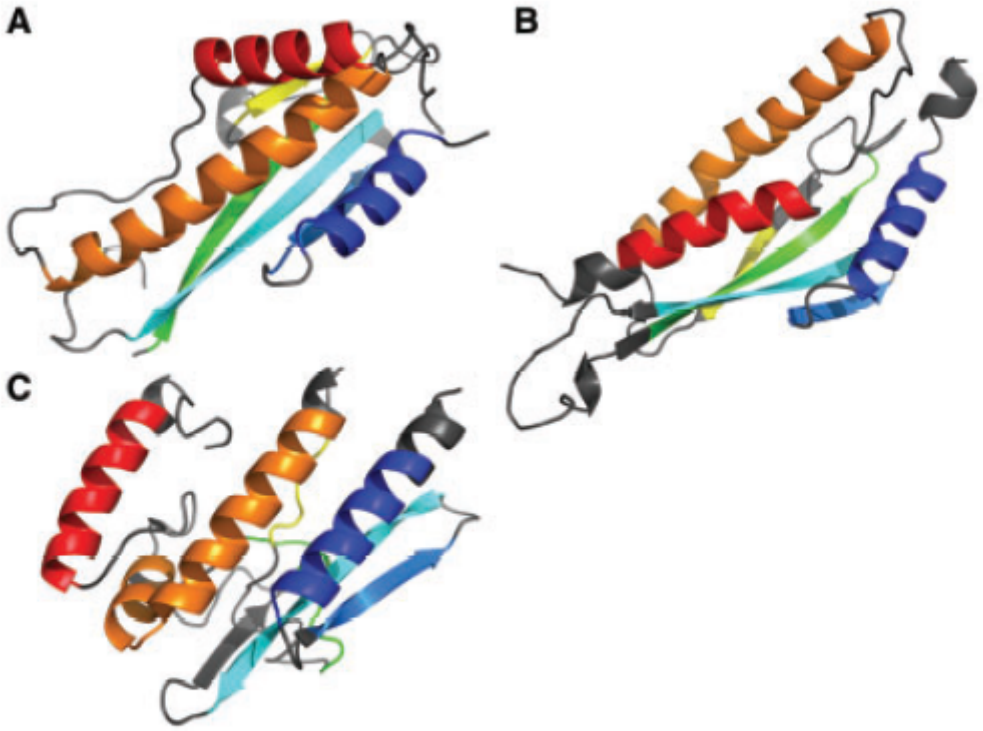


Fig. 8. Example (Target 578) of GDT_TS and QCS revealing models with different advantages. (**A**) The structure of target T0578; (**B**) the structure of best model selected by QCS; (**C**) the structure of best model selected by GDT_TS.

REFERENCES

Ben-David M, Noivirt-Brik O, Paz A, Prilusky J, Sussman JL, Levy Y. (2009) Assessment of CASP8 structure predictions for template free targets. *Proteins.*77 Suppl 9:50-65.

Jauch R, Yeo HC, Kolatkar PR, Clarke ND. (2007) Assessment of CASP7 structure predictions for template free targets. *Proteins.* 69 Suppl 8:57-67.

Kabsch W, Sander C (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22 (12): 2577–637.

Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H, Grishin NV. (2003) CASP5 assessment of fold recognition target predictions. *Proteins.* 53 Suppl 6:395-409.

Kinch LN, Shi S, Cheng H, Cong Q, Pei J, Schwede T, Grishin NV. (2011a) CASP9 Target Classification. *Proteins* (accepted, to be published)

Kinch LN, Shi S, Cong Q, Cheng H, Liao Y, Grishin NV. (2011b) CASP9 Assessment of Free Modeling Target Predictions. *Proteins.* (accepted, to be published)

Kryshtafovych A, Venclovas C, Fidelis K, Moult J. (2005) Progress over the First Decade of CASP Experiments. *Proteins.* 61 Suppl 7:225-36.

Majumdar I, Krishna SS, Grishin NV (2005) PALSSE: a program to delineate linear secondary structural elements from protein structures. *BMC Bioinformatics.* 2005 6:202.

Moult J, Pedersen JT, Judson R, Fidelis K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins.*23(3):ii-v.

Moult J. (2006) Rigorous Performance Evaluation in Protein Structure Modelling and Implication for Computational Biology. *Philos Trans R Soc Lond B Biol Sci.*

Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A. (2009*)* Critical assessment of methods of protein structure prediction - Round VIII. *Proteins.* 77 Suppl 9:1-4.

Ortiz AR, Strauss CE, Olmea O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison*.* *Protein Sci* 11:2606–2621.

Shi S, Pei J, Sadreyev RI, Kinch LN, Majumdar I, Tong J, Cheng H, Kim BH, Grishin NV. (2009) Analysis of CASP8 targets, predictions and assessment methods. *Database (Oxford)*. 2009:bap003.

Tai CH, Lee WJ, Vincent JJ, Lee BK. (2005) Evaluation of domain prediction in CASP6. *Proteins.* 61(Suppl 7):183–192.

Zemla A, Venclovas C, Fidelis K, Rost B. (1999a) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins* 34(2):220–223.

Zemla A, Venclovas C, Moult J, Fidelis K (1999b). Processing and analysis of CASP3 protein structure predictions". Proteins S3: 22–29.

Zemla A, Venclovas C, Moult J, Fidelis K (2001). Processing and evaluation of predictions in CASP4. Proteins 45 (S5): 13–21.

Zemla A (2003). LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Research 31 (13): 3370–3374.

Zhang Y, Skolnick J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*.33(7):2302-9.

CHAPTER FOUR
MESSA: MEta Server for Sequence Analysis

INTRODUCTION

Every research project on a protein should start from computational analysis of its sequence. Well-designed sequence analysis is an efficient way not only to obtain predictive information, but also to prevent potential mistakes in the interpretation of experimental data. The widely known argument about the report of a plant G-protein coupled receptor (GPCR) [1], which subsequently was suggested to be a cytoplasmic lanthionine synthetase-like protein by both computational analysis and experimental verification [2-3], illustrated the value of sequence analysis.

To serve the growing need for computational analysis of protein sequences, many tools have been developed. Such tools typically predict certain local sequence property, spatial structure or function of a query sequence. A reliable prediction usually requires the correct selection of tools and a broad incorporation of predictive information. A consensus based method that derives conclusion based on the consistent judgment among different predictors usually produces better results than individual tools it includes [4-5]. In addition, when independent pieces of information are combined together, errors in a prediction can be revealed, leading to even better performance. For instance, in the last Critical Assessment of Structure Prediction (CASP), even the top performing 3D structure predictors were not able to detect and remove the signal peptides in the target sequences [6], resulting in a negative influence on the prediction quality as a hydrophobic signal peptide would tend to be packed in the middle

of the structure. As a result, in order to generate a reliable hypothesis on the basis of computational analysis, one need to consult many predictors and analyze all the results together, making comprehensive analysis on a given protein a non-trivial task.

Meta severs have been developed to reduce such difficulty by combining various tools and integrating their results. Most meta servers focus on one aspect of sequence analysis, for instance, Jpred for secondary structure [7], metaPrDOS for disordered region [8], metaTM for transmembrane topology [9], Pcons.net [10] and 3D-Jury [11] for 3D structure, JAFA [12] and ProKnow [13] for function prediction. Other web services incorporate more information to further accelerate sequence analysis, such as PredictProtein [14] for predictions of many local sequence properties, SMART [15] for identification of protein domains and special sequence motifs and Genesilico [16] that focuses on secondary and spatial structure predictions.

However, as predictions of local sequence properties, structure and function are usually highly related, it is beneficial to address these questions together thus deriving more reliable conclusions from all information. For instance, the presence of certain local sequence motifs such as a transmembrane helix or a signal peptide and the predicted 3D structures provide essential clue for function interpretation. Thus we developed a MEta Server for Sequence Analysis (MESSA), which balances these predictions and provide results related to subcelluar localization (only for membrane protein and exported proteins), function, 3D structure templates and domain architecture. We tested MESSA on the proteome of *Candidatus Liberibacter asiaticus* (*Ca. L. asiaticus*) [17] and the results showed that MESSA provides extensive information about the structural and functional properties of these proteins, useful for understanding of and designing experiments on certain protein.

RESULTS AND DISCUSSION

Interpretation of Results from MESSA

Upon the submission of a single query protein sequence, MESSA runs a number of top-performing programs and returns a webpage with results conveniently formatted for manual inspection. MESSA contains several time-consuming steps such as PSI-BLAST [18] and HHsearch [19], and thus for proteins from very large families, it might take several hours for the whole process to complete. However, as these time consuming steps are designed to detect remote homologs and the information is useful when closely related protein families or protein structures do not exist. Thus, to avoid the long waiting time, MESSA provides a friendly interface to allow integration and display of the available results at any time after submitting the job upon users' request, while the time-consuming processes are still waiting in the job queue or running in the background. The users will be notified by email once the job is finished and the result of MESSA contains the following sections.

Section I. Prediction of Local Sequence Features (Fig. 1A): Local sequence property predictions, such as secondary structure or disordered region, are helpful for predicting 3D structure, whereas, signal peptide and transmembrane helix predictions are suggestive of the protein localization and function. This section summarizes the predictions of secondary structure, low-complexity region, disordered region, coiled coils, transmembrane helices and signal peptides. Signal peptide is a sequence motif at the N-terminus of proteins characterized as a hydrophobic α-helical region flanked by a positively charged short region at the N-terminus and several polar residues at the C-terminus. The programs used for each prediction

and the explanation of the results are described in details in Table 1. The result from each predictor is represented as one string made up of each residue's predicted status. These strings are all aligned to the original protein sequence for easy comparison.

Section II. Close Homologs: Close homologs or orthlogs usually preserve the same function inherited from the common ancestor, and thus the detection of them is useful for function prediction. MESSA shows the 10 closest confident homologs in the NonrRdundant (NR) and SWISS-PROT database detected by BLAST [20] or 2 iterations of PSI-BLAST [18] (E-value cutoff 0.005). On the one hand, NR database is the most comprehensive database consisting of almost all known sequences; therefore the best hits detected in NR will represent the closest sequences known in the protein sequence space. The taxonomy information of these hits is shown to provide hints to the evolutionary history of the protein and reveal horizontal gene transfer events. On the other hand, SWISS-PROT database contains a subset of the NR and all proteins are manually curated, and the close homologs from this database offers a more reliable resource for annotation transferring.

Section III: Homologous Protein Families (Fig. 1B): Proteins can be classified into families according to similarity in sequences, structures and functions inherited from common ancestors [21-26]. Such classification and the extensive information of each protein family in the databases [21-26] assist in functional annotation greatly. In this section, we listed the homologous protein families and conserved domains identified by RPS-BLAST [27] (E-value cutoff: 0.005) or HHsearch [19] (probability cutoff: 90%) in the NCBI Conserved Domain Database [23]. For each detected protein family or conserved domain, the relevant information and the alignment to the query protein can be easily retrieved for convenient verification. This

section is the most instructive reference for function annotation. Close homology between an unknown protein and certain well characterized protein family usually indicates that the unknown protein should be assigned to this family and share similar functional properties with the protein family.

Section IV. Homologous Structures and structure domains (Fig. 1C): Spatial structure prediction is an important aspect of sequence analysis. First, a 3D view of the protein can disclose the crucial residues and the mechanism for the protein to perform its function. Second, as 3D structure is much more conserved among homologous proteins than function, a reliable structure prediction is achievable for most proteins in a proteome [28], including many cases for which confident function predictions are not feasible. Third, the predicted structure is indicative of protein function: the presence of conserved active-sites and binding surfaces is useful in providing hypothesis about the protein function or validating the predicted function from sequence-based approaches. This section is designed for structure modelling. Homologous structures in the Protein Data Bank [29] and structure domains in the Structure Classification Of Protein (SCOP) database [30] detected by PSI-BLAST (e-value below 0.005), RPS-BLAST (e-value below 0.005) and HHsearch (probability higher than 90.0%) are shown. For each detected protein and protein domain, the alignment and the corresponding structure displayed by Jmol (an open-source Java viewer for chemical structures in 3D, available at: http://www.jmol.org/) can be retrieved. The conservation of protein structures among homologs allows these structures, in most cases, to represent the general topology of the query protein. These evolutionarily related protein structures can be utilized as templates for homology modeling to generate a 3D structure model of the query by MODELLER [31] or

SWISS-MODEL [32]. For structure domains detected in SCOP, we also provide the classification hierarchy of the domain, which gives insights in evolutionary history of the domain and suggests similarities to other proteins.

The extensive information obtained by MESSA can help researchers to acquire knowledge and hypothesis about a protein and help them to interpret experimental results. For instance, part of the result produced by MESSA for the purported GPCR [1] (discussed in **Introduction**, refseq ID: NP_175700) is shown in Fig. 1. Most transmembrane topology predictors implemented by MESSA predict it to be a cytoplasmic protein without transmembrane helices. Only TOPPRED and HMMTOP detected transmembrane helices in the protein. However, TOPPRED and HMMTOP are designed to reveal the topology of a given transmembrane protein rather than distinguishing transmembrane proteins from cytoplasmic proteins, so they might recognize a buried hydrophobic helix in a protein as transmembrane helix and lead to a high false-positive rate in predicting transmembrane proteins. The protein family assignment and the 3D structure templates supported by multiple methods consistently suggest its close relationship to lanthionine synthetase. Moreover, the predicted 3D structure shows that the protein has 14 semi-parallel helices. Although the 7 helices buried in the middle of the structure appear to be hydrophobic, the surface of the protein is largely hydrophilic. MESSA definitively suggests potential problem in the function proposed by the authors [1]. The evidence obtained easily from MESSA could assist with experimental data interpretation and prevent hasty conclusion in such cases.

Comparison between MESSA and other similar meta servers

Except for the broad incorporation of predicted features, MESSA has two additional important features. First, MESSA provides convenient display of the results. For instance, the local sequence feature predictions are all represented as one line and aligned to the sequence and the detected structure templates can be directly displayed on the result page. Second, it relies on confident homology inferred by sequence and profile similarity for structure and function prediction. Structure and function prediction without experimentally studied homologs, such as de novo folding and functional association analysis remains highly challenging. The conservative homology-based approach ensures the confident predictions in most cases. Moreover, the rapid growth in the numbers of experimentally studied proteins and available protein 3D structures has greatly increased the capability of homology-based structure-function annotation and ensures reasonable prediction coverage.

Widely used web services similar to MESSA include PredictProtein [14], SMART [15] and GeneSilico [16]. Instead of focusing on one aspect of protein sequence analysis, such as function prediction or 3D structure prediction, these meta servers also incorporate a large variety of programs and aim at facilitating highly integrated sequence analysis. PredictProtein offers rich information about the local sequence features of a protein, such as the secondary structure, transmembrane helices, protein sorting signals and functional sites. Compared to MESSA, it lacks the function of detecting related protein families and pays limited attention to prediction of 3D structure. Moreover, due to the high volume of usage, PredictProtein only offers 3 free queries for academic users per year. SMART is specialized in annotating domain architecture. Moreover, it offers prediction of signal peptides, transmembrane helices, low

complexity regions and homologous structures detectable by BLAST. Compared to SMART, MESSA is featured by the ability to detect remote homologous protein family and protein structure and thus has higher ability in structure and function prediction at the cost of longer execution time for a query protein. We regard Genesilico meta server as the most similar work to MESSA. Although Genesilico is featured as a Fold Recognition meta server, it also offers information about related protein families, prediction of transmembrane helices and signal peptides. Different from Genesilico that emphasizes on 3D structure prediction, MESSA aims at offering well-balanced information to support integrative analysis of protein local sequence features, 3D structures and function. As a result, MESSA does not include that many tools for structure template identification except most well-performing ones. In addition, MESSA include prediction of signal peptides, conservation index of the protein and information about closely related proteins, which are helpful for function interpretation.

Application of MESSA to the proteome of *Candidatus* Liberibacter asiaticus

We tested MESSA on the proteome of the recently sequenced genome of Candidatus Liberibacter asiaticus [17], and the results are constructed as a website at http://prodata.swmed.edu/liberibacter_asiaticus/. In the genome sequence of C. L. asiaticus, the gene prediction pipeline from NCBI and SEED detected 1233 protein coding genes, with 1046 of them predicted by both methods. In addition, 58 proteins that are identified by either of the single gene prediction pipelines show confident homology to other proteins in the non redundant database. We consider these 1104 proteins to be confidently predicted ones. The remaining 128 proteins exhibit a relatively small size (usually less than 60 residues), comprise

of low complexity sequence, lack similarity to any known protein, and are inconsistently predicted. These genes may represent falsely predicted open reading frames that may not exist in the bacterium.

Based on the MESSA results, we manually analyze every protein. Among the confidently predicted proteins, 63 are likely to be extracellular proteins as they are predicted to harbor a signal peptide by no less than 2 methods, and they are not transmembrane proteins. 197 of all Ca. L. aisaticus proteins are likely to locate to the membrane of the bacterium. Membrane localization is based on the consensus between no less than 3 out of 6 transmembrane-helix predictors, as well as the topology of predicted 3D structures. As shown in **Fig. 2**, we were able to predict the function for 80.1% of these confidently predicted proteins, while NCBI and SEED annotated 68.0% and 70.8% of them respectively, 74.1% taken together. Moreover, out of the 220 proteins without function predictions, 37.3% are predicted to be secreted or transmembrane proteins, indicating their general function in communicating with the environment.

In addition, the information provided by MESSA offers homologous structures for template based structure modeling for Ca. L. asiaticus proteins. The confident structure templates (HHsearch probability above 90%, PSI-BLAST or RPS-BLAST E-value below 0.005) in this website cover 74.3% of all residues in the Ca. L. asiaticus proteome. In addition, regions that are predicted to be disordered by no less than 2 predictors and they appear at the boundaries of protein domains count for another 5.8% of all residues. On the level of individual proteins, 65.9% of all Ca. L. asiaticus proteins exhibit greater than 80% coverage. It is important to note that we adopted conservative criteria for selecting structure templates, which

may underestimate the number of Ca. L. asiaticus proteins that can be confidently predicted by homology modeling. In summary, our results indicate that MESSA can help biologists to efficiently gain understanding about proteins and would be useful for biological studies.

CONCLUSIONS

We developed MESSA, a web service that integrated the results of a dozen of state-of-the-art sequence analysis tools to provide predictions on local sequence properties, 3D structure and function of a given protein. MESSA offers a friendly user interface and display the results convenient for navigation. Our bench-mark study showed that MESSA was able to offer extensive information for most of the proteins in a genome and assist structure and function prediction. We hope MESSA can help biologists to gain understanding about proteins under study.

METHODS

Assemble computational tools

MESSA assembles a dozen of well-established tools to perform analysis for an input protein sequence. First, it predicts the local sequence features (listed in **Table 1**) of a query protein by multiple predictors with default parameters. Second, it detects close homologs of the query in the NR and SWISS-PROT databases by 2 iterations of PSI-BLAST [18] with e-value 0.005 as cutoff. Based on PSI-BLAST alignments, these PSI-BLAST 2nd iteration hits in the NR database filtered by more than 40 percent coverage and less than 90% sequence identity were used to construct the sequence profile and calculate the positional conservation

indexes by AL2CO [33]. Third, related protein families were detected from Conserved Domain Database (CDD) [21-26] by RPS-BLAST [27] and HHsearch [19]. Fourth, to detect evolutionarily related protein structures and reveal domain architectures, we used three protocols: 1) PSI-BLAST against the NR database, starting from the sequence profiles built by the buildali.pl script in the HHsearch package [19], 2) RPS-BLAST and 3) HHsearch against the 70% sequence identity representatives of all PDB entries [29] and SCOP (version 1.75) database [30] starting from the single protein sequence.

Application of MESSA to the proteome of Candidatus Liberibacter asiaticus

All the sequences of *Ca. L. asiaticus* proteins predicted by NCBI gene prediction pipeline [34] were downloaded from the Genbank database (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Candidatus_Liberibacter_asiaticus_psy62_uid29835) and additional proteins that are detected by the SEED (Genome annotation web service on the basis subsystems, http://pseed.theseed.org/seedviewer.cgi) [35-36] but missed by NCBI were added. The relevant information about the *Ca. L. asiaticus* proteome was obtained from NCBI (http://www.ncbi.nlm.nih.gov/nuccore/CP001677), the SEED and Kyoto Encyclopedia of Genes and Genomes (KEEG, http://www.genome.jp/kegg-bin/show_genomemap_top?org_id=las) [37]. Computational analysis by MESSA was performed on each protein and the results were constructed as a website at: http://prodata.swmed.edu/congqian/Candidatus_Liberibacter_genome_home.html.

Based on the information in the website, we manually curated the functional assignment, predict the subcellular localization and selected structure templates for each protein. Functional annotations were mainly based on their close relationship to certain protein families or certain

protein that is curated manually in SWISS-PROT database. This relationship was verified on one hand by the statistical significance, coverage and alignment quality between the *Ca. L. asiaticus* protein and the identified families or domains, and on the other hand by the consensus between different methods and annotations made by other databases. In cases where agreement between methods is lacking or statistical support is marginal, identification of conserved sequence motifs, inspection of predicted structure and clustering of homologous proteins were applied to obtain function predictions.

Table 1 Programs used in MESSA for prediction of local sequence features and their interpretation

Feature	Meaning	Programs used	Output
Secondary structure	Assist three-dimensional structure and domain boundary prediction.	PSIPRED (v2.0) [49] SSPRO (v4.0) [50] DISEMBL (v1.5) [51], coils	PSIPRED and SSPRO predict 3-states secondary structures (H: α-helix, E: β-strand, C: coils); DISEMBL predict coils (lower-case letters highlighted in pink)
Disordered and flexible region	Assist three-dimensional structure prediction.	DISEMBL (v1.5) [51], hot loops	Loops that are likely to have high B factors in the X-ray crystallography (lower-case letters highlighted in pink)
		DISEMBL (v1.5) [51], missing DISPRO (v1.0) [52] DISOPRED (v2.0) [53] IsUnstruct (v2.02) [54]	Residues without a defined structure (represented by star marks and highlighted in red)
Transmembrane helix and Signal Peptide	Predict subcellular localization and transmembrane, reveal topology of transmembrane proteins and provide hints to the protein function.	TMHMM (v2.0) [55] TOPPRED[a] (v2.0) [56] HMMTOP[a] (v2.0) [57] MEMSAT (v3.0) [58]	H: transmembrane helix (colored in blue); h: not confidently predicted transmembrane helix; o: periplasmic loop, i: cytoplasmic loop. x: loop region (not specified as periplasmic or cytoplasmic).
		MEMSATSVM [59] Phobius [60]	H: transmembrane helix (colored in blue); S: signal peptide (colored in green); h: unconfident transmembrane helix; o: periplasmic loop, i: cytoplasmic loop.
		SignalP (v3.0) [61] (HMM mode) SignalP (v3.0) [61] (NN mode)	S: signal peptide (highlighted in green) o: periplasmic region; x: do not have signal peptide
Low-complexity region	Reveal false positive hits of homology search caused by matching of low-complexity region.	SEG [62]	The part with low diversity in amino acid composition (highlighted in pink), likely to be disordered or fold as α helices, such as coiled coil
Coiled coil	Assist three-dimensional structure prediction.	COILS [63]	x: coiled coils, highlighted in yellow
Conservation index	Reveal essential residues for the folding and function of a protein.	BLAST (hits filtered by > 40% coverage and < 90% identity are included in the profile), AL2CO (calculate conservation indices based on profile) [64]	Sequence highlighted by the conservation (highlighted from white, through yellow to dark red as conservation increases)

[a]TOPPRED and HMMTOP are mainly designed to predict the topology of a given membrane protein rather than distinguish transmembrane proteins from cytoplasmic ones. Thus they may recognize the hydrophobic buried helices in cytoplasmic proteins as transmembrane helices, leading to a high false positive rate.

Table 2 Confidence score of homologs from Swiss-Prot database.

Evaluation method	Criteria	Points
BLAST e-value	< 0.001	1
Sequence identity between the query and the hit	identity 30% to 50%, coverage > 40%	1
	identity 50% to 70%, coverage > 40%	2
	identity 70% to 90%, coverage > 40%	3
	identity 90% to 99%, coverage > 40%	4
	identity > 99%, coverage > 40%	5
BLAST alignment coverage for both query and hit	60% to 80%	1
	80% to 100%	2
The query against the proteome associated with the hit	Best hit	2
	N/A	1
The hit against the proteome associated with the query	Best hit	2
	N/A	1

Table 3 Confidence score of predicted gene ontology terms

Evaluation method	Criteria	Points
BLAST e-value	0.001	1
Sequence identity between the query and the hit	identity 30% to 50%, coverage > 40%	1
	identity 50% to 70%, coverage > 40%	2
	identity 70% to 90%, coverage > 40%	3
	identity 90% to 100%, coverage > 40%	4
Alignment coverage for query and hit	60% to 80%	1
	80% to 100%	2
Evidence code of the GO term assigned to the hit	EXP, IDA	3
	IPI, IMP, IGI, IEP, ISO, TAS	2
	ISS, ISA, ISM, IGC, IBA, IBD, IKR, IRD, RCA, NAS, IC, IEA	1
Consensus bonus	Associated with no less than three hits	2

EXP: inferred from experiment; GO: Gene Ontology; IBA: inferred from biological aspect of ancestor; IBD: inferred from biological aspect of descendant; IC: inferred by curator; IDA: inferred from direct assay; IEA: inferred from electronic annotation; IEP: inferred from expression pattern; IGC: inferred from genomic context; IGI: inferred from genetic interaction; IKR: inferred from key residues; IMP: inferred from mutant phenotype; IPI: inferred from physical interaction; IRD: inferred from rapid divergence; ISA: inferred from sequence alignment; ISM: inferred from sequence model; ISO: inferred from sequence orthology; ISS: inferred from sequence or structural similarity; NAS: non-traceable author statement; RCA: inferred from reviewed computational analysis; TAS: traceable author statement.

Table 4 Confidence score of predicted Enzyme Commission numbers

Evaluation method	Criteria	Points
Confidence score of homologous Swiss-Prot hit for EC number transfer	≥ 6 and < 8	1
	≥ 8 and < 10	2
	≥ 10	3
Consensus bonus	If the EC number is assigned for at least three different Swiss-Prot hits	1
Ezypred prediction (no confidence assigned to prediction)	If the EC number agrees with the prediction of Ezypred	2
EFICAz prediction confidence	Low confidence prediction	2
	0.6 to 0.7	2.5
	0.7 to 0.8	3
	0.8 to 0.9	3.5
	0.9 to 1	4

Table 5 Evaluation of homology modeling templates

Evaluation method	Criteria	Points
Sequence identity for BLAST, RPS-BLAST and HHSearch	20% to 40%	1
	40% to 60%	2
	60% to 80%	3
	80% to 90%	4
	90% to100%	5
HHsearch probability	80% to 85%	1
	85% to 90%	2
	90% to 99%	3
	99% to 99.99%	4
	99.99% to 100%	5
BLAST and RPS-BLAST e-value	1e-6 to 1e-2	1
	1e-6 to 1e-18	2
	1e-18 to 1e-54	3
	< 1e-54	4
Consensus bonus	Predicted by two methods	1
	Predicted by three methods	2

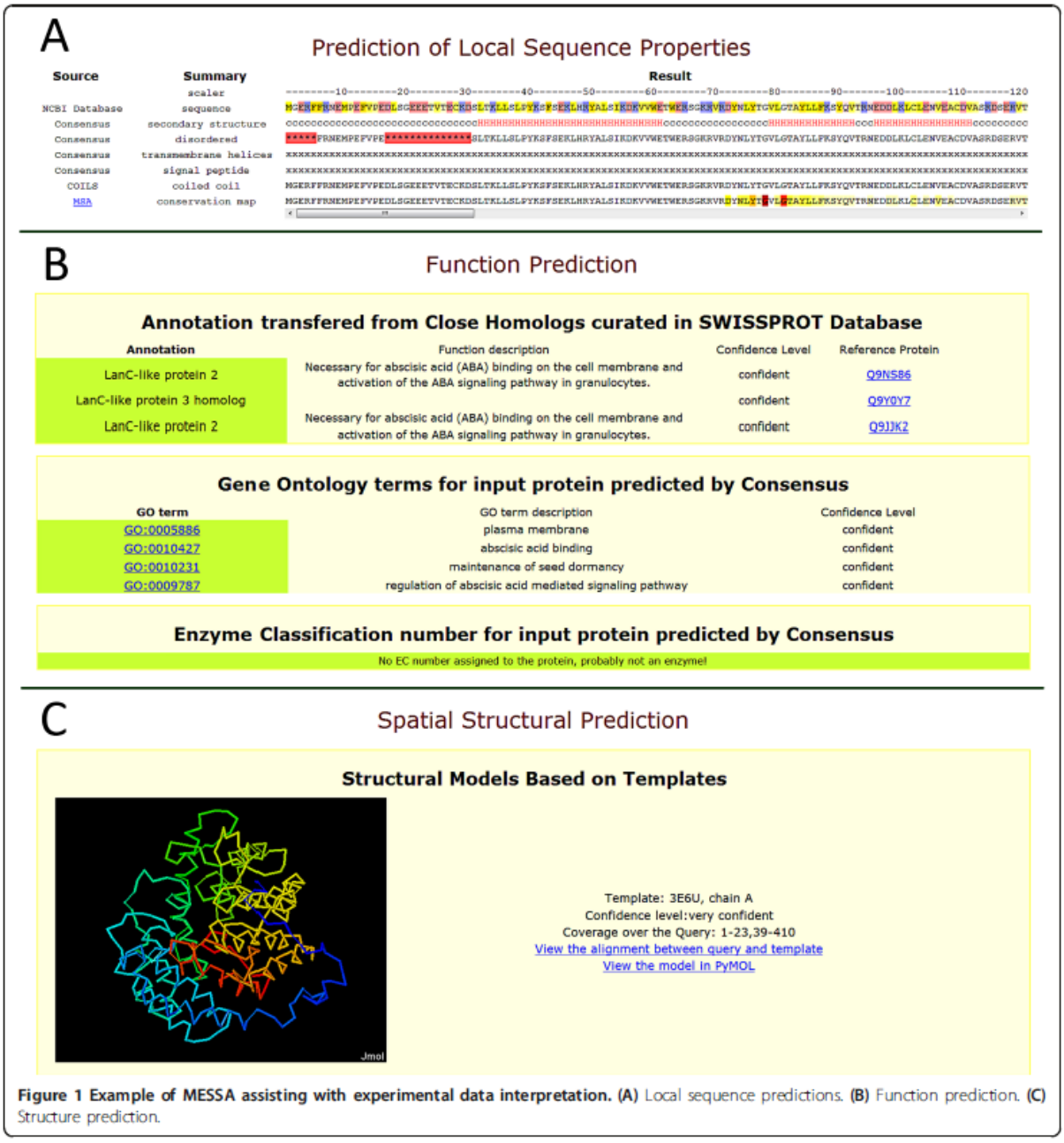


Figure 1 Example of MESSA assisting with experimental data interpretation. (A) Local sequence predictions. **(B)** Function prediction. **(C)** Structure prediction.

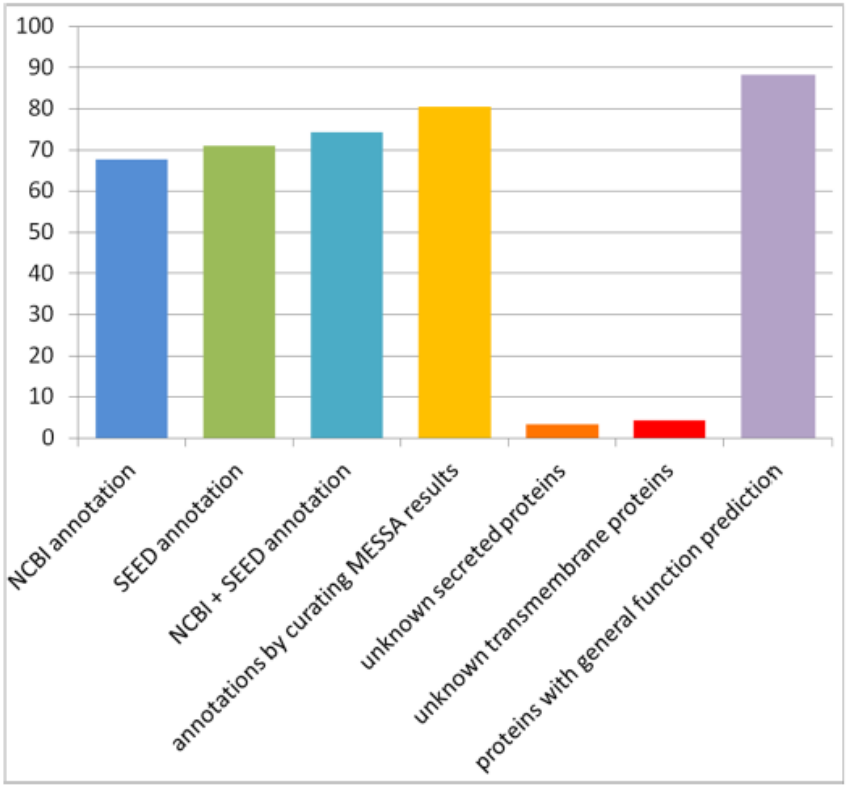


Figure 2 Fractions of proteins in *Ca.* L. asiaticus that are annotated by different methods.

REFERENCES

1. Liu X, Yue Y, Li B, Nie Y, Li W, Wu WH, Ma L: **A G protein-coupled receptor is a plasma membrane receptor for the plant hormone abscisic acid**. *Science* 2007, **315**(5819):1712-1716.
2. Gao Y, Zeng Q, Guo J, Cheng J, Ellis BE, Chen JG: **Genetic characterization reveals no role for the reported ABA receptor, GCR2, in ABA control of seed germination and early seedling development in Arabidopsis**. *Plant J* 2007, **52**(6):1001-1013.
3. Johnston CA, Temple BR, Chen JG, Gao Y, Moriyama EN, Jones AM, Siderovski DP, Willard FS: **Comment on "A G protein coupled receptor is a plasma membrane receptor for the plant hormone abscisic acid"**. *Science* 2007, **318**(5852):914; author reply 914.
4. Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H, Grishin NV: **CASP5 assessment of fold recognition target predictions**. *Proteins* 2003, **53 Suppl 6**:395-409.
5. Fischer D: **Servers for protein structure prediction**. *Curr Opin Struct Biol* 2006, **16**(2):178-182.
6. Kinch L, Yong Shi S, Cong Q, Cheng H, Liao Y, Grishin NV: **CASP9 assessment of free modeling target predictions**. *Proteins* 2011, **79 Suppl 10**:59-73.
7. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ: **JPred: a consensus secondary structure prediction server**. *Bioinformatics* 1998, **14**(10):892-893.
8. Ishida T, Kinoshita K: **Prediction of disordered regions in proteins based on the meta approach**. *Bioinformatics* 2008, **24**(11):1344-1348.

9. Klammer M, Messina DN, Schmitt T, Sonnhammer EL: **MetaTM - a consensus method for transmembrane protein topology prediction**. *BMC Bioinformatics* 2009, **10**:314.

10. Wallner B, Larsson P, Elofsson A: **Pcons.net: protein structure prediction meta server**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W369-374.

11. Ginalski K., Elofsson A., Fischer D., L. R: **3D-Jury: a simple approach to improve protein structure predictions**. *Bioinformatics* 2003, **19**(8):1015-1018.

12. Friedberg I, Harder T, Godzik A: **JAFA: a protein function annotation meta-server**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W379-381.

13. Pal D, Eisenberg D: **Inference of protein function from protein structure**. *Structure* 2005, **13**(1):121-130.

14. Rost B, Liu J: **The PredictProtein server**. *Nucleic Acids Res* 2003, **31**(13):3300-3304.

15. Letunic I, Doerks T, Bork P: **SMART 7: recent updates to the protein domain annotation resource**. *Nucleic Acids Res* 2012, **40**(Database issue):D302-305.

16. Kurowski MA, Bujnicki JM: **GeneSilico protein structure prediction meta-server**. *Nucleic Acids Res* 2003, **31**(13):3305-3307.

17. Duan Y, Zhou L, Hall DG, Li W, Doddapaneni H, Lin H, Liu L, Vahling CM, Gabriel DW, Williams KP *et al*: **Complete genome sequence of citrus huanglongbing bacterium, 'Candidatus Liberibacter asiaticus' obtained through metagenomics**. *Mol Plant Microbe Interact* 2009, **22**(8):1011-1020.

18. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **15**(17):3389-3402.

19. Soding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**(7):951-960.

20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

21. Ouzounis CA, Coulson RM, Enright AJ, Kunin V, Pereira-Leal JB: **Classification schemes for protein structure and function**. *Nat Rev Genet* 2003, **4**(7):508-519.

22. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al*: **The COG database: an updated version includes eukaryotes**. *BMC Bioinformatics* 2003, **4**:41.

23. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR *et al*: **CDD: a Conserved Domain Database for the functional annotation of proteins**. *Nucleic Acids Res* 2011, **39**(Database issue):D225-229.

24. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2012, **40**(D1):D290-D301.

25. Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, Fedorov B, Kiryutin B, O'Neill K, Resch W, Resenchuk S *et al*: **The National Center for Biotechnology**

Information's Protein Clusters Database. *Nucleic Acids Res* 2009, **37**(Database issue):D216-223.

26. Letunic I, Doerks T, Bork P: **SMART 6: recent updates and new developments**. *Nucleic Acids Res* 2009, **37**(Database issue):D229-232.

27. Marchler-Bauer A, Bryant SH: **CD-Search: protein domain annotations on the fly**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W327-331.

28. Zhang Y, Skolnick J: **Automated structure prediction of weakly homologous proteins on a genomic scale**. *Proc Natl Acad Sci U S A* 2004, **101**(20):7594-7599.

29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.

30. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**(4):536-540.

31. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A: **Comparative protein structure modeling using MODELLER**. *Curr Protoc Protein Sci* 2007, **Chapter 2**:Unit 2 9.

32. Bordoli L, Schwede T: **Automated Protein Structure Modeling with SWISS-MODEL Workspace and the Protein Model Portal**. *Methods Mol Biol* 2012, **857**:107-136.

33. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment**. *Bioinformatics* 2001, **17**(8):700-712.

34. Besemer J, Lomsadze A, Borodovsky M: **GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions**. *Nucleic Acids Res* 2001, **29**(12):2607-2618.

35. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M *et al*: **The RAST Server: rapid annotations using subsystems technology**. *BMC Genomics* 2008, **9**:75.

36. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R *et al*: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes**. *Nucleic Acids Res* 2005, **33**(17):5691-5702.

37. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.

38. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices**. *J Mol Biol* 1999, **292**(2):195-202.

39. Pollastri G, Przybylski D, Rost B, Baldi P: **Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles**. *Proteins* 2002, **47**(2):228-235.

40. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB: **Protein disorder prediction: implications for structural proteomics**. *Structure* 2003, **11**(11):1453-1459.

41. Cheng J, Sweredoski M, Baldi P: **Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data**. *Data Mining and Knowledge Discovery* 2005, **11**(3):213-222.

42. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT: **Prediction and functional analysis of native disorder in proteins from the three kingdoms of life**. *J Mol Biol* 2004, **337**(3):635-645.

43. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. *J Mol Biol* 2001, **305**(3):567-580.

44. von Heijne G: **Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule**. *J Mol Biol* 1992, **225**(2):487-494.

45. Tusnady GE, Simon I: **Principles governing amino acid composition of integral membrane proteins: application to topology prediction**. *J Mol Biol* 1998, **283**(2):489-506.

46. Jones DT: **Improving the accuracy of transmembrane protein topology prediction using evolutionary information**. *Bioinformatics* 2007, **23**(5):538-544.

47. Nugent T, Jones DT: **Transmembrane protein topology prediction using support vector machines**. *BMC Bioinformatics* 2009, **10**:159.

48. Kall L, Krogh A, Sonnhammer EL: **A combined transmembrane topology and signal peptide prediction method**. *J Mol Biol* 2004, **338**(5):1027-1036.

49. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0**. *J Mol Biol* 2004, **340**(4):783-795.

50. Wootton JC: **Non-globular domains in protein sequences: automated segmentation using complexity measures**. *Comput Chem* 1994, **18**(3):269-285.

51. Lupas A, Van Dyke M, Stock J: **Predicting coiled coils from protein sequences**. *Science* 1991, **252**(5009):1162-1164.

CHAPTER FIVE
Sequence Analysis of the *Candidatus* Liberibacter asiaticus proteins

INTRODUCTION

Candidatus Liberibacter asiaticus (*Ca. L. asiaticus*) is a Gram negative Alphaproteobacterium. It is closely associated with Citrus Greening (also called HuangLongBing, HLB), one of the most severe worldwide diseases of citrus. The ranking *Candidatus* is assigned to this bacterium as it cannot be maintained in bacterial culture. In nature, the bacterium is transmitted among citrus plants by the piercing-sucking insects, Asian citrus psyllids (*Diaphorina citri Kuwayama*). In the plant, *Ca. L. asiaticus* resides in the phloem tissue [1,2,3]. The infected citrus plants gradually develop symptoms such as yellow leaves, premature defoliation and aborted fruit, followed by the eventual death of the entire plant [4,5]. It is hypothesized that *Ca. L. asiaticus* infection could induce over-accumulation of callose in plant plasmodesmata pore units and sieve pores and inhibits phloem transport, contributing to HLB symptoms [1,2,3].

Ever since HLB was described, efforts have been devoted to understanding the plant response to the infection [1,2,6,7], to diagnosing HLB [8,9] and to controlling the disease [10,11,12]. However, a fundamental understanding of the mechanism of HLB or an ultimate way to save the citrus industry has yet to manifest. This lack of accomplishment is due in part to the limited success in culturing the bacterium [13], which makes carrying out experiments directly on *Ca. L. asiaticus* a challenge.

In 2009, the complete genome sequence of *Ca. L. asiaticus* was obtained [14] and verified [15]. With the sequences available, proteins from *Ca. L. asiaticus* can be studied *in vitro* or through heterologous expression. Such experiments have verified the function of a hypothetical ADP/ATP translocase [16] and identified a moderate inhibitor of the predicted *secA* gene product [17]. These findings demonstrate the possibility of understanding and controlling HLB at the molecular level. Given the genome sequence, computational analysis combined with manual curation can stimulate such research by predicting the structure and function of *Ca. L. asiaticus* proteins, identifying potential virulence factors and selecting drug targets to specifically inhibit the bacterium.

