ToxSci Advance Access published June 30, 2016

Running Title: Cross-Species Extrapolation of Chemical Susceptibility

Title: Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS): A web-based tool for addressing the challenges of cross-species extrapolation of chemical toxicity

[*,1]Carlie A. LaLone, [*]Daniel L. Villeneuve, [†]David Lyons, [‡]Henry W. Helgen, [§,2]Serina L. Robinson, [¶]Joseph A. Swintek, [*]Travis W. Saari, [*]Gerald T. Ankley

[*]U.S. Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Mid-Continent Ecology Division, 6201 Congdon Blvd., Duluth, MN 55804, USA

[†]U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology, Research Triangle Park, NC 27711, USA

[‡]CSC Government Solutions LLC, A CSRA Company, 6201 Congdon Blvd., Duluth, MN 55804, USA

[§]Saint Olaf College; Department of Chemistry; Northfield, MN 55057, USA

[¶]Badger Technical Services, 6201 Congdon Blvd., Duluth, MN 55804, USA

[1]To whom correspondence should be addressed at US EPA Mid-Continent Ecology Division, 6201 Congdon Blvd, Duluth, MN 55804. Fax: (218)529-5003. E-mail: lalone.carlie@epa.gov.

[2]Present Address: University of Minnesota – Twin Cities; Department of Microbiology & Immunology; Minneapolis, MN 55455, USA

Carlie A. LaLone: LaLone.Carlie@epa.gov
Daniel L. Villeneuve: Villeneuve.Dan@epa.gov
Dave Lyons: Lyons.Dave@epa.gov
Henry W. Helgen: henry.helgen@csra.com
Serina L. Robinson: robi0916@umn.edu
Joseph A. Swintek: Swintek.Joe@epa.gov
Travis W. Saari: Saari.Travis@epa.gov
Gerald T. Ankley: Ankley.Gerald@epa.gov

Abstract

Conservation of a molecular target across species can be used as a line-of-evidence to predict the likelihood of chemical susceptibility. The web-based Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS; https://seqapass.epa.gov/seqapass/) application was developed to simplify, streamline, and quantitatively assess protein sequence/structural similarity across taxonomic groups as a means to predict relative intrinsic susceptibility. The intent of the tool is to allow for evaluation of any potential protein target while remaining amenable to variable degrees of protein characterization, in the context of available information about the chemical/protein interaction and the molecular target itself. To accommodate this flexibility in the analysis, three levels of evaluation were developed. The first level of the SeqAPASS analysis compares primary amino acid sequences to a query sequence, calculating a metric for sequence similarity (including detection of orthologs); the second level evaluates sequence similarity within selected functional domains (e.g., ligand-binding domain); and the third level of analysis compares individual amino acid residue positions of importance for protein conformation and/or interaction with the chemical upon binding. Each level of the SeqAPASS analysis provides additional evidence to apply toward rapid, screening-level assessments of probable cross species susceptibility. Such analyses can support prioritization of chemicals for further evaluation, selection of appropriate species for testing, extrapolation of empirical toxicity data, and/or assessment of the cross-species relevance of adverse outcome pathways. Three case studies are described herein to demonstrate application of the SeqAPASS tool: the first two focused on predictions of pollinator susceptibility to molt-accelerating compounds and neonicotinoid insecticides, and the third on evaluation of cross-species susceptibility to strobilurin fungicides.

These analyses illustrate challenges in species extrapolation and demonstrate the broad utility of SeqAPASS for risk-based decision making and research.

Keywords: Sequence similarity; methoxyfenozide; tebufenozide; neonicotinoids; strobilurins; pollinator susceptibility

Introduction

Human and environmental risk assessments for chemicals, by necessity, utilize only a limited number of animal models to generate toxicity test data, which are subsequently extrapolated to species of concern. For ecological assessments this can involve extrapolation of effects from a few (perhaps one) species to many thousands. Substantial challenges exist in extrapolating available toxicity data across species due to variables related to molecular, cellular, anatomical, life-stage, and life-history differences. While it would be optimal to conduct an adequate level of in vivo testing to explicitly address these types of differences, this is not feasible for most chemicals. With decreasing testing resources, an international interest in reducing animal use, and an ever-increasing demand to evaluate more chemicals in a timely manner, predictive approaches that maximize the use of existing data are required.