The *Ca. L. asiaticus* genome is highly reduced relative to other bacteria in the order *Rhizobiales*, likely related to its intracellular lifestyle [18]. Gene prediction and annotation pipeline [19] from National Center for Biotechnology Information (NCBI) and the RAST (Rapid Annotations using Subsystems Technology) server from the SEED (a web service for genome annotation based on subsystem approach) [20,21] have predicted 1233 protein-coding genes in the entire genome. This relatively small genome size allows careful analysis of all the *Ca. L. asiaticus* proteins *in silico* and prediction of their structure and function.

Protein sequence analysis relies heavily on homology inference [22,23,24]. The structures of homologous proteins provide templates for structure modeling, and the function of close homologs can be transferred in most cases to the protein of interest. Meanwhile, in the absence of confident homologs, the presence of certain functional motifs, the predicted 3D structure, the genomic context, the phylogenetic distribution, the known physical or functional

protein-protein interactions and the presence of certain local sequence features (eg. signal peptide and transmembrane helices) can still provide hints to the general function [25,26].

Here we report a database with extensive predictive information for all the 1233 *Ca. L. asiaticus* proteins. Information from various databases was gathered for each protein and essential sequence features, such as signal peptides and transmembrane regions, were predicted. Moreover, the evolutionarily related proteins, protein families, protein structures and protein domains detected by multiple procedures were identified and presented. This website aims to facilitate in-depth manual analysis of the *Ca. L. asiaticus* proteome, such as supporting and modifying function predictions, generating structure models, analyzing domain architectures and more importantly, identifying potential effectors of this pathogen and targets for controlling HLB. To illustrate the potential application of the database, we manually curated the results in the websites to predict the structure and function of each protein. More specifically, we analyzed the duplicated proteins in *Ca. L. asiaticus* proteome and studied the proteins whose closest homologs are from phylogenetically distant organisms instead of Alphaproteobacteria. These proteins with abnormal evolutionary history are candidates of horizontally transferred genes. As a result, we identified several potential virulence factors. Experimental study on these proteins may be helpful for understanding and controlling Citrus greening.

METHODS

Construction of the website

All the sequences of *Ca. L. asiaticus* proteins predicted by NCBI gene prediction pipeline were downloaded from the Genbank database (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Candidatus_Liberibacter_asiaticus_psy62_uid29835) and additional hypothetical proteins that are detected by the SEED (Genome annotation web service on the basis subsystems, http://pseed.theseed.org/seedviewer.cgi) but missed by NCBI were added. The relevant information about the *Ca. L. asiaticus* proteome was obtained from NCBI (http://www.ncbi.nlm.nih.gov/nuccore/CP001677), the SEED and Kyoto Encyclopedia of Genes and Genomes [27] (KEEG, http://www.genome.jp/kegg-bin/show_genomemap_top?org_id=las). For each protein, computational analyses were performed as follows.

First, we predicted the local sequence features (listed in Table 1) of each protein by multiple predictors with default parameters. Second, we detected their close homologs by 2 iterations of PSI-BLAST [28] from Non-redundant database (NR, 05/22/2011) with e-value 0.005 as cutoff. Based on PSI-BLAST alignments, these PSI-BLAST 2nd iteration hits filtered by more than 40 percent coverage and less than 90 (or 70) percent sequence identity were used to construct the sequence profile and calculate the positional conservation indexes by AL2CO [29]. Third, related protein families were detected from Conserved Domain Database (CDD) [30,31,32,33,34,35] by RPS-BLAST [36] and HHsearch [37]. Fourth, to detect evolutionarily related structures and reveal domain architectures, we used three protocols: 1) PSI-BLAST against the NR database (05/22/2011), starting from the sequence profiles built by the

buildali.pl script in the HHsearch package [37], 2) RPS-BLAST and 3) HHsearch against the 70 percent sequence identity representatives of all PDB entries (up to Jan, 2011), Structure Classification of Proteins (SCOP, version 1.75) database [38] and the Molecular Modeling DataBase (MMDB, up to Jan, 2011) from NCBI [39], with the single protein sequence as query. All the results and useful information from other resources (NCBI, SEED and KEGG) were parsed and represented in a web page (details described in **Results and Discussion**). All the web pages were assembled to establish a sequence analyses website for *Ca. L. asiaticus* proteome.

Application of the website

Based on the information in the website, we manually curated functional assignments for each protein and selected a structure template for homology modeling. Functional annotations were mainly based on their close relationship to certain protein families. This relationship was verified on one hand by the statistical significance, coverage and alignment quality between the *Ca. L. asiaticus* protein and the identified families or domains, and on the other hand by the consensus between different methods and annotations made by other databases. In cases where agreement between methods is lacking or statistical support is marginal, identification of conserved sequence motifs, inspection of predicted structure and clustering of homologous proteins were applied to obtain function predictions.

Homologous proteins within the *Ca. L. asiaticus* proteome were identified among BLAST (e-value cutoff 0.005, NR database) hits. *Ca. L. asiaticus* proteins were grouped manually in a single-linkage manner [40] requiring grouped proteins to have similar predicted

function. All the homologous groups with more than one protein were studied manually, from which potential virulence factors were identified and analyzed in detail. The taxonomy information of the top confident BLAST hits (e-value cut off 0.005) were examined, selecting those belonging to organisms other than *Alpharoteobacteria*. These proteins were then investigated carefully with the emphasis on identifying potential virulence factors.

RESULTS AND DISCUSSION

Description of the website

The results of computational analysis on all 1233 *Ca. L. asiaticus* proteins are presented as a website at http://prodata.swmed.edu/liberibacter_asiaticus/. The proteins are sorted by their position in the genome to allow easy analysis on their genomic context. A separate web-page is devoted to each protein and it contains the following information.

Section I. Basic Information (Fig. 1A):

This section provides relevant information from and links to other databases. Several existing annotations were listed, including: gene description (definition line in NCBI Protein Database), COG prediction (from NCBI, based on homologous relationship to COG cluster), Pfam domain (based on the best RPS-BLAST hit from Pfam families), KEGG prediction and the SEED prediction. By connecting our website with external established databases, this section offers an easy reference to all available information. Combining and comparing the annotations from different resources provides the basis for functional assignments.

Section II. Prediction of Local Sequence Features (Fig. 1B):

Local sequence property predictions, such as secondary structures or disordered regions, are helpful for predicting 3D structure, whereas, signal peptide and transmembrane helix predictions are suggestive of the protein localization. This section summarizes the predictions of local sequence features as listed in Table 1. The results from each predictor are represented as a string consisting of predicted status and aligned to the original protein sequence for convenient comparison.

Section III. Close Homologs (shown in Fig. 1C):

Close homologs or orthlogs usually preserve the same function inherited from a common ancestor, which is the basis for function prediction. Moreover, the phylogenetic distribution of these closely related proteins provides hints about the evolutionary history of the protein and reveals horizontal gene transfer (HGF) events. HGF has a profound impact on the evolution of bacterial pathogens and it is a common mechanism to gain virulence-associated genes originated in other organisms [41]. Thus, the 10 closest confident homologs detected by BLAST or 2 iterations of PSI-BLAST (E-value cutoff 0.005) are provided in ranked order. On top of this section, a summery line for each close homolog provides links to relevant information, including the NCBI gi linked to the relevant protein page at NCBI and a bar graph alignment overview linked to the pairwise BLAST or PSI-BLAST result and the taxonomy information, which is on the bottom of this section. Moreover, we specifically detected and reported homologous proteins (if any) from the same organism (*Ca. L. asiaticus*) so that these duplicated genes can be compared and analyzed together (example discussed below).

Section IV: Homologous Protein Family (shown in Fig. 1D):

Protein classification and the extensive information gathered for protein families in databases can assist in functional annotation. In this section, we listed homologous protein families and conserved domains identified by RPS-BLAST (E-value cutoff: 0.005) and HHsearch (probability cutoff: 90%) in ranked order. Information is summarized on top similar to that described in section III, with links to the external databases and the detailed alignments with the identified families listed on the bottom.

Section V. Homologous Structures and structure domains (illustrated in Fig. 1E):

Homology modeling remains the most reliable and effective way of structure prediction, and the detection of a homologous template is the key step for modeling the structure [24,42]. This section is designed for structure modeling. Homologous structures and structure domains detected by PSI-BLAST (e-value below 0.005), RPS-BLAST (e-value below 0.005) and HHsearch (probability higher than 90.0%) are listed in similar format as described in Section III. For each hit, the alignment and the corresponding structure displayed by Jmol (an open-source Java viewer for chemical structures in 3D, available at: http://www.jmol.org/) can be easily retrieved. These protein structures can be utilized as templates for homology modeling to generate a 3D structure model. For structure domains detected in SCOP, we also provide the classification hierarchy of the domain, which gives insights in the evolutionary history of the domain and suggests similarities to other proteins.

Overall prediction statistics on the *Ca. L. asiaticus* proteome.

With the information from the website, in-depth manual analysis can be conveniently carried out to predict the structure and function of each protein. In the genome sequence of *Ca.*

L. asiaticus, the gene prediction pipeline from NCBI and SEED detected 1233 protein coding genes, with 1046 in common. In addition, 63 proteins that are identified by either of the single gene prediction pipelines reveal confident homology to other proteins in the non redundant database. We consider these 1105 sequences to be confidently predicted proteins. The remaining 128 proteins exhibit a relatively small size (usually less than 60 residues), include low complexity sequence, lack similarity to any known protein, and are inconsistently predicted. These genes may represent falsely predicted open reading frames.

Among the confidently predicted proteins, 63 are likely to be extracellular proteins as they are predicted to harbor a signal peptide by no less than 2 methods, and they are not transmembrane proteins. 197 of all *Ca. L. aisaticus* proteins are likely to locate to the membrane of the bacterium, based on the consensus between no less than 3 out of 6 transmembrane-helix predictors, as well as the topology of the 3D structure templates. Confidently identified homology to known proteins or protein families allows us to predict the function for 80.1% of all confidently predicted proteins, while NCBI and SEED annotated 68.0% and 70.8% of them, respectively (74.1% taken together). Moreover, out of the 220 proteins without function predictions, 37.3% are predicted to be secreted or transmembrane proteins, indicating their general function in communicating with the environment.

Another application of the website is to identify homologous structures for template based structure modeling for all *Ca. L. asiaticus* proteins. The confident structure templates identified by programs (HHsearch probability above 90%, PSI-BLAST or RPS-BLAST E-value below 0.005) in this website and confirmed by manual curation cover 74.3% of all residues in the *Ca. L. asiaticus* proteome. In addition, some regions are predicted to be

disordered by no less than 2 predictors and they appear at the boundaries of protein domains. These regions count for another 5.8% of all residues. On the level of individual proteins, 65.9% of all *Ca. L. asiaticus* proteins exhibit greater than 80% coverage by the structure templates or disordered regions (Fig. 3). It is important to note that we adopted conservative criteria for selecting structure templates, which may underestimate the number of *Ca. L. asiaticus* proteins that can ultimately be predicted by homology modeling.

Other application of the website and potential virulence factors

More specific analyses on the proteome can be performed conveniently with the assistance of this website. For instance, we analyzed groups of homologous proteins within the proteome and proteins with abnormal evolutionary history, placing emphasis on the identification of potential virulence factors. We refer to virulence factors as gene products that enable a pathogen to colonize in the host, battle with the defense system and cause damage or inflammation to the host[43]. Plants exhibit pathogen-inducible defense mechanisms and the basal defense could be elicited by the pathogen-associated molecular patterns (PAMPs). Known PAMPs include bacterial lipopolysaccharide, peptidoglycan, and flagellin [44]. The *Ca. L. asiaticus* proteome includes almost all components of the flagellar assembly, including flagellin (FliC: CLIBASIA_02090), which might be able to initiate PAMP-triggered immunity (PTI) responses in Citrus. Common PTI responses include callose deposition, ethylene production and induction of pathogenesis-related proteins that can halt the bacterium from further colonization [44,45]. The detection of accumulated callose in plasmodesmata pore units and sieve pores after *Ca. L. asiaticus* infection supports the existence of PTI in Citrus [3].

Similar to other plant pathogens, *Ca. L. asiaticus* might produce virulence factors to interfere with PTI and escape from the plant immune responses. These pathogenic factors are the key to understanding the mechanism of HLB.

Homologous protein groups within the genome

Only 22% of *Ca. L. asiaticus* proteins have detectable homologs by BLAST within the same proteome, which is lower than the average (31%) for bacteria proteomes of similar size. Based on detected sequence relationships, we identified all the close homologous clusters within *Ca. L. asiaticus* proteome. The distribution of cluster size is shown in Fig. 4 (trivial clusters consisting of just one protein excluded). We further studied clusters of homologs with more than one protein and classified them into 3 categories according to our interpretation of the duplication events.

The first category is **Ancient Duplication events during the functional divergence of proteins**. They represent either paralogs with similar function but different specificity and partners (such as ABC-transporters, GTP-binding proteins, amino acid-tRNA synthetases) or evolutionarily related proteins that cooperate with each other in the same pathway or complex (such as pilin component proteins or NADH dehydrogenase subunits). Such phenomena where paralogous proteins either cooperate with each other in the same process or participate in similar steps of different pathways are common during the evolution of protein function [46]. The largest cluster is the ABC-type P-loop ATPases. The ABC-type ATPase is the largest protein family in bacteria [47] and they mainly work together with a transmembrane permease to function as ATP-binding cassette transporters (ABC transporters) [48]. In this parasitic

bacterium, their roles of gaining nutrition, resisting harmful compounds in the environment and constructing outer membrane are crucial for the survival of the bacterium.

The second category of duplicated genes are **recent duplication likely caused by the integration of bacteriophage**. This category includes protein pairs with very high identity (more than 90% and even 100%), indicating recent duplication events. Recently, the sequence of the *SC1 Liberibacter phage* and *SC2 Liberibacter phage* [49], which coincides with *Ca. L. asiaticus* and can integrate into the bacterial genome, reveals that the current sequence of *Ca. L. asiaticus* str. psy62 (GenBank ID: CP001677.5) harbors an integrated *SC1 Liberibacter* prophage. The *SC1 Liberibacter phage* genome sequence can be aligned to the *Ca. L. asiaticus* genome with over 98% identity in nucleotide sequence. Moreover, 42 of *Ca. L. asiaticus* protein coding genes consecutively located on the chromosome match exactly all the proteins in *SC1 Liberibacter* phage. Moreover, one *SC1 Liberibacter phage* protein can be aligned to two duplicated *Ca. L. asiaticus* proteins at their N- and C-terminal halves respectively. We hypothesize that these two proteins contain the sites where the phage integrated into the bacterial genome. Many proteins in the prophage region, such as SNF2 family DNA/RNA helicases, NAD-dependant DNA ligase and guanylate kinases, have close homologs in the bacterial proteome. And these are likely homologous recombination events caused by the integration of the bacteriophage. The proteins in the integrated *SC1 Liberibacter phage* region, especially proteins that are not related to the life cycle of the phage deserve special interest, as the bacteriophage is a common vector for transmitting pathogenicity islands among bacteria [50].

Proteins that may contribute to the virulence of *Ca. L. asiaticus* are of primary interest, and thus we list them in a special category. Their suspected role in bacterial pathogenicity is supported by some of the following criteria: (1) presence of signal peptide, (2) lack of detectable homologs in other organisms, likely resulting from fast evolution, (3) homology to known virulence factor. The existence of multiple copies of similar virulence proteins may intensify the pathogenicity.

One of the most unusual homologous groups is the von Willebrand factor type A (shown in Fig. 5) (vWFA) domain containing proteins. There are 5 copies of such proteins in the *Ca. L. asiaticus* proteome. Only von Willebrand factor type A CLIBASIA_05050 (gi: 254781108) and CLIBASIA_05060 (gi: 254781110) are annotated as vWFA, however, all evidence suggests the hypothetical proteins CLIBASIA_01365 (gi: 254780388), CLIBASIA_03630 (gi: 254780833) and CLIBASIA_04165 (gi: 254780934) to include vWFA domains. Starting from any protein in this group, all homology detection methods we applied detect vWFA domain at the C-terminus of the protein with confident statistics (e-value below 1e-5 for RPS-BLAST and HHsearch probability above 99.8%). Moreover, every protein in this group preserves a metal ion dependent adhesion site (MIDAS, shown in Fig. 5B), which is the signature motif of vWFA domain. Transmembrane helices were detected at the N-termini of these proteins and the vWFA domains are predicted to be on the extracellular side by most predictors. vWFA domains are mainly found as extracellular eukaryotic domains involved in cell adhesion, migration, homing, pattern formation, and signal transduction [51,52]. Although the function of vWFA domains in bacteria is still unclear, they have been detected in some repeat in toxins (rtx, typical virulence factors secreted by Type I secretion system) and are

proposed to be involved in the virulence of the human pathogen, *Legionella pneumophila.* [53] Similarly, these vWFA domains that are predicted to be exposed on the surface of *Ca. L. asiaticus* may utilize their MIDAS motif to interact with host proteins and contribute to the virulence of *Ca. L. asiaticus.*

Another potentially harmful for the plant group contains four hypothetical proteins that are all predicted to harbor signal peptides (gi: 254780135, 254781007, 254780914 and 254780929). These four proteins share above 90 percent sequence identity with each other, and they are highly likely to preserve the same function. No confident homologs can be detected for them from organisms outside the *Candidatus Liberibacter* genus, indicating that they are fast-evolving proteins that may have unique functions. The other bacterium in this genus, *Candidatus Liberibacter solanacearum,* which has one copy of this unknown protein, is the pathogen of "zebra chip" disease in potatoes [54]. Due to the lack of homologs outside the *Candidatus Liberibacter* genus, we cannot predict the structure or exact function of these proteins, but the fact that they are duplicated secreted proteins unique to two plant pathogens already suggests their possible virulence role in HLB.

Interestingly, 1.0% of the *Ca. L. asiaticus* proteins (listed in Table 2) have detectable homologs (by BLAST or PSI-BLAST) only within this proteome (up to 05/22/2011, after the closely related *Candidatus Liberibacter solanacearum* was sequenced). Despite the possibility that they are "novel" genes originating in this bacterium, it is more likely that these genes have diverged from their homologs so fast that the relationship is hardly detectable. Fast evolution is considered to be an important character of virulence factors [55], and thus these "redundant" and fast-evolving proteins in a small genome might be related to the virulence of *Ca. L.*

asiaticus. Moreover, the surprising prediction result that many of them are either predicted to be secreted proteins or membrane proteins further signifies the possibility for some of them to be virulence factors associated with HLB.

Analysis of proteins with abnormal evolutionary history

The taxonomy information of the close homologs of a protein is an indicator of its evolutionary history. Thus, we inspected the taxonomic information for each protein's first hit in NR database detected by BLAST. We excluded the closest bacterium, *Ca. L. solanacearum* as some HGF events we are interested in might happen before their divergence. As expected, most (77%) of *Ca. L. asiaticus* proteins have closest homologs from the same *Alphaproteobacterium* phylogenetic class. Only 11% of all proteins display closest homologs from other classes (including bacteria, viruses and eukaryotes) and the other 12% appear only in the *Candidatus Liberibacter* genus. (shown in Fig. 6). However, careful manual analysis of proteins with close eukaryotic homologs reveals that these proteins are more likely to be horizontally transferred to the certain eukaryotic proteomes from *Alphaproteobacteria* or simply be a contamination in the sequencing of certain eukaryotic protein. For example, the best BLAST hit (proteins from *Liberibacter* genus excluded.) of the flagellar biosynthesis protein FliQ (CLIBASIA_02030) is from *Gossypium hirsutum* (upland cotton). Given all other close homologs are from *Alphaproteobacteria* and FliQ is a clearly bacterial gene, it is likely this hit is a contamination or a HGF events from *Alphaproteobacteria* to *Gossypium hirsutum.*

Proteins with closest homologs from viruses provide hints to the integration of bacteriophage into the genome. Most of them are from the recently integrated *SC1 liberibacter*

phage. It is important to note that the proteins from the integrated phage might be the product of bacterial genes captured by the phage. In addition, our analysis revealed 13 other phage-related proteins that do not belong to the *SC1 liberibacter phage*. This result indicates that another prophage might have integrated into this genome, but its genome has been reduced greatly during long time of evolution.

Out of these proteins with potentially abnormal evolutionary history, we identified several potential virulence factors. As an example, hypothetical protein CLIBASIA_03975 (GI: 254780898), was analyzed in detail with the assistance of information from our website. Homologous families of CLIBASLA_03975 detected by BLAST, RPS-BLAST and HHserach consensually suggest its close relationship to the dual specificity phosphatases (DSP, protein serine/threonine and tyrosine phosphatase) protein family. Structure prediction also reveals phosphotyrosine protein phosphatases II fold proteins (shown in Fig. 7), with the functional motifs for DSP preserved and located in a shallow cleft on the surface of the structure. Protein Ser/Thr and Tyr phosphatase functions as typical components of eukaryotic signaling pathways [53], while bacteria usually utilize histidine kinase for signal transduction. Although these phosphatases can participate in a bacteria's own signaling pathway [56], they likely act as virulence factors since they can easily interact with the signaling system of the host [53,57]. There is no clearly predicted protein Ser/Thr or Tyr kinase in the *Ca. L. asiaticus* proteome that could function as a counterpart of a DSP, suggesting that this predicted DSP actually participates in a signaling pathway of the plant host, potentially interfering with immune reactions that involve protein kinase signaling cascades [45]. More strikingly, local sequence feature prediction reveals a signal peptide at its N-terminus of this protein by several methods,

suggesting this is a secreted protein and further increasing the possibility of it being a virulence factor.

CONCLUSIONS

We carried out computational analysis on all *Ca. L. asiaticus* proteins and presented the results as a website that shows computational analyses for each protein. With the assistance of this website, we performed manual curation to predict the function, selected structure template and identified potential virulence factors and drug targets. The website serves as an encyclopedia of the *Ca. L. asiaticus* proteome to help researchers characterize the bacterial proteins, understand the mechanism of Citrus Greening and guide the development of methods to control the disease.

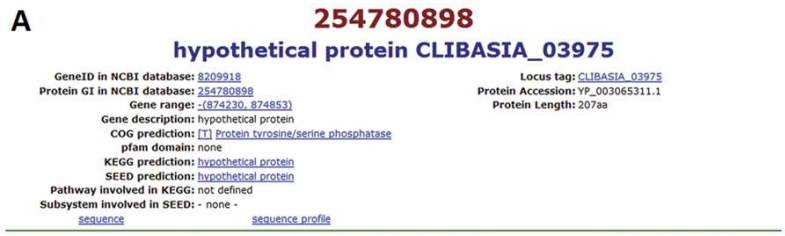
Table 1. The Predicted Local Sequence Features.

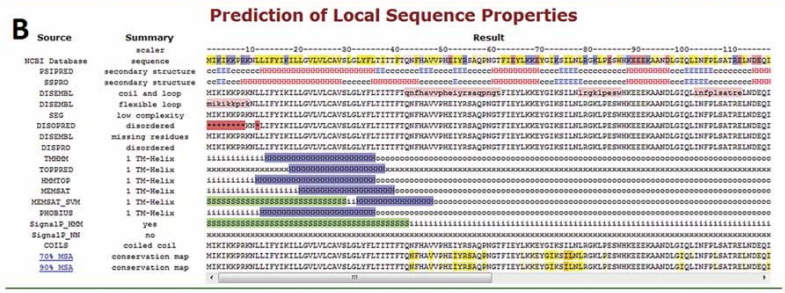
Feature	Programs Used For The Prediction	Implication
Secondary structure	PSIPRED (v2.0) [81] and SSPRO (v4.0) [82]	assist 3D structure and domain boundary prediction
Disordered or flexible region	DISEMBL (v1.5) [83], DISPRO (v1.0) [84] and DISOPRED (v2.0) [85]	assist 3D structure modeling and indicate the domain boundaries
Transmembrane helix	TMHMM (v2.0), TOPPRED (v2.0), HMMTOP (v2.0), MEMSAT (v3.0), MEMSATSVM and Phobius	predict subcellular localization; provide hints to the protein function. predict the topology of membrane proteins
Signal peptide	SignalP (v3.0), Phobius and MEMSATSVM	predict secreted proteins that could potentially be virulence factors
Low-complexity	SEG [86]	Reveal false positive hits of homology search caused by matching of low-complexity region
Coiled coil	COILS [87]	reveal false positive hits of homology search caused by matching of non-homologous coiled coils
Conservation	PSI-BLAST, AL2CO	reveal essential residues for the folding and function of a protein

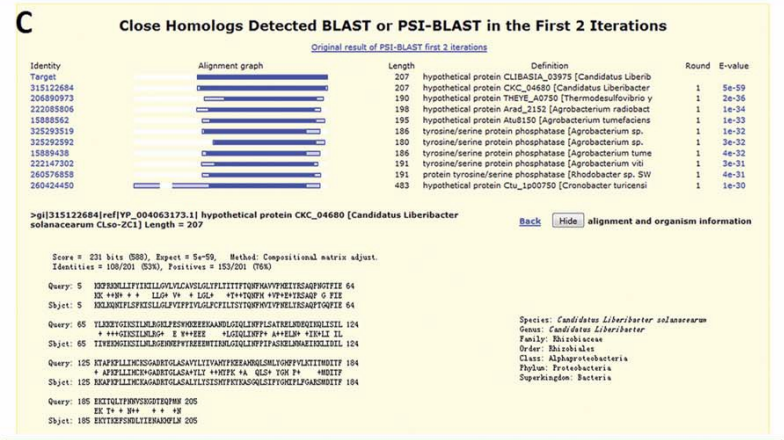
doi:10.1371/journal.pone.0041071.t001

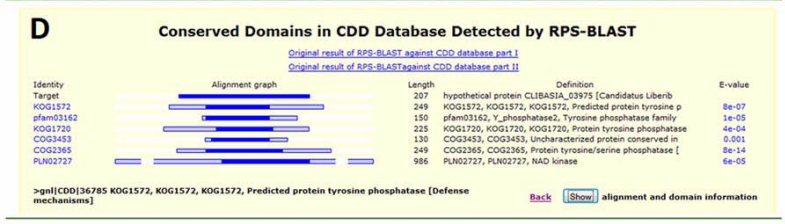
Table 2. Duplicated proteins that are unique to *Ca.* L. asiaticus proteome.

Group	Locus	gi	Comments
I	CLIBASIA_02215	254780556	with SP, potential virulence factor
	CLIBASIA_04405	254780980	with SP, potential virulence factor
II	CLIBASIA_03915	254780886	with SP, potential virulence factor
	CLIBASIA_04530	254781005	with SP, potential virulence factor
III	CLIBASIA_04425	254780984	with SP, potential virulence factor
	CLIBASIA_05140	254781126	do not have the SP part
	CLIBASIA_04410	254780981	with SP, potential virulence factor
IV	CLIBASIA_00440+ CLIBASIA_00445	254780203+254780204	Two neighboring proteins both aligned to part of CLIBASIA_05480. It is possible they are psuedogenes
	CLIBASIA_05130+ CLIBASIA_05135	254781124+254781125	Two neighboring proteins both aligned to part of CLIBASIA_05480. It is possible they are psuedogenes
	CLIBASIA_05480	254781189	Transmembrane protein


A
254780898
hypothetical protein CLIBASIA_03975
GeneID in NCBI database: 8209918
Protein GI in NCBI database: 254780898
Gene range: -(874230, 874853)
Gene description: hypothetical protein
COG prediction: [T] Protein tyrosine/serine phosphatase
pfam domain: none
KEGG prediction: hypothetical protein
SEED prediction: hypothetical protein
Pathway involved in KEGG: not defined
Subsystem involved in SEED: - none -
Locus tag: CLIBASIA_03975
Protein Accession: YP_003065311.1
Protein Length: 207aa
sequence
sequence profile


B
Prediction of Local Sequence Properties
Source
Summary
Result
NCBI Database
sequence
PSIPRED
secondary structure
SSPRO
secondary structure
DISEMBL
coil and loop
DISEMBL
flexible loop
SEG
low complexity
DISOPRED
disordered
DISEMBL
missing residues
DISPRO
disordered
TMPRM
1 TM-Helix
TOPPRED
1 TM-Helix
HMMTOP
1 TM-Helix
MEMSAT
1 TM-Helix
MEMSAT_SVM
1 TM-Helix
PHOBIUS
1 TM-Helix
Signal_HMM
yes
Signal_NN
no
COILS
coiled coil
70% MSA
conservation map
90% MSA
conservation map


C
Close Homologs Detected BLAST or PSI-BLAST in the First 2 Iterations
Original result of PSI-BLAST first 2 iterations
Identity
Alignment graph
Length
Definition
Round
E-value
Target
315122684
206890973
22085806
15888562
325293519
325292592
15889438
222147302
260576858
260424450
>gi|315122684|ref|YP_004063173.1| hypothetical protein CKC_04680 [Candidatus Liberibacter solanacearum CLso-ZC1] Length = 207
Back
Hide
alignment and organism information
Species: Candidatus Liberibacter solanacearum
Genus: Candidatus Liberibacter
Family: Rhizobiaceae
Order: Rhizobiales
Class: Alphaproteobacteria
Phylum: Proteobacteria
Superkingdom: Bacteria


D
Conserved Domains in CDD Database Detected by RPS-BLAST
Original result of RPS-BLAST against CDD database part I
Original result of RPS-BLASTagainst CDD database part II
Identity
Alignment graph
Length
Definition
E-value
Target
KOG1572
pfam03162
KOG1720
COG3453
COG2365
PLN02727
>gnl|CDD|36785 KOG1572, KOG1572, Predicted protein tyrosine phosphatase [Defense mechanisms]
Back
Show
alignment and domain information

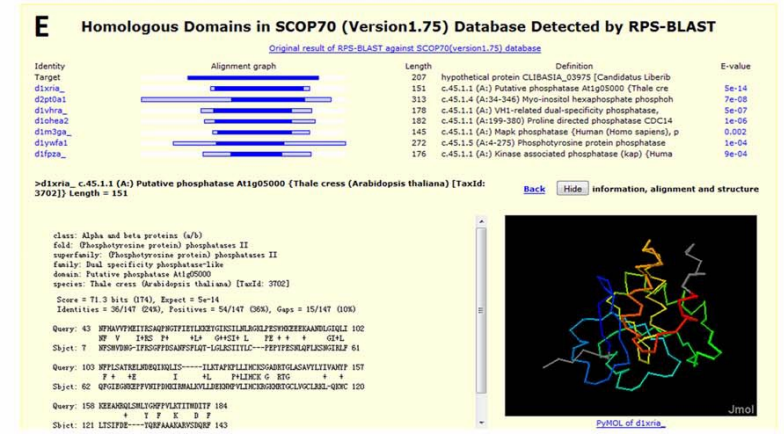

E
Homologous Domains in SCOP70 (Version1.75) Database Detected by RPS-BLAST
Original result of RPS-BLAST against SCOP70(version1.75) database
Identity
Alignment graph
Length
Definition
E-value
Target
d1xria_
d2ot0a1
d1vhra_
d1ohea2
d1m3ga_
d1vwfa1
d1fpza_
>d1xria_ c.45.1.1 (A:) Putative phosphatase At1g05000 {Thale cress (Arabidopsis thaliana) [TaxId: 3702]} Length = 151
Back
Hide
information, alignment and structure
class: Alpha and beta proteins (a/b)
fold: (Phosphotyrosine protein) phosphatases II
superfamily: (Phosphotyrosine protein) phosphatases II
family: Dual specificity phosphatase-like
domain: Putative phosphatase At1g05000
species: Thale cress (Arabidopsis thaliana) [TaxId: 3702]
Jmol
PyMOL of d1xria_

Figure 1 (see previous page) Illustration of the webpage

(A) Section I: basic information, function predictions from different resources and links to other databases.

(B) Section II: local sequence feature prediction. It contains the following information: (1) sequence (highlighted according to the property of amino acid) from NCBI database; (2),(3) secondary structure prediction by PSIPRED and SSPRO (H: α helix, E: β strand, C: coils); (4) Coil and loop (highlighted in pink) predicted by DISEMBL; (5) Flexible loop (highlighted in pink) predicted by DISEMBL; (6) Low complexity region (highlighted in light red) predicted by SEG; (7)-(9): Disordered region (highlighted in red) prediction by DISPRED, DISEMBL and DISPRO; (10)-(15) Transmembrane helix (highlighted in blue) prediction by TMHMM, TOPPRED2, HMMTOP, MEMSAT, MEMSATSVM, Phobius; (14)-(17) Signal Peptide (highlighted in green) prediction by MEMSATSVM, Phobius, SignalP Hidden Markov Model mode and SignalP Neural Network mode; (18) Coiled coils (highlighted in yellow) prediction by COILS; (19),(20) Sequence colored by conservation (highlighted from white, through yellow to dark red as the level of conservation increases) computed on the Multiple Sequence Alignment of homologous proteins filtered by 70% or 90% sequence identity.

(C) Section III: top 10 homologs detected by BLAST or 2 iterations of PSIBLAST are listed. For each hit, the alignment and taxonomy information are provided.

(D) Section IV: homologous protein family and conserved domains detected by RPS-BLAST. The confident hits detected by certain method are listed and the relative information of each protein family and its alignment to the C. L. asiaticus protein can be retrieved

(E) Section VI: evolutionary related protein domains detected by RPS-BLAST in SCOP database. It includes a table summarizing all confident hits, followed by is the relative information, the alignment and the 3D structure for each detected structure domain.

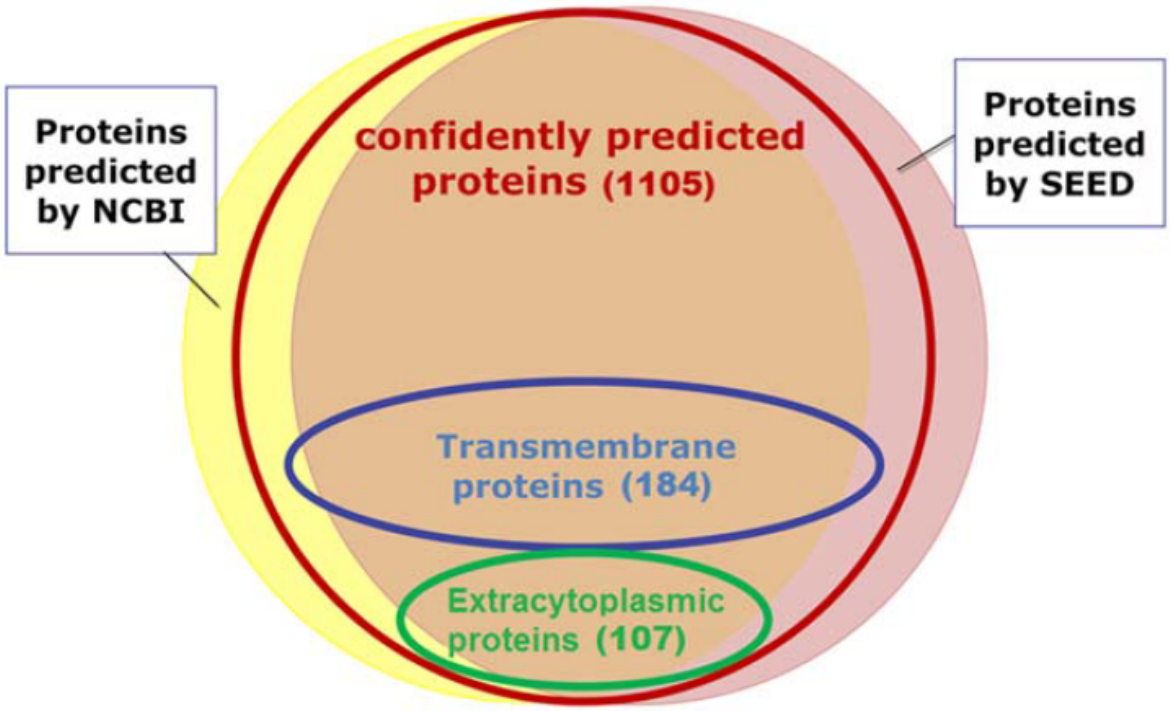


Figure 2. Venn diagram of the predicted protein coding genes by different methods in the *Ca.* L. asiaticus genome. The yellow disk represents the set of protein coding genes identified by NCBI and the pink disk stands for the set of protein coding genes predicted by the SEED. The red, blue and green circle includes all confidently predicted protein coding genes, transmembrane proteins and secreted proteins via Sec in the proteome after manual inspection.

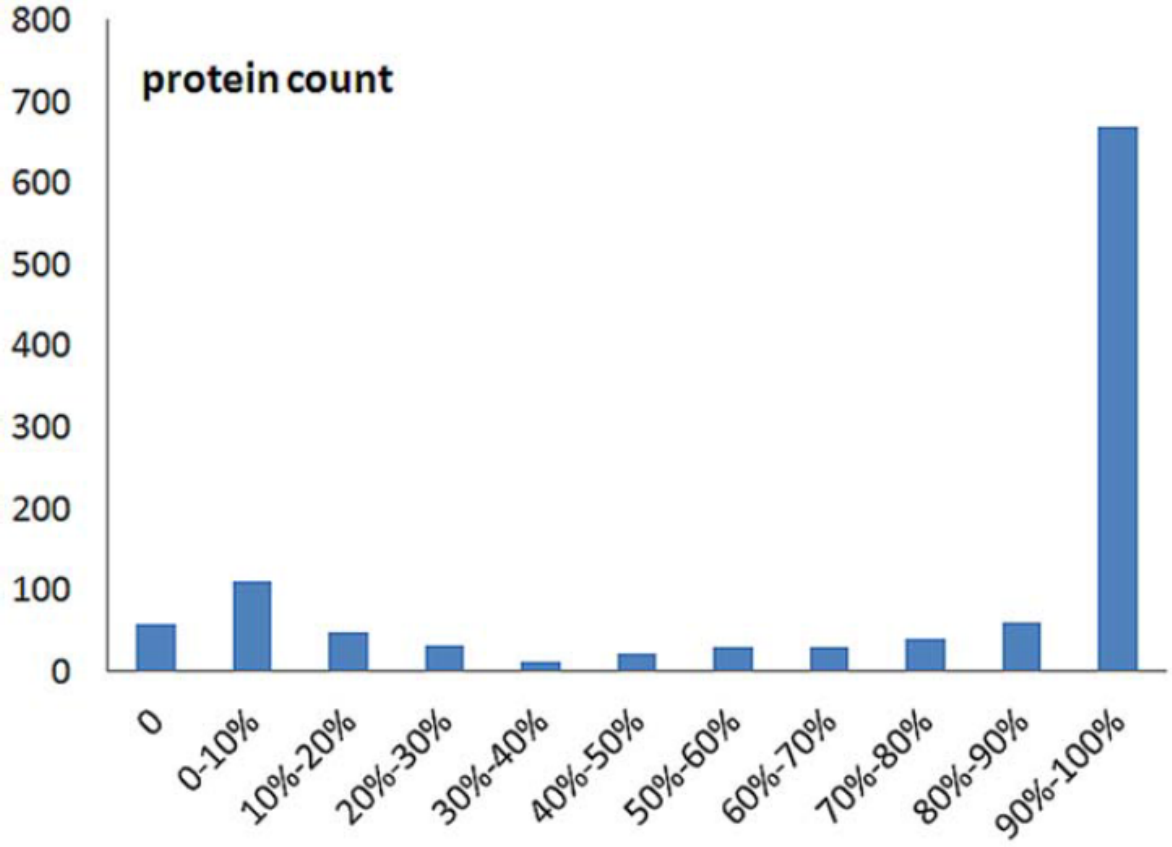


Figure 3. The distribution of 3D structure prediction coverage for each protein.

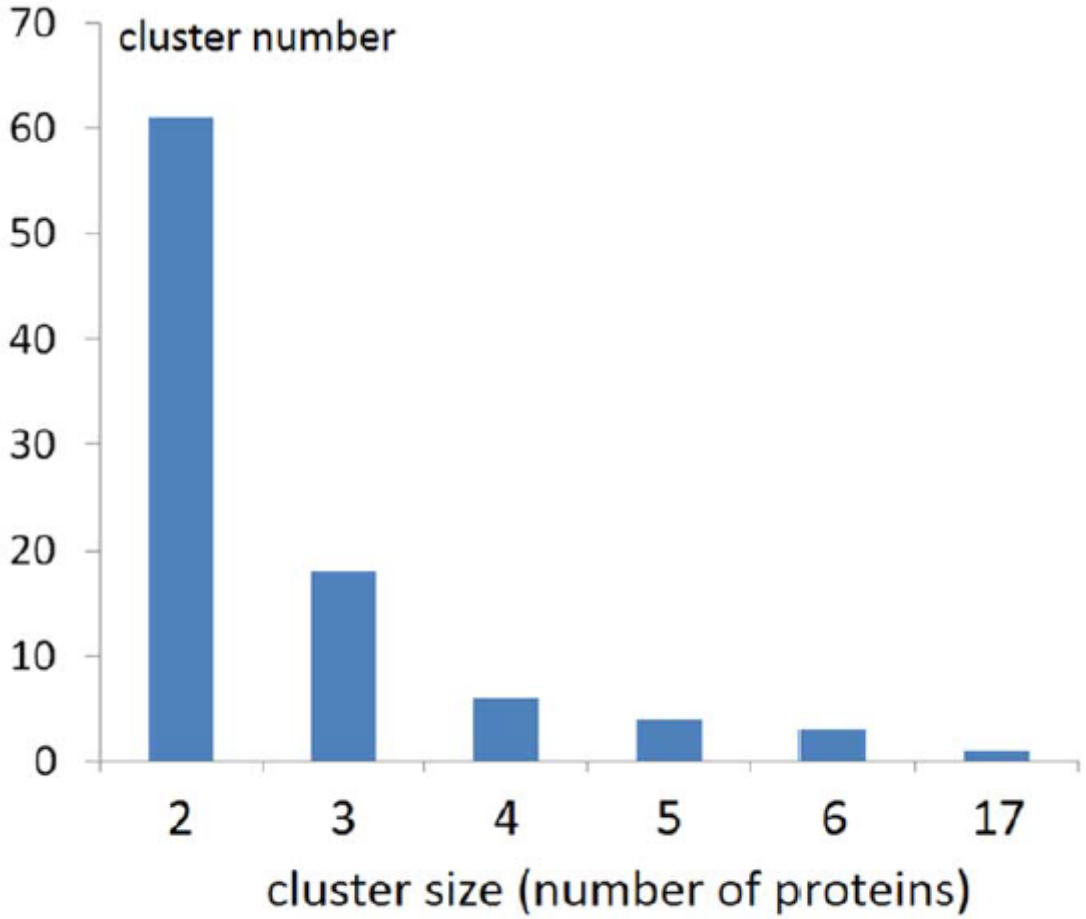


Figure 4. Distribution of homologous protein cluster sizes (clusters of single proteins excluded) within the *Ca.* L. asiaticus proteome.

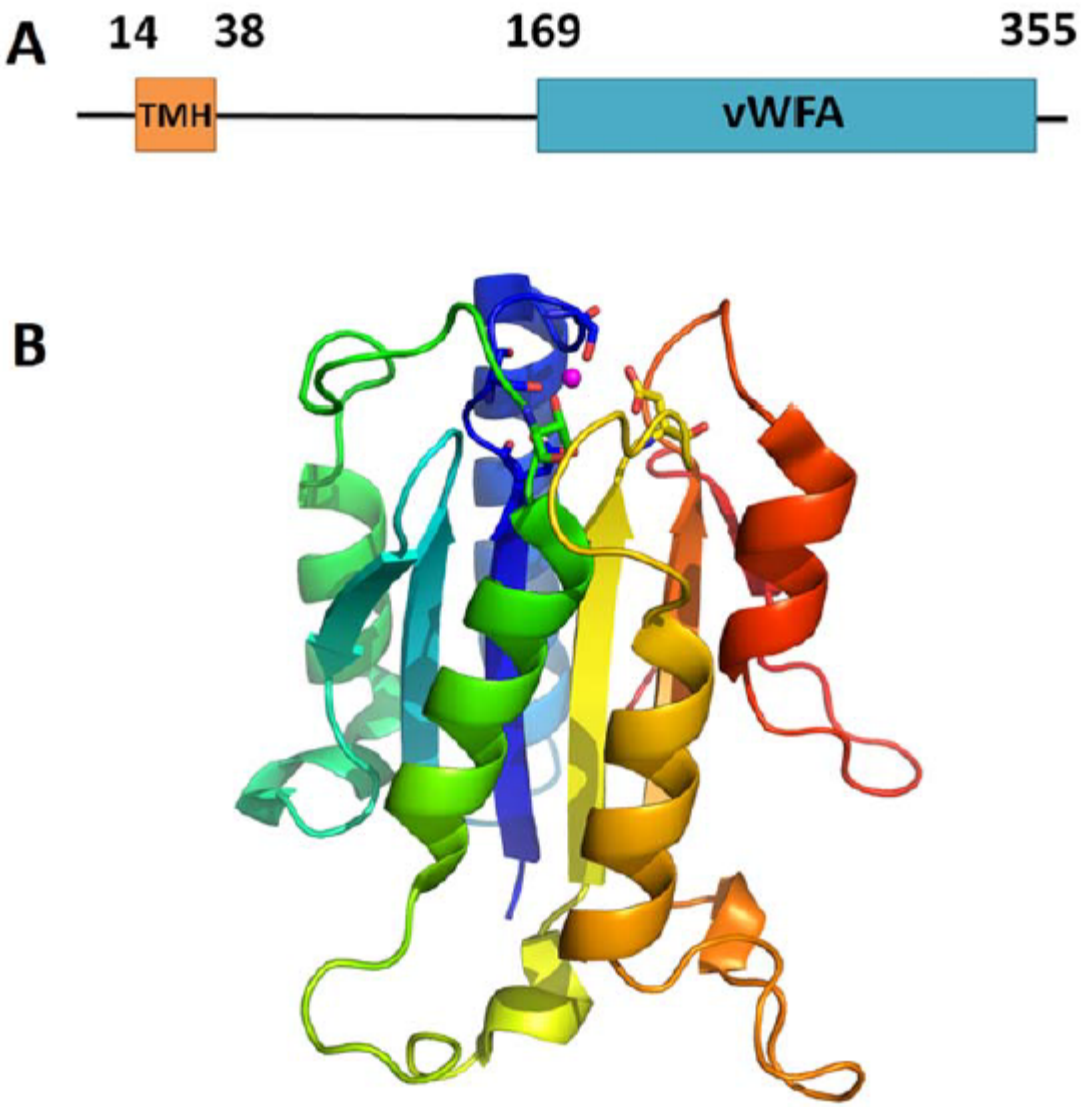


Figure 5. Potential virulence factor, von Willebrand factor type A domain containing protein. (locus: CLIBASIA_03630, gi: 254780833). (A) Domain diagram of the protein (B) Predicted structure of the protein colored in rainbow. The side-chains of the conserved residues for metal binding are shown.

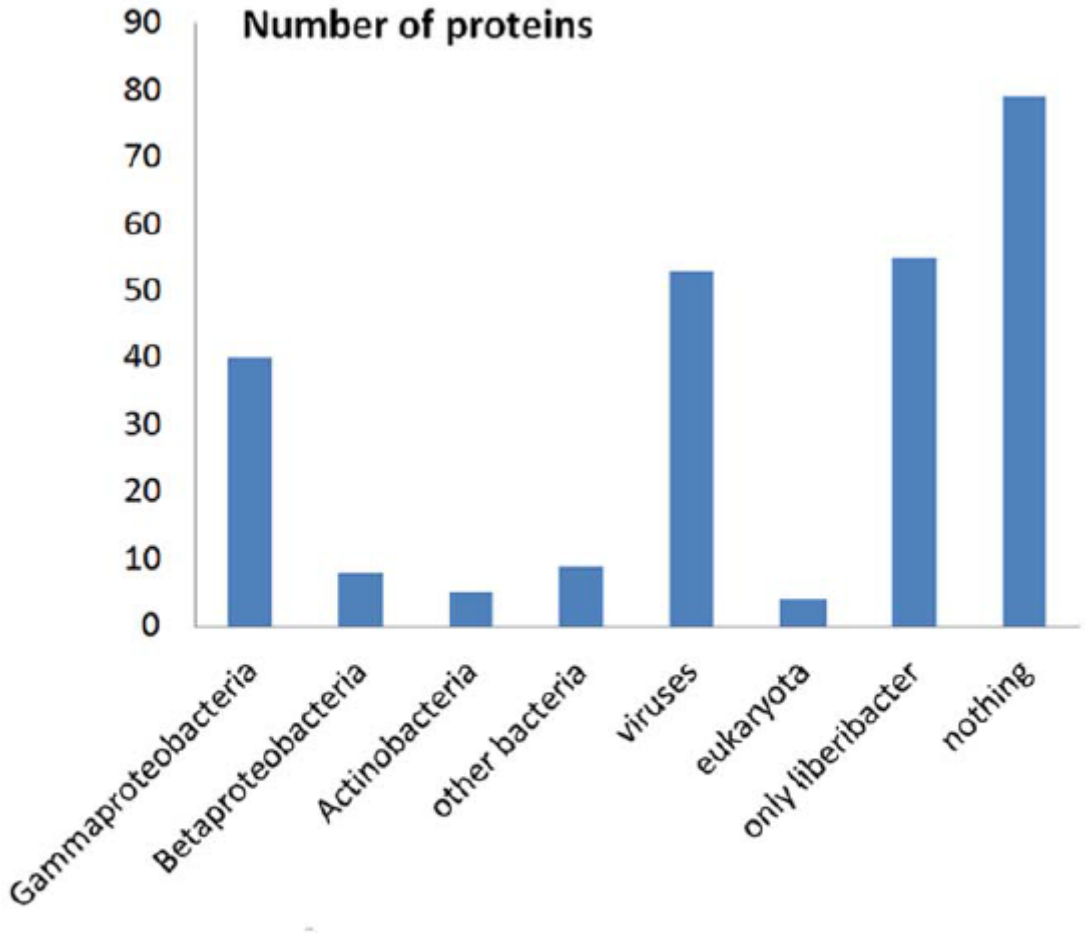


Figure 6. The distribution of *Ca.* L. asiaticus proteins' closest homologs among organisms. (proteins from Liberibacter genus excluded).

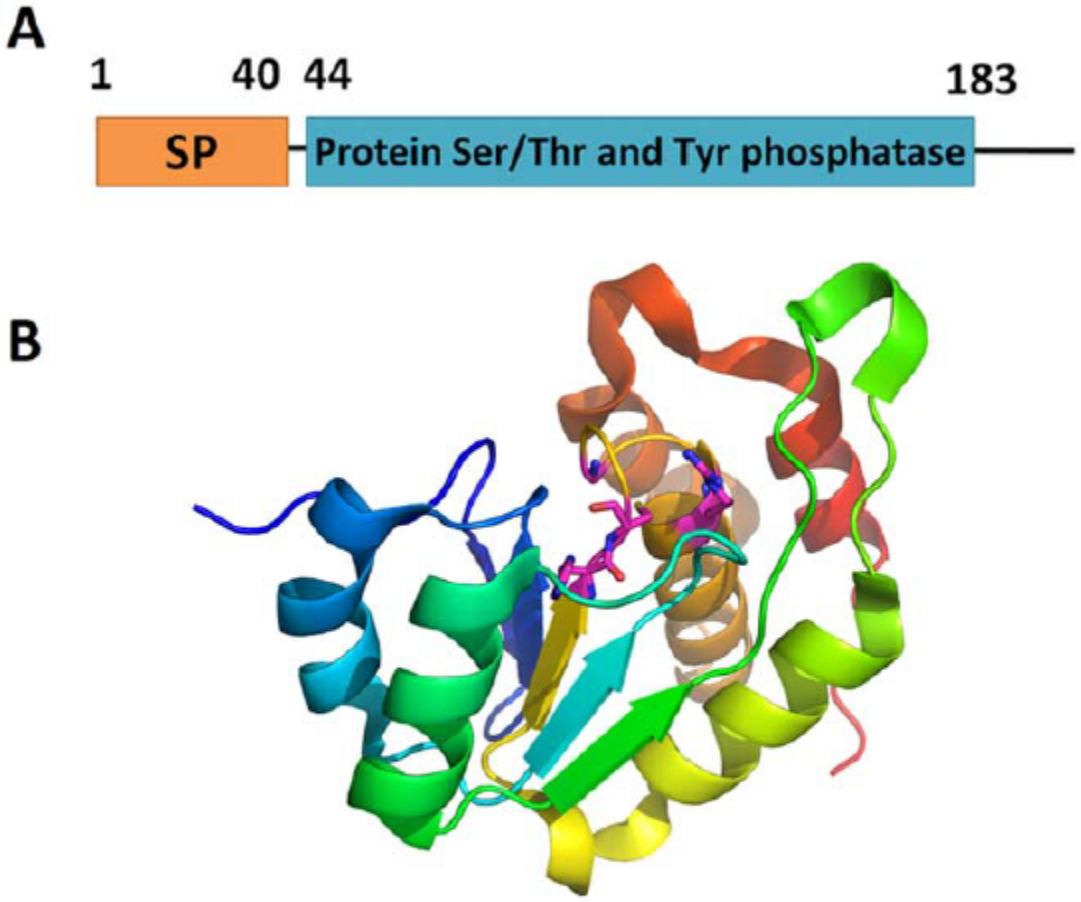


Figure 7. Potential virulence factor, protein serine/tyrosine phosphatase. (locus: CLIBASIA_03975, gi: 254780898). (A) Domain diagram of the protein (B) Predicted structure of the protein. The side-chains of the active site residues are shown.

REFERENCES

1. Folimonova SY, Achor DS (2010) Early events of citrus greening (Huanglongbing) disease development at the ultrastructural level. Phytopathology 100: 949-958.

2. Kim JS, Sagaram US, Burns JK, Li JL, Wang N (2009) Response of sweet orange (Citrus sinensis) to 'Candidatus Liberibacter asiaticus' infection: microscopy and microarray analyses. Phytopathology 99: 50-57.

3. Koh EJ, Zhou L, Williams DS, Park J, Ding N, et al. (2011) Callose deposition in the phloem plasmodesmata and inhibition of phloem transport in citrus leaves infected with "Candidatus Liberibacter asiaticus". Protoplasma.

4. Bove JM, Ayres AJ (2007) Etiology of three recent diseases of citrus in Sao Paulo State: sudden death, variegated chlorosis and huanglongbing. IUBMB Life 59: 346-354.

5. Gottwald TR (2010) Current epidemiological understanding of citrus Huanglongbing. Annu Rev Phytopathol 48: 119-139.

6. Cevallos-Cevallos JM, Rouseff R, Reyes-De-Corcuera JI (2009) Untargeted metabolite analysis of healthy and Huanglongbing-infected orange leaves by CE-DAD. Electrophoresis 30: 1240-1247.

7. Fan J, Chen C, Yu Q, Brlansky RH, Li ZG, et al. (2011) Comparative iTRAQ proteome and transcriptome analyses of sweet orange infected by "Candidatus Liberibacter asiaticus". Physiol Plant.

8. Lin H, Chen C, Doddapaneni H, Duan Y, Civerolo EL, et al. (2010) A new diagnostic system for ultra-sensitive and specific detection and quantification of Candidatus Liberibacter

asiaticus, the bacterium associated with citrus Huanglongbing. J Microbiol Methods 81: 17-25.

9. Sankaran S, Ehsani R, Etxeberria E (2010) Mid-infrared spectroscopy for detection of Huanglongbing (greening) in citrus leaves. Talanta 83: 574-581.

10. Ding F, Jin S, Hong N, Zhong Y, Cao Q, et al. (2008) Vitrification-cryopreservation, an efficient method for eliminating Candidatus Liberobacter asiaticus, the citrus Huanglongbing pathogen, from in vitro adult shoot tips. Plant Cell Rep 27: 241-250.

11. Zhang M, Duan Y, Zhou L, Turechek WW, Stover E, et al. (2010) Screening molecules for control of citrus huanglongbing using an optimized regeneration system for 'Candidatus Liberibacter asiaticus'-infected periwinkle (Catharanthus roseus) cuttings. Phytopathology 100: 239-245.

12. Zhang M, Powell CA, Zhou L, He Z, Stover E, et al. (2011) Chemical compounds effective against the citrus Huanglongbing bacterium 'Candidatus Liberibacter asiaticus' in planta. Phytopathology 101: 1097-1103.

13. Sechler A, Schuenzel EL, Cooke P, Donnua S, Thaveechai N, et al. (2009) Cultivation of 'Candidatus Liberibacter asiaticus', 'Ca. L. africanus', and 'Ca. L. americanus' associated with huanglongbing. Phytopathology 99: 480-486.

14. Duan Y, Zhou L, Hall DG, Li W, Doddapaneni H, et al. (2009) Complete genome sequence of citrus huanglongbing bacterium, 'Candidatus Liberibacter asiaticus' obtained through metagenomics. Mol Plant Microbe Interact 22: 1011-1020.

15. Tyler HL, Roesch LF, Gowda S, Dawson WO, Triplett EW (2009) Confirmation of the sequence of 'Candidatus Liberibacter asiaticus' and assessment of microbial diversity

in Huanglongbing-infected citrus phloem using a metagenomic approach. Mol Plant Microbe Interact 22: 1624-1634.

16. Vahling CM, Duan Y, Lin H (2010) Characterization of an ATP translocase identified in the destructive plant pathogen "Candidatus Liberibacter asiaticus". J Bacteriol 192: 834-840.

17. Akula N, Zheng H, Han FQ, Wang N (2011) Discovery of novel SecA inhibitors of Candidatus Liberibacter asiaticus by structure based design. Bioorg Med Chem Lett 21: 4183-4188.

18. Hartung JS, Shao J, Kuykendall LD (2011) Comparison of the 'Ca. Liberibacter asiaticus' Genome Adapted for an Intracellular Lifestyle with Other Members of the Rhizobiales. PLoS One 6: e23289.

19. Besemer J, Lomsadze A, Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29: 2607-2618.

20. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, et al. (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9: 75.

21. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33: 5691-5702.

22. Erdin S, Lisewski AM, Lichtarge O (2011) Protein function prediction: towards integration of similarity metrics. Curr Opin Struct Biol 21: 180-188.

23. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, et al. (2009) Protein function annotation by homology-based inference. Genome Biol 10: 207.

24. Soding J, Remmert M (2011) Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Curr Opin Struct Biol 21: 404-411.

25. Salavati R, Najafabadi HS (2010) Sequence-based functional annotation: what if most of the genes are unique to a genome? Trends Parasitol 26: 225-229.

26. Sleator RD, Walsh P (2010) An overview of in silico protein function prediction. Arch Microbiol 192: 151-155.

27. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30.

28. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

29. Pei J, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17: 700-712.

30. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, et al. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39: D225-229.

31. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211-222.

32. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, et al. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 29: 41-43.

33. Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, et al. (2009) The National Center for Biotechnology Information's Protein Clusters Database. Nucleic Acids Res 37: D216-223.

34. Letunic I, Doerks T, Bork P (2009) SMART 6: recent updates and new developments. Nucleic Acids Res 37: D229-232.

35. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.

36. Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Res 32: W327-331.

37. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21: 951-960.

38. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.

39. Wang Y, Addess KJ, Chen J, Geer LY, He J, et al. (2007) MMDB: annotating protein sequences with Entrez's 3D-structure database. Nucleic Acids Res 35: D298-300.

40. Gower JC, Ross GJS (1969) Minimum Spanning Trees and Single Linkage Cluster Analysis. Journal of the Royal Statistical Society Series C (Applied Statistics) 18: 54-64.

41. Kado CI (2009) Horizontal gene transfer: sustaining pathogenicity and optimizing host-pathogen interactions. Mol Plant Pathol 10: 143-150.

42. Zhang Y (2009) Protein structure prediction: when is it useful? Curr Opin Struct Biol 19: 145-155.

43. Chen L, Yang J, Yu J, Yao Z, Sun L, et al. (2005) VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res 33: D325-328.

44. Gomez-Gomez L, Boller T (2002) Flagellin perception: a paradigm for innate immunity. Trends Plant Sci 7: 251-256.

45. Jones JD, Dangl JL (2006) The plant immune system. Nature 444: 323-329.

46. Taylor JS, Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 38: 615-643.

47. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Res 40: D290-D301.

48. Jones PM, George AM (2004) The ABC transporter structure and mechanism: perspectives on recent research. Cell Mol Life Sci 61: 682-699.

49. Zhang S, Flores-Cruz Z, Zhou L, Kang BH, Fleites LA, et al. (2011) 'Ca. Liberibacter asiaticus' carries an excision plasmid prophage and a chromosomally integrated prophage that becomes lytic in plant infections. Mol Plant Microbe Interact 24: 458-468.

50. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 23: 1089-1097.

51. Colombatti A, Bonaldo P, Doliana R (1993) Type A modules: interacting domains found in several non-fibrillar collagens and in other extracellular matrix proteins. Matrix 13: 297-306.

52. Ruggeri ZM, Ware J (1993) von Willebrand factor. FASEB J 7: 308-316.

53. Cozzone AJ (2005) Role of protein phosphorylation on serine/threonine and tyrosine in the virulence of bacterial pathogens. J Mol Microbiol Biotechnol 9: 198-213.

54. Lin H, Lou B, Glynn JM, Doddapaneni H, Civerolo EL, et al. (2011) The complete genome sequence of 'Candidatus Liberibacter solanacearum', the bacterium associated with potato zebra chip disease. PLoS One 6: e19135.

55. Lederberg J (1997) Infectious disease as an evolutionary paradigm. Emerg Infect Dis 3: 417-423.

56. Pereira SF, Goss L, Dworkin J (2011) Eukaryote-like serine/threonine kinases and phosphatases in bacteria. Microbiol Mol Biol Rev 75: 192-212.

57. Guan KL, Dixon JE (1993) Bacterial and viral protein tyrosine phosphatases. Semin Cell Biol 4: 389-396.

58. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195-202.

59. Pollastri G, Przybylski D, Rost B, Baldi P (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47: 228-235.

60. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, et al. (2003) Protein disorder prediction: implications for structural proteomics. Structure 11: 1453-1459.

61. Cheng J, Sweredoski M, Baldi P (2005) Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. Data Mining and Knowledge Discovery 11: 213-222.

62. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337: 635-645.

63. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580.

64. von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225: 487-494.

65. Tusnady GE, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283: 489-506.

66. Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics 23: 538-544.

67. Nugent T, Jones DT (2009) Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics 10: 159.

68. Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338: 1027-1036.

69. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783-795.

70. Wootton JC (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 18: 269-285.

71. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science 252: 1162-1164.

CHAPTER SIX
Predictive and comparative analysis of *Ebolavirus* proteins

INTRODUCTION

Zaire Ebolavirus, the pathogen for Ebola Hemorrhagic Fever (EHF) with a 25-90% fatality rate(1), continues to threaten people's lives. The current (2013 - Jun. 2015) West African outbreak of EVD has infected more than 27,000 people and caused 11,000 deaths(2). The genus *Ebolavirus* contains five known species: *Bundibugyo* (BDBV), *Reston* (RESTV), *Sudan* (SUDV), *Taï Forest* (TAFV) and *Zaire ebolavirus* (ZEBOV)(3). The current outbreak is associated with ZEBOV(4). Four *Ebolavirus* species cause EHF in human, with the sole exception being RESTV(5). RESTV can cause EHF to long-tailed macaque (*Macaca fascicularis*). People who had contact with RESTV-infected monkeys tested positive for RESTV antibodies but did not develop symptoms associated with hemorrhagic fevers(5).

Ebolavirus belongs to the order *Mononegavirales* and the family *Filoviridae*(3). Its genome contains seven protein-coding genes that encode the following products: Envelope glycoprotein (GP), Nucleoprotein (NP), RNA-dependent RNA polymerase L (L), Membrane-associated protein VP24 (VP24), Minor nucleoprotein VP30 (VP30), Polymerase cofactor VP35 (VP35), and matrix protein VP40 (VP40). The GP transcript can be edited(6), and the gene product can be processed by host protease, giving rise to four alternative forms of gene products: GP1,2; GP1,2delta; sGP and ssGP. Host furin can cleave the longest product translated from edited mRNA of GP and generate GP1,2, which consists of two peptide chains connected by a disulfide bond(7, 8), GP1 and GP2. GP1,2 is assembled on the membrane of

Ebolavirus and mediates the cell entry. GP1,2delta is the processed product after removal of the C-terminal transmembrane region of GP1,2 by host ADAM17(9). Other products of the GP gene, sGP and ssGP are translated from the unedited mRNA and alternatively edited mRNA, respectively (10, 11). These products share the N-terminal 295 residues with GP1,2, but differ in their short tails (69 and 3 residues, respectively).

In addition to serving as structural components, the *Ebolavirus* proteins play multiple roles in the virus life cycle. GP mediates cell entry(12, 13) and membrane fusion(14, 15) between the virus and host cell. NP encapsidates the genome and protects it from nucleases(16, 17). VP30 is a transcription anti-terminator(18, 19) and regulates the switch between transcription and replication(20, 21). VP35 acts as a cofactor of the polymerase(22, 23), and VP40 may also play a role in genome replication and transcription(24). VP24 and VP35 participate in viral nucleocapsid assembly(17), and VP40 is essential for virus budding and assembly(25-27). In addition, GP, VP24, VP30, VP35 and VP40 interact with multiple host proteins to complete the viral life cycle and to suppress the host immune response.

Three-dimensional (3D) structures are available for a number of *Ebolavirus* proteins. Interpreting available experimental data and sequence variation among *Ebolavirus* species in the context of the 3D structures not only allows researchers to understand detailed mechanisms for cell entry, virus assembly and immune suppression, but also provides promising leads for structure-based drug design. In the current study, we predict the 3D structure and functional sites for *Ebolavirus* protein domains that are not yet characterized. In addition, we compare sequences of *Ebolavirus* proteins' interacting partners from RESTV-resistant primates with those from RESTV-susceptible monkeys. Elevated sequence divergence for GP and VP35's

interaction partners suggests that these two viral proteins may be responsible for host specificity in RESTV. Finally, we compare the protein sequences from different *Ebolavirus* species to detect positions that are conserved among human pathogenic species but variable in non-pathogenic RESTV (RESTV-specific mutations). Mapping of these RESTV-specific mutations and known functional sites to the 3D structures reveals clusters of RESTV-specific mutations on the surfaces of GP, VP35 and VP24. These clusters do not overlap with the known functional sites and may suggest novel interaction sites with host proteins.

MATERIALS AND METHODS

Sequence analysis of Ebolavirus proteins

The protein sequences of *Zaire ebolavirus* were downloaded from the UniProt database(28) and submitted to the MESSA web server(29) to predict the secondary structure(30, 31), disordered regions(32-35), transmembrane helices(36-40), signal peptides(37, 38, 41), coiled coils(42) and structure templates(43, 44). The 3D structures are mostly known, except for protein L, the N-terminal zinc-finger domain of VP30 and the coiled-coil region of VP35. For proteins and domains without known structure, we considered putative structural templates detected by HHpred(44), iTASSER(45, 46) and known structures for proteins of similar function from other families of RNA virus in PDB and ECOD databases(47). Once a candidate structural template was detected, we further validated its relationship to the *Ebolavirus* protein by similarity in function, compatibility between the predicted secondary structure(48) of the *Ebolavirus* protein and the 3D structure of the template, conservation of residues in the

Ebolavirus protein that were aligned to the active sites of the template, and the consistency among multiple structural templates. Sequences of the structural templates and the ZEBOV protein were aligned by Promals3D(49, 50) and the alignments were manually adjusted to ensure that the corresponding secondary structure elements in different templates were aligned together. Based on these alignments and knowledge about functional sites in the template structures from literature, the active sites of uncharacterized *Ebolavirus* domains of were predicted.

Identification of positions associated with human pathogenicity

We downloaded protein sequences of 124 *Ebolavirus* samples from 5 *Ebolavirus* species(4) at www.sciencemag.org/content/345/6202/1369/suppl/DC1, aligned them using MAFFT (51), and evaluated the similarity between amino acids at a certain position using BLOSUM62 scores(52). We considered a position in the sequence alignment to be associated with the loss of human pathogenicity if it satisfies the following two criteria. First, the similarity in amino acids at this position from pathogenic species is always higher than the similarity between RESTV and a pathogenic species. Second, the average similarity in amino acids at this position from four pathogenic species (BDBV, TAFV, SUDV and ZEBOV) is significantly (p-value < 0.05) higher than that between RESTV and pathogenic species. In order to calculate the p-value for each position, we obtained an estimate of the background distribution for the positional difference between average sequence similarity within a group of any four *Ebolavirus* species (all possible combination except the one with all four pathogenic species)

and the average sequence similarity between a fifth species and those in the group. This distribution suggests that a difference larger than 2 is associated with p-value less than 0.05. Enrichment of these pathogenicity-associated positions in each protein was measured by a binomial test (p = total number of positions/total length of all proteins, m = number of selected positions in this protein, N = length of this protein). These pathogenicity-associated positions and the functional sites reported in literature were further mapped to the known 3D structures of *Ebolavirus* proteins.

RESULTS AND DISCUSSION

3D structure and functional sites prediction for Ebolavirus proteins

The domain diagrams of all the *Ebolavirus* proteins are shown in Fig. 1. The positions that are variable between different *Ebolavirus* species are marked as a black line above the domain diagram. The average sequence identity of these proteins between different *Ebolavirus* species ranges from 60% to 80%. *Ebolavirus* proteins contain a significant fraction (20%) of structurally disordered regions, and the fraction of variable positions in these regions is significantly higher ($p < 0.01$) than the structurally ordered regions. The 3D structures of globular regions are mostly known (53-70) except for the N-terminal zinc-finger domain of VP30, the coiled-coil domain of VP35, and protein L. Identification and analysis of structurally characterized homologs allowed us to predict the structure of the zinc-finger domain in VP30, the overall topology of NP, and the structure and catalytic sites for the catalytic domains of protein L.

The zinc-finger domain of VP30

The zinc-finger domain of VP30 was shown to coordinate zinc (71) and it contains a conserved C-x8-C-x4-C-x3-H motif. A search using the VP30 zinc finger motif (residues 70-95) as a query against SUPERFAMILY(72) database with HHpred web server (MSA generation method: HHblits, Maximal MSA Generation iterations: 3, Score secondary structure: yes, Alignment mode: local) reveals similarity (Probability: 52.4; Identity: 35%; E-value: 2.2) to CCCH zinc finger superfamily (seed: SCOP domain d1m9oa_). This hit has the highest coverage and it is the only one (probability cutoff: 20) that contains all the zinc-binding residues. In addition, a scan of the PDB sequences with the conserved pattern C-x(8)-C-x(4)-C-x(3)-H using ScanProsite(73) reveals exactly the same motif in CCCH zinc fingers (PDB id: 2d9n). The confident hits detected by both methods belong to the "CCCH zinc finger" family in the ECOD database(47), and this family contains the N-terminal domain of the transcription antiterminator M2-1 from another *Mononegavirales*, *Pneumovirus*. In addition to their common function, the C-terminal domain of M2-1 and *Ebolavirus* VP30 share the same topology (Fig. 2a,b). M2-1 uses a C-x7-C-x5-C-x3-H motif to bind zinc, which is connected to an alpha-helix at its C-terminus. The VP30 zinc-finger domain very likely adopts a similar structure (Fig. 2c,d), as supported by the presence of a similar C-x8-C-x4-C-x3-H motif and a predicted alpha-helix following the motif.

The N-terminal domain of NP

NP has two structural domains that are connected by a long disordered linker of about 240 amino acids. The C-terminal domain (PDB id: 4QAZ) is shared among *Filoviridae* and is involved in protein-protein interaction(53). The N-terminal domain is likely shared among *Mononegavirales*. The known 3D structures of NP from several virus families(76-80) in this order possess the same topology (Fig. 3a-d). Structures from *Rhabdoviridae* and *Paramyxoviridae* families are determined in complex with ssRNA. (Fig. 3a, c), and both clamp around the RNA using positively charged grooves (Fig. 3g, h) between the two subdomains after a remarkable conformational change compared to the RNA-free form (Fig. 3c, d). The RNA-bound NPs oligomerize to form a ring (Fig. 3i, j), but the oligmerization interface could vary: *Rhabdoviridae* pack the single-stranded RNA (ssRNA) inside the ring formed by NPs while ssRNA binds on the outside of the NP oligomer in *Paramyxoviridae*.

We predicted that the N-terminal domain of *Ebolavirus* NP would adopt the same conserved topology and suggested that its structure is similar to the NP from *Nipah virus* (PDB id: 4CO6(80)). The 3D structure of this domain was released while our manuscript was under review and it supported our prediction (Fig. 3e, f). The available 3D structures for *Ebolavirus* NP (70, 81) were all determined in the absence of RNA. But its similarity to the NPs of other *Mononegavirales* and the presence of a positively charged groove between the two subdomains suggest a similar manner for RNA binding.

The RNA-dependent RNA polymerase catalytic domain of protein L

Sequence analysis suggests that the N-terminal half of protein L functions as a RNA-dependent RNA polymerase (RdRP), and is responsible for both DNA replication and transcription.

HHpred (44) detects a the *Bunyavirus* RdRP (PDB id: 5AMR(82)) as a structural template (Probability: 84%). The alignment between *Ebolavirus* RdRP and *Bunyavirus* RdRP includes both the adenylyl and guanylyl cyclase-like catalytic domain (palm domain) and a helical bundle connected to its C-terminus, and these two domains are conserved among known structures of RdRPs from RNA viruses (83-87) (Fig. 4a-c). Known RdRPs from RNA viruses share the same topology except for *Birnavirus* RdRP, which has a circular permutation in the catalytic domain. This structural conservation of RdRPs across different groups of RNA virus suggests that the RdRP of *Ebolavirus* also adopts the same topology. Secondary structure prediction for the *Ebolavirus* RdRP is consistent with the topology adopted by most RNA viruses, but not with the circular permuted structure from *Birnavirus* (Fig. 4e).

Multiple sequence alignment and 3D structures suggest a conserved catalytic mechanism of RdRP from dsRNA virus and ss(+)RNA virus. Two conserved Asp residues that are used to coordinate Magnesium ions in the catalytic site are in the same position in the 3D structures(88) (Fig. 4). A sequence alignment of these RdRPs allows us to predict the catalytic sites for *Ebolavirus* RdRPs: D632D and D742. These two positions are conserved among close homologs of *Ebolavirus* RdRP detected by PSI-BLAST(89). The second conserved Asp residue immediately follows a conserved Gly residue, forming a GD motif. Another Asp residue after the GD motif also participates in coordinating Mg^{2+} in most of the templates (Fig. 4d). However, this residue is not conserved in *Ebolavirus* and *Birnavirus* RdRPs. Alternatively, *Birnavirus* RdRP has a Glu residue after the first conserved Asp (Fig. 4f), which is in the correct position to bind Mg^{2+}. Similarly, a conserved Glu residue (634E) in the same position

in the *Ebolavirus* RdRP may participate in Mg^{2+} binding, and the arrangement of these active site residues likely resembles that in *Birnavirus* RdRP.

The methyltransferase domain of protein L for mRNA capping

Addition of a 7-methylguanosine cap to the 5' end of mRNA is essential for its subsequent translation and stability in eukaryotic cells(90). The C-terminal half of protein L is responsible for mRNA capping, and it contains an S-adenosyl-L-methionine-dependent methyltransferase domain that likely works in this process. HHpred detects several structural templates (Fig. 5) for this domain with probabilities above 95%. A sequence alignment between the *Ebolavirus* methyltransferase domain and the detected templates reveals that three residues, K1816, D1927, and K1962, are aligned to the conserved catalytic residues in the templates(91). In addition, the "GEGAGA" motif at positions 1836-1841 of *Ebolavirus* protein L is aligned to the conserved S-adenosyl-L-methionine-binding motif in the templates. This motif is also conserved in sequences from *Filoviridae*, suggesting a similar function in co-factor binding.

Interaction between Ebolavirus proteins and host proteins

RESTV causes EHF symptoms in Asian cynomolgus monkeys (*Macaca fascicularis*), but not human and African green monkeys (*Chlorocebus aethiops*)(5). This difference in susceptibility between closely related hosts is likely due to the sequence divergence in the host proteins that interact with virus proteins. Therefore, comparing the interacting partners of virus proteins from different hosts may provide insight into how host specificity is determined and further

suggest the mechanism for RESTV's loss of human pathogenicity. The known interacting partners in the host for each *Ebolavirus* protein are summarized in Table 1.

The known host proteins that interact with VP24, VP30, and VP40 are highly similar between the RESTV-resistant (*Chlorocebus* and human) and RESTV-susceptible species (*Macaca*), suggesting that they may not be responsible for the loss of human pathogenicity in RESTV. In contrast, seven most divergent host partners interact either with GP or VP35. Three of them (marked in Table 1) show significantly ($P<0.05$) elevated divergence between the susceptible and resistant species, including Hepatitis A virus cellular receptor 1 (TIM-1) and pathogen-recognition receptor CD209 that interact with GP and facilitates cell entry, as well as the interferon-induced, dsRNA-activated kinase PKR that is inhibited by VP35.

The elevated divergence level for interacting partners of GP and VP35 in the host suggests that VP35 and GP may play important roles in determining host specificity. This is consistent with some indirect experimental data. RESTV GP pseudotyped viruses show significantly lower ability to infect human cells and damage human endothelial cells than that of ZEBOV GP pseudotyped viruses(92). In addition, RESTV GP shows lower ability to deplete T cells and down-regulate interferon-stimulated gene expression compared to ZEBOV GP(93, 94). Meanwhile, ZEBOV VP35 shows stronger Interferon inhibition than RESTV VP35 in human cells(67). However, direct studies of all RESTV proteins' effect in cells from both RESTV-susceptible and RESTV-resistant species are needed to prove our hypothesis.

Interpreting residues associated with RESTV's loss of human pathogenicity in the context of 3D structure and known functional sites

We define positions that are associated with the loss of pathogenicity in RESTV as those that are always and significantly more similar among pathogenic species (BDBV, TAFV, SUDV and ZEBOV) than between RESTV and the pathogenic species. We referred to them as "RESTV-specific mutations". We identified 215 such positions (Table 2), and VP30 and VP35 are significantly enriched in such mutations.

43 of the RESTV-specific mutations can be mapped to known 3D structures of *Ebolavirus* proteins. None of them overlap with functional sites that are proved to be crucial by mutagenesis and six of them overlap with interaction surfaces (summarized in Table 3) on these structures. They may affect the binding affinities but would not likely abolish the interactions. One loop (129-141) of VP24 at the boundary of the interacting surface between VP24 and KPNA5(50) contains four RESTV-specific mutations (T131S, N132T, M136L, Q139R, within the red circle in Fig. 6d). These mutations may affect the binding affinity between RESTV and KPNA5 in RESTV, resulting in a poorer immune suppression by RESTV. One mutation to GP (N514D) is at the boundary of its interacting surface with neutralizing antibodies from human survivor and this may affect the efficiency of the ZEBOV antibodies to antagonize the RESTV.

Mapping of the RESTV-specific mutations to the 3D structures revealed a couple of mutation clusters in GP and VP35, which may be related to RESTV's difference in pathogenicity (Fig. 6). A first cluster is in the C-terminal subdomain of the GP *Filovirus* glycoprotein domain. The cluster consists of three mutations on the surface: Y261R, T269S,

and S307H (inside the blue circle in Fig. 6a). The functional role of this subdomain is not clear, and the cell entry of ZEBOV is mostly mediated by the interaction between N-terminal 150 residues of GP and cell receptors like NPC1 and TIM-1. One possibility is that it may interact with other host proteins, such as lectins, that facilitate the infection of *Ebolavirus*. In contrast, another cluster of mutations (Q44K, and V45A, inside the magenta circle in Fig. 6b) may affect the interaction between GP and the receptors. Even more, mutation E156N is close to the functional sites that are shown by mutagenesis to be important to maintain the infectivity of ZEBOV. Therefore, they may cause a significantly lower infectivity in RESTV and contribute to the loss of human pathogenicity.

RESTV-specific mutations (A290V, A291P, V314A, and Q329K) in VP35 form a cluster (inside the pink circle in Fig. 6f) on the opposite side of the dsRNA-binding surface of VP35. Host immune suppression by VP35 is mainly related to its interaction with dsRNA, but the loss of dsRNA-binding ability does not completely abolish VP35-mediated immune suppression(95). This observation indicates the existence of other mechanisms for immune suppression by VP35, where the surface enriched in RESTV-specific mutations may play a role. One RESTV-specific mutation (T226A) is adjacent to the position in VP24 that is mutated (T50I) during adaptation to mice(96, 97) (orange circles in Fig. 6c). This adaptation site is not close to any known functional sites. But the clustering of the adaptation site and RESTV-specific mutation suggests the possibility that they are at the interface of some uncharacterized interaction with other host proteins.

Table 1 Host proteins that functionally interact with *Ebolavirus* proteins, and their divergence level between RESTV-susceptible and RESTV-resistant species

Name	Host protein	Functional implication	*Chlorocebus* vs *Macaca*	*Homo* vs *Macaca*
GP	NPC1(12, 13)	Receptor for the virus	6 (99.5%)	28 (97.8%)
GP	TIM-1(98)		11 (96.5%)***	56 (80.2%)***
GP	CD209 (99, 100)	Facilitate cell entry in specific cell types	15 (95.9%)***	31 (92.1%)***
GP	CLEC4M (99, 100)		Not available in *Macaca* and *Chlorocebus*	
GP	CLEC10A(92)		5 (98.4%)	42 (86.7%)
GP	FOLR1(101)		3 (98.8%)	8 (96.9%)
GP	FURIN(7, 8)	Process GP to GP1,2	1 (99.9%)	9 (98.9%)
GP	CTSB(102)	Process GP1,2 and initiate membrane fusion	2 (99.4%)	10 (97.0%)
GP	CTSL(102)		3 (99.1%)	14 (95.8%)
GP	ADAM17(9)	Process GP1,2 to GP1,2delta	1 (99.9%)	3 (99.6%)
GP	Dynamin (multiple)(103)	Activates endothelial cells, reduces their barrier function	0~1 (99.9~100%)	2~6 (99.3~99.8%)
GP	ITGAV(103)		2 (99.8%)	9 (99.1%)
VP24	STAT1(60)	Inhibit JAK-STAT pathway for interferon sensing	0 (100%)	5 (99.3%)
VP24	KPNA5(55)		0 (100%)	2 (99.6%)
VP24	MAPK14 (p38)(104)	Prevents phosphorylation and	0 (100%)	1 (99.6%)
VP30	PPP1C(20)	Dephosphorylate VP30, control replication-transcription switch	0 (100%)	0 (100%)
VP30	PPP2C(20)		0 (100%)	0 (100%)
VP30	Dicer(105)	Antagonize the RNAi machinery that could target	5 (99.7%)	10 (99.5%)
VP30	TRBP(105)		2 (99.5%)	3 (99.2%)
VP35	Dicer(105)	Antagonize the RNAi machinery that could target	5 (99.7%)	10 (99.5%)
VP35	TRBP(105)		2 (99.5%)	3 (99.2%)
VP35	ILF3 (DRBP76)(106)	Inhibit the effect of interferon	0 (100%)	3 (99.7%)
VP35	IKBKε(107)	Block phosphorylation of IRF-3 by TBK-1 and IKBKε, inhibiting interferon production	4 (99.4%)	15 (97.9%)
VP35	TBK-1(107)		2 (99.7%)	8 (98.9%)
VP35	IRF-3(107)		2 (99.5%)	17 (96.0%)
VP35	PACT(108)	Inhibit its role as RIG-I	0 (100%)	0 (100%)
VP35	PKR(109)	Inhibit the effect of interferon	42 (92.4%)***	110 (80%)***
VP35	UBE2I(110)	Use SUMO E2 enzyme (UBE21) and E3 ligase (PIAS1) to modify IRF7 and inhibits its function	0 (100%)	0 (100%)
VP35	PIAS1(110)		0 (100%)	0 (100%)
VP35	IRF-7(110)		9 (98.2%)	35 (92.9%)
VP35	DLC8(111)	May regulate viral life cycle	1 (98.9%)	0 (100%)
VP40	Sec24C(112)	Virus utilize COPII vesicular	7 (99.4%)	24 (97.8%)
VP40	TSG101(27)	Virus uses multi-vesicular body biogenesis pathway for budding	0 (100%)	0 (100%)
VP40	ABL1(113)	ABL1 controls budding/release by phosphorylating VP40	5 (99.6%)	13 (98.8%)
VP40	NEDD4(114)	NEDD4 facilitates budding by adding ubiquitin to VP40	5 (99.5%)	28 (97.8%)
VP40	Tubulin (multiple)(115)	Virus utilize cytoskeleton in its life cycle	0 (100%)	0 (100%)
VP40	Actin (1 and 2)(116)		0 (100%)	0 (100%)
VP40	IQGAP1(117)		2 (99.9%)	9 (99.4%)

*** significantly ($p<0.05$) elevated divergence level

Table 2. Positions in ZEBOV that are likely associated with the loss of human pathogenicity by RESTV

Name	ID	Length	P-value	Mutations associated with the loss of human pathogenicity
GP	Q05320	676	0.457	F31I, Q44K, V45A, E156N, S196A, L199A, S210T, Y261R, T269S, T283P, S307H, T335P, E337T, H339N, E345T, H354L, E359T, A361E, A427M, G488K, R498K, R500K, N514D, D607S, K622E, I627K, Q638H, D642L, W644L, T659I
L	Q05318	2212	0.690	V66T, E93T, Q109H, N120A, V128T, E130I, F132T, L146V, L179F, N201T, T202I, A221S, Q223L, H227Q, V229L, P262V, V263D, S274L, L283V, Y312F, A326S, T330D, S343Y, E350D, T361S, L365F, I402N, Q447H, P450S, D465N, R654H, E689S, S847A, S868A, F896Y, L925F, A954S, S995T, T1024N, R1073K, A1119S, Q1149P, S1154L, P1163A, K1171D, D1189S, A1214S, R1217K, D1237E, Q1253N, Y1322L, R1354K, T1366A, I1408M, S1436N, K1461Q, S1473C, L1488Y, S1506A, A1538S, V1562L, E1564S, T1571K, Q1608I, H1619L, L1624Y, C1628S, D1744G, E1752P, S1769G, Q1782L, R1792H, W1822L, V1850T, R1916N, K1938Q, E1941R, V1955Y, Q2024G, P2038V, S2077T, K2078G, R2079L, E2098D, Q2105L, Q2108E, Y2131F, L2157V, R2168H, R2175K, L2177F, M2186L, L2203F
NP	P18272	739	0.587	R4G, T15G, S30T, R39K, I52M, R105K, M137L, F212Y, K274R, S279A, K373R, K374R, A411L, K416N, Y421Q, D426E, D435N, Q442L, D443E, T453I, V458A, D492E, Q507S, S511I, N551R, T563S, E633L, S647K, A705R, T714Y, D716N
VP24	Q05322	251	0.932	T131S, N132T, M136L, Q139R, T226A, S248L
VP30	Q05323	288	0.010	G20P, V25S, Y39R, T52N, V53L, T63I, E93D, T96N, R98H, K107R, S111I, L116S, N117Q, A120S, Q135S, T150I, Q157R, R196H, E205D, R262A, S268Q
VP35	Q05127	340	0.019	T5L, L25T, S26T, E48D, D76E, C79Y, N80V, E85K, S92M, V97T, Q98S, S106A, A154S, T159V, E160D, G167K, S174A, I258T, E269D, A290V, A291P, V314A, Q329K
VP40	Q05128	326	0.786	M14N, T46V, P85T, A128I, G201N, F209L, Q245P, H269Q, T277Q, V323H, E325D

P-value: binomial test for enrichment of residues that may be associated with RESTV's loss of human pathogenicity in each protein.