In 2007 the National Research Council (NRC) outlined a vision and strategy for toxicity testing in the 21st century (National Research Council, 2007). The premise of this vision was to shift focus away from whole-animal tests evaluating apical outcomes as the primary means of assessing potential for adverse effects. Instead the committee advocated focusing on the chemicals ability to initiate toxicity through perturbation of biological pathways (National Research Council, 2007). Predictive in silico, and in vitro methods (where appropriate) were emphasized as future directions for toxicity testing.

This shifting paradigm has subsequently led to a number of ongoing activities, including an internationally-harmonized effort to develop adverse outcome pathways (AOPs), which describe the linkage of a molecular initiating event (defined as the point of chemical/biomolecule interaction that initiates a biological perturbation) through key events at various levels of biological organization (molecular, cellular, tissue, organ levels) to an adverse outcome (e.g., reduced fecundity, mortality) relevant to risk assessment at the individual or population level (http://www.oecd.org/env/ehs/testing/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm (Ankley *et al.*, 2010; Organisation of Economic Cooperation and Development, 2013). The AOP framework captures existing knowledge and outlines the weight of evidence supporting the linkages between key events (key event relationships) that can be used for identifying molecular or cellular indicators that are predictive of potential to cause adverse apical responses, thereby facilitating use of mechanistic data and focused testing strategies for risk-based decision making.

High-throughput screening (HTS) efforts, such as US Environmental Protection Agency's ToxCast™ Program, have started to put the NRC's vision to the test and have generated pathway-specific data on thousands of chemicals using a suite of automated in vitro assays based largely on mammalian systems (Kavlock *et al.*, 2012). Effective use of pathway-based (e.g., AOP) approaches and HTS data for risk assessment requires an understanding of cross-species extrapolation of chemical effects. Fortunately, advances and efficiencies in gene/protein sequencing techniques, genomics, transcriptomics, proteomics, and metabolomics, have facilitated innovative strategies for addressing the challenge of species extrapolation, particularly in ecotoxicology. These technologies allow for more comprehensive comparisons between species at molecular and biochemical levels than formerly achievable. Further,

computational tools, that incorporate modern bioinformatic approaches, permit rapid evaluation of large and complex data-sets as a means to address focused research and regulatory questions.

Previously we described a computational method for the assessment of primary amino acid sequences, and a cursory evaluation of common conserved domains across species as a basis for predicting relative intrinsic susceptibility of different phyla to chemicals with known molecular targets (LaLone *et al.*, 2013a). Employing case studies with a human pharmaceutical, a veterinary drug, and a pesticide, sequence-based predictions of susceptibility were compared to empirical toxicity data, illustrating the value of understanding protein sequence conservation across species for such predictions. The previous work addressed in detail the limitations and challenges associated with protein sequence-based cross-species extrapolation, and established the foundation for the development of the Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) application presented herein. The current version of SeqAPASS has considerably expanded protein analysis capabilities compared to initial methods described by LaLone et al. (LaLone *et al.*, 2013b). Notably, the program now runs through a web-based user interface and includes a data visualization component to aid interpretation and presentation of results. Added features also allow for greater flexibility in performing the analyses and greater taxonomic and sequence resolution for susceptibility predictions.

Application of this tool is predicated on the recognition that conservation of a molecular target is but one key component for predicting susceptibility across species; there are multiple other contributing factors related to the exposure route, intensity (dose) and timing, as well as absorption, distribution, metabolism, and elimination that also play key roles in potential susceptibility. The SeqAPASS application does not account for these additional factors in

making predictions of susceptibility; rather, SeqAPASS results are intended to be integrated with other data sources to inform risk assessment or study design.