Table 3. Experimentally characterized functional sites in *Ebolavirus* proteins

Name	Residues	Function	Experimental evidence
GP	40	Glycosylated by host	N40D loss ability to infect(118)
GP	41-43, 503-511, 513, 514	Interact with antibody	On the interacting surface with neutralizing antibody (54)
GP	51, 68, 86, 99, 109, 111, 113, 122, 139, 154, 159, 161, 162, 171,176,183-185	Maintain the hydrophobic core structure	W86A, Y99A, Y109A, H139A, H154A, F159A, L161A, Y162A, Y171A, F176A, F183A reduce expression, reduce viral incorporation and abolish infectivity; L111A, I113A, L122A reduce viral incorporation and abolish infectivity; L51A, L68A, L184A, I185A abolish infectivity(119)
GP	53, 108, 121, 135, 147, 511, 556, 601, 608, 609	Disulfide bond	C53G, C108A, C121G, C135S, C147S, C511G, C556S, C601S, C608G, C609G reduce expression and abolish infectivity(118)
GP	55, 85, 103, 117, 178	Hydrophilic to maintain the structure	E85A, E103A, E178A reduce expression; E85A, E103A, D117A, E178A reduce viral incorporation; D55A, E103A, D117A, E178A loss ability to infect(119)
GP	529, 531-533, 535-537	Fusion peptide	I529A, W531R, W531A, I532R, P533R, F535R, G536R, G536A, P537R loss ability to infect(14)
GP	57, 63, 64, 88, 95, 170	Cell entry	L63K, L63A reduce expression; L57A, L57F, L57I, L57K, L63K, L63A, L63F, R64E, R64A, F88E, F88A, K95E, K95A, I170A, I170E loss ability to infect(119)
VP24	96-98, 106-121	Interact with STAT1	Show reduced hydrogen exchange rate upon binding(60)
VP24	113, 115, 117, 121, 124, 125, 128-131, 134-141, 184-186, 201-205, 218	Interact with KPNA5	On the interacting surface in crystal structure with KPNA5 (PDB id: 4U2X)(55)
VP24	50, 71, 147, 187	Adapt to new host	T50I mouse adaptation(97); M71I, L147P, and T187I guinea pig adaptation(96)
VP30	179, 180, 183, 197	Activate transcription	Mutation to Ala reduces interaction with nucleocapsid; K180A, K183A, E197A block transcription activation(58)
VP30	143, 146	Phosphorylation	T143A, T143D, T146A, T146D inhibit transcription(20)
VP35	239, 240, 309, 312, 319, 322, 339	Bind dsRNA	K309A, K319A reduce dsRNA binding; F239A, H240A, R312A, R322A, K339A abolishes dsRNA binding(69)
VP35	239, 240, 309, 312, 319, 322, 339	IRF-3 inhibition	K309A, K319A reduce IRF-3 inhibition; F239A, H240A, R312A, R322A, K339A greatly reduce IRF-3 inhibition(69)
VP35	235, 240	Polymerase cofactor	F235A, H240A impair replication of mini-genome(23)
VP35	312, 322, 339	Bind DRPB76	Mutation to alanine reduce ability to bind DRPB76(106)
VP35	309, 312	Inhibit RNAi	K309A and R312A lost the inhibition effect(120)
VP35	305, 309, 312	Inhibit PKR	Mutant any two to alanine abolish the inhibition(121)
VP40	303-308	Interact with Sec24C	303-306A and 305-308A cannot interact with Sec24C, and reduce virus-like particles(112)
VP40	51-54, 96-101, 212-214, 286-291, 303-308, 314-316	Release of virus-like particles	51-52A, 53-54A, deletion of 96-101, K212A, L213A, R214A, 286-288A, 289-291A, 303-306A, 305-308A reduce the release of virus-like particles(112, 122, 123)
VP40	127, 129, 130, 283, 286, 293, 295, 298, 309-317	Membrane localization	K127A, T129A, N130A, P283L, P286L, I293A, L295A, V298A and deletion of 309-317 reduce membrane localization(26, 124)
VP40	226-255	Interaction with microtubules	Deletion of 226-240 or 241-255 abolish ability to protect microtubules from depolymerization(115)
VP40	213, 293, 295, 298	Penetrate membrane	Mutation to alanine reduces membrane localization(125)

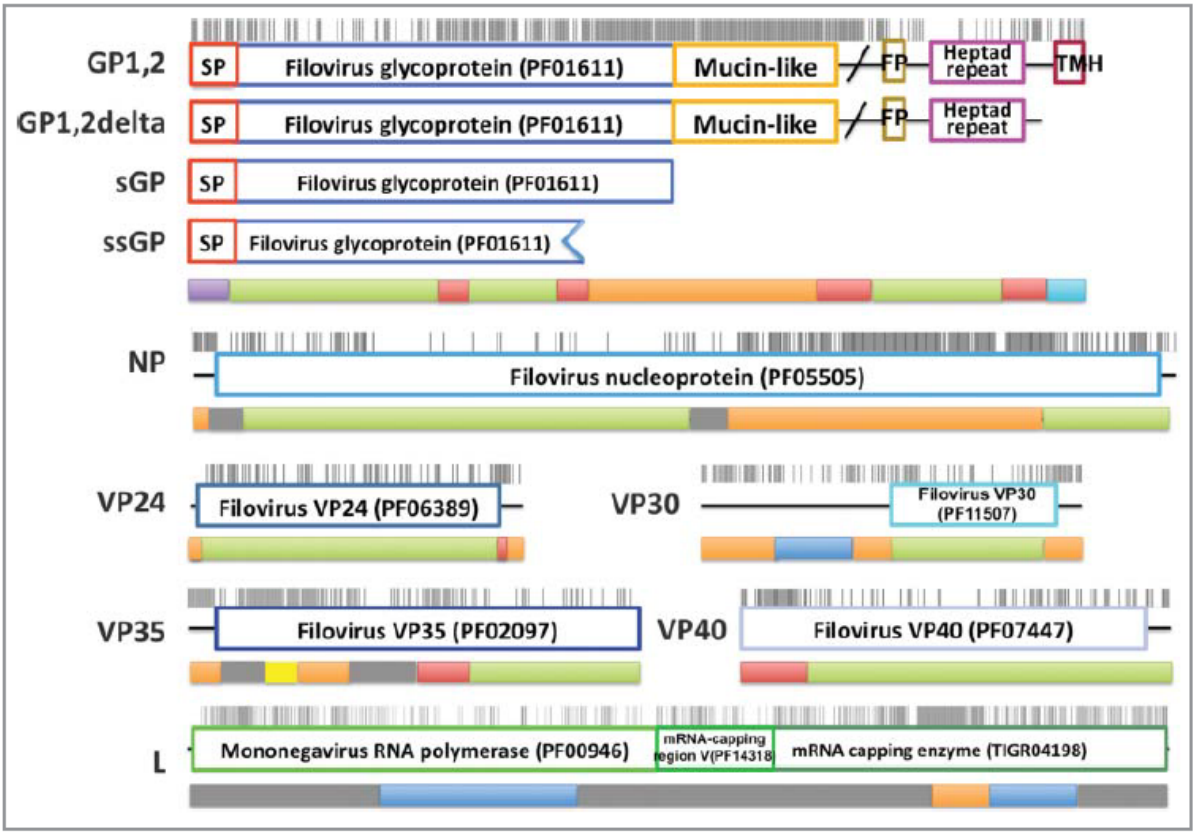


Figure 1. Domain diagrams for *Ebolavirus* proteins and coverage of the proteins by experimentally determined and predicted structures. The domains of each protein are represented by boxes on a thread and the positions that are variable among different species are marked by black sticks above the domain diagrams. The band below is aligned to the domain diagram and the color of this band indicates the prediction status of the corresponding region. The color codes are: green, regions that are structurally characterized and adopt globular structure; red, regions that are experimentally determined but intrinsically disordered; blue, regions with predicted 3D structure; yellow, coiled coil; cyan, transmembrane helix; purple, signal peptide; orange, predicted intrinsically disordered regions; grey, predicted regions that have a propensity to adopt secondary structure but 3D structure cannot be predicted;. Abbreviations: SP, signal peptide; FP, fusion peptide; TMH, Transmembrane helix.

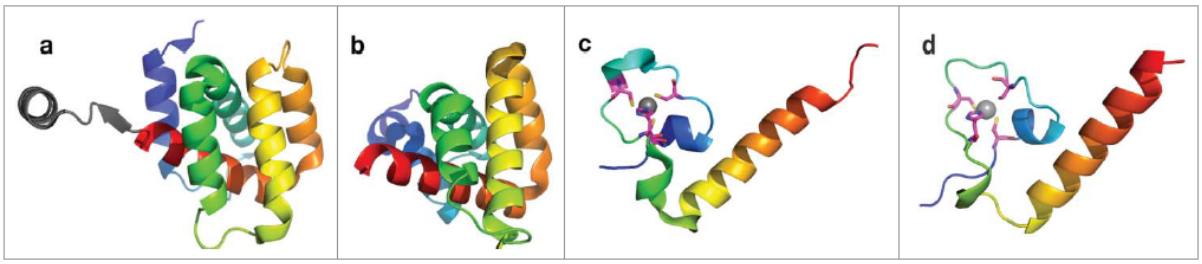


Figure 2. Structure prediction for N-terminal domain of VP30. (a) 3D structure (PDB id: 2I8B) for VP30 C-terminal domain; (b) 3D structure (PDB ID: 4C3B) for *Pneumovirus* M2-1 C-terminal domain; (c) 3D structure (PDB ID: 4C3B) for *Pneumovirus* M2-1 N-terminal domain, which can be used as template to predict the structure for the VP30 N-terminal domain; (d) structure model for the *Ebolavirus* VP30 N-terminal domain.

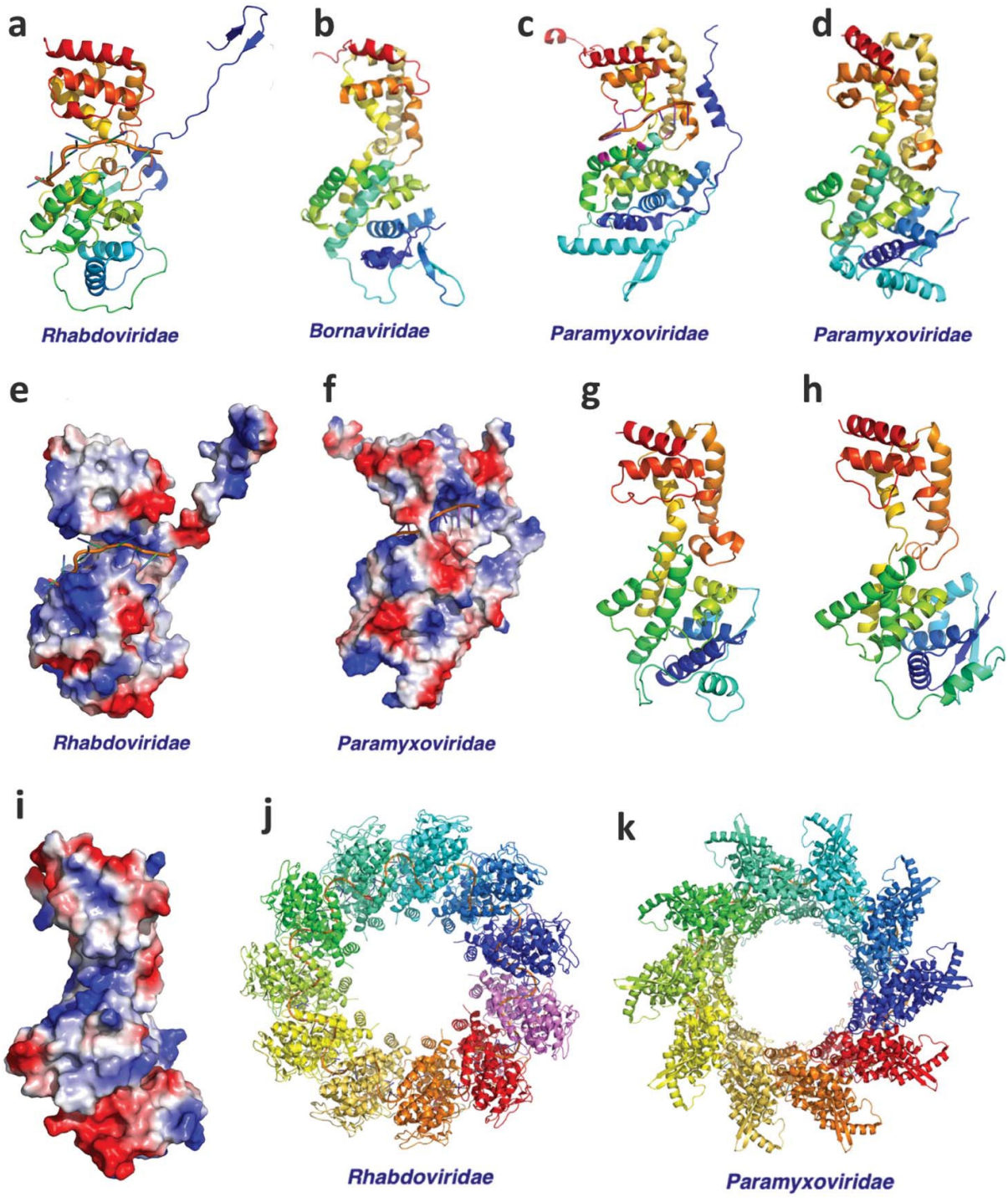


Figure 3. Structures of *Mononegavirales* Nucleoproteins. The virus family is labeled below. (a-d) Monomeric structures (PDB IDs: 2GIC, 1N93, 2WJ8, and 4CO6) of Nucleoproteins from

Mononegavirales. The structures are colored in rainbow; (e, f) The electrostatic potential mapped on to the surface of Nucleoprotein structures (PDB ids: 2GIC and 2WJ8). Blue area corresponds to positively charged surface and the red area corresponds to negatively charged surface; (g) Structure model for the N-terminal domain of *Ebolavirus* NP; (h) Real structure of the N-terminal domain of *Ebolavirus* NP; (i) The electrostatic potential mapped on to the surface of experimentally determined N-terminal domain of *Ebolavirus* NP; (j, k) Structure complex of RNA and Nucleoproteins from *Rhabdoviridae* and *Paramyxoviridae* (PDB ids: 2GIC and 2WJ8).

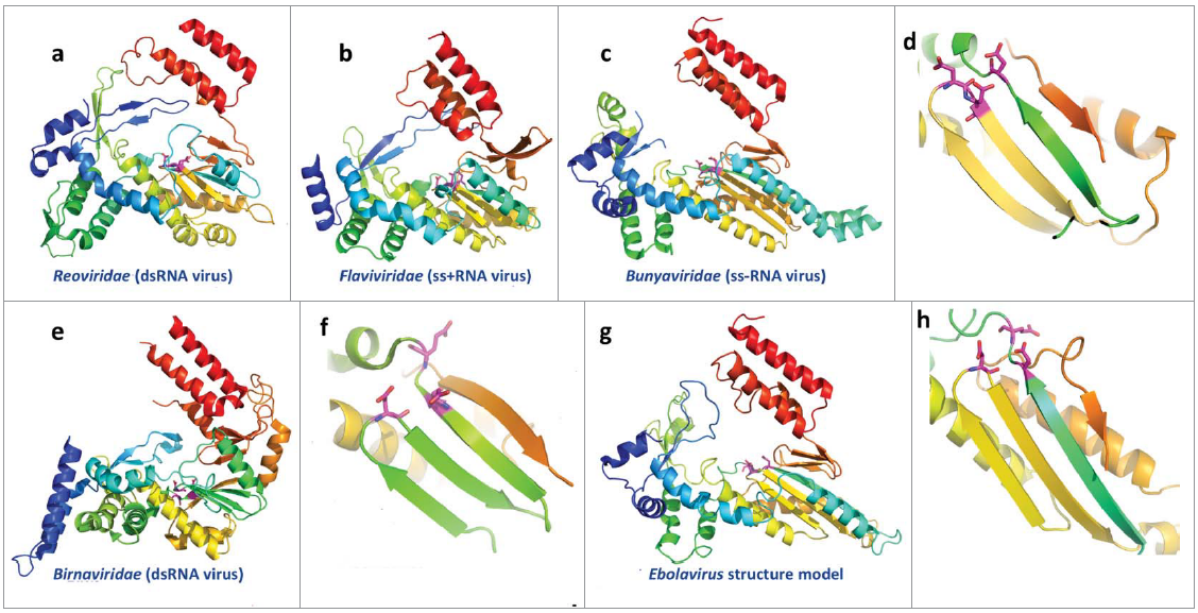


Figure 4. Structures of the catalytic domains of RNA-dependent RNA polymerases (RdRP) from RNA viruses and the structure model for *Ebolavirus* RdRP. The virus family is labeled below. The structures are colored in rainbow, with equivalent secondary structure elements from different structures colored similarly, except for the *Birnaviridae* RdRP, which has a circularly permutated topology. The functional sites used to coordinate Mg^{2+} are shown as sticks and colored in magenta. (a-c) Overall structure of the core domains of RdRPs from RNA viruses (PDB IDs: 2R7O, 1GX5, and 5AMR); (d) close up view of the classic arrangement of functional sites for the core domains of RdRPs from RNA viruses; (e-f) overall structure and close up view of the functional sites for the core domains of RdRPs from *Birnaviridae* (dsRNA virus); PDB id: 2PGG; (g-h) structure model for the core domains of ZEBOV RdRP and close up view of the predicted active sites.

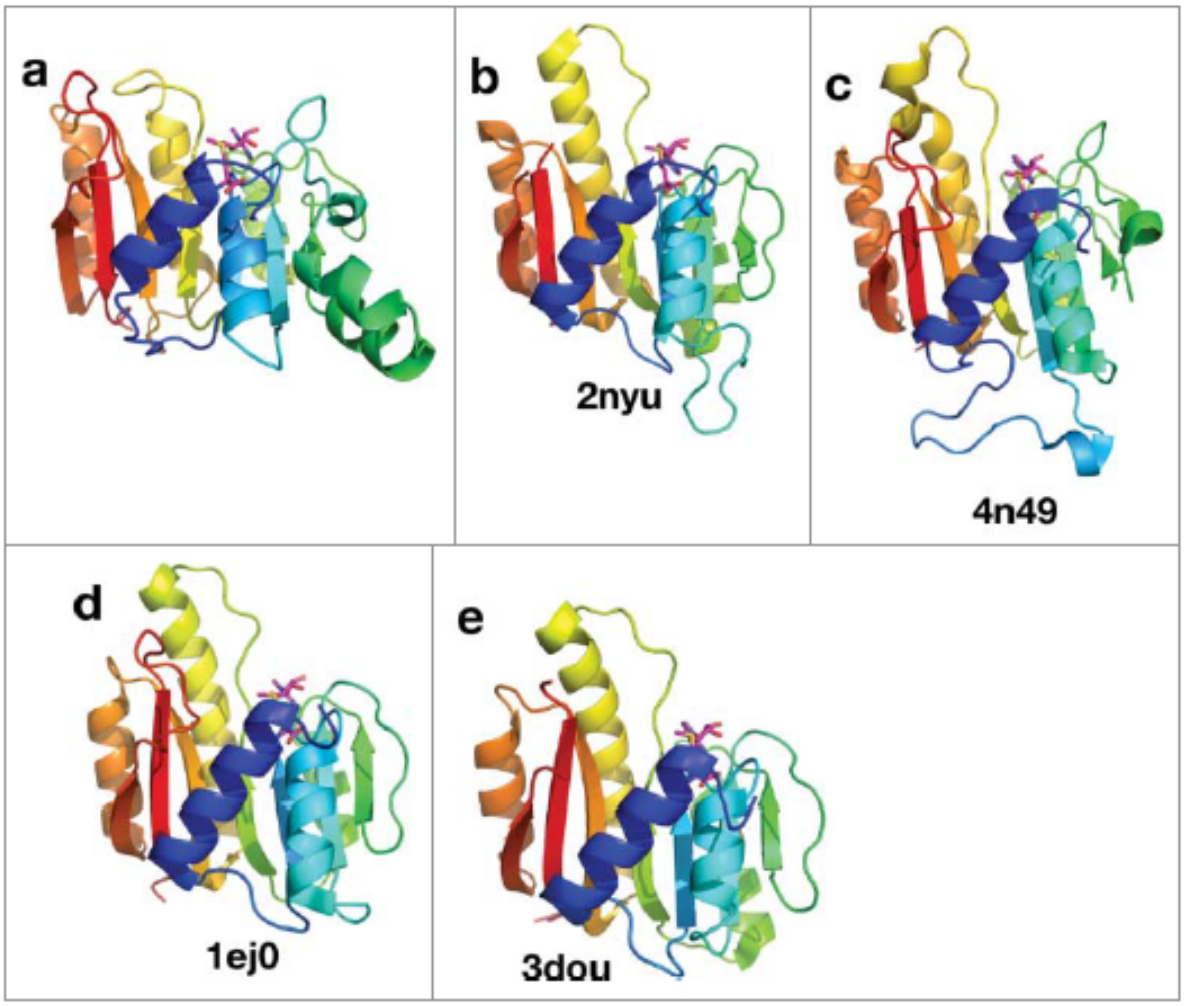


Figure 5. Structural model and templates for the mRNA capping methyltransferase domain in *Ebolavirus* protein L. The structures are colored in rainbow. Equivalent secondary structure elements from different structures are colored in the same color. The co-factor, S-adenosyl-L-methionine, is shown as stick. (a) structure model for the mRNA capping methyltransferase domain of ZEBOV; (b-e) other methyltransferase domains.

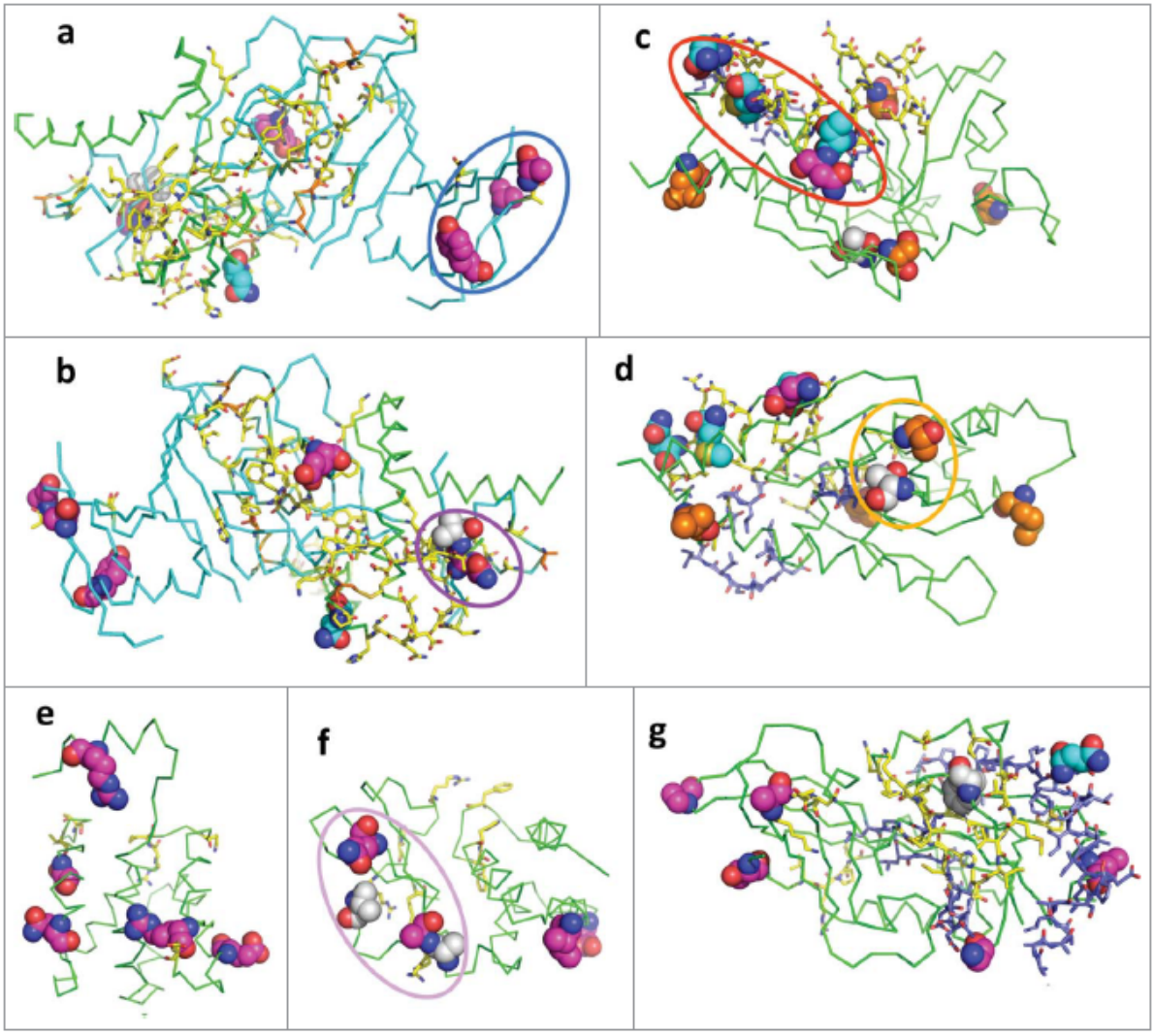


Figure 6. Mapping of RESTV-specific residues, functional sites and interaction surfaces to known 3D structures of *Ebolavirus* proteins. The structure is shown in ribbon; the functional sites are shown as sticks; and positions with RESTV-specific mutations and alternate host (rodent) adaptation residues are shown as spheres. Carbon atoms of the functional sites and sites with RESTV-specific mutations are colored to show the property of that residue: RESTV-specific surface residues are in magenta; RESTV-specific buried residues are in white; RESTV-specific residues that belong to interaction surfaces are in cyan; known

functional residues are in yellow; disulfide bonded and alternate host (rodent) adaptation residues are in orange; predicted or indirectly shown functional residues are blue. Other atoms are colored as follows: oxygen (red); nitrogen (blue) and sulfur (orange). Circles highlight sites with RESTV-specific residue clusters that are discussed in the text. (a,b) GP (PDB id: 3CSY); (c,d) VP24 (PDB id: 4U2X); (e) VP30 (PDB id: 2I8B); (f) VP35 (PDB id: 3L26); (g) VP40 (PDB id: 1ES6).

REFERENCES

1. **Organization WH.** 2014. Ebola virus disease Fact Sheets: http://www.who.int/mediacentre/factsheets/fs103/en/.
2. **Organization WH.** 2015. Ebola data and statistics (published on 18 February 2015): http://apps.who.int/gho/data/view.ebola-sitrep.ebola-summary-20150218?lang=en.
3. **Kuhn JH, Becker S, Ebihara H, Geisbert TW, Johnson KM, Kawaoka Y, Lipkin WI, Negredo AI, Netesov SV, Nichol ST, Palacios G, Peters CJ, Tenorio A, Volchkov VE, Jahrling PB.** 2010. Proposal for a revised taxonomy of the family Filoviridae: classification, names of taxa and viruses, and virus abbreviations. Archives of virology **155:**2083-2103.
4. **Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang PP, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A,**

Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally JL, Chapman SB, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheiffelin JS, Lander ES, Happi C, Gevao SM, Gnirke A, Rambaut A, Garry RF, Khan SH, Sabeti PC. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science **345:**1369-1372.

5. **Morikawa S, Saijo M, Kurane I.** 2007. Current knowledge on lower virulence of Reston Ebola virus (in French: Connaissances actuelles sur la moindre virulence du virus Ebola Reston). Comparative immunology, microbiology and infectious diseases **30:**391-398.

6. **Sanchez A, Trappier SG, Mahy BW, Peters CJ, Nichol ST.** 1996. The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. Proceedings of the National Academy of Sciences of the United States of America **93:**3602-3607.

7. **Volchkov VE, Feldmann H, Volchkova VA, Klenk HD.** 1998. Processing of the Ebola virus glycoprotein by the proprotein convertase furin. Proceedings of the National Academy of Sciences of the United States of America **95:**5762-5767.

8. **Wool-Lewis RJ, Bates P.** 1999. Endoproteolytic processing of the ebola virus envelope glycoprotein: cleavage is not required for function. Journal of virology **73:**1419-1426.

9. **Dolnik O, Volchkova V, Garten W, Carbonnelle C, Becker S, Kahnt J, Stroher U, Klenk HD, Volchkov V.** 2004. Ectodomain shedding of the glycoprotein GP of Ebola virus. The EMBO journal **23:**2175-2184.

10. **Mehedi M, Falzarano D, Seebach J, Hu X, Carpenter MS, Schnittler HJ, Feldmann H.** 2011. A new Ebola virus nonstructural glycoprotein expressed through RNA editing. Journal of virology **85:**5406-5414.

11. **Volchkova VA, Feldmann H, Klenk HD, Volchkov VE.** 1998. The nonstructural small glycoprotein sGP of Ebola virus is secreted as an antiparallel-orientated homodimer. Virology **250:**408-414.

12. **Carette JE, Raaben M, Wong AC, Herbert AS, Obernosterer G, Mulherkar N, Kuehne AI, Kranzusch PJ, Griffin AM, Ruthel G, Dal Cin P, Dye JM, Whelan SP, Chandran K, Brummelkamp TR.** 2011. Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. Nature **477:**340-343.

13. **Miller EH, Obernosterer G, Raaben M, Herbert AS, Deffieu MS, Krishnan A, Ndungo E, Sandesara RG, Carette JE, Kuehne AI, Ruthel G, Pfeffer SR, Dye JM, Whelan SP, Brummelkamp TR, Chandran K.** 2012. Ebola virus entry requires the host-programmed recognition of an intracellular receptor. The EMBO journal **31:**1947-1960.

14. **Ito H, Watanabe S, Sanchez A, Whitt MA, Kawaoka Y.** 1999. Mutational analysis of the putative fusion domain of Ebola virus glycoprotein. Journal of virology **73:**8907-8912.

15. **Gomara MJ, Mora P, Mingarro I, Nieva JL.** 2004. Roles of a conserved proline in the internal fusion peptide of Ebola glycoprotein. FEBS letters **569:**261-266.

16. **Watanabe S, Noda T, Kawaoka Y.** 2006. Functional mapping of the nucleoprotein of Ebola virus. Journal of virology **80:**3743-3751.

17. **Huang Y, Xu L, Sun Y, Nabel GJ.** 2002. The assembly of Ebola virus nucleocapsid requires virion-associated proteins 35 and 24 and posttranslational modification of nucleoprotein. Molecular cell **10:**307-316.

18. **Weik M, Modrof J, Klenk HD, Becker S, Muhlberger E.** 2002. Ebola virus VP30-mediated transcription is regulated by RNA secondary structure formation. Journal of virology **76:**8532-8539.

19. **Martinez MJ, Biedenkopf N, Volchkova V, Hartlieb B, Alazard-Dany N, Reynard O, Becker S, Volchkov V.** 2008. Role of Ebola virus VP30 in transcription reinitiation. Journal of virology **82:**12569-12573.

20. **Ilinykh PA, Tigabu B, Ivanov A, Ammosova T, Obukhov Y, Garron T, Kumari N, Kovalskyy D, Platonov MO, Naumchik VS, Freiberg AN, Nekhai S, Bukreyev A.** 2014. Role of protein phosphatase 1 in dephosphorylation of Ebola virus VP30 protein and its targeting for the inhibition of viral transcription. The Journal of biological chemistry **289:**22723-22738.

21. **Biedenkopf N, Hartlieb B, Hoenen T, Becker S.** 2013. Phosphorylation of Ebola virus VP30 influences the composition of the viral nucleocapsid complex: impact on viral transcription and replication. The Journal of biological chemistry **288:**11165-11174.

22. **Muhlberger E, Weik M, Volchkov VE, Klenk HD, Becker S.** 1999. Comparison of the transcription and replication strategies of marburg virus and Ebola virus by using artificial replication systems. Journal of virology **73:**2333-2342.

23. **Prins KC, Binning JM, Shabman RS, Leung DW, Amarasinghe GK, Basler CF.** 2010. Basic residues within the ebolavirus VP35 protein are required for its viral polymerase cofactor function. Journal of virology **84:**10581-10591.

24. **Hoenen T, Jung S, Herwig A, Groseth A, Becker S.** 2010. Both matrix proteins of Ebola virus contribute to the regulation of viral genome replication and transcription. Virology **403:**56-66.

25. **Han Z, Boshra H, Sunyer JO, Zwiers SH, Paragas J, Harty RN.** 2003. Biochemical and functional characterization of the Ebola virus VP24 protein: implications for a role in virus assembly and budding. Journal of virology **77:**1793-1800.

26. **Panchal RG, Ruthel G, Kenny TA, Kallstrom GH, Lane D, Badie SS, Li L, Bavari S, Aman MJ.** 2003. In vivo oligomerization and raft localization of Ebola virus protein VP40 during vesicular budding. Proceedings of the National Academy of Sciences of the United States of America **100:**15936-15941.

27. **Martin-Serrano J, Zang T, Bieniasz PD.** 2001. HIV-1 and Ebola virus encode small peptide motifs that recruit Tsg101 to sites of particle assembly to facilitate egress. Nature medicine **7:**1313-1319.

28. **UniProt C.** 2015. UniProt: a hub for protein information. Nucleic acids research **43:**D204-212.

29. **Cong Q, Grishin NV.** 2012. MESSA: MEta-Server for protein Sequence Analysis. BMC biology **10:**82.

30. **Pollastri G, Przybylski D, Rost B, Baldi P.** 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins **47:**228-235.

31. **Jones DT.** 1999. Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology **292:**195-202.

32. **Cheng J, Sweredoski M, Baldi P.** 2005. Accurate prediction of protein disordered regions by mining protein structure data. Data Mining and Knowledge Discovery **11:**213-222.

33. **Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB.** 2003. Protein disorder prediction: implications for structural proteomics. Structure **11:**1453-1459.

34. **Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT.** 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. Journal of molecular biology **337:**635-645.

35. **Lobanov MY, Galzitskaya OV.** 2011. The Ising model for prediction of disordered residues from protein sequence alone. Physical biology **8:**035004.

36. **Tusnady GE, Simon I.** 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. Journal of molecular biology **283:**489-506.

37. **Jones DT.** 2007. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics **23:**538-544.

38. **Kall L, Krogh A, Sonnhammer EL.** 2004. A combined transmembrane topology and signal peptide prediction method. Journal of molecular biology **338:**1027-1036.

39. **Krogh A, Larsson B, von Heijne G, Sonnhammer EL.** 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology **305:**567-580.

40. **von Heijne G.** 1992. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. Journal of molecular biology **225:**487-494.

41. **Bendtsen JD, Nielsen H, von Heijne G, Brunak S.** 2004. Improved prediction of signal peptides: SignalP 3.0. Journal of molecular biology **340:**783-795.

42. **Lupas A, Van Dyke M, Stock J.** 1991. Predicting coiled coils from protein sequences. Science **252:**1162-1164.

43. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. Journal of molecular biology **215:**403-410.

44. **Remmert M, Biegert A, Hauser A, Soding J.** 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods **9:**173-175.

45. **Zhang Y.** 2008. I-TASSER server for protein 3D structure prediction. BMC bioinformatics **9:**40.

46. **Roy A, Kucukural A, Zhang Y.** 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nature protocols **5:**725-738.

47. **Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV.** 2014. ECOD: an evolutionary classification of protein domains. PLoS computational biology **10:**e1003926.

48. **Cole C, Barber JD, Barton GJ.** 2008. The Jpred 3 secondary structure prediction server. Nucleic acids research **36:**W197-201.

49. **Pei J, Kim BH, Grishin NV.** 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic acids research **36:**2295-2300.

50. **Pei J, Grishin NV.** 2014. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. Methods in molecular biology **1079:**263-271.

51. **Katoh K, Standley DM.** 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution **30:**772-780.

52. **Henikoff S, Henikoff JG.** 1992. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America **89:**10915-10919.

53. **Dziubanska PJ, Derewenda U, Ellena JF, Engel DA, Derewenda ZS.** 2014. The structure of the C-terminal domain of the Zaire ebolavirus nucleoprotein. Acta crystallographica. Section D, Biological crystallography **70:**2420-2429.

54. **Lee JE, Fusco ML, Hessell AJ, Oswald WB, Burton DR, Saphire EO.** 2008. Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. Nature **454:**177-182.

55. **Xu W, Edwards MR, Borek DM, Feagins AR, Mittal A, Alinger JB, Berry KN, Yen B, Hamilton J, Brett TJ, Pappu RV, Leung DW, Basler CF, Amarasinghe GK.** 2014. Ebola virus VP24 targets a unique NLS binding site on karyopherin alpha 5 to selectively compete with nuclear import of phosphorylated STAT1. Cell host & microbe **16:**187-200.

56. **Edwards MR, Johnson B, Mire CE, Xu W, Shabman RS, Speller LN, Leung DW, Geisbert TW, Amarasinghe GK, Basler CF.** 2014. The Marburg virus VP24 protein interacts with Keap1 to activate the cytoprotective antioxidant response pathway. Cell reports **6:**1017-1025.

57. **Kimberlin CR, Bornholdt ZA, Li S, Woods VL, Jr., MacRae IJ, Saphire EO.** 2010. Ebolavirus VP35 uses a bimodal strategy to bind dsRNA for innate immune suppression. Proceedings of the National Academy of Sciences of the United States of America **107:**314-319.

58. **Hartlieb B, Muziol T, Weissenhorn W, Becker S.** 2007. Crystal structure of the C-terminal domain of Ebola virus VP30 reveals a role in transcription and nucleocapsid association. Proceedings of the National Academy of Sciences of the United States of America **104:**624-629.

59. **Bale S, Julien JP, Bornholdt ZA, Krois AS, Wilson IA, Saphire EO.** 2013. Ebolavirus VP35 coats the backbone of double-stranded RNA for interferon antagonism. Journal of virology **87:**10385-10388.

60. **Zhang AP, Bornholdt ZA, Liu T, Abelson DM, Lee DE, Li S, Woods VL, Jr., Saphire EO.** 2012. The ebola virus interferon antagonist VP24 directly binds STAT1 and has a novel, pyramidal fold. PLoS pathogens **8:**e1002550.

61. **Bale S, Dias JM, Fusco ML, Hashiguchi T, Wong AC, Liu T, Keuhne AI, Li S, Woods VL, Jr., Chandran K, Dye JM, Saphire EO.** 2012. Structural basis for differential neutralization of ebolaviruses. Viruses **4:**447-470.

62. **Dias JM, Kuehne AI, Abelson DM, Bale S, Wong AC, Halfmann P, Muhammad MA, Fusco ML, Zak SE, Kang E, Kawaoka Y, Chandran K, Dye JM, Saphire EO.** 2011. A shared structural solution for neutralizing ebolaviruses. Nature structural & molecular biology **18:**1424-1427.

63. **Binning JM, Wang T, Luthra P, Shabman RS, Borek DM, Liu G, Xu W, Leung DW, Basler CF, Amarasinghe GK.** 2013. Development of RNA aptamers targeting Ebola virus VP35. Biochemistry **52:**8406-8419.

64. **Malashkevich VN, Schneider BJ, McNally ML, Milhollen MA, Pang JX, Kim PS.** 1999. Core structure of the envelope glycoprotein GP2 from Ebola virus at 1.9-A resolution. Proceedings of the National Academy of Sciences of the United States of America **96:**2662-2667.

65. **Bornholdt ZA, Noda T, Abelson DM, Halfmann P, Wood MR, Kawaoka Y, Saphire EO.** 2013. Structural rearrangement of ebola virus VP40 begets multiple functions in the virus life cycle. Cell **154:**763-774.

66. **Prins KC, Delpeut S, Leung DW, Reynard O, Volchkova VA, Reid SP, Ramanan P, Cardenas WB, Amarasinghe GK, Volchkov VE, Basler CF.** 2010. Mutations abrogating VP35 interaction with double-stranded RNA render Ebola virus avirulent in guinea pigs. Journal of virology **84:**3004-3015.

67. **Leung DW, Shabman RS, Farahbakhsh M, Prins KC, Borek DM, Wang T, Muhlberger E, Basler CF, Amarasinghe GK.** 2010. Structural and functional characterization of Reston Ebola virus VP35 interferon inhibitory domain. Journal of molecular biology **399:**347-357.

68. **Leung DW, Ginder ND, Fulton DB, Nix J, Basler CF, Honzatko RB, Amarasinghe GK.** 2009. Structure of the Ebola VP35 interferon inhibitory domain. Proceedings of the National Academy of Sciences of the United States of America **106:**411-416.

69. **Leung DW, Prins KC, Borek DM, Farahbakhsh M, Tufariello JM, Ramanan P, Nix JC, Helgeson LA, Otwinowski Z, Honzatko RB, Basler CF, Amarasinghe GK.** 2010. Structural basis for dsRNA recognition and interferon antagonism by Ebola VP35. Nature structural & molecular biology **17:**165-172.

70. **Leung DW, Borek D, Luthra P, Binning JM, Anantpadma M, Liu G, Harvey IB, Su Z, Endlich-Frazier A, Pan J, Shabman RS, Chiu W, Davey RA, Otwinowski Z, Basler CF, Amarasinghe GK.** 2015. An Intrinsically Disordered Peptide from Ebola Virus VP35 Controls Viral RNA Synthesis by Modulating Nucleoprotein-RNA Interactions. Cell reports **11:**376-389.

71. **Modrof J, Becker S, Muhlberger E.** 2003. Ebola virus transcription activator VP30 is a zinc-binding protein. Journal of virology **77:**3334-3338.

72. **Gough J, Karplus K, Hughey R, Chothia C.** 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. Journal of molecular biology **313:**903-919.

73. **de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N.** 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic acids research **34:**W362-365.

74. **Tanner SJ, Ariza A, Richard CA, Kyle HF, Dods RL, Blondot ML, Wu W, Trincao J, Trinh CH, Hiscox JA, Carroll MW, Silman NJ, Eleouet JF, Edwards TA, Barr JN.** 2014. Crystal structure of the essential transcription antiterminator M2-1 protein of human respiratory syncytial virus and implications of its phosphorylation. Proceedings of the National Academy of Sciences of the United States of America **111:**1580-1585.

75. **Leyrat C, Renner M, Harlos K, Huiskonen JT, Grimes JM.** 2014. Drastic changes in conformational dynamics of the antiterminator M2-1 regulate transcription efficiency in Pneumovirinae. eLife **3:**e02674.

76. **Rudolph MG, Kraus I, Dickmanns A, Eickmann M, Garten W, Ficner R.** 2003. Crystal structure of the borna disease virus nucleoprotein. Structure **11:**1219-1226.

77. **Green TJ, Zhang X, Wertz GW, Luo M.** 2006. Structure of the vesicular stomatitis virus nucleoprotein-RNA complex. Science **313:**357-360.

78. **Albertini AA, Wernimont AK, Muziol T, Ravelli RB, Clapier CR, Schoehn G, Weissenhorn W, Ruigrok RW.** 2006. Crystal structure of the rabies virus nucleoprotein-RNA complex. Science **313:**360-363.

79. **Tawar RG, Duquerroy S, Vonrhein C, Varela PF, Damier-Piolle L, Castagne N, MacLellan K, Bedouelle H, Bricogne G, Bhella D, Eleouet JF, Rey FA.** 2009. Crystal structure of a nucleocapsid-like nucleoprotein-RNA complex of respiratory syncytial virus. Science **326:**1279-1283.

80. **Yabukarski F, Lawrence P, Tarbouriech N, Bourhis JM, Delaforge E, Jensen MR, Ruigrok RW, Blackledge M, Volchkov V, Jamin M.** 2014. Structure of Nipah virus

unassembled nucleoprotein in complex with its viral chaperone. Nature structural & molecular biology **21:**754-759.

81. **Dong S, Yang P, Li G, Liu B, Wang W, Liu X, Xia B, Yang C, Lou Z, Guo Y, Rao Z.** 2015. Insight into the Ebola virus nucleocapsid assembly mechanism: crystal structure of Ebola virus nucleoprotein core domain at 1.8 A resolution. Protein & cell **6:**351-362.

82. **Gerlach P, Malet H, Cusack S, Reguera J.** 2015. Structural Insights into Bunyavirus Replication and Its Regulation by the vRNA Promoter. Cell **161:**1267-1279.

83. **Lu X, McDonald SM, Tortorici MA, Tao YJ, Vasquez-Del Carpio R, Nibert ML, Patton JT, Harrison SC.** 2008. Mechanism for coordinated RNA packaging and genome replication by rotavirus polymerase VP1. Structure **16:**1678-1688.

84. **Tao Y, Farsetta DL, Nibert ML, Harrison SC.** 2002. RNA synthesis in a cage--structural studies of reovirus polymerase lambda3. Cell **111:**733-745.

85. **Pan J, Vakharia VN, Tao YJ.** 2007. The structure of a birnavirus polymerase reveals a distinct active site topology. Proceedings of the National Academy of Sciences of the United States of America **104:**7385-7390.

86. **Bressanelli S, Tomei L, Rey FA, De Francesco R.** 2002. Structural analysis of the hepatitis C virus RNA polymerase in complex with ribonucleotides. Journal of virology **76:**3482-3492.

87. **Ferrer-Orta C, Arias A, Perez-Luque R, Escarmis C, Domingo E, Verdaguer N.** 2004. Structure of foot-and-mouth disease virus RNA-dependent RNA polymerase and

its complex with a template-primer RNA. The Journal of biological chemistry **279:**47212-47221.

88. **te Velthuis AJ.** 2014. Common and unique features of viral RNA-dependent polymerases. Cellular and molecular life sciences : CMLS **71:**4403-4420.

89. **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research **25:**3389-3402.

90. **Bouvet M, Ferron F, Imbert I, Gluais L, Selisko B, Coutard B, Canard B, Decroly E.** 2012. [Capping strategies in RNA viruses]. Medecine sciences : M/S **28:**423-429.

91. **Smietanski M, Werner M, Purta E, Kaminska KH, Stepinski J, Darzynkiewicz E, Nowotny M, Bujnicki JM.** 2014. Structural analysis of human 2'-O-ribose methyltransferases involved in mRNA cap structure formation. Nature communications **5:**3004.

92. **Takada A, Fujioka K, Tsuiji M, Morikawa A, Higashi N, Ebihara H, Kobasa D, Feldmann H, Irimura T, Kawaoka Y.** 2004. Human macrophage C-type lectin specific for galactose and N-acetylgalactosamine promotes filovirus entry. Journal of virology **78:**2943-2947.

93. **Yaddanapudi K, Palacios G, Towner JS, Chen I, Sariol CA, Nichol ST, Lipkin WI.** 2006. Implication of a retrovirus-like glycoprotein peptide in the immunopathogenesis of Ebola and Marburg viruses. FASEB journal : official publication of the Federation of American Societies for Experimental Biology **20:**2519-2530.

94. **Kash JC, Muhlberger E, Carter V, Grosch M, Perwitasari O, Proll SC, Thomas MJ, Weber F, Klenk HD, Katze MG.** 2006. Global suppression of the host antiviral response by Ebola- and Marburgviruses: increased antagonism of the type I interferon response is associated with enhanced virulence. Journal of virology **80:**3009-3020.

95. **Cardenas WB, Loo YM, Gale M, Jr., Hartman AL, Kimberlin CR, Martinez-Sobrido L, Saphire EO, Basler CF.** 2006. Ebola virus VP35 protein binds double-stranded RNA and inhibits alpha/beta interferon production induced by RIG-I signaling. Journal of virology **80:**5168-5178.

96. **Volchkov VE, Chepurnov AA, Volchkova VA, Ternovoj VA, Klenk HD.** 2000. Molecular characterization of guinea pig-adapted variants of Ebola virus. Virology **277:**147-155.

97. **Ebihara H, Takada A, Kobasa D, Jones S, Neumann G, Theriault S, Bray M, Feldmann H, Kawaoka Y.** 2006. Molecular determinants of Ebola virus virulence in mice. PLoS pathogens **2:**e73.

98. **Kondratowicz AS, Lennemann NJ, Sinn PL, Davey RA, Hunt CL, Moller-Tank S, Meyerholz DK, Rennert P, Mullins RF, Brindley M, Sandersfeld LM, Quinn K, Weller M, McCray PB, Jr., Chiorini J, Maury W.** 2011. T-cell immunoglobulin and mucin domain 1 (TIM-1) is a receptor for Zaire Ebolavirus and Lake Victoria Marburgvirus. Proceedings of the National Academy of Sciences of the United States of America **108:**8426-8431.

99. **Alvarez CP, Lasala F, Carrillo J, Muniz O, Corbi AL, Delgado R.** 2002. C-type lectins DC-SIGN and L-SIGN mediate cellular entry by Ebola virus in cis and in trans. Journal of virology **76:**6841-6844.

100. **Simmons G, Reeves JD, Grogan CC, Vandenberghe LH, Baribaud F, Whitbeck JC, Burke E, Buchmeier MJ, Soilleux EJ, Riley JL, Doms RW, Bates P, Pohlmann S.** 2003. DC-SIGN and DC-SIGNR bind ebola glycoproteins and enhance infection of macrophages and endothelial cells. Virology **305:**115-123.

101. **Chan SY, Empig CJ, Welte FJ, Speck RF, Schmaljohn A, Kreisberg JF, Goldsmith MA.** 2001. Folate receptor-alpha is a cofactor for cellular entry by Marburg and Ebola viruses. Cell **106:**117-126.

102. **Chandran K, Sullivan NJ, Felbor U, Whelan SP, Cunningham JM.** 2005. Endosomal proteolysis of the Ebola virus glycoprotein is necessary for infection. Science **308:**1643-1645.

103. **Sullivan NJ, Peterson M, Yang ZY, Kong WP, Duckers H, Nabel E, Nabel GJ.** 2005. Ebola virus glycoprotein toxicity is mediated by a dynamin-dependent protein-trafficking pathway. Journal of virology **79:**547-553.

104. **Halfmann P, Neumann G, Kawaoka Y.** 2011. The Ebolavirus VP24 protein blocks phosphorylation of p38 mitogen-activated protein kinase. The Journal of infectious diseases **204 Suppl 3:**S953-956.

105. **Fabozzi G, Nabel CS, Dolan MA, Sullivan NJ.** 2011. Ebolavirus proteins suppress the effects of small interfering RNA by direct interaction with the mammalian RNA interference pathway. Journal of virology **85:**2512-2523.

106. **Shabman RS, Leung DW, Johnson J, Glennon N, Gulcicek EE, Stone KL, Leung L, Hensley L, Amarasinghe GK, Basler CF.** 2011. DRBP76 associates with Ebola virus VP35 and suppresses viral polymerase function. The Journal of infectious diseases **204 Suppl 3:**S911-918.

107. **Prins KC, Cardenas WB, Basler CF.** 2009. Ebola virus protein VP35 impairs the function of interferon regulatory factor-activating kinases IKKepsilon and TBK-1. Journal of virology **83:**3069-3077.

108. **Luthra P, Ramanan P, Mire CE, Weisend C, Tsuda Y, Yen B, Liu G, Leung DW, Geisbert TW, Ebihara H, Amarasinghe GK, Basler CF.** 2013. Mutual antagonism between the Ebola virus VP35 protein and the RIG-I activator PACT determines infection outcome. Cell host & microbe **14:**74-84.

109. **Feng Z, Cerveny M, Yan Z, He B.** 2007. The VP35 protein of Ebola virus inhibits the antiviral effect mediated by double-stranded RNA-dependent protein kinase PKR. Journal of virology **81:**182-192.

110. **Chang TH, Kubota T, Matsuoka M, Jones S, Bradfute SB, Bray M, Ozato K.** 2009. Ebola Zaire virus blocks type I interferon production by exploiting the host SUMO modification machinery. PLoS pathogens **5:**e1000493.

111. **Kubota T, Matsuoka M, Chang TH, Bray M, Jones S, Tashiro M, Kato A, Ozato K.** 2009. Ebolavirus VP35 interacts with the cytoplasmic dynein light chain 8. Journal of virology **83:**6952-6956.

112. **Yamayoshi S, Noda T, Ebihara H, Goto H, Morikawa Y, Lukashevich IS, Neumann G, Feldmann H, Kawaoka Y.** 2008. Ebola virus matrix protein VP40 uses

the COPII transport system for its intracellular transport. Cell host & microbe **3:**168-177.

113. **Garcia M, Cooper A, Shi W, Bornmann W, Carrion R, Kalman D, Nabel GJ.** 2012. Productive replication of Ebola virus is regulated by the c-Abl1 tyrosine kinase. Science translational medicine **4:**123ra124.

114. **Harty RN, Brown ME, Wang G, Huibregtse J, Hayes FP.** 2000. A PPxY motif within the VP40 protein of Ebola virus interacts physically and functionally with a ubiquitin ligase: implications for filovirus budding. Proceedings of the National Academy of Sciences of the United States of America **97:**13871-13876.

115. **Ruthel G, Demmin GL, Kallstrom G, Javid MP, Badie SS, Will AB, Nelle T, Schokman R, Nguyen TL, Carra JH, Bavari S, Aman MJ.** 2005. Association of ebola virus matrix protein VP40 with microtubules. Journal of virology **79:**4709-4719.

116. **Han Z, Harty RN.** 2005. Packaging of actin into Ebola virus VLPs. Virology journal **2:**92.

117. **Lu J, Qu Y, Liu Y, Jambusaria R, Han Z, Ruthel G, Freedman BD, Harty RN.** 2013. Host IQGAP1 and Ebola virus VP40 interactions facilitate virus-like particle egress. Journal of virology **87:**7777-7780.

118. **Jeffers SA, Sanders DA, Sanchez A.** 2002. Covalent modifications of the ebola virus glycoprotein. Journal of virology **76:**12463-12472.

119. **Manicassamy B, Wang J, Jiang H, Rong L.** 2005. Comprehensive analysis of ebola virus GP1 in viral entry. Journal of virology **79:**4793-4805.

120. **Haasnoot J, de Vries W, Geutjes EJ, Prins M, de Haan P, Berkhout B.** 2007. The Ebola virus VP35 protein is a suppressor of RNA silencing. PLoS pathogens **3:**e86.

121. **Schumann M, Gantke T, Muhlberger E.** 2009. Ebola virus VP35 antagonizes PKR activity through its C-terminal interferon inhibitory domain. Journal of virology **83:**8993-8997.

122. **McCarthy SE, Johnson RF, Zhang YA, Sunyer JO, Harty RN.** 2007. Role for amino acids 212KLR214 of Ebola virus VP40 in assembly and budding. Journal of virology **81:**11452-11460.

123. **Yamayoshi S, Kawaoka Y.** 2007. Mapping of a region of Ebola virus VP40 that is important in the production of virus-like particles. The Journal of infectious diseases **196 Suppl 2:**S291-295.

124. **Adu-Gyamfi E, Soni SP, Jee CS, Digman MA, Gratton E, Stahelin RV.** 2014. A loop region in the N-terminal domain of Ebola virus VP40 is important in viral assembly, budding, and egress. Viruses **6:**3837-3854.

125. **Adu-Gyamfi E, Soni SP, Xue Y, Digman MA, Gratton E, Stahelin RV.** 2013. The Ebola virus matrix protein penetrates into the plasma membrane: a key step in viral protein 40 (VP40) oligomerization and viral egress. The Journal of biological chemistry **288:**5779-5789.

CHAPTER SEVEN
Tiger Swallowtail genome reveals hotspots for speciation and molecular basis for predator defense in caterpillars

INTRODUCTION

An organism in all its complexity of morphological and behavioral traits develops through interaction of its genetic makeup with the environment. Unraveling and predicting these traits from genotype chart the future of biological research. Success in such prediction depends on an ability to routinely sequence and analyze genomes of thousands of individuals from selected model organisms. In this quest, butterflies and moths with relatively small genomes but complex life cycles and diverse wing patterns, are emerging as powerful models. A new paradigm that gene exchange between species is pivotal in evolution of adaptation[7], and anticipation of using comparative genomics to uncover molecular mechanisms responsible for complex traits are fueling excitement in the field[8-10].

A showy North American butterfly, the Eastern Tiger Swallowtail, *Papilio glaucus* (*Pgl,* Fig. 1A,B), is honored as the state insect in five USA states. *Pgl* has remarkable morphological and behavioral features at all stages of development. Like other swallowtails, the caterpillar of *Pgl* possesses a fleshy fork-shaped osmeterium. Upon threat, this organ everts to emit malodorous predator-repelling terpenes[1]. Combined with the tongue-like osmeterium, two eyespots on the thorax complete the snake mimicry of the caterpillar. The *Pgl* chrysalis undergoes conditional diapause[2]. Female adults of *Pgl* are dimorphic between a yellow form and a melanic form to mimic the unpalatable Pipevine Swallowtail, *Battus philenor*[3].

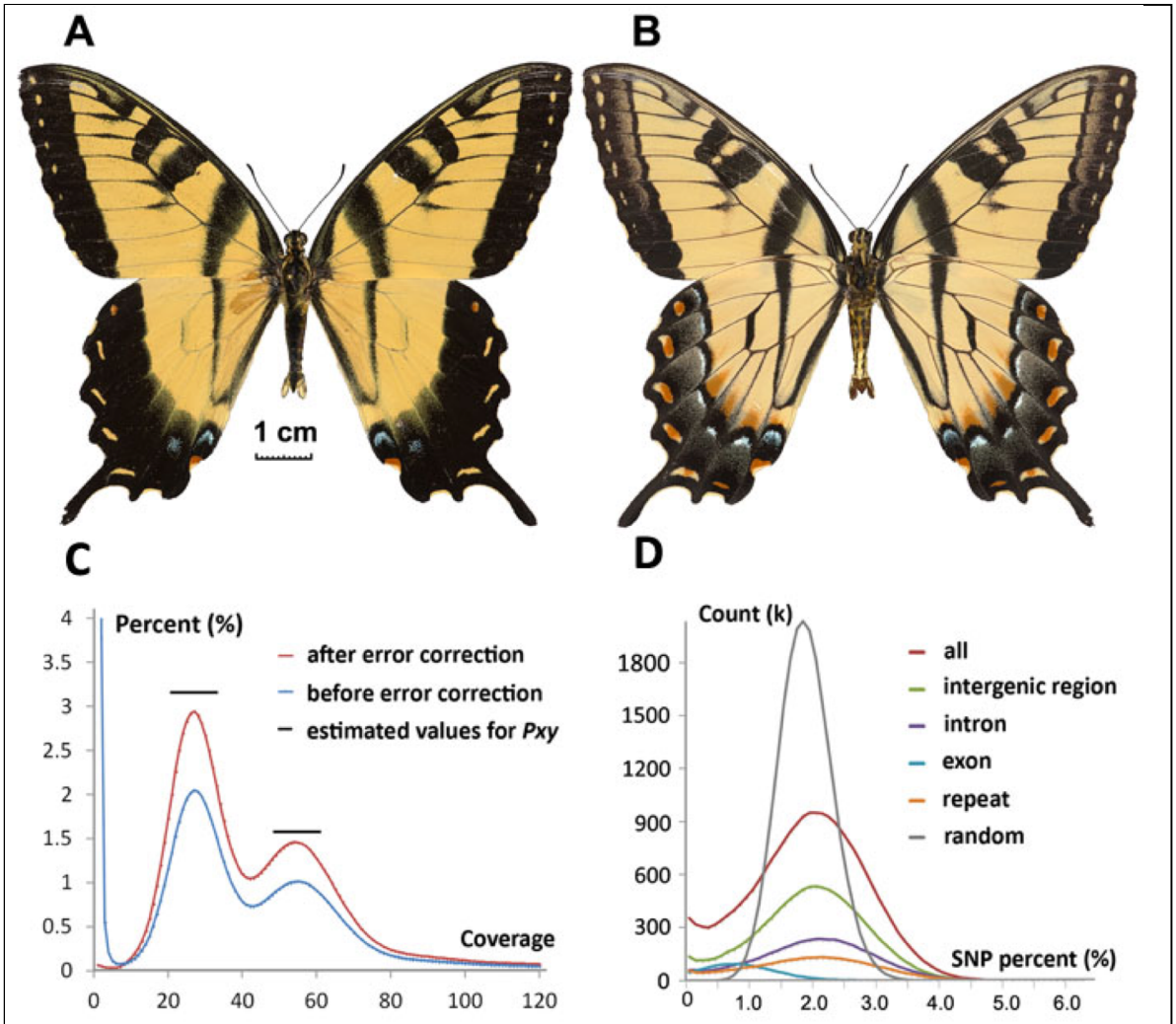


Fig. 1 The *Pgl* genome is highly heterozygous. (A) dorsal and (B) ventral aspects of the sequenced *Pgl* specimen preserved after tissue sampling; (C) 17-mer coverage before and after error correction. The height of the two peaks in a similar graph for *Pxy* is estimated from Fig S3 in [9]. (D) Percent of SNP in 1000 bp overlapping windows.

Pgl and its sister species, *Papilio canadensis* (*Pca*), diverged just 0.6 million years ago[4], and yet developed substantial differences in thermal preference, caterpillar food plants, body size, and female mimicry[2]. Unlike *Pgl*, *Pca* undergoes obligate pupal diapause. However, *Pgl* and *Pca* hybridize in a narrow zone where they meet. A hybrid species from the Appalachian Mountains, *Papilio appalachiensis* (*Pap*), was described recently[5]. These three species offer a

model system to study evolution, hybridization and speciation, and these studies will benefit from decoding a *Papilio* genome.

RESULTS AND DISCUSSION

A cost-effective protocol for *de novo* sequencing and assembly of highly heterozygous genomes

Despite rapid development of next generation sequencing techniques, assembly of highly heterozygous genomes remains a challenge. Many insects have large, widespread and morphologically variable populations with high heterozygosity[11]. Extensive, laboratory inbreeding was used to overcome this problem in the *Heliconius melpomene* (*Hme*)[2] and *Bombyx mori* (*Bmo*)[12,13] genome projects, while the highly hetergozygous *Plutella xylostella* (*Pxy*)[9] genome was cloned into over 100,000 fosmids and required 114 Illumina lanes to sequence. These laborious and expensive procedures impede acquisition of numerous eukaryotic genomes for data-driven discoveries.

The *Pgl* genome is comparable to *Pxy* in heterozygosity (Fig. 1C) and size. However, our protocol allowed us to obtain the *Pgl* genome with quality comparable to other Lepidoptera using genetic material from a single wild-caught specimen and sequence data from a single Illumina lane. Briefly, we extracted DNA from a piece of *Pgl* adult thoracic muscle (Fig 1A,B). Pair-end libraries (250 bp and 500 bp) and three mate-pair libraries (2 kb, 6 kb, and 15 kb) made with a modified Cre-lox protocol[14], were sequenced at both ends for 150 bp. After removal of low quality sequences and error correction, we used Platanus[15] software designed for highly heterozygous genomes to assemble the reads. The primary assembly was larger than

expected and contained a number of shorter scaffolds with significantly lower coverage (Figs. S2, S3) representing divergent alleles in homologous chromosomes. Using in-house scripts, these scaffolds were merged to obtain the final assembly.

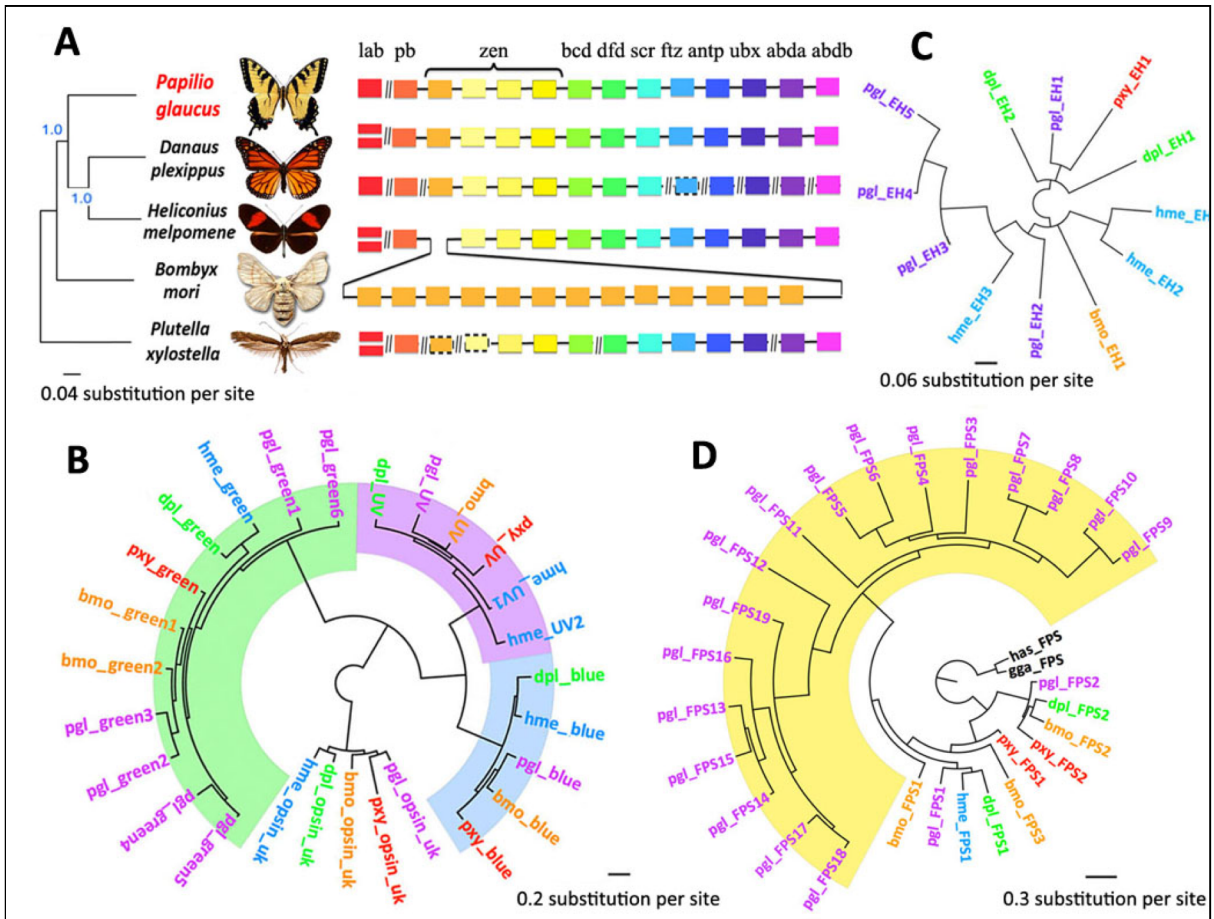


Fig. 2 Comparative analysis of *Pgl* and other Lepidoptera genomes. (A) Evolutionary tree based on the concatenated alignment of universal single-copy orthologs and arrangement of *hox* genes in draft genomes. Orthologs are shown in the same color; double boxes in the same position indicate duplications and "//" marks the boundaries between different scaffolds (B-D). Phylogenetic tree for expanded protein families in *Pgl.* Abbreviation of the species names and protein names are used as tip labels**.** (B) Opsins. (C) Eclosion hormones. (D) Farnesyl pyrophosphate synthase homologs.

Genome quality assessment and gene annotation

We assembled a 376 Mb genome draft of *Pgl* and compared its quality and content with published Lepidoptera genomes (Table 1). The scaffold N50 of *Pgl* is 230 Mb, comparable to

other butterfly genomes, but shorter than the *Pxy* genome. However, despite a larger N50, the *Pxy* genome assembly is incomplete as measured by presence of CEGMA (Core Eukaryotic Genes Mapping Approach) genes[16], Cytoplasmic Ribosomal Proteins (CRP) and independently assembled transcripts, while the *Pgl* genome is among the best in completeness. The residue coverage of CEGMA genes by single *Pgl* scaffolds is the same as the current *Bmo* assembly with an N50 of 3.7 Mb, indicating that the quality of the *Pgl* draft is sufficient for protein annotation and comparative analysis.

The *Pgl* genome is highly heterozygous with an overall SNP rate of 2%. The distribution of SNPs in the genome is prominently non-random (Figs. 1D, S4), with coding regions having an average of 0.8% SNPs. 505 protein coding genes have significantly more (false discovery rate < 0.1) SNPs than the average. Enriched GO-terms show that these genes mostly encode enzymes and proteins involved in the detection of stimuli. Repeats constitute 23% of the *Pgl* genome, which is similar to other butterflies, but less than in moth genomes. We predicted 15,695 protein-coding genes in the *Pgl* and annotated the function for 11,975 of them.

Table 1. Quality and composition of Lepidoptera genomes.

feature	***Pgl***	***Dpl****	***Hme***	***Bmo***	***Pxy***
genome size (Mb)	376	249	274	480	394
Heterozygosity (%)	1.80	0.55	ND	ND	~2†
Scaffold N50 (kb)	230	207	277	27(3700‡)	734
CEGMA (%)	99.3	99.3	98.0	99.3	98.0
CEGMA coverage by single scaffold (%)	85.6	85.6	85.9	85.6	81.7
CRP (%)	100	100	95.7	98.9	94.6
De novo assembled transcripts (%)	98	96	ND	98	83
Repeat content (%)	22.8	16.3	24.9	44.1	34.0
number of proteins (k)	15.7	15.1	12.8	14.3	18.1

* *Dpl: Danaus plexippus*; † Estimated by comparing the distribution of K-mer coverage, as shown in Fig. 1C; ‡ The N50 for the improved assembly.

Comparative analysis of Lepidoptera genomes reveals genetic bases for morphological traits

We compared the *Pgl* protein set with other Lepidoptera. Both phylogenetic trees built from

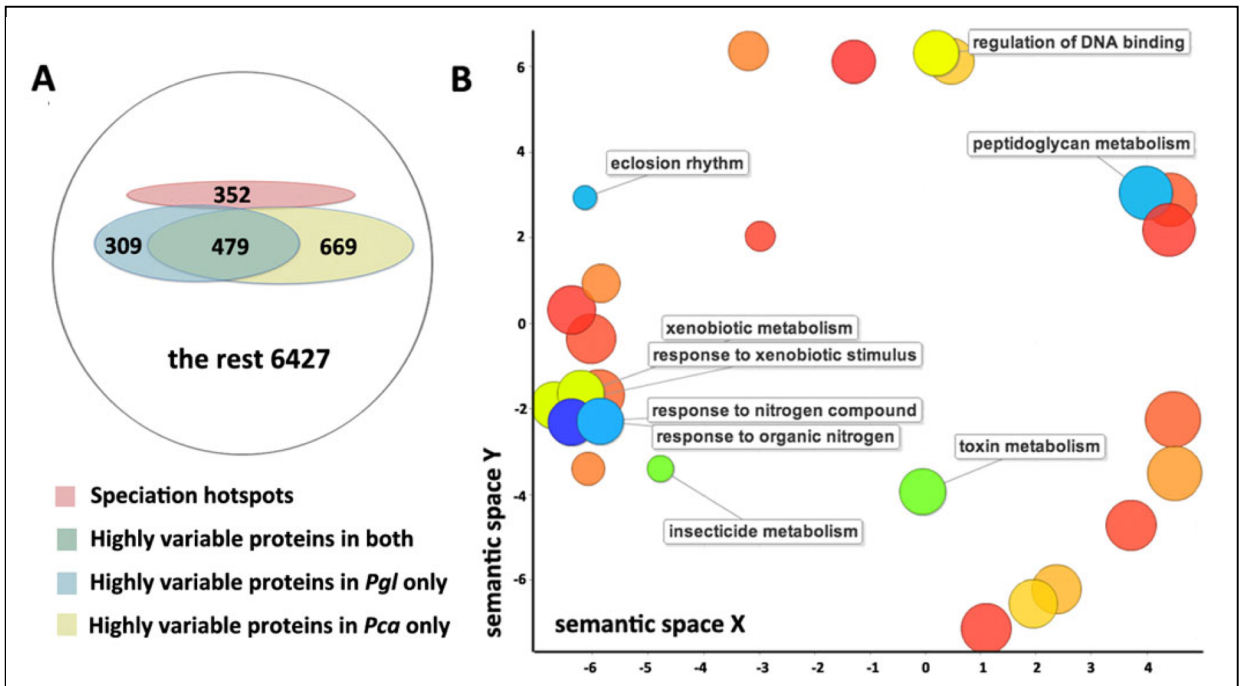


Fig. 3 Speciation hotspots and associated GO-terms. (A) Venn diagram showing speciation hotspots and highly variable proteins within species. (B) Enriched GO terms (biological processes) associated with speciation hotspots. GO terms are grouped in space by similarity in meaning and colored by the significant level. Annotations are shown for the most significantly enriched terms that passed the false discovery rate test.

alignment of the 3,858 universal single copy orthologs (Fig. 2A) and synteny of genes, group *Pgl* with other butterflies. Except *Pxy*, the other four species share high synteny, with over 85% of proteins in micro-syntenic blocks. All the *Pgl* Hox genes that are expected to be linked are on the same scaffold (Fig 2A), indicating the good quality of *Pgl* assembly.

Pgl genome revealed expansion in several protein families. Previous studies characterized six opsins from *Pgl*[17], while the genome assembly suggests nine. *Pgl* has more green light-sensitive opsins, which may indicate a more advanced color perception. The

identified opsins cluster into four groups (Fig. 2B): in addition to previously reported UV, blue and green light sensitive opsins, we find another group of putative UV-sensitive opsins similar to *Drosophila* Rh7[18]. Other notable gene expansions include the eclosion hormone (Fig. 2C) that triggers the emergence of adults[19], and circadian clock-controlled proteins that are involved in timing of the eclosion[20]. Different eclosion hormone copies may vary in their temporal-spatial distribution and impart complex regulation of eclosion, allowing *Pgl* to diapause conditionally in response to external stimuli.

The largest expansion (Fig. 2D) involves farnesyl pyrophosphate synthase (FPPS) homologs belonging to a family of isoprenoid biosynthesis enzymes that synthesize steroids and terpenes[21]. The 24 *Pgl* FPPS genes cluster at several genomic loci. The FPPS proteins are predicted to adopt an isprenoid synthase fold (Fig S11) with fully preserved catalytic sites in 19 of them[22] (Figs. S12, S13). Amino acids lining the FPPS substrate-binding sites are less conserved, implying diverse substrate specificity. RNA-seq data indicates that this gene expansion occurs in other Papilionidae species. The *Papilio*-specific FPPS enzymes form a clade in the evolutionary tree, and they could function in a pathway to synthesize predator-repelling terpenes secreted by the osmeterium, a Papilionidae specific organ among butterflies.

Speciation between *Papilio glaucus* and *Papilio canadensis*[23]

We have built an isolation-with-migration model[24] for *Pgl* and *Pca*. The model predicts that they diverged about half a million years ago and have undergone dramatic increases in the effective population size, resulting in high DNA variability. The model also suggests high gene

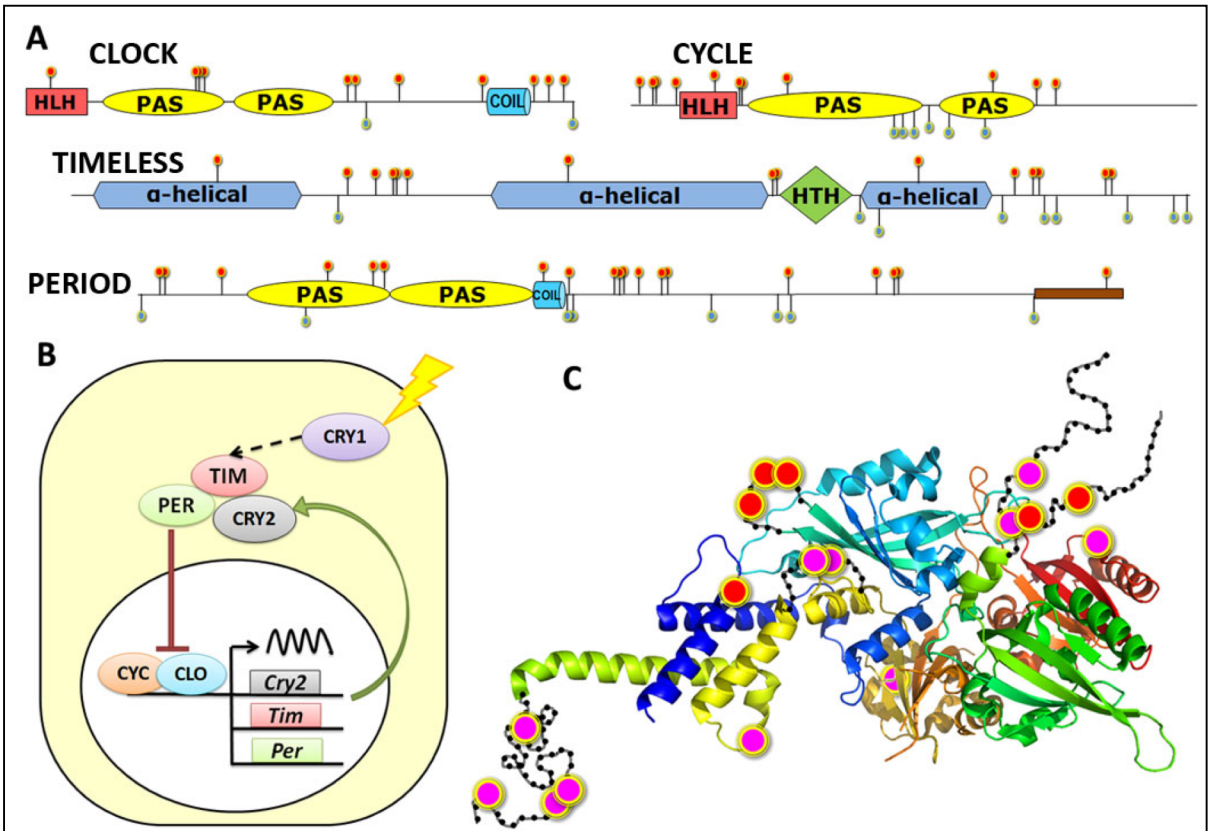


Fig. 4 Circadian clock system may explain differences in diapause between *Pgl* and *Pca*. (A) Domain diagram of CLOCK, CYCLE, PERIOD, and TIMELESS. Mutations within species are marked by green flags and positions that are conserved within but differ between species are marked by red flags. (B) Circadian clock system. CRY: cryptochrome proteins. (C) Map of inter-species mutations on the spatial structure template (PDB id: 4F3L) of CLOCK/CYCLE complex. The mutations are marked by red (CLOCK) and pink (CYCLE) dots and the approximate position of disordered loops is shown as black beads on threads.

flow between *Pgl* and *Pca*, consistent with their successful mating in the lab and the discovery of a hybrid species, *Pap*.

At the whole transcriptome level, *Pgl* and *Pca* are clearly distinguishable. Combining all 8,230 transcripts shared among the *Pgl* and *Pca* specimens, the average variation rate and

dN/dS ratio within species are significantly lower ($P<10^{-4}$) than the inter-species values. However, due to the closeness of intra- and inter- species variation rates, these two species are indistinguishable by most individual genes. Only 351 (4.3%) transcripts display higher inter-species variation in both protein and DNA. Therefore, only a small fraction of genes, termed "speciation hotspots", dictate speciation and adaptation to different environments. The speciation hotspots are mostly conserved within species (Fig. 3A). Overlap between speciation hotspot and positively selected loci in *Pgl* and *Pca* is small (11.1%), since 97.5% of these loci reflect adaptive evolution within either *Pgl* or *Pca*.

The speciation hotspots show a significant ($P<10^{-2}$) enrichment in 57 GO-terms (Fig. 3B). The GO-terms suggest that *Pgl* and *Pca* differ in defense against xenobiotics (GO:0009410 et al.), insecticides (GO:0017143 et al.), and bacteria (GO:009617 et al.), which agrees with that they are exposed to different food plants, insecticides, and bacteria.

The GO term, "eclosion rhythm" is among the most significantly enriched. This GO term is associated with four speciation hotspots that are the central players in the circadian clock system: CLOCK, CYCLE, PEROID, and TIMELESS[25] (Fig. 4A,B). These proteins regulate the timing for adults to hatch from pupae (opposite to diapause) and the temperature preference rhythm in *Drosophila*[26,27]. Mapping amino acid differences between species to 3D structure templates shows that these mutations concentrate on one side of the CLOCK/CYCLE complex[28], forming clusters on the surface (Fig. 4C). Similar mutation site distribution is observed in PERIOD. The surface clustering of mutations suggest that they likely modify interactions between circadian clock proteins and other regulators. Differences in modulation of this timing system could determine obligate diapause *vs.* conditional diapause.

Proteins involved in lipid metabolism (e.g. GO:0006629 et al.) and regulation of transcription factors (GO:2000678 et al.) are also among speciation hotspots. Many insect pheromones are derivatives of metabolites along the fatty acid synthesis and degradation pathways and they play essential roles in insect social and mating behavior[29]. Therefore, differences in enzymes of lipid metabolism could result in pheromone divergence and have a profound impact on speciation. Similarly, differences in the regulators of transcription factors affect many downstream genes and have a significant influence on an organism.