Our work with SeqAPASS complements existing and evolving efforts in both drug discovery and toxicology relating molecular target structure to biological activity. Receptor homology modeling, ligand docking, and molecular dynamic simulation studies are common in drug design and development (Antes, 2010; Carlson, 2002; Wieman *et al.*, 2004). These techniques also have been adapted to assess protein conservation across species in the context of environmental risk assessment. For example, to efficiently design target-specific assays for potential endocrine disrupting chemicals, McRobb et al. (2014) computationally evaluated ligand binding pocket similarity, including ligand contact strength fingerprints for 28 human protein targets across three small fish and two amphibian species, and proposed that there was sufficient cross-species similarity to support use of non-mammalian models for evaluating potential effects in human. In another computational study, Walker and McEldowney (2013) explored the utility of molecular docking experiments using homologs from six commonly studied ecotoxicologically-relevant model organisms and focused on the pharmaceuticals diclofenac, ibuprofen, and levonorgestrel interacting with cyclooxygenase 2 and the progesterone receptor. That work suggested that docking approaches offer valuable insights for the identification of conserved subpockets and amino acid residues important for the chemical/protein interaction across-species. In combination with expansion of protein structural knowledge from x-ray crystallography studies, three-dimensional homology models, and site-directed mutagenesis, these computational approaches can enhance predictions for the likelihood of a chemical to interact with a receptor/enzyme/transporter and therefore provide key information, in the context of species similarities/differences and potential for chemical susceptibility. Furthermore, within

the rapidly evolving science of computational protein sequence/structural analyses, new and improved databases, tools, and algorithms for protein comparisons continue to develop (e.g., BLASTp (Altschul *et al.*, 1990), InParanoid (Remm *et al.*, 2001), OrthoMCL (Li *et al.*, 2003), Clustal W (Larkin *et al.*, 2007), COBALT (Papadopoulos and Agarwala, 2007), SWISS-MODEL (Arnold *et al.*, 2006), Protein Data Bank (Berman *et al.*, 2000), Auto-Dock (Morris *et al.*, 2009), UniProtKB (Magrane and Consortium, 2011)); however, most require the user to knowledgeably link together a number of different complex analyses to understand conservation of protein targets for a few model organisms. Therefore, an important (and novel) objective in developing the SeqAPASS tool, was to make use of all available sequence information for as many species as possible, and simplify and streamline the process to transparently evaluate protein sequence similarity at multiple levels of organization.

Since initial work described by LaLone et al. (LaLone et al., 2013a), we have expanded upon the development of this in silico approach to species extrapolation, for example further integrating stepwise processes to examine protein similarity at the level of conserved functional domains and individual amino acid residue positions. This multi-level investigation of the protein sequence/structure can be adapted, at the discretion of the SeqAPASS user, to the goal of the analysis, degree of protein characterization, and available knowledge of the chemical/protein interaction. The SeqAPASS tool is focused on minimizing the complexity of protein sequence/structural comparisons for species extrapolation, making the process more rapid and less daunting for scientists and regulators, alike, as a means to guide research and inform risk assessments.

To demonstrate the capabilities and utility of the SeqAPASS tool (version 1.0) we describe case studies to predict pollinator susceptibility to common insecticides and aid in

hypothesis generation for focused research on the potential for toxic effects. Other envisioned applications for the SeqAPASS tool could include use in green chemistry design, AOP development, extrapolation of toxicity data, informed selection of appropriate test organisms for focused toxicity testing, and risk assessment.

Material and Methods

Sequence Alignment to Predict Across Species Susceptibility

The SeqAPASS application captures data from multiple well-established and publically-available National Center for Biotechnology Information (NCBI) platforms, tabulates database information, and performs sequence similarity calculations in a streamlined manner useful for consistent and transparent cross-species predictions that can be performed and interpreted by both novice and advanced users. The NCBI databases utilized include the protein database (http://www.ncbi.nlm.nih.gov/protein/), conserved domain database (CDD; http://www.ncbi.nlm.nih.gov/cdd/), and taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy/). The SeqAPASS tool also incorporates the protein Basic Local Alignment Search Tool (BLASTp; http://www.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download; using default settings including the BLOSUM-62 matrix) and the Constraint-based Multiple Alignment Tool (COBALT; http://www.ncbi.nlm.nih.gov/tools/cobalt/) alignment, with an automated extraction of information pertinent to the cross-species evaluation of susceptibility.

Query Formulation Stage: Molecular target identification and query selection (Figure 1a)

Each analysis conducted with SeqAPASS requires the selection of a query species (scientific or common name) and a query protein (molecular target based on chemical of interest). Although these steps may seem intuitive, there are some important considerations that

can help ensure that the results generated by SeqAPASS are appropriate and useful to a given assessment scenario. For example, during the query species formulation stage, if the intent is to determine cross-species susceptibility to a human pharmaceutical and the drug is not known to adversely affect other non-target organisms, then the therapeutic protein target for humans (*Homo sapiens*) would be an appropriate query sequence. However, if instead there is knowledge that the same human pharmaceutical causes toxicity in a fish species through the same molecular target as in humans, and the interest is in the potential for toxic effects in other fish, then the known susceptible fish species would serve as an appropriate query species/sequence. This situation occurred with a previous analysis conducted by our group, where a human pharmaceutical, spironolactone, was known to activate the androgen receptor (AR) in western mosquito fish (*Gambusia affinis*), causing adverse effects (LaLone et al., 2013b). Therefore, to make predictions of susceptibility to other small fish species, the western mosquito fish AR (not the human AR) was used as the query sequence to examine protein conservation.

Another example, common for pesticides, is when the chemical of interest is known to target numerous species. For this scenario, one approach is to arbitrarily select one of the targeted species to serve as the query species. In this regard, one may want to select target species for which the most complete sequence information and/or supporting toxicity data are available. Alternatively, if there is concern for a particular species (e.g., for pesticide effects, threatened or endangered species), a pragmatic approach would be to use the species of concern as the query species and compare its protein sequence with those species known to be susceptible (or resistant) to the pesticide of interest. Yet another application scenario involves extrapolation of specific toxicity data from one species to others. In this case, the species used to generate the empirical data should serve as the query species. Overall, there are a number of ways to logically

select the query species meaningful to the susceptibility question being posed, thus highlighting the need for well-considered problem formulation.

Another critical aspect of formulating a SeqAPASS query involves gathering information about the target protein. For example, SeqAPASS analyses can be conducted differently if the protein is made up of numerous subunits or has multiple isoforms. Here, one could evaluate all subunits from a select query species, or if various isoforms exist, could either query each isoform individually or gather knowledge about the isoform known to be most sensitive to the chemical of interest for the susceptibility prediction. Other necessary information would include an understanding of how well the protein has been characterized, e.g., which functional domains (if known) are important for interaction with the chemical(s) of interest, as well as whether tertiary structural information exists, like that gained from x-ray crystallography, homology models, site-directed mutagenesis studies, or from reported field resistance to a pesticide due to identified mutation(s).

There are many instances where the molecular target of a given chemical may be unknown; this is the case for most personal care products and the majority of industrial chemicals. When this occurs, it may be possible to use available in vitro data (e.g., from the USEPA ToxCast™ program (http://actor.epa.gov/dashboard/, (Kavlock et al., 2012)) to identify putative protein molecular targets that could be queried in SeqAPASS. Even when chemical-specific in vitro data do not exist, it may be possible to assign tentative molecular targets based on information for other structurally-similar chemicals that have been tested.

Although the approaches for analyzing protein similarity incorporated in the SeqAPASS tool are meant to be used by the general scientific community, it is anticipated that the more sophisticated ones' knowledge is about the protein and chemical/protein interaction of interest,

the more complete the evaluation will be for predicting cross-species susceptibility using SeqAPASS. Hence, thoughtful formulation of the susceptibility question to be answered and an understanding of the degree of characterization for the chemical protein target are important components of any analysis.