New nuclear DNA barcodes for insect identification

The widely used mitochondrial DNA barcode encoding part of Cytochrome c oxidase subunit 1 (COI) is routinely used for insect identification and cryptic species discovery. However, maternally inherited mitochondrial DNA may have history different from the whole organism and can be transferred between species via cellular symbionts[30]. Consequently, tests with COI barcodes need to be supplemented with work based on nuclear barcodes. Commonly used nuclear markers for insects include 18s rRNA, wingless, EF1a genes and non-coding ITS1 and ITS2. However, these genes fail to distinguish closely-related species such as *Pgl* and *Pca*, which are cleanly separated by COI barcodes. In a quest for nuclear barcodes, we searched for long (> 150 bp) exons that: (1) are present in most genomes as confidently identifiable and alignable single-copy orthologs; (2) differ between many pairs of closely related insect species, but are less variable within species.

Out of 22,731 long exons shared by *Pgl* and *Pca* specimens, only 236 can confidently ($p<0.05$) distinguish the two species, and only 41 have higher discriminating power than COI barcode in either binomial tests or inter-species divergence level. We used 56 insect genomes

forming 460 close species pairs to further reduce the candidate list. Finally, 11 nuclear barcodes were selected (Table 2). In addition to their ability to distinguish sister species, most of them represent phylogeny of insects better than the COI barcode.

Reference genome provides new insights into hybrid species, *Papilio appalachiensis*. Using the *Pgl* genome as a reference, we compared transcriptomes of *Pap* and its parental species (Fig. 5A), *Pgl* and *Pca*. Based on the 7410 shared transcripts, *Pap* is more similar to *Pca* than to *Pgl*. The two *Pap* specimens differ less from each other than homologous

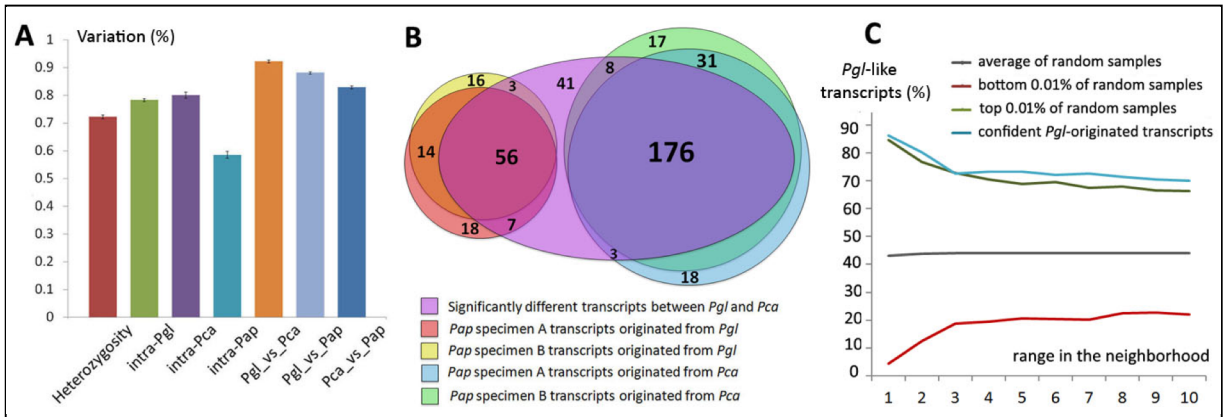


Fig. 5 Reference genome supports *Pap* as a hybrid species. (A) Variation within and between species over all common transcripts; (B) Venn diagram of *Pap* proteins originated from *Pca* or *Pap* and proteins that are significantly different between *Pca* and *Pgl*; (C) Percent of *Pgl*-like proteins (0.2% or more similar) in the neighborhood of confident *Pgl*-originated proteins is significantly (P<0.01) higher than those near randomly selected samples

chromosomes of a single *Pgl* specimen, not to mention different specimens of either parental species. Low variability of *Pap* specimens agrees with it being a distinct species with a smaller effective population size, rather than a result of presently continuing hybridizations between *Pgl* and *Pca*.

High intra-species variation hinders attribution of a *Pap* gene to its parental species by marginally higher sequence identity alone. We attribute a *Pap* gene origin to a particular species if its sequence is similar to those from this species but is different ($p<0.05$) from the other species, and detected 207 *Pca*-originated and 70 *Pgl*-originated transcripts. Stringent

tests with Bonferroni correction show similar hybrid composition of *Pap* genes. Despite the small number of confidently assigned genes, they represent a majority (86%) of genes that significantly differ between *Pgl* and *Pca* (Fig. 5B). Only 8 *Pap* genes are significantly ($p<0.05$) different from both parental species.

We used the *Pgl* assembly to analyze the distribution of statistically supported *Pgl*-originated genes in the genome. These genes are significantly ($p<10^{-7}$) more likely to be clustered compared to randomly selected gene sets of the same size. Neighborhoods of these *Pgl* genes are enriched in genes with higher similarity to *Pgl* (Fig. 5C). Clustering is even more prominent for *Pca* genes (Figs. S30, S31), which agrees with a model of hybridization followed by limited gene recombination.

The results support a hybrid origin of *Pap* with about 72% genes inherited from *Pca*, explaining the higher morphological and behavioral similarity between *Pap* and *Pca*. For instance, *Pap* with obligate diapause has all four speciation hotspots involved in the circadian clock system inherited from *Pca*, offering additional evidence for the proposed functional role of these genes. In contrast, the *Pgl* 6-phosphogluconate dehydrogenase (6PGD), which was shown to be closely linked to the melanic female-enabling gene on the Z chromosome[31], is inherited from *Pgl.* This link could be relevant to the observed *Pgl*-like black females in *Pap* from West Virginia, where the two *Pap* specimens were collected[23]. Two putative transcriptional factors in the neighborhood of 6PGD are candidate regulators of melanic female phenotype.

REFERENCES

1 Eisner, T. & Meinwald, Y. C. Defensive Secretion of a Caterpillar (*Papilio*). *Science* **150**, 1733-1735 (1965).

2 Hagen, R. H., Lederhouse, R. C., Bossart, J. L. & Scriber, J. M. *Papilio canadensis* and *P. glaucus* are distinct species. *Journal of the Lepidopterists' Society* **45**, 245-258 (1991).

3 Brower, J. V. Z. Experimental studies of mimicry in some North American butterflies: Part II. Battus philenor and Papilio troilus, P. polyxenes and P. glaucus. *Evolution* **12**, 123-136 (1958).

4 Kunte, K. *et al.* Sex chromosome mosaicism and hybrid speciation among tiger swallowtail butterflies. *PLoS genetics* **7**, e1002274, doi:10.1371/journal.pgen.1002274 (2011).

5 Pavulaan, H. & Wright, D. M. *Pterourus appalachiensis* (*Papilionidae: Papilioninae*), a new swallowtail butterfly from the Appalachian region of the United States. *Taxonomic Report of the International Lepidoptera Survery* **3**, 1-10 (2002).

6 Via, S. & Lande, L. Genotype-Environment Interaction and the Evolution of Phenotypic Plasticity. *Evolution* **39**, 505-522 (1985).

7 Heliconius Genome, C. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94-98, doi:10.1038/nature11041 (2012).

8 Zhan, S., Merlin, C., Boore, J. L. & Reppert, S. M. The monarch butterfly genome yields insights into long-distance migration. *Cell* **147**, 1171-1185, doi:10.1016/j.cell.2011.09.052 (2011).

9 You, M. *et al.* A heterozygous moth genome provides insights into herbivory and detoxification. *Nature genetics* **45**, 220-225, doi:10.1038/ng.2524 (2013).

10 Kunte, K. *et al.* doublesex is a mimicry supergene. *Nature* **507**, 229-232, doi:10.1038/nature13112 (2014).

11 Allendorf, F. W. Genetic drift and the loss of alleles versus heterozygosity. *Zoo biology* **5**, 181-190 (1986).

12 Xia, Q. *et al.* A draft sequence for the genome of the domesticated silkworm (Bombyx mori). *Science* **306**, 1937-1940, doi:10.1126/science.1102210 (2004).

13 International Silkworm Genome, C. The genome of a lepidopteran model insect, the silkworm Bombyx mori. *Insect biochemistry and molecular biology* **38**, 1036-1045, doi:10.1016/j.ibmb.2008.11.004 (2008).

14 Van Nieuwerburgh, F. *et al.* Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic acids research* **40**, e24, doi:10.1093/nar/gkr1000 (2012).

15 Kajitani, R. *et al.* Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research* **24**, 1384-1395, doi:10.1101/gr.170720.113 (2014).

16 Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067, doi:10.1093/bioinformatics/btm071 (2007).

17 Briscoe, A. D. Six opsins from the butterfly Papilio glaucus: molecular phylogenetic evidence for paralogous origins of red-sensitive visual pigments in insects. *Journal of molecular evolution* **51**, 110-121 (2000).

18 Brody, T. & Cravchik, A. Drosophila melanogaster G protein-coupled receptors. *The Journal of cell biology* **150**, F83-88 (2000).

19 Truman, J. W. Hormonal control of insect ecdysis: endocrine cascades for coordinating behavior with physiology. *Vitamins and hormones* **73**, 1-30, doi:10.1016/S0083-6729(05)73001-6 (2005).

20 Myers, E. M., Yu, J. & Sehgal, A. Circadian control of eclosion: interaction between a central and peripheral clock in Drosophila melanogaster. *Current biology : CB* **13**, 526-533 (2003).

21 Dhar, M. K., Koul, A. & Kaul, S. Farnesyl pyrophosphate synthase: a key enzyme in isoprenoid biosynthetic pathway and potential molecular target for drug development. *New biotechnology* **30**, 114-123, doi:10.1016/j.nbt.2012.07.001 (2013).

22 Zhang, Y. *et al.* Chemo-Immunotherapeutic Anti-Malarials Targeting Isoprenoid Biosynthesis. *ACS medicinal chemistry letters* **4**, 423-427, doi:10.1021/ml4000436 (2013).

23 Zhang, W., Kunte, K. & Kronforst, M. R. Genome-wide characterization of adaptation and speciation in tiger swallowtail butterflies using de novo transcriptome assemblies. *Genome biology and evolution* **5**, 1233-1245, doi:10.1093/gbe/evt090 (2013).

24 Hey, J. & Nielsen, R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 2785-2790, doi:10.1073/pnas.0611164104 (2007).

25 Zhu, H. *et al.* Cryptochromes define a novel circadian clock mechanism in monarch butterflies that may underlie sun compass navigation. *PLoS biology* **6**, e4, doi:10.1371/journal.pbio.0060004 (2008).

26 Blanchardon, E. *et al.* Defining the role of Drosophila lateral neurons in the control of circadian rhythms in motor activity and eclosion by targeted genetic ablation and PERIOD protein overexpression. *The European journal of neuroscience* **13**, 871-888 (2001).

27 Kaneko, H. *et al.* Circadian rhythm of temperature preference and its neural control in Drosophila. *Current biology : CB* **22**, 1851-1857, doi:10.1016/j.cub.2012.08.006 (2012).

28 Huang, N. *et al.* Crystal structure of the heterodimeric CLOCK:BMAL1 transcriptional activator complex. *Science* **337**, 189-194, doi:10.1126/science.1222804 (2012).

29 Matsumoto, S. Molecular mechanisms underlying sex pheromone production in moths. *Bioscience, biotechnology, and biochemistry* **74**, 223-231, doi:10.1271/bbb.90756 (2010).

30 Whitworth, T. L., Dawson, R. D., Magalon, H. & Baudry, E. DNA barcoding cannot reliably identify species of the blowfly genus Protocalliphora (Diptera: Calliphoridae). *Proceedings. Biological sciences / The Royal Society* **274**, 1731-1739, doi:10.1098/rspb.2007.0062 (2007).

31 Hagen, R. H. & Scriber, J. M. Sex-Linked Diapause, Color, and Allozyme Loci in Papilio glaucus: Linkage Analysis and Significance in a Hybrid Zone. *Journal of Heredity* **80**, 179-185 (1989).

CHAPTER EIGHT
Speciation in Cloudless Sulphurs gleaned from complete genomes

INTRODUCTION

Butterflies and moths (Lepidoptera) are some of the best-known and best-studied insects. Their colorful wings and complex life cycles attract wide attention from both researchers and the public. Despite this popularity, little is known about the genetic makeup of Lepidoptera, and compete genomes are available for less than a dozen species [1-11]. However, small genome sizes and extensive knowledge about the morphology and life histories of Lepidoptera offer a promise to further our understanding in genetics, molecular evolution, and speciation by comparative genomics. Among butterflies, representative genomes are currently known for only three families: the swallowtails (Papilionidae), the brushfoots (Nymphalidae), and the skippers (Hesperiidae). The brushfoots have been prevalent in genomics studies, with research on *Heliconius* and the Monarch (*Danaus plexippus*) leading the field [12, 13]. For comparative genomics of butterflies, it is essential to sequence complete genomes of all major phylogenetic groups.

The family Pieridae (Whites and Sulphurs) may be the prototype for the name "butterfly". A common yellow European species, the Brimstone (*Gonepteryx rhamni*), was called the "butter-colored fly" by early naturalists [14]. This family includes some of the very few butterflies known as crop pests, such as the Cabbage Whites (*Pieris rapae* and *Pieris brassicae*) and Alfalfa Sulphur (*Colias eurytheme*). Pierids are particularly well known for using pterins as pigments on their wings [15]. While most swallowtails diapause as pupae,

many Pierids overwinter as adults and enter reproductive diapause in the fall. Due to similarities in pupae, Pierids were previously hypothesized to be a sister family to the swallowtails (can you put ref. 16 here instead of ???), a view not supported by recent molecular studies [16, 17]. To help understand genetic bases for morphological traits of Pieridae and to clarify its phylogenetic placement, we sequenced the first complete genome from this family. We chose a large and showy American species, the Cloudless Sulphur (*Phoebis sennae*), which is similar in size and color to the European Brimstone butterfly.

The Cloudless Sulphur is a large yellow butterfly distributed from the southern regions of the United States through the Neotropics. Its caterpillars feed on Senna plants and close relatives from the Pea family (Fabaceae). Adults are highly vagile but do not survive cold winters. Eastern USA populations are known as subspecies *Phoebis sennae eubule*, and southwestern populations that range throughout Central and most of South America are attributed to subspecies *Phoebis sennae marcellina* [18]. Both subspecies are present in Texas. The two subspecies are morphologically distinct, with *P. s. eubule* being typically less patterned on the underside of the wings and *P. s. marcellina* females characterized by pronounced dark spots along the margin of hindwings above (Fig. 1). In addition, their caterpillars show somewhat different foodplant preferences. *P. s. eubule* mostly feeds on partridge pea (*Chamaecrista fasciculata*), while *P. s. marcellina* prefers Senna species. However, their COI mitochondrial DNA sequences show small divergence, no more than 0.6% [19]. The divergence in nuclear genes, that likely cause the morphological differences, has remained unclear.

We obtained a complete reference genome of *P. s. eubule* from a single male collected in southeast Texas. To compare genetic divergence between the North American *Phoebis sennae* subspecies, we sequenced genomes of two more *P. s. eubule* specimens (from north Texas and Oklahoma) and of three *P. s. marcellina* specimens from south Texas. In contrast to mitochondrial DNA, their nuclear genomes revealed unexpectedly large divergence (nearly 2%), larger than that between the two sister species of Tiger Swallowtails (*Pterourus canadensis* and *Pterourus glaucus*), suggesting that the two subspecies of *Phoebis sennae* are better treated as species-level taxa.

RESULTS AND DISCUSSION

Genome quality assessment and gene annotation of the reference genome

We assembled a 406 Mb reference genome of *Phoebis sennae* (*Pse*) and compared its quality and composition (Table 1) with genomes of the following Lepidoptera species: *Plutella xylostella* (Pxy), *Bombyx mori* (*Bmo*), *Manduca sexta* (*Mse*), *Lerema accius* (*Lac*), *Pterourus glaucus* (*Pgl*), *Papilio polytes* (*Ppo*), *Papilio xuthus* (*Pxu*), *Melitaea cinxia* (*Mci*), *Heliconius melpomene* (*Hme*), and *Danaus plexippus* (*Dpl*) [1-11]. The scaffold N50 of *Pse* genome assembly is 257 kb. The genome assembly is better than many other Lepidoptera genomes in terms of completeness measured by the presence of Core Eukaryotic Genes Mapping Approach (CEGMA) genes [20], cytoplasmic ribosomal proteins and independently assembled transcripts. The average coverage (87.4%) of CEGMA genes by single *Pse* scaffolds is comparable to the coverage by the current *Bmo* assembly with an N50 of about 4.0 Mb, indicating that the quality of the *Pse* draft is sufficient for protein annotation and comparative

analysis. The genome sequence has been deposited at DDBJ/EMBL/GenBank under the accession XXX. The version described in this paper is version YYY. In addition, the main results from genome assembly, annotation and analysis can be downloaded at http://prodata.swmed.edu/LepDB/.

We assembled the transcriptome of *Phoebis sennae* from the same specimen. Based on the transcriptome, homologs from other Lepidoptera and *Drosophila melanogaster*, *de novo* gene predictions, and repeat identification, we predicted 16,493 protein-coding genes in the *Phoebis sennae* genome. 67% of these genes are likely expressed in the adult, as they fully or partially overlap with the transcripts. We annotated the putative functions for 12,584 protein-coding genes.

Phylogeny of Lepidoptera

We identified orthologous proteins encoded by 11 Lepidoptera genomes (*Plutella xylostella*, *Bombyx mori*, *Manduca sexta*, *Lerema accius*, *Pterourus glaucus*, *Papilio polytes*, *Papilio xuthus*, *Melitaea cinxia*, *Heliconius melpomene*, *Danaus plexippus*, and *Phoebis sennae*) and detected 5143 universal orthologous groups, from which 2106 consist of a single-copy gene in each of the species. A phylogenetic tree built on the concatenated alignment of the single-copy orthologous groups using RAxML placed *Pheobis* as the sister to Nymphalidae clade. This placement is consistent with the previously published results based on molecular data [11, 17], as expected in the absence of genomes from the families Lycaenidae and Riodinidae.

In addition, our analysis placed Papilionidae as a sister to all other butterflies, including skippers (Hesperiidae). Such placement contradicts the traditional view based on

morphological studies, but is indeed reproduced in all maximum-likelihood and Bayesian trees published recently [11, 21]. All nodes received 100% bootstrap support when the alignment of all the single-copy orthologous groups was used. To find the weakest nodes we reduced the amount of data by randomly splitting the concatenated alignment into 100 alignments (about 3670 positions in each alignment). The consensus tree based on these alignments revealed that the node referring to the relative position of skippers and swallowtails has much lower support (68%) compared to all other nodes (above 90%). Thus, the placement of swallowtails and skippers within Lepidoptera tree remains to be investigated further when better taxon sampling by complete genomes is achieved.

Six genomes of *Phoebis sennae*

In addition to the reference genome of *P. s. eubule* from southeast Texas, we sequenced the genomes of five *Phoebis sennae* specimens and mapped the reads to the reference. Two specimens were *P. s. eubule* from north-central Texas and southern Oklahoma and three were *P. s. marcellina* from south Texas. The coverage by the reads and the completeness of these genomes are summarized in Table 2. The sequencing reads for all the specimens are expected to cover the genome 10-12 times, and about 97% of coding regions in the reference genome can be mapped by reads from each specimen. However, fractions of the noncoding region that can be mapped differ significantly ($p < 0.001$) between specimens. Reads from specimens of the same subspecies as the reference genome can map to 88% of the positions in the reference genome while reads from the specimens of a different subspecies can map to only 83% of the

positions. This indicates a higher divergence in the non-coding region and a substantial difference between the two subspecies in the non-coding region.

We identified SNPs in these genomes compared to the reference genome using Genome Analysis Toolkit (GATK) [22]. There are 1.2% heterozygous positions in the reference genome, and the heterozygosity levels (about 1.4%) for two other *P. s. eubule* specimens are comparable to the reference genome. The southwestern population shows a higher heterozygosity level (about 2.2%), which agrees with the expected larger population size of *P. s. marcellina*. In all the genomes, the percentage of SNP in the coding regions (0.91% ~ 1.00% for *P. s. eubule* and 1.45% ~ 1.56% for *P. s. marcellina*) is lower than that for the non-coding regions (1.23% ~ 1.46% for *P. s. eubule* and 2.21% ~ 2.32% for *P. s. marcellina*), which is likely due to the potential deleterious effect of SNPs in the coding regions.

We clustered all 6 specimens based on their genotype in positions with two possible nucleotides. The three *P. s. eubule* specimens formed a tight cluster, indicating high similarity between them. The three *P. s. marcellina* specimens were more divergent, but they still clustered closer to each other than to the *P. s. eubule* specimens. In addition, analysis of the same data using fastStructure [23] also confirmed this population structure by likelihood calculation: the three *P. s eubule* specimens represent one population while the three *P. s. marcellina* specimens are from another population.

Incongruence between the divergence in nuclear and mitochondrial genes

COI mitochondrial DNA barcode sequences have been determined for a number of *Phoebis sennae* specimens across its distribution range [19], and they show very little divergence

between subspecies. The eastern subspecies in United States, *P. s. eubule* and the southwestern subspecies, *P. s. marcellina* differ by only 0.6% (4 positions) in their barcode sequences. Barcode differences of 2% and above likely correspond to species-level divergence [24]. For example, tiger swallowtails *Pterourus glaucus* and *Pterourus canadensis* differ by 2.2% in their barcode sequences. To understand the reasons for apparent morphological and life history differences in the absence of substantial barcode divergence, we compared the nuclear and mitochondrial genomes of all 6 *Phoebis sennae* specimens and correlated the results with the complete transcriptome data for *Pterourus canadensis* and *Pterourus glaucus*.

P. s. eubule and *P. s. marcellina* show low divergence (about 0.5%) not only in the COI barcode, but also for all the mitochondrial genes. The mitochondrial genes are very conserved (divergence 0.02% ~ 0.11%) within each subspecies, and thus the phylogenetic tree based on them clearly separates the two subspecies into clades with branch length between them indicating 0.42% difference (Fig. 4c). In contrast, nuclear genes show much higher divergence both within (1.17% for *P. s. eubule*, 1.78% for *P. s. marcellina*) and between (1.86%) subspecies. In the phylogenetic tree based on nuclear genes (16,137 genes, 18,877,324 base pairs), the branch length between the two subspecies (branches colored in green and orange in Fig 4a) is 0.7%, twice of that for mitochondrial genes.

The higher divergence in nuclear genes compared to mitochondrial genes is unexpected. Mitochondrial DNA usually evolves faster than the nuclear DNA, and thus it is frequently used to resolve relationships of closely-related taxa [25]. Indeed, the divergence level in mitochondrial DNA (about 2.0%) between two *Pterourus* species is twice that seen in nuclear DNA (about 1.0%). Both nuclear genes (9,622 transcripts, 13,525,930 base pairs) and

mitochondrial genes clearly separate the two species in phylogenetic trees, but the internal branch length between the two taxa in the tree based on nuclear DNA (0.18%) is about 10 times smaller than that for mitochondrial DNA (1.8%). The clear incongruence between divergence in nuclear and mitochondrial DNA in *Phoebis* and *Pterourus* reiterates the need for inclusion of nuclear DNA in phylogenetic studies. Based on the divergence in the nuclear genes, along with the morphological differences, *P. s. eubule* and *P. s. marcellina* may be better treated as two species-level taxa.

We speculate that high nuclear divergence in *Phoebis* is related to its fast development. While *Pterourus canadensis* breeds only once each year, *P. s. marcellina* can have up to 15 generations per year. Low divergence in mitochondrial DNA of *Phoebis* remains a mystery. It might be due to more accurate error-correction machinery during the replication of mitochondrial DNA, keeping the mutation rate very low. Alternatively, a more mundane view is that introgression, population bottlenecks and mitochondria selective sweeps [26-29] might result in transfer of mitochondria between taxa or spread of a certain mitochondrial haplotype across all *P. sennae* populations throughout their vast distribution range.

Interestingly, southern taxa of both *Phoebis* and *Pterourus* display larger internal divergence than northern taxa (Fig. 5). The difference between three specimens of *P. s. marcellina* (1.80%) collected from the same locality is larger than that between three *P. s. eubule* specimens (1.12%) collected from different localities that are separated by several hundred miles. The lower sequence variation of *P. s. eubule* specimens suggests smaller population size and possible bottlenecks. Such bottlenecks for northern populations are more

likely because *Phoebis* has low tolerance to subzero temperatures and most individuals do not survive cold winters.

Molecular processes differentiating *P. s. eubule* and *P. s. marcellina*

Phoebis s. eubule and *P. s. marcellina* can be clearly distinguished based on the whole-genome data. The average inter-taxa divergence for protein coding genes is significantly (p = 5.8e-58) higher than the intra-taxa divergence (Fig. 5a,b). However, the two taxa are not diverged in most individual genes, and only 20% of genes can confidently (bootstrap >= 75%) distinguish them (Fig. 5e). The situation is very similar to that of *Pterourus glaucus* and *Pterourus canadensis* (Fig. 5c-e).

To further investigate the possible phenotypic consequences caused by genetic divergence between the two *Phoebis* taxa, we focused on the genes that can clearly distinguish them both by their sequences and by the proteins they encode (i.e., separate the two taxa into clades with bootstrap support no less than 75%). We identified 924 (5.7%) such proteins, but they were significantly enriched (p = 4.6e-24) in non-conserved proteins within each taxon. Out of 710 such proteins, 314 are enzymes. The functional sites of enzymes are constrained to several catalytically important residues, and therefore the rest of their sequence is likely to be more tolerant to mutations and can undergo faster divergence.

In contrast, the remaining 214 proteins are conserved within each taxon, but can clearly distinguish the two taxa. We term these divergence hotspots. The presence of such proteins could cause Dobzhansky-Muller hybrid incompatibility between the two taxa, as the proteins from *P. s. eubule* may not work well with proteins and genetic materials from *P. s. marcellina*

when functioning in the same pathway. GO-term analysis of these divergence hotspots revealed a prevalence of epigenetic mechanisms including histone modification enzymes and chromatin organization (Table 3). Variations in epigenetics-related proteins might be an easy source of hybrid incompatibility because these proteins directly interact with the genetic materials, especially the non-coding regions that could evolve rapidly [30]. Epigenetic variation has been shown to be a speciation mechanism in several organisms [31, 32]. Among the genomic regions covered in the mapping results of all 6 specimens, the non-coding region differ by 3.5% between the two taxa while the coding region differ by only 1.8%. The actual divergence in the non-coding region should be even larger as the most divergent regions would fail to map to the reference genome (discussed above). Therefore, proteins involved in epigenetic mechanisms from one taxon may not be compatible with the binding sites in the DNA of another taxon, resulting in lower fitness of the hybrids.

Another group of GO terms that are significantly enriched are related to the circadian sleep/wake cycle. The divergence hotspots for the two *Pterourus* species are also enriched in circadian clock related proteins, and in particular, those related to eclosion rhythm. This is consistent with their observed phenotypic divergence in pupal diapause (i.e., the timing of eclosion). The two *Phoebis sennae* taxa mostly show divergence in the sleep/wake cycle, but not the eclosion rhythm. This might be related to the lack of pupal diapause in *Phoebis sennae*. However, proteins related to the sleep/wake cycle could have diverged adaptively since the two taxa were partly separated into different latitudes with different levels of sunlight and average temperatures. In addition, proteins associated with early development and cell differentiation are also enriched in the divergence hotspots. Divergence in these proteins may

have a profound impact on the morphology and biology of an organism, driving speciation and adaptation.

Nuclear DNA markers to identify *P. s. eubule* and *P. s. marcellina*

Eleven out of 13 mitochondrial protein-coding genes can clearly separate *P. s. eubule* and *P. s. marcellina* as the maximal intra-taxa divergence is smaller than the minimal inter-taxa divergence. The only two exceptions are the ND4L and ATP8 coding genes, which are identical between the two subspecies. The low divergence in the mitochondrial genes within one taxon could be a result of going through narrower bottlenecks when the population size goes down due to their maternal inheritance, and strong selection pressure to function together with the nuclear-encoded proteins and maintain the high efficiency of the mitochondrial electron transport chain.

However, the two taxa cannot be clearly identified using the nuclear markers (Fig. 5) previously selected for phylogenetic studies of butterflies. This situation is very similar to that of *Pterourus glaucus* and *Pterourus canadensis*. Out of the 16,137 well-covered nuclear genes, only 92 always show higher divergence between *P. s. eubule* and *P. s. marcellina* than within either taxon. Eleven of them are associated with GO terms that are enriched in the divergence hotspots. These likely participate in the biological processes that have diverged between the two taxa and we suggest them to be possible nuclear markers (Table 4) to identify the two taxa. For example, two of them are related to chromatin remodeling, and they are orthologous to the *Drosophila* genes Grunge (CG6964) and Nucleoplasmin (CG7917), respectively. Both proteins directly interact with the chromatin and could contribute to a certain level of

reproductive isolation as they may not interact well with the genetic material of a different taxon.

Should *P. s. eubule* and *P. s. marcellina* be treated as species-level taxa?

Comparative analysis of complete genomes of six *Phoebis sennae* specimens revealed an unexpectedly large divergence between subspecies *P. s. eubule* and *P. s. marcellina* in nuclear genes compared to that of mitochondrial genes. This divergence appears more prominent than that between the two swallowtails species *Pterourus canadensis* and *P. glaucus*. The two *Phoebis* subspecies show significant divergence in epigenetic mechanisms, regulation of the sleep/wake cycle and early development. Multiple proteins participating in each of these processes show clear divergence between the two taxa. It is possible that protein from one taxon may show reduced compatibility with a partner from another taxon, leading to Dobzhansky-Muller hybrid incompatibility.

In addition, both *Phoebis* subspecies occur in Texas and their ranges partly overlap in central Texas around Austin and San Antonio, where specimens of both subspecies can be found, and *P. s. marcellina* can stray north into Oklahoma. However, in areas of sympatry they remain morphologically distinct. It is apparent that these butterflies are strong flyers and are known to migrate. A single individual can fly a hundred miles or more, so there should be ample opportunities for the two taxa to mix. Nevertheless, they remain morphologically and genetically distinct, which indicates a certain level of reproductive isolation and thus possible genetic incompatibilities. Taken together, the profound genomic divergence, morphological differences and maintenance of distinctness between eastern and southern populations in Texas,

we suggest that it is more meaningful to treat both *P. s. eubule* and *P. s. marcellina* as species-level taxa. However, the relationship of each to nominotypical *Phoebis sennae sennae* from the Caribbean Islands remains to be elucidated.

CONCLUSIONS

We report six genomes of the Cloudless Sulphur, three of *P. s. eubule* and three of *P. s. marcellina*. Being the first sequenced genomes from the family Pieridae, they offer a rich dataset for comparative genomics and phylogenetic studies of Lepidoptera. Comparative analyses of *Phoebis* genomes and *Pterourus* transcriptomes reveal a remarkable incongruence between relative rates of nuclear and mitochondrial divergence. *Phoebis* species show low mitochondrial divergence (0.5%) and high nuclear divergence (1.8%). The situation is reversed in *Pterourus* species. *P. s. marcellina* and *P. s. eubule* differ from each other in histone methylation regulators, chromatin associated proteins, circadian clock, and some early developmental proteins. The divergence in these processes, taken together with the unexpectedly high divergence in nuclear genes, suggests a certain level of reproductive isolation between the two taxa, and both *P. s. eubule* and *P. s. marcellina* are best treated as species-level taxa.

MATERIALS AND METHODS

Library preparation and sequencing

We removed and preserved the wings and genitalia of six freshly caught *Phoebis sennae* specimens (three *P. s. eubule*: NVG-3314, male, Texas: San Jacinto Co., Sam Houston

National Forest, 30.50596, -95.08868, 12-Apr-2015; NVG-4452, female, Texas: Wise Co., LBJ National Grassland, 33.38401, -97.57381, 9-Aug-2015; NVG-4541, male, Oklahoma: Atoka Co., McGee Creek Recreation Area, 34.41040, -95.91059, 22-Aug-2015; and three *P. s. marcellina*: Hidalgo Co., 1.5 air mi southeast of Relampago, 26.07093, -97.89131: NVG-3356, male, 23-May-2015; NVG-3377, female, 24-May-2015; NVG-3393, male, 30-May-2015), and the rest of the bodies were stored in *RNAlater* solution.

We used specimen NVG-3314 for the reference genome. We extracted approximately 20 µg genomic DNA from about 4/5 of the specimen NVG-3314 with the ChargeSwitch gDNA mini tissue kit. 250 bp and 500 bp paired-end libraries were prepared using enzymes from NEBNext Modules and following the Illumina TruSeq DNA sample preparation guide. 2 kb, 6 kb and 15 kb mate pair libraries were prepared using a protocol similar to previously published Cre-Lox-based method [33]. For the 250 bp, 500 bp, 2 kbp, 6 kbp and 15 kbp libraries, approximately 500 ng, 500 ng, 1.5 µg, 3 µg and 6 µg of DNA were used, respectively. We quantified the amount of DNA from all the libraries with the KAPA Library Quantification Kit, and mixed 250 bp, 500 bp, 2 kbp, 6 kbp, 15 kbp libraries at relative molar concentration 40:20:8:4:3. The mixed library was sent to the genomics core facility at UT Southwestern Medical Center to sequence 150 bp at both ends (PE150) using one lane in Illumina HiSeq2500.

The remaining 1/5 of specimen NVG-3314 was used to extract RNA using QIAGEN RNeasy Mini Kit. We further isolated mRNA using NEBNext Poly(A) mRNA Magnetic Isolation Module and RNA-seq libraries for both specimens were prepared with NEBNext Ultra Directional RNA Library Prep Kit for Illumina following manufactory's protocol. The RNA-seq library was sequenced for 150 bp from both ends using 1/8 of an Illumina lane.

The other five specimens were used to prepare paired-end libraries to map to the reference genome. For each of them, we extracted about 5 μg genomic DNA and used about 500 ng genomic DNA to prepare a 400 bp paired-end library. These paired-end libraries were mixed at equal ratio and sequenced using similar strategy (PE150) using half of an Illumina lane. The sequencing reads for all the specimens have been deposited in NCBI SRA database under accession numbers: XXX.

Genome and transcriptome assembly

We removed sequence reads that did not pass the purity filter and classified the pass-filter reads according to their TruSeq adapter indices to get individual sequencing libraries. Mate pair libraries were processed by the Delox script [33] to remove the loxP sequences and to separate true mate pair from paired-end reads. All reads were processed by mirabait [34] to remove contamination from the TruSeq adapters, an in-house script to remove low quality portions (quality scale < 20) at both ends, JELLYFISH [35] to obtain k-mer frequencies in all the libraries, and QUAKE [36] to correct sequencing errors. The data processing resulted in nine libraries that were supplied to Platanus [37] for genome assembly: 250 bp and 500 bp paired-end libraries, three paired-end and three mate pair libraries from 2 kb, 6 kb and 15 kb libraries and a single-end library containing all reads whose pairs were removed in the process.

We mapped these reads to the initial assembly with Bowtie2 [38] and calculated the coverage of each scaffold with the help of SAMtools [39]. Many short scaffolds in the assembly showed coverage that was about half of the expected value; they likely came from highly heterozygous regions that were not merged to the equivalent segments in the

homologous chromosomes. We merged them into other scaffolds if they could be fully aligned to another significantly less covered region (coverage > 90% and uncovered region < 500 bp) in a longer scaffold with high sequence identity (>95%). Similar problems occurred in the *Heliconius melpomene*, *Pterourus glaucus* and *Lerema accius* genome projects, and similar strategies were used to improve the assemblies [3, 8, 11].

The RNA-seq reads were processed using a similar procedure as the genomic DNA reads to remove contamination from TruSeq adapters and the low quality portion of the reads. Afterwards, we applied three methods to assemble the transcriptomes: (1) *de novo* assembly by Trinity [40], (2) reference-based assembly by TopHat [41] (v2.0.10) and Cufflinks [42] (v2.2.1), and (3) reference-guided assembly by Trinity. The results from all three methods were then integrated by Program to Assemble Spliced Alignment (PASA) [43].

Identification of repeats and gene annotation

Two approaches were used to identify repeats in the genome: the RepeatModeler [44] pipeline and in-house scripts that extracted regions with coverage 4 times higher than expected. These repeats were submitted to the CENSOR [45] server to assign them to the repeat classification hierarchy. The species-specific repeat library and repeats classified in RepBase [46] (V18.12) were used to mask repeats in the genome by RepeatMasker [47].

We obtained two sets of transcript-based annotations from two pipelines: TopHat followed by Cufflinks and Trinity followed by PASA. In addition, we obtained five sets of homology-based annotations by aligning protein sets from *Drosophila melanogaster* [48] and four published Lepidoptera genomes (*Plutella xylostella*, *Bombyx mori*, *Heliconius melpomene*, and *Danaus*

plexippus) to the *Phoebis sennae* genome with exonerate [49]. Proteins from Invertebrate in the entire UniRef90 [50] database were used to generate another set of gene predictions by genblastG [51]. We manually curated and selected 1152 confident gene models by integrating the evidence from transcripts and homologs to train *de novo* gene predictors: AUGUSTUS [52], SNAP [53] and GlimmerHMM [54]. These trained predictors, the self-trained Genemark [55] and a consensus based pipeline, Maker [56] were used to generate another five sets of gene models. Transcript-based and homology-based annotations were supplied to AUGUSTUS, SNAP and Maker to boost their performance. In total, we generated 13 sets of gene predictions and integrated them with EvidenceModeller [43] to generate the final gene models.

We predicted the function of *Phoebis sennae* proteins by transferring annotations and GO-terms from the closest BLAST [57] hits (E-value < 10^{-5}) in both the Swissprot [58] database and Flybase [59]. Finally, we performed InterproScan [60] to identify conserved protein domains and functional motifs, to predict coiled coils, transmembrane helices and signal peptides, to detect homologous 3D structures, to assign proteins to protein families and to map them to metabolic pathways.

Identification of orthologous proteins and phylogenetic tree reconstruction

We identified the orthologous groups from all 11 Lepidoptera genomes using OrthoMCL [61]. 2106 orthologous groups consisted of single-copy genes from each species, and they were used for phylogenetic analysis. An alignment was built for each universal single-copy orthologous group using both global sequence aligner MAFFT [62] and local sequence aligner BLASTP.

Positions that were consistently aligned by both aligners were extracted from each individual alignment and concatenated to obtain an alignment containing 362,743 positions. The concatenated alignment was used to obtain a phylogenetic tree using RAxML [63]. Bootstrap was performed to assign the confidence level of each node in the tree. In addition, in order to detect the weakest nodes in the tree, we reduced the amount of data by randomly splitting the concatenated alignment into 100 alignments (about 3630 positions in each alignment) and applied RAxML to each alignment. We obtained a 50% majority rule consensus tree and assigned confidence level to each node based on the percent of individual trees supporting this node.

Assembly and annotation of mitochondrial genomes

The mitogenomes of several closely related species, including *Catopsilia pomona* [64], *Colias erate* [65], and *Gonepteryx mahaguru* [66] were used as reference. Based on these mitogenomes, we applied mitochondrial baiting and iterative mapping (MITObim) v1.6 [67] software to extract the sequencing reads of the mitogenome in the paired-end libraries for specimen NVG-3314. About 4.3 million reads for the mitogenome were extracted, and they were expected to cover the mitogenome 40 thousand times. We used JELLYFISH to obtain the frequencies of 15-mers in these reads, and applied QUAKE to correct errors in 15-mers with frequencies lower than 1,000 and excluded reads containing low-frequency 15-mers after error correction. We used the error-corrected reads to assemble into contigs *de novo* with Platanus. We manually selected the contig corresponding to the mitogenome (it is the longest one with highest coverage), and manually extended its sequence based on the sequencing reads

to obtain a complete circular DNA. In addition, by aligning the protein coding sequences from the mitogenomes of closely related species mentioned above to the *Phoebis sennae* mitogenome, we annotated the 13 protein coding genes.

Obtaining the genomes of six *Phoebis sennae* specimens and phylogenetic analysis

We mapped the sequencing reads of all 6 *Phoebis sennae* specimens to the reference genome using BWA [68] and detected SNPs using the Genome Analysis Toolkit (GATK) [69]. We deduced the genomic sequences for each specimen based on the result of GATK. We used two sequences to represent the paternal and maternal DNA in each specimen. For heterozygous positions, each possible nucleotide was randomly assigned to either paternal or maternal DNA. Based on the gene annotation of the reference genome, we further deduced the protein-coding sequences of each gene in each specimen.

In study the population structure, we selected bi-allelic loci (two nucleotide types in a position of alignment covering all 6 specimens, coding and non-coding regions). First, we encoded each specimen by a vector consisting of the frequency of a certain nucleotide in each position. For example, if a position is occupied by A and T in all 6 specimens, then their possible genotypes AA, AT and TT were represented as 0, 0.5 and 1, respectively. We calculated the covariance between each pair of specimens and obtained a covariance matrix. We performed singular value decomposition on the covariance matrix and visualized the clustering of the 6 specimens in two dimensional space defined by the first two singular vectors. Second, we applied fastStructure software [23] to analyze the same SNP genotype data. We

tested all the possible number of model components (from 1 to 6) and selected the population structure with the maximal likelihood.

In order to quantify the divergence between the two *Phoebis sennae* subspecies, we compared their divergence level in the protein-coding regions to that for a pair of sister species, *Pterourus glaucus* and *Pterourus canadensis*. The transcripts of *Pterourus* specimens were mapped to the *Pterourus glaucus* reference genome using methods described before [8]. From alignments of *Pterourus* transcripts to the reference genome, we selected 9,622 nuclear genes for which there are at least 60 aligned positions from at least two *Pt. canadensis* and two *Pt. glaucus* specimens. Similarly, we selected 16,137 nuclear genes of *Phoebis sennae*, requiring the selected genes to have at least 50% coverage for the coding regions in two *P. s. marcellina* and two *P. s. eubule* specimens. We extracted the coding regions in the alignments of individual nuclear genes and concatenated them for both *Pterourus* and *Phoebis*. The concatenated alignment was used to build both a neighbor-joining tree with PHYLIP [70] based on the percentage of different positions between specimens and a maximal-likelihood tree with RAxML (model: GTRGAMMA). Bootstrap resampling was performed to assign confidence levels for nodes in the maximal-likelihood tree.