Level 1: Primary sequence alignment, common conserved domain count, and ortholog detection (Figure 1b)

The SeqAPASS tool was developed to make predictions of susceptibility for any protein target, regardless of how much or how little is known about the protein or protein/chemical interaction. Hence, there will be cases where primary amino acid sequence alignments will be the sole predictor of susceptibility. This is generally the case when limited or no information exists relative to functional domains or important tertiary structural features. Primary sequence analyses of proteins that have not been well characterized allow for use of existing knowledge, with the ability to expand upon an analysis as the science advances, though it must be recognized that at this level, a relatively higher degree of uncertainty is associated with the susceptibility predictions. Therefore, this first level of analysis may be most appropriately applied to screening-level assessments of probable susceptibility to a chemical that targets the protein without a great deal of phylogenetic resolution (e.g., vertebrates versus invertebrates). Methods for aligning primary amino acid sequences and identifying ortholog candidates across species are described in detail elsewhere (LaLone et al., 2013a). Briefly, the Level 1 evaluation for a given query species/protein is the alignment of available primary amino acid sequences to the query sequence and automated extraction of specific species, aligned protein(s) (i.e., hit protein(s)), and sequence alignment information. Included in the primary report from SeqAPASS is the NCBI protein accession (i.e., NCBI's protein ID), protein count (number of protein sequences known

for each species), NCBI taxonomy ID, the taxonomic lineage information for each species (based on availability of classification in NCBI's taxonomy database selecting one group, sorting by organism Class, Subclass, Superorder, Order, Suborder, Family, Genus, and finally Subgenus), species common name, protein name, BLASTp bit score, and SeqAPASS-calculated percent similarity (hit sequence bit score normalized to the query species (maximum) bit score and multiplied by 100). The full report from SeqAPASS includes other alignment output from BLASTp such as protein length, identity (same residue at the same position), positives (similar residue at the same position), and E-value (expect value). The default E-value for the alignments included in the SeqAPASS output has been set to exclude those ≥ 0.01, however, if desired this default can be re-assigned by the user.

When key conserved domains are not known or well-defined for a protein target/chemical of interest, the SeqAPASS algorithms compare all identified curated domains found in the NCBI conserved domain database for the query protein to all domains identified for the hit proteins, and reports the number of common domains. For some protein targets or chemical protein interactions it is unknown how many domains must be preserved to maintain function, so SeqAPASS only excludes hit proteins with zero common conserved domains from the overall analysis in the primary report.

Finally, in this first level of the SeqAPASS evaluation, ortholog candidates are identified using a reciprocal best hit BLAST (LaLone et al., 2013a; Tatusov et al, 1997). Briefly, each hit protein having sequence similarity to the original query protein is used to reciprocally query all proteins available for the original query species. If the original query protein is identified as the best match (or maintains the same bitscore), then the hit protein is considered an ortholog

candidate. However, if a different protein is identified as the best match then the hit protein is not an ortholog.

To support internally consistent cross-phyla predictions of potential susceptibility, we developed a method for setting a susceptibility "cut-off" that incorporated knowledge of evolutionary biology, integrating the assumption that similar function would be conserved among orthologs (i.e., sequences that diverged from a speciation event and tend to maintain similar function) (Altenhoff *et al.,* 2012). Susceptibility predictions at this level are incorporated into the SeqAPASS report using methods modified from LaLone et al (LaLone et al., 2013a). An additional automated evaluation was incorporated as a means to further exclude poorly aligned and less meaningful sequence comparisons from the analysis. Once ortholog candidates have been identified, SeqAPASS algorithms utilize R code (R Core Team, 2014) to plot the distribution of percent similarities calculated for each hit protein (Russom *et al.*, 2014). From this plot, critical points are identified, including the local minimums and maximums. The first local minimum percent similarity is set as the default value for predictions of susceptibility. However, based on existing knowledge of sensitive species from empirical toxicity data for a given chemical, the SeqAPASS user has the option to enter another local minimum from the density plot for setting a cut-off. SeqAPASS results are ranked and reported in order from the greatest to least percent similarity and identified as ortholog candidates (Y = yes; N = no). Using the ortholog candidate data, a susceptibility cut-off is automatically determined by identifying the first ortholog candidate at an equal or higher percent similarity than the first local minimum percent similarity (or that local minimum selected by the SeqAPASS user). In recent work we utilized this method as a means to define the taxonomic domain of applicability for an AOP describing acetylcholinesterase inhibition leading to acute mortality (Russom et al., 2014).

Level 2: Across species sequence comparisons of functional domains (Figure 1c)

For many human therapeutic molecular targets, as well as some pesticide targets, the functional domains involved in the chemical interaction, such as the ligand binding domain (LBD), have been defined. If this type of knowledge is available, the SeqAPASS user can query the next level of the protein sequence comparison to evaluate those specific domains. The SeqAPASS algorithms automatically and specifically evaluate domains in the NCBI conserved domain database and extract the portions of the primary amino acid sequence that best align with the curated conserved domain for each sequence (and, therefore, species). SeqAPASS then compares the extracted domain sequences between the query and hit proteins using reverse position specific (RPS) BLAST (Marchler-Bauer *et al.*, 2011), producing alignment output, such as E-values and bit scores that are used to calculate a percent similarity at the domain level for each query/hit protein pair.

The same process described above for setting the susceptibility "cut-off" can be applied at this level of the analysis as well. Here, the ortholog candidate at an equal or higher percent similarity than the first local minimum (or user defined minimum) is automatically identified as the "cut-off". These data can be incorporated as further evidence for susceptibility (or lack thereof) at the conserved domain level and provide additional taxonomic resolution, potentially applicable at lower ranks of the taxonomic hierarchy (e.g., class, order, family).

Level 3: Individual amino acid residue position queries across species (Figure 1d)

The SeqAPASS tool was developed to employ all potentially useful protein sequence/structural knowledge relative to chemical interactions. However, due to our desire to minimize the complexity of SeqAPASS evaluations and extrapolate toxicity information across a diversity of species, we did not focus on developing methodology to assess homology models,

conduct docking studies, or evaluate 3D structures, though, as the science evolves, future versions of SeqAPASS may explore the utility of such methods for cross species comparisons. Instead, we focused on developing a method for comparisons across species at individual amino acid residue positions shown to be important for chemical interaction and/or optimal protein binding conformation (e.g., identified through x-ray crystallography, docking models, and/or site-directed mutagenesis studies). Evaluation of the available literature for the chemical(s) of interest, including knowledge about how it/they bind(s) to the protein target can be conducted while formulating a SeqAPASS query. These data, if available, can then be incorporated into susceptibility evaluations. It is at this level of sequence/structural detail that individual species differences, or even differences among strains or subpopulations within a species, are most likely to occur and contribute to innate differences in susceptibility.

For this level of analysis, the NCBI COBALT is incorporated to align multiple user selected sequences and request individual amino acid positions from a defined template protein sequence (Papadopoulos and Agarwala, 2007). In some cases key amino acid residue positions may be identified specifically for the query species used in Level 1. However, if the literature provides information based on a different species/sequence, then that sequence can be used as the template for identifying amino acid residue positions for the evaluation. When setting up a Level 3 SeqAPASS evaluation, the template sequence can be entered as an NCBI protein accession or in FASTA format and aligned with selected hit sequences. Following this alignment, the SeqAPASS user can then select which amino acid residue(s) at specified position(s) in the template sequence are to be aligned with the hit sequences. The output from SeqAPASS displays the aligned amino acid and position in each hit sequence and whether the amino acid(s) is an exact match or not between the template and hit proteins. The output also

reports the NCBI protein accession, protein name, species scientific name and taxonomic information.

The simple, logical, and transparent multi-level format of the SeqAPASS application provides flexibility in the protein analysis depending on available knowledge and intended applications of the comparative data. We anticipate that evaluations of cross-species susceptibility will continue to evolve as more protein sequence data become available for diverse species and as more protein structural studies are published, making these highly informative individual amino acid residue position queries more common. Further, the overarching goal of this project is the integration of all components of the evaluation to the single, openly accessible SeqAPASS interface.

SeqAPASS user guide and data visualization using an R workspace

SeqAPASS programming details, user guide, and additional documentation can be found in Supplemental Data, SeqAPASS S1 and is available upon accessing the SeqAPASS application 'Home' page. Briefly, the tool was built with a SQL database and Java PrimeFaces user interface. Further, a data visualization method and tutorial was developed to complement the SeqAPASS tool. The method utilizes an R workspace and the ggplot2 package (Wickham, 2009; R Core Team, 2014) to generate box-plots and customizable views that enhance interpretation of data contained in the SeqAPASS reports (Supplemental Data, SeqAPASS S2).