Identification of divergence hotspots and selection of nuclear barcodes

We defined "divergence hotspots" as genes that satisfied the following two criteria: (1) can confidently (bootstrap > 75) separate *P. s. eubule* and *P. s. marcellina* specimens into clades in phylogenetic trees by both the DNA sequence and the protein sequence encoded by them; (2) the divergence within both *P. s. eubule* and *P. s. marcellina* specimens is lower than the

median divergence level over all the genes. We identified the enriched GO terms associated with these "divergence hotspots" using binomial tests (m = the number of "divergence hotspots" that were associated with this GO term, N = number of "speciation hotspots", p = the probability for this GO term to be associated with any gene). GO terms with P-values lower than 0.01 were considered enriched. We further identified genes that are always more divergent between taxa than within taxa. These genes could be used as nuclear markers to distinguish *P. s. eubule* and *P. s. marcellina*.

Table 1
Quality and Composition of Lepidoptera Genomes

Feature	*Pgl*	*Ppo*	*Pxu*	*Dpl*	*Hme*	*Mci*	*Lac*	*Bmo*	*Mse*	*Pxy*	*Pse*
Size w/o gap (Mb)	361	218	238	242	270	361	290	432	400	387	347
GC content (%)	35.4	34.0	33.8	31.6	32.8	32.6	34.4	37.7	35.3	38.3	39.0
Repeat (%)	22.2	NA	NA	16.3	24.9	28.0	15.5	44.1	24.9	34.0	17.2
Exon (%)	5.11	7.79	8.59	8.41	6.19	4.34	7.24	4.07	5.34	6.47	6.20
Intron (%)	24.8	51.6	45.5	26.6	24.1	31.2	32.3	16.1	38.3	31.3	25.5
Genome size (Mb)	375	227	244	249	274	390	298	481	419	394	406
Heterozygosity (%)	2.3	NA	NA	0.55	NA	NA	1.5	NA	NA	~2[a]	1.2
Scaffold N50 (kb)	231	3,672	6,199	207 (716)	194	119	525	27(3,999)	664	734	257
CEGMA (%)	99.6	99.3	99.6	99.6	98.2	98.9	99.6	99.6	99.8	98.7	99.3
CEGMA coverage by single scaffold (%)	86.9	85.8	88.8	87.4	86.5	79.2	86.8	86.8	86.4	84.1	87.4
Ribosomal Proteins (%)	98.9	98.9	97.8	98.9	94.6	94.6	98.9	98.9	98.9	93.5	98.9
De novo assembled transcripts (%)	98	NA	NA	96	NA	97	97~99	98	NA	83	97
number of proteins (k)	15.7	12.3	13.1	15.1	12.8	16.7	17.4	14.3	15.6	18.1	16.5

NOTE.—NA, data not available.
[a]Estimated from k-mer frequency histogram.

Table 2
Quality of *Phoebis* Genomes

Specimen (NVG-)	3314	4452	4541	3356	3377	3393
Coverage	103	11.8	10.4	12.5	12	10.8
% Mapped position noncoding region	99.99	87.33	87.20	82.41	82.42	81.60
% Mapped position coding region (exon)	99.98	96.78	96.47	97.41	97.33	96.92
100% covered genes	16,493	13,080	13,161	12,846	12,577	12,189
90% covered genes	16,493	14,824	14,735	14,917	14,850	14,625
50% covered genes	16,493	15,987	15,897	16,146	16,158	16,067
Heterozygosity	1.23%	1.46%	1.38%	2.32%	2.23%	2.21%
Heterozygosity in coding region (exon)	0.91%	1.00%	0.96%	1.56%	1.48%	1.45%
Heterozygosity in noncoding region	1.25%	1.49%	1.41%	2.38%	2.28%	2.28%

Table 3
Enriched GO Terms for the Divergent Hotspots that are Conserved within Each Subspecies

GO Term	*P*	Category	Annotation of the GO Term	Summary
GO:0051574	7.0E-05	BP	Positive regulation of histone H3-K9 methylation	Histone methylation and chromatin associated proteins
GO:0051570	2.8E-04	BP	Regulation of histone H3-K9 methylation	
GO:1900112	1.0E-03	BP	Regulation of histone H3-K9 trimethylation	
GO:1900114	1.0E-03	BP	Positive regulation of histone H3-K9 trimethylation	
GO:0031062	2.0E-03	BP	Positive regulation of histone methylation	
GO:0051571	8.7E-03	BP	Positive regulation of histone H3-K4 methylation	
GO:0006325	2.6E-03	BP	Chromatin organization	
GO:0042393	8.3E-03	MF	Histone binding	
GO:0000791	8.6E-03	CC	Euchromatin	
GO:0031519	3.0E-03	CC	PcG protein complex	
GO:0044666	3.8E-03	CC	MLL3/4 complex	
GO:0035097	2.5E-03	CC	Histone methyltransferase complex	
GO:0034708	3.5E-03	CC	Methyltransferase complex	
GO:0008607	5.4E-03	MF	phosphorylase kinase regulator activity	
GO:2000044	5.4E-03	BP	Negative regulation of cardiac cell fate specification	Early development and cell fate specification
GO:2000043	8.7E-03	BP	Regulation of cardiac cell fate specification	
GO:0045611	6.9E-03	BP	Negative regulation of hemocyte differentiation	
GO:0009997	5.4E-03	BP	Negative regulation of cardioblast cell fate specification	
GO:0042686	8.7E-03	BP	Regulation of cardioblast cell fate specification	
GO:0061351	8.7E-03	BP	Neural precursor cell proliferation	
GO:0045177	3.8E-03	CC	Apical part of cell	
GO:0008158	8.7E-03	MF	Hedgehog receptor activity	
GO:0090102	5.4E-03	BP	Cochlea development	
GO:0042745	7.2E-03	BP	Circadian sleep/wake cycle	Circadian clock
GO:0022410	6.3E-03	BP	Circadian sleep/wake cycle process	
GO:0050802	6.9E-03	BP	Circadian sleep/wake cycle, sleep	
GO:0016469	9.5E-03	CC	Proton-transporting two-sector ATPase complex	Transporter
GO:0015399	5.9E-03	MF	Primary active transmembrane transporter activity	
GO:0015405	5.9E-03	MF	P-P-bond-hydrolysis-driven transmembrane transporter	
GO:0015078	6.3E-03	MF	Hydrogen ion transmembrane transporter activity	
GO:0044283	5.9E-03	BP	Small molecule biosynthetic process	Metabolic
GO:0016053	6.1E-03	BP	Organic acid biosynthetic process	
GO:0046394	6.1E-03	BP	Carboxylic acid biosynthetic process	
GO:0008206	8.7E-03	BP	Bile acid metabolic process	
GO:0009314	6.8E-03	BP	Response to radiation	Adaptation to sunlight
GO:0034644	2.0E-03	BP	Cellular response to UV	
GO:0019233	6.9E-03	BP	Sensory perception of pain	Other
GO:0005606	8.7E-03	CC	Laminin-1 complex	
GO:0043256	8.7E-03	CC	Laminin complex	
GO:0009881	7.0E-03	MF	Photoreceptor activity	

BP: biological process; CC: cellular components; MF: molecular function.

Table 4. Information about the selected nuclear markers

Flybase ID	*Phoebis sennae* ID	Function
CG3731	pse1226.10	Mitochondrial-processing peptidase subunit beta
CG9138	pse132.8	Regulator of tracheal tube development
CG10731	pse1425.14	ATP synthase subunit s, mitochondrial
CG43388	pse35.12	Voltage-gated potassium channel
CG7917	pse730.7	Nucleoplasmin
CG6964	pse9575.4	Transcriptional repressor
CG31548	pse1095.3	3-oxoacyl-[acyl-carrier-protein] reductase FabG
CG2488	pse1216.5	Cryptochrome-1
CG7675	pse1218.21	Retinol dehydrogenase 11
CG5722	pse243.10	Niemann-Pick C1 protein
CG1753	pse42.13	Bifunctional L-3-cyanoalanine synthase

Figure 1. Specimens of *Phoebis*.

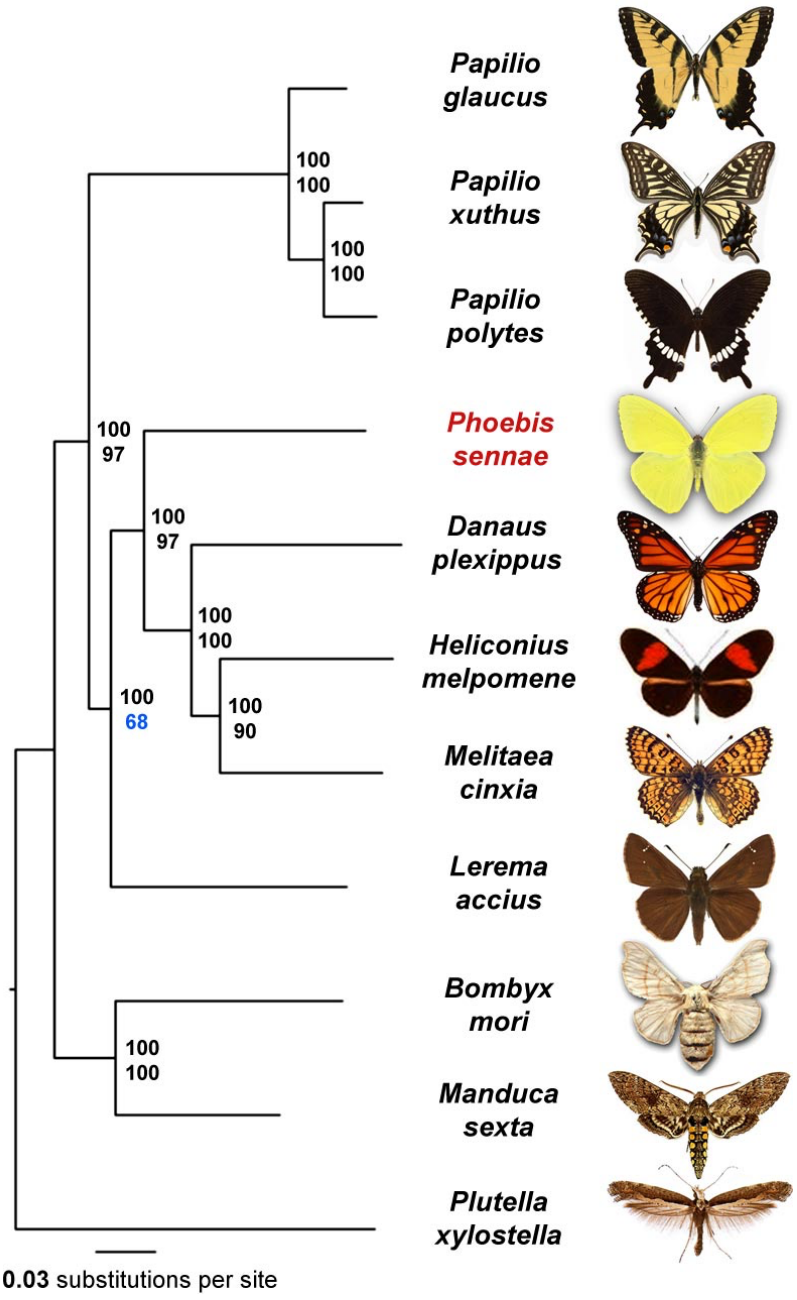


Figure 2. Phylogenetic tree of the Lepidoptera species with complete genome sequences. Majority-rule consensus tree of the maximal likelihood trees constructed by RAxML on the concatenated alignment of universal single-copy orthologous proteins.

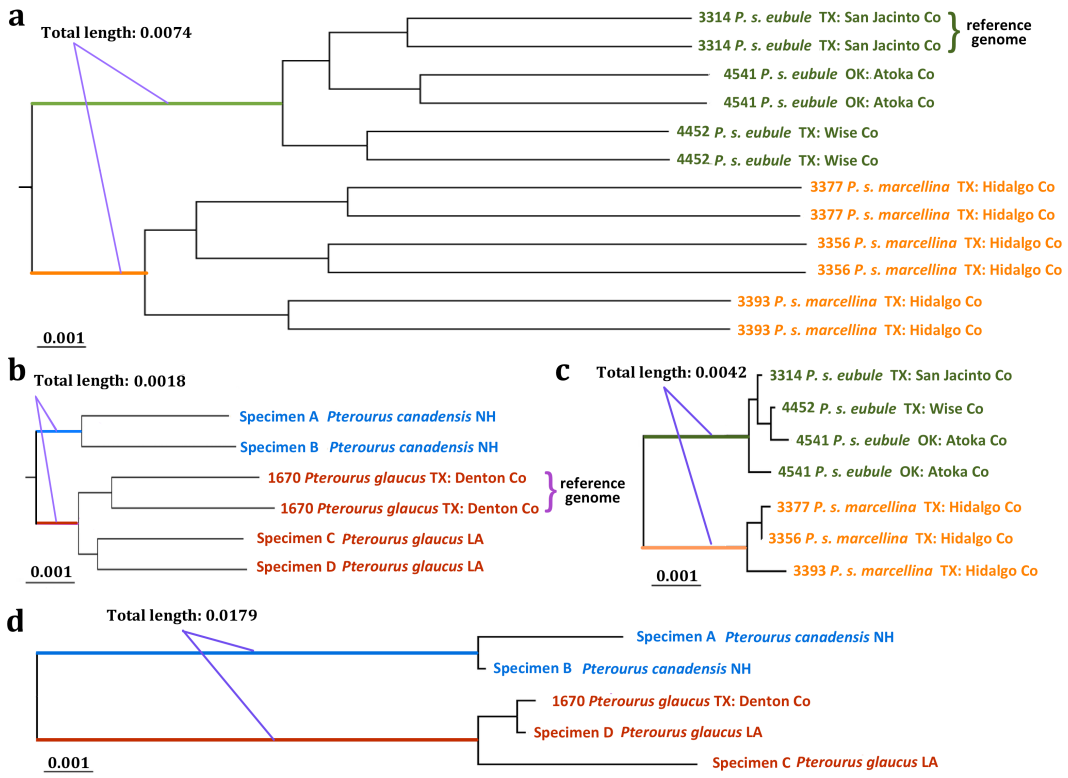


Figure 3. Incongruence between the speed of evolution for mitochondrial and genomic DNA.

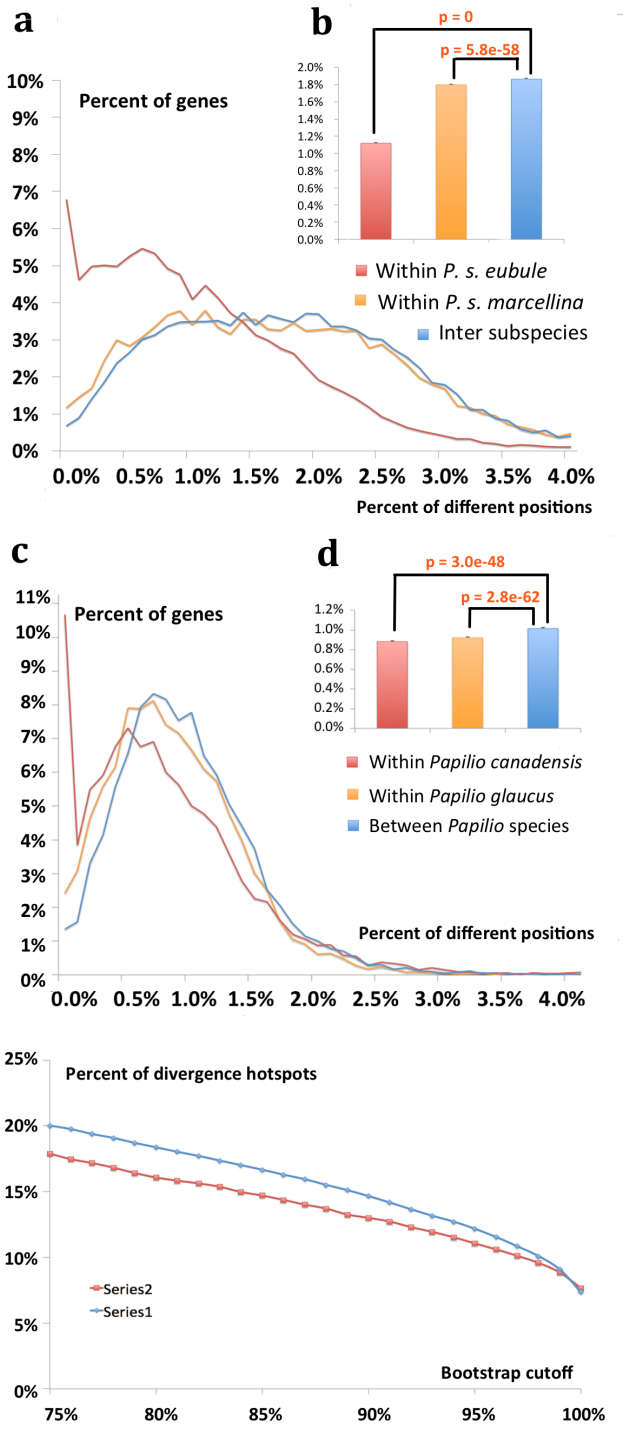


Figure 4. Genomic divergences within and between taxa shown as histograms: percent of genes for each level of divergence.

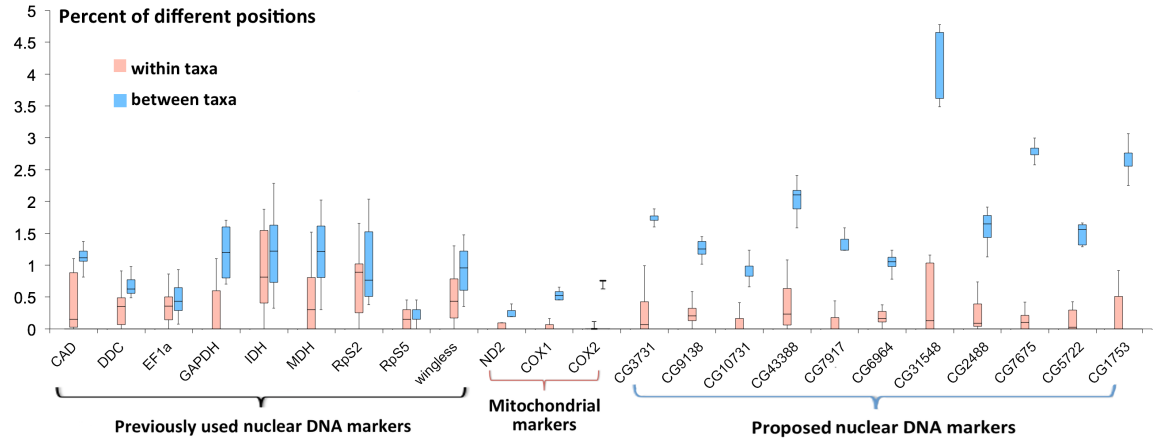


Figure 5. Divergence of selected genes (=markers) within (red) and between (blue) *Phoebis* taxa. Nuclear genes commonly used in phylogenetic analysis of Lepidoptera (General nuclear markers) are shown on the left and nuclear genes that discriminate best between the taxa based on this study are shown on the right (Specific nuclear markers). See Table 4 for information about these markers.

REFERENCES

1. International Silkworm Genome C: **The genome of a lepidopteran model insect, the silkworm Bombyx mori.** *Insect Biochem Mol Biol* 2008, **38:**1036-1045.

2. You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, et al: **A heterozygous moth genome provides insights into herbivory and detoxification.** *Nat Genet* 2013, **45:**220-225.

3. Heliconius Genome C: **Butterfly genome reveals promiscuous exchange of mimicry adaptations among species.** *Nature* 2012, **487:**94-98.

4. Zhan S, Merlin C, Boore JL, Reppert SM: **The monarch butterfly genome yields insights into long-distance migration.** *Cell* 2011, **147:**1171-1185.

5. Tang W, Yu L, He W, Yang G, Ke F, Baxter SW, You S, Douglas CJ, You M: **DBM-DB: the diamondback moth genome database.** *Database (Oxford)* 2014, **2014:**bat087.

6. Zhan S, Reppert SM: **MonarchBase: the monarch butterfly genome database.** *Nucleic Acids Res* 2013, **41:**D758-763.

7. Duan J, Li R, Cheng D, Fan W, Zha X, Cheng T, Wu Y, Wang J, Mita K, Xiang Z, Xia Q: **SilkDB v2.0: a platform for silkworm (Bombyx mori) genome biology.** *Nucleic Acids Res* 2010, **38:**D453-456.

8. Cong Q, Borek D, Otwinowski Z, Grishin NV: **Tiger Swallowtail Genome Reveals Mechanisms for Speciation and Caterpillar Chemical Defense.** *Cell Rep* 2015.

9. Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Valimaki N, Paulin L, Kvist J, Wahlberg N, et al: **The Glanville fritillary genome retains an ancient**

karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat Commun* 2014, **5:**4737.

10. Nishikawa H, Iijima T, Kajitani R, Yamaguchi J, Ando T, Suzuki Y, Sugano S, Fujiyama A, Kosugi S, Hirakawa H, et al: **A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly.** *Nat Genet* 2015, **47:**405-409.

11. Cong Q, Borek D, Otwinowski Z, Grishin NV: **Skipper genome sheds light on unique phenotypic traits and phylogeny.** *BMC Genomics* 2015, **16:**639.

12. Zhan S, Zhang W, Niitepold K, Hsu J, Haeger JF, Zalucki MP, Altizer S, de Roode JC, Reppert SM, Kronforst MR: **The genetics of monarch butterfly migration and warning colouration.** *Nature* 2014, **514:**317-321.

13. Nadeau NJ, Ruiz M, Salazar P, Counterman B, Medina JA, Ortiz-Zuazaga H, Morrison A, McMillan WO, Jiggins CD, Papa R: **Population genomics of parallel hybrid zones in the mimetic butterflies, H. melpomene and H. erato.** *Genome Res* 2014, **24:**1316-1333.

14. Asher J, Warren M, Fox R, Harding P, Jeffcoate G, Jeffcoate S: **The Millennium Atlas of Butterflies in Britain and Ireland.** *Oxford University Press* 2001**:**xx + 433pp. .

15. Pfeiler EJ: **The effect of pterin pigments on wing coloration of four species of Pieridae (Lepidoptera).** *J Research on the Lepidoptera* 1968, **7:**183-189.

16. Ehrlich PR: **The comparative morphology, phylogeny and higher classification of the butterflies (Lepidoptera: Papilionoidea).** *The University of Kansas Science Bulletin* 1958, **39:**305-370.

17. Weller SJ, Pashley DP, Martin JA: **Reassessment of Butterfly Family Relationships Using Independent Genes and Morphology.** *Annals of the Entomological Society of America* 1996, **89:**184-192.

18. Brown FM: **A revision of the genus Phoebis (Lepidoptera).** *American museum novitates* 1929, **368:**1-22.

19. Ratnasingham S, Hebert PD: **bold: The Barcode of Life Data System (**http://www.barcodinglife.org**).** *Mol Ecol Notes* 2007, **7:**355-364.

20. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23:**1061-1067.

21. Kawahara AY, Breinholt JW: **Phylogenomics provides strong evidence for relationships of butterflies and moths.** *Proc Biol Sci* 2014, **281:**20140970.

22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.

23. Raj A, Stephens M, Pritchard JK: **fastSTRUCTURE: variational inference of population structure in large SNP data sets.** *Genetics* 2014, **197:**573-589.

24. Aliabadian M, Beentjes KK, Roselaar CS, van Brandwijk H, Nijman V, Vonk R: **DNA barcoding of Dutch birds.** *Zookeys* 2013**:**25-48.

25. Brown WM, George M, Jr., Wilson AC: **Rapid evolution of animal mitochondrial DNA.** *Proc Natl Acad Sci U S A* 1979, **76:**1967-1971.

26. Pons JM, Sonsthagen S, Dove C, Crochet PA: **Extensive mitochondrial introgression in North American Great Black-backed Gulls (Larus marinus) from the American Herring Gull (Larus smithsonianus) with little nuclear DNA impact.** *Heredity (Edinb)* 2014, **112:**226-239.

27. Bazin E, Glemin S, Galtier N: **Population size does not influence mitochondrial genetic diversity in animals.** *Science* 2006, **312:**570-572.

28. Graham RI, Wilson K: **Male-killing Wolbachia and mitochondrial selective sweep in a migratory African insect.** *BMC Evol Biol* 2012, **12:**204.

29. Ballard JW, Whitlock MC: **The incomplete natural history of mitochondria.** *Mol Ecol* 2004, **13:**729-744.

30. Sawamura K: **Chromatin evolution and molecular drive in speciation.** *Int J Evol Biol* 2012, **2012:**301894.

31. Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J: **A mouse speciation gene encodes a meiotic histone H3 methyltransferase.** *Science* 2009, **323:**373-375.

32. Durand S, Bouche N, Perez Strand E, Loudet O, Camilleri C: **Rapid establishment of genetic incompatibility through natural epigenetic variation.** *Curr Biol* 2012, **22:**326-331.

33. Van Nieuwerburgh F, Thompson RC, Ledesma J, Deforce D, Gaasterland T, Ordoukhanian P, Head SR: **Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination.** *Nucleic Acids Res* 2012, **40:**e24.

34. Chevreux B, Wetter T, Suhai S: **Genome Sequence Assembly Using Trace Signals and Additional Sequence Information.** *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics* 1999, **99:**45-56.

35. Marcais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics* 2011, **27:**764-770.

36. Kelley DR, Schatz MC, Salzberg SL: **Quake: quality-aware detection and correction of sequencing errors.** *Genome Biol* 2010, **11:**R116.

37. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, et al: **Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads.** *Genome Res* 2014, **24:**1384-1395.

38. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.

39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

40. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.** *Nat Protoc* 2013, **8:**1494-1512.

41. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol* 2013, **14:**R36.

42. Roberts A, Pimentel H, Trapnell C, Pachter L: **Identification of novel transcripts in annotated genomes using RNA-Seq.** *Bioinformatics* 2011, **27:**2325-2329.

43. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.** *Genome Biol* 2008, **9:**R7.

44. Smit AFA, Hubley R: **(**http://www.repeatmasker.org**) RepeatModeler Open-1.0.** 2008-2010.

45. Jurka J, Klonowski P, Dagman V, Pelton P: **CENSOR--a program for identification and elimination of repetitive elements from DNA sequences.** *Comput Chem* 1996, **20:**119-121.

46. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110:**462-467.

47. Smit AFA, Hubley R, Green P: **(**http://www.repeatmasker.org**) RepeatMasker Open-3.0.** 1996-2010.

48. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, et al: **Annotation of the Drosophila melanogaster euchromatic genome: a systematic review.** *Genome Biol* 2002, **3:**RESEARCH0083.

49. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6:**31.

50. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics* 2007, **23:**1282-1288.

51. She R, Chu JS, Uyar B, Wang J, Wang K, Chen N: **genBlastG: using BLAST searches to build homologous gene models.** *Bioinformatics* 2011, **27:**2141-2143.

52. Stanke M, Schoffmann O, Morgenstern B, Waack S: **Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources.** *BMC Bioinformatics* 2006, **7:**62.

53. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5:**59.

54. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.** *Bioinformatics* 2004, **20:**2878-2879.

55. Besemer J, Borodovsky M: **GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses.** *Nucleic Acids Res* 2005, **33:**W451-454.

56. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Res* 2008, **18:**188-196.

57. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

58. UniProt C: **Activities at the Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2014, **42:**D191-198.

59. St Pierre SE, Ponting L, Stefancsik R, McQuilton P, FlyBase C: **FlyBase 102--advanced approaches to interrogating FlyBase.** *Nucleic Acids Res* 2014, **42:**D780-788.

60. 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.

61. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13:**2178-2189.

62. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol* 2013, **30:**772-780.

63. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30:**1312-1313.

64. Hao JJ, Hao JS, Sun XY, Zhang LL, Yang Q: **The complete mitochondrial genomes of the Fenton's wood white, Leptidea morsei, and the lemon emigrant, Catopsilia pomona.** *J Insect Sci* 2014, **14:**130.

65. Wu Y, Fang J, Li W, Han D, Wang H, Zhang B: **The complete mitochondrial genome of Colias erate (Lepidoptera: pieridae).** *Mitochondrial DNA* 2015**:**1-2.

66. Yang J, Xu C, Li J, Lei Y, Fan C, Gao Y, Xu C, Wang R: **The complete mitochondrial genome of Gonepteryx mahaguru (Lepidoptera: Pieridae).** *Mitochondrial DNA* 2014**:**1-2.

67. Hahn C, Bachmann L, Chevreux B: **Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads--a baiting and iterative mapping approach.** *Nucleic Acids Res* 2013, **41:**e129.

68. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26:**589-595.

69. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43:**491-498.

70. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** 1989, **5:**164-166.

CHAPTER NINE
Skipper genome sheds light on unique phenotypic traits and phylogeny

INTRODUCTION

Butterflies and moths (Lepidoptera) have relatively small genomes compared to other eukaryotes, yet they display complex life cycles and diverse wing patterns. These characteristics have contributed to their emergence as powerful models for genetics and evolutionary studies. A new paradigm that gene exchange between species being a driver in the evolution of adaptation in Heliconius butterflies, has increased excitement in the field [1]. Additional interest in the Lepidoptera models has resulted from discovering molecular mechanisms responsible for complex traits, such as sexual dimorphism [2-5].

Despite the wealth of life cycle, habit and morphological data available for butterflies (Rhopalocera), their phylogeny is uncertain. Traditionally, the Papilionidae (swallowtails), Pieridae, Nymphalidae, Lycaenidae and Riodinidae families were grouped into a single superfamily, Papilionoidea, that represents typical butterflies. A sister superfamily Hesperioidea contained the single Hesperiidae family [6]. Hesperiidae are similar to many typical butterflies in the egg, larval and pupal stages, however, adults are morphologically distinct, and are characterized by reflexed antennal clubs, larger heads, and several moth-like characteristics such as stockier bodies, stronger wing muscles and darting flight with faster wing beats [6]. Their ability to fly rapidly gained them the common name “skippers”. Skippers were traditionally considered to be the basal branch of butterflies based on morphological characters [6]. Phylogenetic reconstructions of 57 butterfly and skipper species that combined

DNA sequences of three phylogenetic markers with morphological characters agreed with the basal placement of skippers [7]. However, a purely DNA-based phylogeny presented in the same study contradicted this view and placed Papilionidae at the base with Hesperiidae as a sister to other butterfly families. Similarly, a recent larger-scale study that included transcriptomes of 9 butterflies and skippers reported a highly confident phylogeny with Papilionidae in the basal position [8]. Therefore, the reconciliation of the discrepancy between these morphology-based and DNA-based phylogenies requires further studies and the phylogeny of major families of butterflies remain an open question.

Decoding the skipper genomes could help the reconstruction of Lepidoptera tree and provide information that is essential for understanding the evolution of their moth-like morphological features, which are either inherited from their ancestor or are character reversals. Here we report the assembly and gene annotations for the highly heterozygous genome of the Clouded Skipper *Lerema accius* (J. E. Smith, 1797), abbreviated as *Lac*, shown on Figure 1. *Lac* belongs to the subfamily Hesperiinae, commonly known as Grass Skippers, the most species-rich subfamily of skippers. Caterpillars of most Hesperiinae feed on grasses and sedges. Hesperiinae adults typically hold wings erect over the thorax and abdomen when feeding and resting. They adopt a “jet plane” pose when basking: partially open the wings and hold the fore- and hindwings at different angles.

Comparative analysis of this first genome from the family Hesperiidae with other Lepidoptera genomes provides hypotheses about bases for unique morphological traits of skippers, such as their fast flight. Phylogenetic analyses of *Lac* with Lepidoptera species with available complete genomes fail to resolve the position of Hesperiidae. Maximum likelihood

tree constructed by RAxML [9] using the most suitable evolutionary model (JTTDCMUT model) selected by the program place swallowtails at the base of the tree, consistent with published DNA phylogenies, while Bayesian inference [10] with an evolutionary model that accounts for site-heterogeneity [11], weakly supports the traditional morphology-based phylogeny in which skippers are the basal branch of butterflies. More extensive taxon sampling and/or more advanced methods of phylogenetic analysis are needed to resolve the position of Hesperiidae conclusively, and the first Hesperiidae genome provides a starting point for these studies.

RESULTS AND DISCUSSION

Genome quality assessment and gene annotation

We assembled a 310 Mb genome of *Lac* and compared its quality with genomes (Table 1) of the following Lepidoptera species: *Plutella xylostella* (Pxy), *Bombyx mori* (*Bmo*), *Papilio glaucus* (*Pgl*), *Melitaea cinxia* (*Mci*), *Heliconius melpomene* (*Hme*), and *Danaus plexippus* (*Dpl*) [1-3, 12-17]. The scaffold N50 of *Lac* is 513 kb, which is longer than several other butterfly genomes. The genome is among the best in terms of completeness measured by the presence of CEGMA (Core Eukaryotic Genes Mapping Approach) genes [18], cytoplasmic ribosomal proteins and independently assembled transcripts. The residue coverage (86.6%) of CEGMA genes by single *Lac* scaffolds is comparable to the residue coverage by the current *Bmo* assembly with an N50 of about 4.0 Mb, indicating that the quality of the *Lac* draft is sufficient for protein annotation and comparative analysis. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession LGAG00000000.

The version described in this paper is version LGAG01000000. In addition, the main results from genome assembly, annotation and analysis can be downloaded at http://prodata.swmed.edu/LepDB/.

We assembled the transcriptomes from two other *Lac* specimens, a pupa and an adult. Based on the transcriptomes, homologs from other insects, *de novo* predictions and repeat identification, we predicted 17,416 protein-coding genes in *Lac*. 79% of these genes are likely expressed, as they fully or partially overlap with the transcripts. We annotated the putative function for 12,283 protein-coding genes.

Comparison of Lepidoptera genomes

We compared the composition of the *Lac* genome with that of other Lepidoptera (Table 1). Although the genome sizes of Lepidoptera range from 250 to 500 Mbp, the total lengths of coding regions are comparable. The reported repeat content of these genomes varies significantly, and repeat content is positively correlated with the genome size. We identified orthologous proteins encoded by these genomes and detected 5770 universal orthologous groups where 2940 consist of a single-copy gene in each of the species (Fig. 2a). We compared two protein families: Hox genes that are crucial for development and Odorant Receptors (OR) that are particularly important for the feeding and mating behaviors of insects. *Lac* had the same set of Hox genes as other Lepidoptera. All the *Lac* Hox genes that are expected to be linked are located on the same scaffold in the order, typical for Lepidoptera (Fig. 2b). The *Lac* genome encodes 56 ORs, which is comparable to *Pgl* but less than *Hme*, *Dpl* and moths. The *Mci* genome appears to encode the smallest number of ORs (48), but this number is likely

underestimated because of the poor continuity of the current *Mci* genome assembly (119 kbp). Clustering analysis shows that ORs in Lepidoptera can be classified into several subfamilies, and the *Lac* genome encodes ORs from each of these subfamilies.

Functional implication of Non-random Single-nucleotide polymorphism (SNP) distribution

The *Lac* genome is highly heterozygous, as suggested by the distribution of k-mer frequencies (Fig. 3a). Here, we compare and describe heterozygosity properties of the *Lac* genome and the highly heterozygous *Pgl* genome that we previously assembled [16]. Approximately 2.3% of the positions in the *Pgl* genome and 1.6% of the positions in the *Lac* genome are different between the two homologous chromosomes. In both genomes, the SNP rate in the coding regions (0.91% for *Pgl* and 0.96% for *Lac*) is much lower than that for the non-coding regions (2.4% for *Pgl* and 1.7% for *Lac*), which is likely due to the potential deleterious effect of SNPs in the coding regions.

Both the *Pgl* and *Lac* genome contain long segments (>1,000) that are free of SNPs. However, the SNP-free segments in the *Pgl* genome are significantly longer than those in *Lac*. The longest SNP-free segments in *Pgl* and *Lac* are about 734.8 kbp and 13.5 kbp, respectively. One possible explanation for the presence of SNP-free regions is that the high heterozygosity of these regions prevents the mapping of reads from the alternative homologous chromosomes, resulting in failures to detect SNPs. But this is not likely the dominant reason, since we included only regions with the expected coverage by the reads in this analysis. Another potential reason for SNP-free regions is that insects frequently inbreed in nature and that the

parents of the sequenced specimen could share a recent common ancestor, from which they inherited the same alleles.

Omitting the SNP-free regions, the distribution of SNP rates in the *Pgl* genome can be approximated by a single normal distribution (Fig. 3b). In contrast, the distribution of SNPs rates (Fig. 3c) in the *Lac* genome can be represented by a mixture of two Gaussian distributions: one centered around 0.3-0.4% and a second centered at 2.5%. We speculate that a SNP rate of 0.3-0.4% corresponds to the variation accumulated within the local population of *Lac*, whereas the higher SNP rates in certain regions reflect gene flow from other populations or even from other species. Human activities might have an impact on the high SNP rates of *Lac*. For example, *Lac* feeds on widely planted grasses (*Poaceae* family). Expansion of this common food source by humans might cause previously isolated *Lac* populations to meet.

A quarter (22% for *Lac* and 26% for *Pgl*) of the SNPs are non-synonymous and result in amino acid substitutions in proteins. Protein regions that are predicted to be structurally disordered are significantly more enriched in substitutions. This enrichment is likely due to higher tolerance of disordered regions to substitutions [19]. To help understand the functional consequence of SNPs in the *Pgl* and *Lac* genomes, we identified proteins that are significantly enriched (false discovery rate < 0.1) in substitutions in their structurally ordered regions.

The enriched GO terms associated with substitution-enriched proteins in both genomes show a significant ($p < 1e-15$) overlap (Fig. 3d). Among the enriched biological process (Fig. 3e) and molecular function (Fig. 3f) GO terms shared by both species, the molecular function "catalytic activity" is among the most significant ($p < 1e-4$). Approximately 40% of the substitution-enriched proteins are enzymes in both species. The most significantly enriched

GO ($p < 1e-8$) terms for *Lac* (GO:0045931, GO:0031935, GO:0060968, GO:0045787 and GO:0030178) can each be attributed solely to a single substitution-rich protein family: C2H2 zinc fingers. Both insect and mammalian genomes encode large numbers of C2H2 zinc fingers, and their exact function is not fully understood [20]. However, C2H2 zinc fingers were implicated in transcriptional silencing of exogenous DNA [21, 22]. We hypothesize that the C2H2 zinc fingers evolved adaptively as the population was exposed to exogenous DNA sources, such as retrovirus or gene flow from other species.

Phylogenetic analysis with whole-genome data

The morphology-based view of butterfly evolution suggests a tree topology (((((*Mci*, *Hme*), *Dpl*), *Pgl*), *Lac*), *Bmo*, *Pxy*) [6, 7], whereas recent DNA-based phylogenetic analyses supports an alternative topology (((((*Mci*, *Hme*), *Dpl*), *Lac*), *Pgl*), *Bmo*, *Pxy*) [7, 8]. We refer to these two topologies as the traditional topology and the alternate topology, respectively.

Whole-genome sequences of these species allowed us to model their phylogeny using both the alignments of universal single-copy orthologs and the synteny of genes. However, both the traditional and the alternate tree topologies can be supported by the data depending on which evolutionary models and tree construction methods are selected. The 50% majority rule consensus tree of maximum likelihood trees constructed with RAxML [9] on the alignments of individual proteins failed to completely resolve the phylogeny due to short lengths (median length: 209 amino acids) of individual alignments. Instead, a similar consensus tree built on 1000 random samples of long alignments (> 5,000 aligned positions)

from concatenated alignments agreed with the alternate topology (Fig. 4a). However, the clade that groups skippers with other butterflies is only the best solution in 72% of random samples.

To further test which topology is better supported, we used the Bayesian phylogenetic analysis software PhyloBayes [10] with the CAT model [11] that accounts for site heterogeneity in amino acid substitutions by dividing the sites into 4 categories. We constrained the tree topology to either the traditional or the alternative one. This analysis supported the traditional topology in 66% of the 1000 random samples. A consensus tree summarizing the tree topologies with higher likelihood based on each data set is shown in Fig. 4b.

Similarly, the phylogeny inferred from gene rearrangement events produced different results depending on the selection of evolutionary model. While a simple neighbor-joining tree based on the frequency of gene arrangement supported the alternative topology (Fig. 4c), Bayesian interference with the CAT model supported the traditional topology with a higher likelihood (Fig. 4d). Although the traditional tree topology based on morphological features is not contradicted by our genomic data analysis, the uncertainty of reconstructions is too high to conclusively determine the evolutionary history of butterflies.

The discrepancy between morphological and molecular phylogeny has been a long-standing problem in evolutionary biology [23]. The incongruence between molecular trees obtained with different methods or different data sets is also frequently encountered [23, 24], and studies on several other systems reveal similar uncertainty as we observed in our analysis [25, 26]. This uncertainty in butterfly phylogeny may also result from incomplete lineage sorting [27]. Trees built from different orthologous groups support different topologies with

high bootstrap values. Out of the 522 maximum likelihood trees of individual orthologous groups with bootstrap support above 80%, a significant portion of them supports the traditional topology (24.7%) and a third possible topology (33.7%) that groups *Pgl* and *Lac* in a clade, although the alternate topology is overall better supported by 41.6% of them. In addition, the limited number of available butterfly genomes impedes a better taxon sampling for the phylogenetic reconstruction of butterflies. Genome sequences of species that represents the early branches in each family of butterflies could help to resolve the uncertainty in the phylogenetic tree of butterflies.

Expanded gene families in *Lac* suggest possible genetic bases for phenotypic traits

Compared to other Lepidoptera species, the *Lac* genome contains expansions in several protein families. Endochitinase-like proteins are uniquely expanded (Fig. 5a) and cluster on the same scaffold in the genome, which indicates that they originated from recent gene duplication events. As shown in the phylogenetic tree (Fig. 5a), these duplicated endochitinase-like proteins diverged rapidly and only one copy retained high sequence similarity to the orthologous proteins in other Lepidoptera and *Drosophila melanogaster* genomes. While this single conserved copy likely preserved the function of endochitinase, we hypothesize that the other divergent endochitinase-like proteins could have adopted new functions to digest cellulose. This hypothesis is based on the following three facts: (1) *Lac* and most skippers in the Hesperiinae subfamily feed on the cellulose-rich grasses; (2) the *Lac* genome and other Lepidoptera genomes do not encode proteins that belong to the families of known cellulases; (3) endochitinases are homologs of cellulases and they are structurally very similar [28]; (4)

cellulose and chitin are structurally similar and they are both glycoside hydrolases. Therefore, these endochitinases-like proteins in *Lac* may have evolved to digest cellulose, allowing Lac, and possibly other grass-feeding skippers in the Hesperiinae subfamily, to feed on grasses that are rich in it. It remains to be explored if other Monocot feeders, such as Satyrinae (Nymphalidae), have a similar expansion or use different enzymes.