Results and Discussion

SeqAPASS for species extrapolation

LaLone et al. (2013a) showed that primary amino acid sequence comparisons across species, together with identification of ortholog candidate sequences, and a consideration of the shared number of functional domains was useful for predicting relative intrinsic susceptibility to

drugs and a pesticide with well-defined molecular targets. The current capabilities of the SeqAPASS application substantially expand upon that introductory method. To demonstrate application of the more comprehensive multi-level approach described herein for assessing potential species differences in susceptibility related to target protein structure, we employ three case studies that illustrate SeqAPASS capabilities for addressing some of the challenges associated with species extrapolation. These examples focus on protein sequence/structural comparisons to predict susceptibility of beneficial insects to molt-accelerating pesticides, potential honey bee susceptibility to neonicotinoid insecticides, and demonstrate how SeqAPASS data can be evaluated to develop focused research hypotheses relevant to cross-species susceptibility to strobilurin fungicides. Although these basic approaches are relevant to any class of chemicals, here we chose to focus on pesticides of current regulatory relevance.

Case example 1: Susceptibility of beneficial insects to molt-accelerating compounds

Background

Diacylhydrazines, including methoxyfenozide, were discovered in the 1980s, followed by the commercial analog, tebufenozide (Carlson et al., 2001). Methoxyfenozide and tebufenozide have been registered in the U.S. for use on a number of crops, including fruit trees, berries, nut trees, vines, vegetables, pastures, and cotton (The Dow Chemical Company, 2014; USEPA Label Review ID: EPA-HQ-OPP-2008-0824-0039). With the ability of these insecticides to act effectively and selectively to control Lepidopteran pests, chemical manufacturer Rohm and Haas Co. was awarded a 'Presidential Green Chemistry Award' from the U.S. Government in 1998 (https://www.epa.gov/greenchemistry, accessed April 2016).

Query formulation: Protein characterization and targeted species

Diacylhydrazine (DAH) and bisacylhydrazine (BAH) chemicals (e.g., methoxyfenozide and tebufenozide, respectively), also known as molt–accelerating compounds, act as nonsteroidal agonists of the ecdysteroid receptor (EcR) (Amor et al., 2012). Molting is a necessary process for normal growth and development in insects and other invertebrates. This process is regulated by 20-hydroxyecdysone (20E) binding to a heterodimer of ecdysone receptor (EcR) and ultraspiracle protein (USP), leading to regulation of ecdysone complex responsive genes which play a role in cessation of feeding, restricted mobility, removal of old exoskeleton (apolysis), escape behaviors for emergence from old exoskeleton, and synthesis of new exoskeleton (Carlson et al., 2001). Methoxyfenozide and tebufenozide are potent mimics of the natural hormone 20E, binding to the EcR with an affinity 400 or 70 times greater than that of the endogenous ligand, respectively (Carlson 2000 and Carlson et al., 2001). These synthetic chemicals have been recognized for their specific toxicity to larval Lepidopteran pests, such as armyworm, budworm, moth, and corn borer and low toxicity to non-targeted organisms including vertebrates and invertebrates such as honey bees, earthworms, and crustaceans (Carlson et al., 2001). A number of reasons for the specificity of these chemicals to Lepidopterans have been suggested, including how well the chemical is absorbed, distributed, metabolized, and excreted. However a major focus has been placed on the specific binding of these chemicals to the EcR ligand binding domain and key amino acid residues. Therefore, in the context of the SeqAPASS analysis, querying all possible Lepidopteran target insects for cross-species sequence similarity could be performed. However this volume of data would likely be challenging to interpret for transparent predictions of susceptibility. Therefore, the tobacco budworm (*Heliothis virescens*), from the Lepidopteran class, EcR was selected as the query sequence (NCBI protein accession CAA70212.1) for the SeqAPASS evaluation because this

species was consistently identified as a target pest both known to be sensitive to these chemicals and for which, crystal structure data exist for the ligand bound receptor and heterodimer complex.