Another expanded protein family is geranylgeranyl pyrophosphate synthase (GGPPS, Fig. 5b) homologs. GGPPSs are used in the biosynthesis of terpenes and terpenoids, which are frequently used as an intermediate product for pheromone biosynthesis. 13 copies of GGPPS homologs in *Lac* form a clade in the phylogenetic tree (highlighted in magenta in Fig. 5b) and their sequences have diverged from the Drosophila GGPPS. It is possible that these homologs have adopted slightly different functions and gained the ability either to catalyze different steps to synthesize one type of pheromone or to produce a wide range of different pheromone molecules. In addition, *Lac* encodes a much larger number of pheromone-binding proteins (PBPs) than other Lepidoptera species and these PBPs form a clade in the phylogenetic tree of Lepidoptera PBPs (Fig. 5c). Both gene expansion events suggest a more advanced pheromone production and sensing system in *Lac*. Butterflies can select their mates both by using visual cues and by sensing pheromones at close range. However, many skipper species have similar wing colors and patterns, which might confuse recognition by the mates of the same species. Therefore, a stronger pheromone system in *Lac* might allow individuals to efficiently detect mates of the same species.

The phylogenetic tree of GGPPS homologs reveals two copies in *Lac* (annotated as *Lac*_GPS5 and *Lac*_GPS6) that clustered closely to the *Drosophila* GGPPS, rather than in the

clade of other divergent GGPPS homologs. We speculate that these two copies are orthologs of the *Drosophila* GGPPS and retain similar function. *Drosophila* GGPPS was shown to be crucial for heart formation. It works in the mevalonate pathway and directly synthesizes GGPP, which can be transferred to G protein Gγ1. The geranylgeranylation of Gγ1 is required for heart formation [29], and the duplication of GGPPS may be related to heart development for efficient energy supply to sustain the rapid wing beats of *Lac*. In addition, several mitochondria targeted genes encoded by the nuclear genome are also duplicated in the *Lac* genome (Table 2), including components of the NADH dehydrogenase [uniquinone] complex, which is directly linked to energy production. The *Lac* genome is significantly ($p < 1e-7$) enriched in mitochondria targeted genes compared to other Lepidoptera as reflected by the GO terms. Taken together, we propose that the observed enrichment and duplications of mitochondrial proteins allow for dynamic adaptation of mitochondrial functions depending on type of organ, tissue, or life stage and ensure efficient energy supply for rapid wing beats in adults of *Lac*.

CONCLUSIONS

We report the draft genome of Clouded Skipper. Being the first sequenced genome from the Hesperiidae family, it offers a rich dataset for comparative genomics and phylogenetic studies of Lepidoptera. We devised a cost-efficient protocol that overcomes the difficulty in assembling highly heterozygous genome. Despite the high level of heterozygosity (1.5%), the quality of our genome assembly is nearly the best among published Lepidoptera genomes. This protocol should stimulate and enable sequencing of other insect genomes. Comparative analyses of Lepidoptera genomes suggest possible genetic bases for the unique phenotypic

traits of skippers, including fast flight with rapid wing beats, ability to feed on grasses in larval stage, and recognize mates efficiently in spite of the similarity in wing patters of many species. These new data should facilitate experimental studies of skippers and contribute to the understanding of how diverse phenotypes are encoded by the genomes.

METHODS

Library preparation and sequencing

We removed and preserved the wings and abdomen of a freshly caught and frozen male *Lac* specimen (USA: Texas: Dallas County, Dallas, White Rock Lake, Olive Shapiro Park, 10-Nov-2013, GPS: 32.8621, -96.7305, elevation: 141 m), and extracted approximately 15 µg genomic DNA from the rest of its body with the ChargeSwitch gDNA mini tissue kit. 250 bp and 500 bp paired-end libraries were prepared using enzymes from NEBNext Modules and following the Illumina TruSeq DNA sample preparation guide. 2 kb, 6 kb and 15 kb mate pair libraries were prepared using a protocol that was modified from a previously published Cre-Lox-based method [30]. For the 250 bp, 500 bp, 2 kb, 6 kb and 15 kb libraries, approximately 500 ng, 500 ng, 1.5 µg, 3 µg and 6 µg of DNA were used, respectively. A *Lac* adult and a pupa reared from a caterpillar collected at the same locality (White Rock Lake) were preserved in *RNAlater* solution and total RNA was extracted from them using QIAGEN RNeasy Plus Mini Kit. We further isolated mRNA using NEBNext Poly(A) mRNA Magnetic Isolation Module and RNA-seq libraries for both specimens were prepared with NEBNext Ultra Directional RNA Library Prep Kit for Illumina following manufactory's protocol.

We quantified the amount of DNA from all the libraries with the KAPA Library Quantification Kit, and mixed 250 bp, 500 bp, 2 kb, 6 kb, 15 kb genomic DNA pupal RNA-seq and adult RNA-seq libraries to get the final library with relative molar concentration 40:20:8:4:3:20:10. The final library was sent to the genomics core facility at UT Southwestern Medical Center for 150 bp paired-end sequencing on Illumina HiSeq2000. The sequencing reads have been deposited in NCBI SRA database under accession numbers: SRR2089769-SRR2089775. The sequence reads to assemble the genome and transcriptome have been deposited at the same database under accession numbers: SRR2089776 and SRR2089777.

Genome assembly

We removed sequence reads that did not pass the Illumina purity filter and classified the remainder according to their TruSeq adapter indices. Mate pair libraries were processed by the Delox script [30] to remove the LoxP sequences and to separate true mate pair from paired-end reads. All reads were processed by mirabait [31] to remove contamination from the TruSeq adapters, fastq_quality_trimmer (Kondratowicz et al., 2011) to remove low quality portions at both ends, JELLYFISH [32] to obtain k-mer frequencies in all the libraries, and QUAKE [33] to correct sequencing errors. The data processing resulted in nine libraries that were supplied to Platanus [34] for genome assembly: 250 bp and 500 bp paired-end libraries, three paired-end and three mate pair libraries from 2 kb, 6 kb and 15 kb libraries and a single-end library containing all reads whose pairs were removed in the process.

We mapped these reads to the initial assembly with Bowtie2 [35] and calculated the coverage of each scaffold with the help of SAMtools [36]. Many short scaffolds in the

assembly showed coverage that was about half of the expected value, which likely resulted from highly heterozygous regions that were not merged to the equivalent segments in the homologous chromosomes. We merged them into other scaffolds if they could be fully aligned to another significantly less covered region (coverage > 90% and uncovered region < 500 bp) in a longer scaffold with high sequence identity (>95%). Similar problems occurred in the *Heliconius melpomene* and *Papilio glaucus* genome projects, and similar strategies were used to improve the assemblies [1, 16].

Transcriptome assembly

After removing contamination from TruSeq adapters and the low quality portion of the reads using the methods mentioned above, we applied three methods to assemble the transcriptomes: (1) *de novo* assembly by Trinity [37], (2) reference-based assembly by TopHat [38] (v2.0.10) and Cufflinks [39] (v2.2.1), and (3) reference-guided assembly by Trinity. The results from all three methods were then integrated by Program to Assemble Spliced Alignment (PASA) [40].

Identification of repeats and gene annotation

Two approaches were used to identify repeats in *Lac* genome: the RepeatModeler [41] pipeline and in-house scripts that extracted regions with coverage 4 times higher than expected. These repeats were submitted to the CENSOR [42] server to assign them to the repeat classification hierarchy. The species-specific repeat library and repeats classified in RepBase [43] (V18.12) were used to mask repeats in the genome by RepeatMasker [44].

From the transcripts of both specimens in the pupal and adult stages, we obtained two sets of transcript-based annotations from two pipelines: TopHat followed by Cufflinks and Trinity followed by PASA. In addition, we obtained five sets of homology-based annotations by aligning protein sets from *Drosophila melanogaster* [45] and four published Lepidoptera genomes to the *Lac* genome with exonerate [46]. Proteins from the entire UniRef90 [47] database were used to generate another set of gene predictions by genblastG [48]. We manually curated and selected 1427 confident gene models by integrating the evidence from transcripts and homologs to train *de novo* gene predictors: AUGUSTUS [49], SNAP [50] and GlimmerHMM [51]. These trained predictors, the self-trained Genemark [52] and a consensus based pipeline, Maker [53] were used to generate another five sets of gene models. Transcript-based and homology-based annotations were supplied to AUGUSTUS, SNAP and Maker to boost their performance. In total, we generated 15 sets of gene predictions and integrated them with EvidenceModeller [40] to generate the final gene models.

We predicted the function of *Lac* proteins by transferring annotations and GO-terms from the closest BLAST [54] hits (E-value<10^{-5}) in both the Swissprot [55] database and Flybase [56]. Finally, we performed InterproScan [57] to identify conserved protein domains and functional motifs, to predict coiled coils, transmembrane helices and signal peptides, to detect homologous 3D structures, to assign *Lac* proteins to protein families and to map them to metabolic pathways.

Assembly quality assessment and comparison to other Lepidoptera genomes

We obtained the most recent versions of other published Lepidoptera genomes, including *Bombyx mori*, *Danaus plexippus*, *Heliconius melpomene*, *Melitaea cinxia*, *Papilio glaucus*, and *Plutella xylostella* [1-3, 12-17]. Using the criteria applied in the Monarch butterfly genome paper [2], we estimated the completeness of these genomes based on their coverage of independently obtained transcripts, CEGMA [18] genes and the Cytoplasmic Ribosomal Proteins.

We compared various properties of these published genomes and clustered the proteins annotated in them using OrthoMCL [58]. We identified the Hox genes using homeodomains from *Drosophila* in the HomeoDB [59] as reference, and relationship among them were detected using a phylogenetic tree built by RAxML [9] with automatically selected model on the MAFFT [60] alignment. Starting from the annotated odorant receptors from the *Bmo*, *Hme* and *Dpl* genomes, we identified all the odorant receptors in the annotated protein sets from these Lepidoptera genomes using reciprocal BLAST. Odorant receptors encoded by the genome but missed in the protein sets were predicted with the help of genblastG. All the candidates identified by the automatic programs were further curated to remove short fragments (<200 aa) and false positive hits that do not detect odorant receptors as the top hit in a BLAST search against Flybase entries. Sequences of these odorant receptors were compared and clustered using CLANS [61].

Detection and analysis of SNPs

We analyzed the SNPs in *Lac* and *Pgl* genomes using the same protocol, in which we mapped each of the sequence reads to the genomes and detected SNPs using the Genome Analysis

Toolkit [62]. For both *Pgl* and *Lac* genomes, this distribution shows two peaks. In addition to the main peak centered at the expected coverage for a diploid genome, there is an additional peak to the left that corresponds to highly divergent regions between the two homologous chromosomes. Owing to this sequence divergence, only the reads corresponding to the sequence of one of the homologous chromosomes can be mapped, which results in the lower-than-expected coverage. To analyze the distribution of SNPs, we focused on the regions, in which coverage by the reads falls within the diploid peak. We divided these regions into exons, introns, repeats and intergenic regions. The percent of SNPs in overlapping 1000 bp windows in the genome was used to reflect their distributions. We detected non-synonymous SNPs that will cause substitutions in proteins and predicted structurally disordered regions in proteins with ESpritz server [63].

We identified proteins with significantly more substitutions with binomial tests (p = average percent of substitutions in all proteins, m = number of substitutions in a protein, N = length of a protein) followed by False Discovery Rate (FDR) tests [64]. We considered proteins with Q-values (maximal FDR level) smaller than 0.1 to be significantly enriched in substitutions. We excluded the regions that were predicted to be structurally disordered and performed similar tests. Enriched GO terms associated with these substitution-enriched proteins were identified with another binomial test (P = probability of this GO-term being associated with any protein, m = number of substitution-enriched proteins associated with this GO-term, N = number of substitution-enriched proteins). The significantly enriched GO terms were submitted to the REVIGO [65] web server to cluster similar GO terms and visualize them.

Phylogenetic tree reconstruction

We performed the phylogenetic analysis based on the 2940 universal single-copy orthologs in the Lepidoptera genomes (*Lac*, *Bmo*, *Pxy*, *Dpl*, *Hme*, *Mci*, and *Pgl*) detected by OrthoMCL. We built alignment for each orthologous group using both global sequence aligner MAFFT and local sequence aligner BLASTP. 570,686 positions that were consistently aligned by both aligners were extracted. All the alignments were concatenated and the aligned positions were randomly divided to 100 groups, so that each group contained about 5,706 or 5,707 aligned positions. We repeated this procedure 10 times to obtain a total of 1,000 representative alignments for phylogenetic analysis. In addition, the 1,991 alignments of individual orthologous groups containing more than 100 aligned positions were used as a separate data set in the phylogenetic analysis.

For the phylogenetic analysis we used two methods: a maximum likelihood method RAxML, in which the evolutionary model is automatically selected by the program based on the data and a Bayesian inference method PhyloBayes [10] with CAT model that divide sites into categories and account for site-heterogeneities [11]. In addition to allowing the program to search for the best tree topologies, we further constrained the Bayesian analysis to two previously observed topologies: (((((*Mci*, *Hme*), *Dpl*), *Lac*), *Pgl*), *Bmo*, *Pxy*) and (((((*Mci*, *Hme*), *Dpl*), *Pgl*), *Lac*), *Bmo*, *Pxy*). We compared the posterior probabilities given the two topologies imposed as priors to select the tree topology that is better supported by the data for each alignment.

In addition, we used the frequencies of gene rearrangements to construct phylogenetic trees. We started from the 5770 orthologous families present in each of the 7 species and

removed families with extensive gene duplications (more than 4 copies of a gene in any species), which resulted in 5639 families. In each species, we determined the relative genomic orientation for every pair of gene families on the same scaffold. There are four possible relative orientations: [a+, b+]; [a-, b-]; [a+, b-]; [a-, b+], where a and b are genes from two families and "+" and "-" indicate the DNA strand they are encoded on. Due to the limited continuity of draft genomes, relative orientations in all 7 species could be determined for 2120 such gene pairs. Then, we restricted the analysis to 1121 such pairs so that each family participated in only one pair. We used four letters (A, B, C, and D) to denote the relative orientations of family pairs, and expressed the arrangement of the 1121 pairs in each species by a string of these letters. These strings were used as input for PhyloBayes for tree construction. The numbers of differences between these strings were used as evolutionary distances between species to construct phylogenetic tree with BioNJ [66].

Analysis of gene expansion in *Lac*

We identified the closest homolog (BLASTP e-value < 0.00001) of each Lepidoptera protein in Flybase. If two OrthoMCL-defined orthologous families in Lepidoptera shared a common Flybase entry as their closest homolog, we merged them into one family. We considered *Lac* to have undergone gene expansion in a family if both the number and total length of *Lac* proteins in this family are more than 1.5 times of the average number and total length for other Lepidoptera species. The most significantly expanded gene families with well-defined functions were further investigated using reciprocal BLAST results and function annotations to include all relevant proteins. Proteins encoded by the genome but missed in the protein sets

were predicted with the help of genblastG. Protein sequences from each family were aligned with MAFFT. Evolutionary trees were built with RAxML and visualized in FigTree.

Table 1 Quality and composition of Lepidoptera genomes

Feature	*Lac*	*Pgl*	*Dpl*	*Hme*	*Mci*	*Bmo*	*Pxy*
Genome size (Mb)	310	376	249	274	390	480	394
Genome size without gap (Mb)	292	362	242	270	361	432	387
Heterozygosity (%)	1.6	2.3	0.55	n.a.	n.a.	n.a.	~2
Scaffold N50 (kb)	513	230	716	194	119	3999	737
CEGMA (%)	99.3	99.3	99.3	98.0	98.7	99.3	98.0
Average CEGMA coverage by single scaffold (%)	86.6	86.8	87.3	86.4	79.1	86.7	84.0
Cytoplasmic Ribosomal Proteins (%)	98.9	98.9	98.9	94.6	94.6	97.8	93.5
De novo assembled transcripts (%)	97 ~ 99	98	96	n.a.	~97	98	83
GC content (%)	34.4	35.4	31.6	32.8	32.6	37.7	38.3
Repeat (%)	15.5	22.0	16.3	24.9	28.0	44.1	34.0
Exon (%)	6.96	5.07	8.40	6.38	6.36	4.03	6.35
Intron (%)	31.6	25.6	28.1	25.4	30.7	15.9	30.7
Number of proteins (thousands)	17.4	15.7	15.1	12.8	16.7	14.3	18.1
Number of universal ortholog lost	153	114	47	354	521	394	1188
Number of species specific genes	4586	3172	2361	1526	4691	2486	5260

n.a. Data not available

Table 2 Mitochondria-targeted proteins that are duplicated in *Lerema accius*

Lerema accius proteins	Function
lac1604.25, lac1604.24, lac947.51	NADH dehydrogenase [ubiquinone] 1 α subcomplex subunit 6
lac3140.17, lac2615.10, lac570.16	NADH dehydrogenase [ubiquinone] 1 α subcomplex subunit 11
lac279.21, lac34153.1	Heat shock protein 75 kDa
lac151.15, lac151.16	Acetyl-CoA acetyltransferase A
lac492.55, lac676.35	28S ribosomal protein S18b
lac5129.10, lac6133.18	2-oxoisovalerate dehydrogenase subunit β

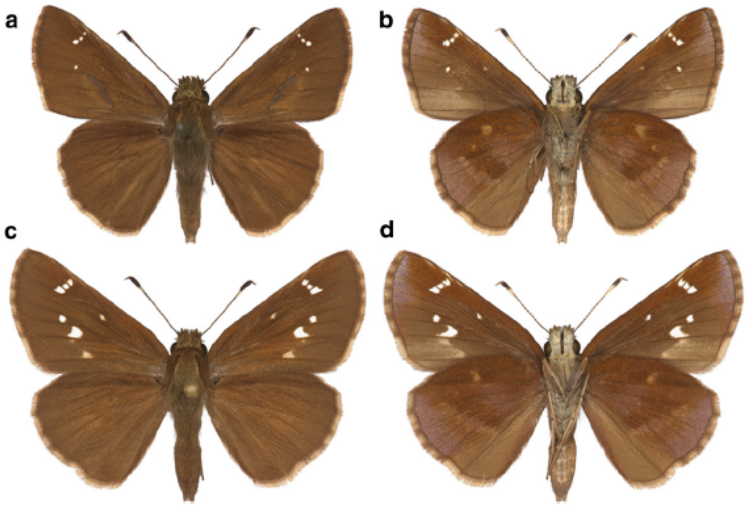


Fig. 1 Photographs of *Lac* specimens. The specimens were reared from caterpillars collected near the Grapevine Lake (USA: Texas, Denton County, Flower Mound). **a** Dorsal and **b** ventral aspects of a male specimen, eclosed on 31-Jul-1997; **c** dorsal and **d** ventral aspects of a female specimen, eclosed on 29-Sep-1997

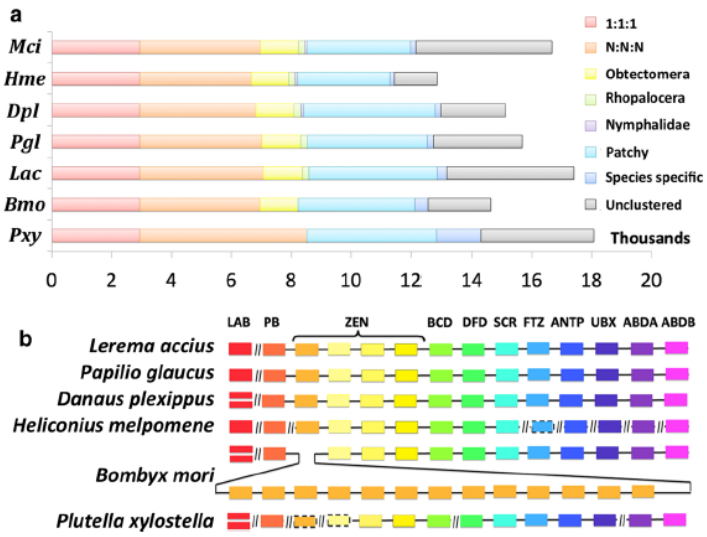


Fig. 2 Comparative analysis of Lepidoptera genomes. **a** Number of different types of orthologs in each Lepidoptera species with published genomes. 1:1:1: single-copy orthologs shared among all species; N:N:N: multiple-copy orthologs shared among all species, i.e. more than one copy in at least one species; Obtectomera: orthologs specific to Obtectomera, i.e. all other six species except *Pxy*; Rhopalocera: orthologs specific to Rhopalocera, i.e. all other five species except *Bmo* and *Pxy*; Nymphalidae: orthologs specific to Nymphalidae, i.e. *Dpl*, *Mci* and *Hme*; Patchy: orthologs that are shared between more than one, but not all species (excluding those belongs to previous categories); Species-specific: specific to only one species and having close homologs within that species; Unclustered: proteins that do not belong to any of the orthologous groups. **b** Arrangements of Hox genes in Lepidoptera genomes. Orthologs are shown as boxes of the same color; double boxes in the same position indicate gene duplications, dashed-line around a box implies that this gene is missing in the genome assembly but present in the transcriptome; "//" marks the boundaries between different scaffolds

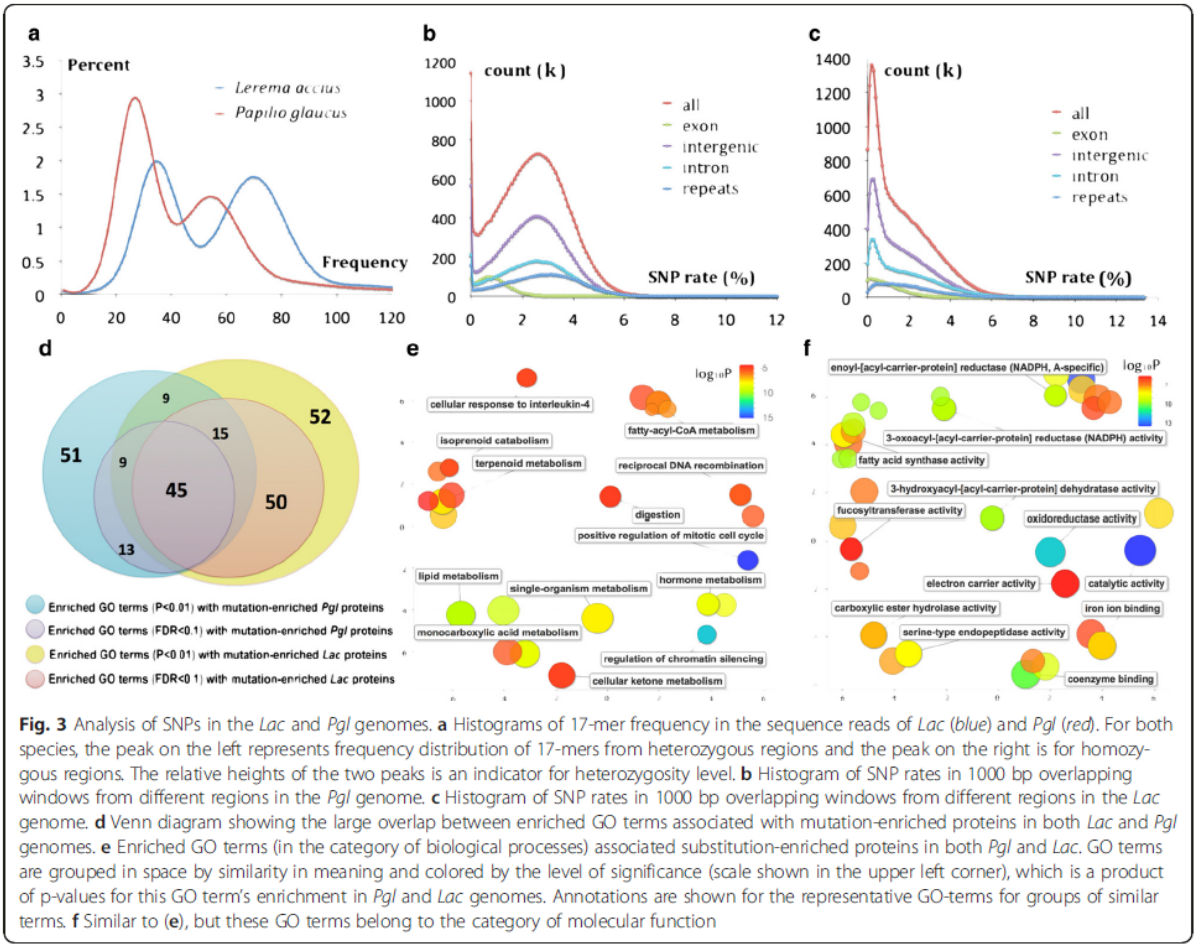


Fig. 3 Analysis of SNPs in the *Lac* and *Pgl* genomes. **a** Histograms of 17-mer frequency in the sequence reads of *Lac* (*blue*) and *Pgl* (*red*). For both species, the peak on the left represents frequency distribution of 17-mers from heterozygous regions and the peak on the right is for homozygous regions. The relative heights of the two peaks is an indicator for heterozygosity level. **b** Histogram of SNP rates in 1000 bp overlapping windows from different regions in the *Pgl* genome. **c** Histogram of SNP rates in 1000 bp overlapping windows from different regions in the *Lac* genome. **d** Venn diagram showing the large overlap between enriched GO terms associated with mutation-enriched proteins in both *Lac* and *Pgl* genomes. **e** Enriched GO terms (in the category of biological processes) associated substitution-enriched proteins in both *Pgl* and *Lac*. GO terms are grouped in space by similarity in meaning and colored by the level of significance (scale shown in the upper left corner), which is a product of p-values for this GO term's enrichment in *Pgl* and *Lac* genomes. Annotations are shown for the representative GO-terms for groups of similar terms. **f** Similar to (**e**), but these GO terms belong to the category of molecular function

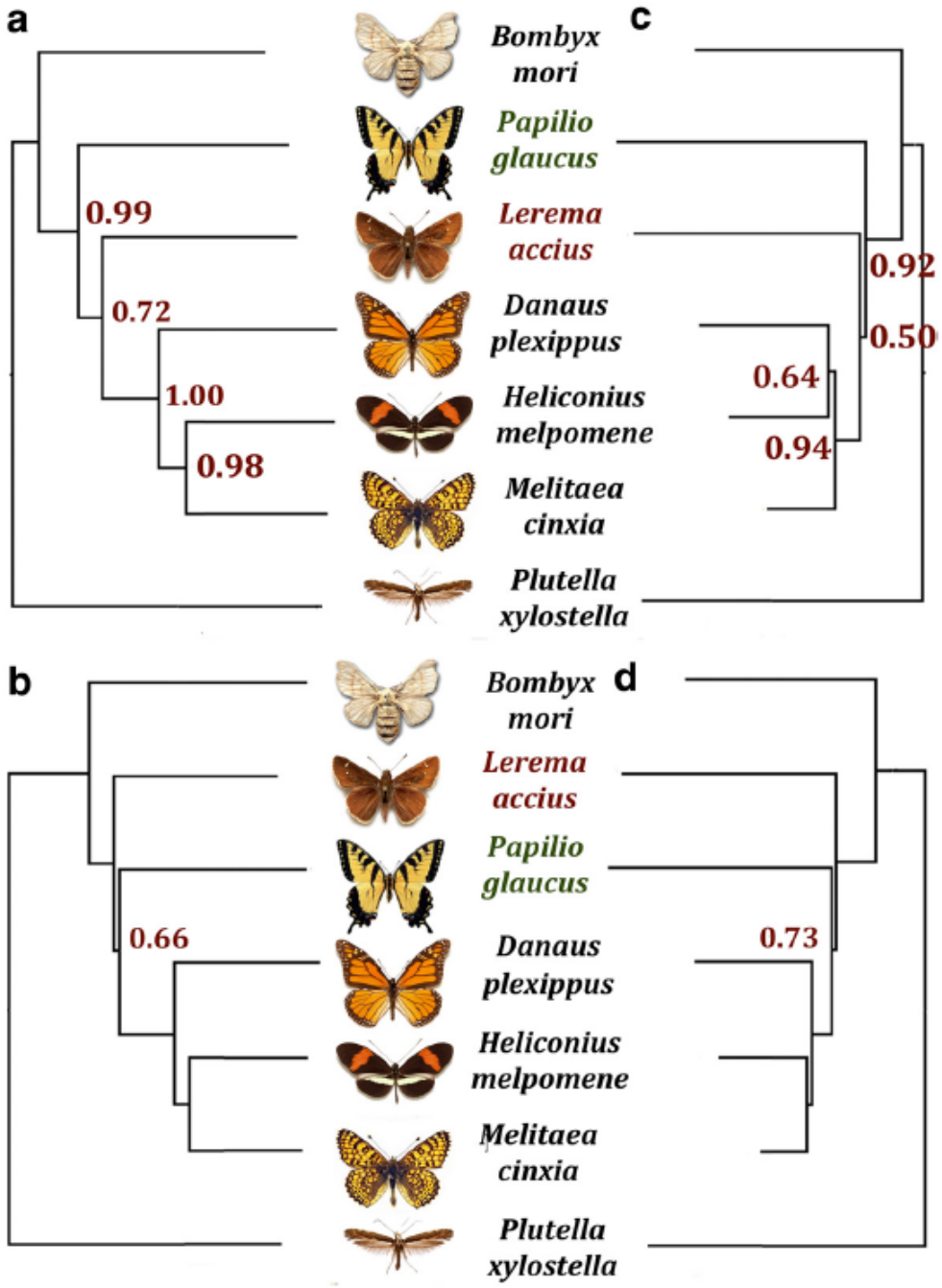


Fig. 4 Phylogenetic trees of the butterflies based on whole-genome data. **a** Majority-rule consensus tree of the maximal likelihood trees constructed by RAxML on 1000 random samples from the concatenated alignment of universal single-copy orthologs. **b** Neighbor-joining tree based on the frequency of gene rearrangement events between species. **c** Consensus tree of the better-supported trees inferred from PhyloBayes analyses on 1000 random samples from the concatenated alignment of universal single-copy orthologs. The tree topology was constrained to either of the two reported topologies: (((((*Mci*, *Hme*), *Dpl*), *Lac*), *Pgl*), *Bmo*, *Pxy*) or (((((*Mci*, *Hme*), *Dpl*), *Pgl*), *Lac*), *Bmo*, *Pxy*). **d** Phylogenetic tree using the gene-rearrangement data inferred by PhyloBayes with CAT model

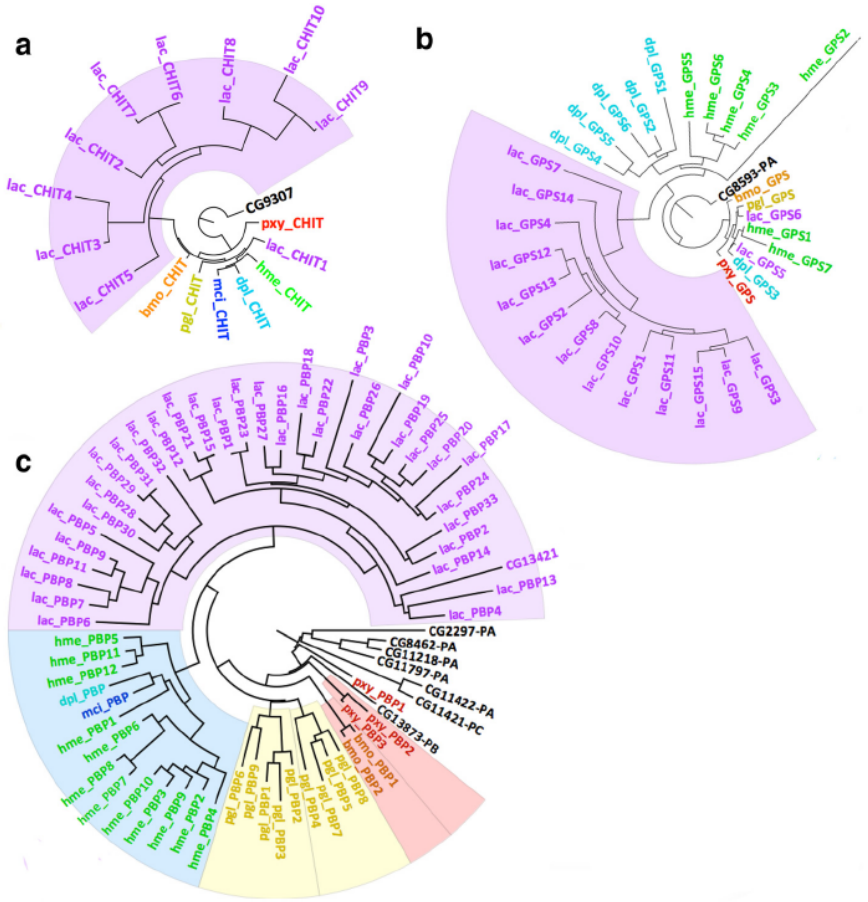


Fig. 5 Phylogenetic trees for expanded protein families in *Lac*. Abbreviation of the species and protein names are used as labels in the phylogenetic trees. We colored the labels to indicate which species the protein is from: *Lac* (*purple*), *Pgl* (*dark yellow*) *Dpl* (*cyan*), *Hme* (*green*), *Mci* (*blue*), *Bmo* (*orange*), *Pxy* (*red*) and *Drosophila melanogaster* (*black*). The clades corresponding to the unique gene expansion events in *Lac* are highlighted in light magenta. **a** Phylogenetic tree of endochitinases. **b** Phylogenetic tree of Geranylgeranyl pyrophosphate synthases. **c** Phylogenetic tree of Pheromone-binding proteins.

REFERENCES

1. Heliconius Genome C: **Butterfly genome reveals promiscuous exchange of mimicry adaptations among species.** *Nature* 2012, **487:**94-98.

2. Zhan S, Merlin C, Boore JL, Reppert SM: **The monarch butterfly genome yields insights into long-distance migration.** *Cell* 2011, **147:**1171-1185.

3. You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, et al: **A heterozygous moth genome provides insights into herbivory and detoxification.** *Nat Genet* 2013, **45:**220-225.

4. Kunte K, Zhang W, Tenger-Trolander A, Palmer DH, Martin A, Reed RD, Mullen SP, Kronforst MR: **doublesex is a mimicry supergene.** *Nature* 2014, **507:**229-232.

5. Nishikawa H, Iijima T, Kajitani R, Yamaguchi J, Ando T, Suzuki Y, Sugano S, Fujiyama A, Kosugi S, Hirakawa H, et al: **A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly.** *Nat Genet* 2015, **47:**405-409.

6. Ackery PR, de Jong R, Vane-Wright RI: **The butterflies: Hedyloidea, Hesperioidea and Papilionoidae.** *Handbook of Zoology A Natural History of the phyla of the Animal Kingdom* 1999, **IV Arthropoda: Insecta:**263-300.

7. Wahlberg N, Braby MF, Brower AV, de Jong R, Lee MM, Nylin S, Pierce NE, Sperling FA, Vila R, Warren AD, Zakharov E: **Synergistic effects of combining morphological and molecular data in resolving the phylogeny of butterflies and skippers.** *Proc Biol Sci* 2005, **272:**1577-1586.

8. Kawahara AY, Breinholt JW: **Phylogenomics provides strong evidence for relationships of butterflies and moths.** *Proc Biol Sci* 2014, **281:**20140970.

9. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30:**1312-1313.

10. Lartillot N, Lepage T, Blanquart S: **PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating.** *Bioinformatics* 2009, **25:**2286-2288.

11. Lartillot N, Philippe H: **A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process.** *Mol Biol Evol* 2004, **21:**1095-1109.

12. International Silkworm Genome C: **The genome of a lepidopteran model insect, the silkworm Bombyx mori.** *Insect Biochem Mol Biol* 2008, **38:**1036-1045.

13. Tang W, Yu L, He W, Yang G, Ke F, Baxter SW, You S, Douglas CJ, You M: **DBM-DB: the diamondback moth genome database.** *Database (Oxford)* 2014, **2014:**bat087.

14. Zhan S, Reppert SM: **MonarchBase: the monarch butterfly genome database.** *Nucleic Acids Res* 2013, **41:**D758-763.

15. Duan J, Li R, Cheng D, Fan W, Zha X, Cheng T, Wu Y, Wang J, Mita K, Xiang Z, Xia Q: **SilkDB v2.0: a platform for silkworm (Bombyx mori) genome biology.** *Nucleic Acids Res* 2010, **38:**D453-456.

16. Cong Q, Borek D, Otwinowski Z, Grishin NV: **Tiger Swallowtail Genome Reveals Mechanisms for Speciation and Caterpillar Chemical Defense.** *Cell Rep* 2015.

17. Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Valimaki N, Paulin L, Kvist J, Wahlberg N, et al: **The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera.** *Nat Commun* 2014, **5:**4737.

18. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23:**1061-1067.

19. Macossay-Castillo M, Kosol S, Tompa P, Pancsa R: **Synonymous constraint elements show a tendency to encode intrinsically disordered protein segments.** *PLoS Comput Biol* 2014, **10:**e1003607.

20. Knight RD, Shimeld SM: **Identification of conserved C2H2 zinc-finger gene families in the Bilateria.** *Genome Biol* 2001, **2:**RESEARCH0016.

21. Thomas JH, Schneider S: **Coevolution of retroelements and tandem zinc finger genes.** *Genome Res* 2011, **21:**1800-1812.

22. Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, Albu M, Weirauch MT, Radovani E, Kim PM, et al: **C2H2 zinc finger proteins greatly expand the human regulatory lexicon.** *Nat Biotechnol* 2015.

23. Patterson C, William DM, Humpries CJ: **Congruence of Morphological and Molecular Phylogenies.** *Annual Review of Ecology and Systematics* 1993, **24:**153-188.

24. Pisani D, Benton MJ, Wilkinson M: **Congruence of morphological and molecular phylogenies.** *Acta Biotheor* 2007, **55:**269-281.

25. Jekely G, Paps J, Nielsen C: **The phylogenetic position of ctenophores and the origin(s) of nervous systems.** *Evodevo* 2015, **6:**1.

26. Talavera G, Vila R: **What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny.** *BMC Evol Biol* 2011, **11:**315.

27. 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.

28. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV: **ECOD: an evolutionary classification of protein domains.** *PLoS Comput Biol* 2014, **10:**e1003926.

29. Yi P, Han Z, Li X, Olson EN: **The mevalonate pathway controls heart formation in Drosophila by isoprenylation of Ggamma1.** *Science* 2006, **313:**1301-1303.

30. Van Nieuwerburgh F, Thompson RC, Ledesma J, Deforce D, Gaasterland T, Ordoukhanian P, Head SR: **Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination.** *Nucleic Acids Res* 2012, **40:**e24.

31. Chevreux B, Wetter T, Suhai S: **Genome Sequence Assembly Using Trace Signals and Additional Sequence Information.** *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics* 1999, **99:**45-56.

32. Marcais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics* 2011, **27:**764-770.

33. Kelley DR, Schatz MC, Salzberg SL: **Quake: quality-aware detection and correction of sequencing errors.** *Genome Biol* 2010, **11:**R116.

34. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, et al: **Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads.** *Genome Res* 2014, **24:**1384-1395.

35. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.

36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

37. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.** *Nat Protoc* 2013, **8:**1494-1512.

38. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol* 2013, **14:**R36.

39. Roberts A, Pimentel H, Trapnell C, Pachter L: **Identification of novel transcripts in annotated genomes using RNA-Seq.** *Bioinformatics* 2011, **27:**2325-2329.

40. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.** *Genome Biol* 2008, **9:**R7.

41. Smit AFA, Hubley R: **(**http://www.repeatmasker.org**) RepeatModeler Open-1.0.** 2008-2010.

42. Jurka J, Klonowski P, Dagman V, Pelton P: **CENSOR--a program for identification and elimination of repetitive elements from DNA sequences.** *Comput Chem* 1996, **20:**119-121.

43. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110:**462-467.

44. Smit AFA, Hubley R, Green P: **(**http://www.repeatmasker.org**) RepeatMasker Open-3.0.** 1996-2010.

45. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, et al: **Annotation of the Drosophila melanogaster euchromatic genome: a systematic review.** *Genome Biol* 2002, **3:**RESEARCH0083.

46. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6:**31.

47. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics* 2007, **23:**1282-1288.

48. She R, Chu JS, Uyar B, Wang J, Wang K, Chen N: **genBlastG: using BLAST searches to build homologous gene models.** *Bioinformatics* 2011, **27:**2141-2143.

49. Stanke M, Schoffmann O, Morgenstern B, Waack S: **Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources.** *BMC Bioinformatics* 2006, **7:**62.

50. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5:**59.

51. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.** *Bioinformatics* 2004, **20:**2878-2879.

52. Besemer J, Borodovsky M: **GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses.** *Nucleic Acids Res* 2005, **33:**W451-454.

53. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Res* 2008, **18:**188-196.

54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

55. UniProt C: **Activities at the Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2014, **42:**D191-198.

56. St Pierre SE, Ponting L, Stefancsik R, McQuilton P, FlyBase C: **FlyBase 102--advanced approaches to interrogating FlyBase.** *Nucleic Acids Res* 2014, **42:**D780-788.

57. 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.

58. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13:**2178-2189.

59. Zhong YF, Holland PW: **HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology.** *Evol Dev* 2011, **13:**567-568.

60. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol* 2013, **30:**772-780.

61. Frickey T, Lupas A: **CLANS: a Java application for visualizing protein families based on pairwise similarity.** *Bioinformatics* 2004, **20:**3702-3704.

62. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43:**491-498.

63. Walsh I, Martin AJ, Di Domenico T, Tosatto SC: **ESpritz: accurate and fast prediction of protein disorder.** *Bioinformatics* 2012, **28:**503-509.

64. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100:**9440-9445.

65. Supek F, Bosnjak M, Skunca N, Smuc T: **REVIGO summarizes and visualizes long lists of gene ontology terms.** *PLoS One* 2011, **6:**e21800.

66. Gascuel O: **BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data.** *Mol Biol Evol* 1997, **14:**685-695.