Level 1: Primary amino acid sequence comparisons

The initial evaluation of the primary amino acid sequence using the default settings placed the susceptibility cut-off at 17.3%, with included vertebrates from Actinopteri (bony fish), Lepidosauria (lizards, snakes), Mammalia (mammals), Amphibia (frogs, salamander, newt), Chondrichthyes (shark, skate), Testudines (turtle), and Aves (bird) taxonomic groups with alignments to other nuclear receptors (e.g., oxysterols receptor LXR, thyroid hormone receptor, estrogen receptor) but not the EcR. Empirical data and phylogenetic evidence suggest that the EcR is present only in invertebrate species and that invertebrates are more sensitive than vertebrates to molt-accelerating compounds, providing evidence to justify adjustment to the default susceptibility cut-off settings (Carlson et al., 2001; Bonneton et al., 2003). Therefore, the 2nd local minimum was selected from the density plot, identifying the next ortholog candidate and placing the susceptibility cut-off at 33.9%. The Level 1 SeqAPASS evaluation predicted that species of the Insecta (insects), Malacostraca (prawn, lobster, crab), Branchiopoda (water flea), Arachnida (spider, tick, scorpion), Merostomata (horseshoe crabs), Maxillopoda (copepod), and Chilopoda (centipedes) classes would likely be susceptible to chemicals known to act on the tobacco budworm EcR (**Figure 2a**).

Level 2: Functional domain analysis

For the Level 2 assessment of functional domain(s), the NCBI CDD displays two specific hits for the tobacco budworm EcR, the ligand binding domain (accession cd06938) and the DNA-binding domain (accession cd07161). The EcR ligand binding domain (LBD) was selected

for cross species evaluation as both methoxyfenozide and tebufenozide have been shown to interact via the LBD (Carmichael *et al*., 2005; Amor et al., 2012). From this evaluation, selecting the 2nd local minimum for the same reason described above and therefore placing the susceptibility cut-off at 45.8, SeqAPASS data predict that species within the Insecta, Merostomata, Malacostraca, Arachnida, Branchiopoda, Chilopoda, Maxillopoda, and Priapulidae (marine worm) classes would likely be susceptible to chemicals that interact with the LBD of the tobacco budworm (**Figure 2b**).

Level 3: Individual amino acid residue comparison

The greatest taxonomic resolution for species specific susceptibility predictions can be gleaned from evaluation of similarities and differences at the individual amino acid residue level. This can be critical information in discerning the potential differential susceptibility of more closely-related species, for example, in this case Lepidopterans considered to be pests versus those identified as beneficial species. Crystal structures have been reported for the tobacco budworm EcR complexed with the natural ligand pontasterone A and the nonsteroidal BAH, BYI06830 (Billas et al., 2003). From comparisons between the structures complexed with these different ligands, valine 384 (V384) was identified as essential for species sensitivity to BAH and therefore was selected as an amino acid residue of interest for Level 3 evaluation. Therefore, NCBI accession O18473.1 was used as the template sequence where V402 aligned with V384 from Billas et al. (2003) (Table 1).

Amino acid residues linked to species insensitivity to a given chemical can also provide valuable insights for predicting susceptibility (or lack thereof). BAH insecticides are ineffective against species from the Hemiptera (cicada, aphid, planthopper, leafhopper) order, which include

the sweet potato whitefly (*Bemisia tabaci*), a pest insect, and the Insidious flower bug (*Orius laevigatus*) a beneficial insect that preys on agricultural pests (Carmichael et al., 2005).

The crystal structure of the sweet potato whitefly EcR ligand binding domain heterodimer has been elucidated and was compared that of the tobacco budworm (sensitive pest Lepidopteran), identifying residues that contribute to the pocket surface of the LBD which lead to a different architecture in the insensitive whitefly, severing the pocket for BAH binding (Carmichael et al., 2005). The residues that contribute to the altered conformation of the Hemiptera (*B. tabaci*) LBD include histidine 200 (H200), threonine 304 (T304), methionine 389 (M389), threonine 393 (T393), and valine 404 (V404) (Carmichael et al., 2005), all of which were subsequently evaluated using SeqAPASS (Table 1).

Toxicity assays exposing Insidious flower bug to molt-accelerating compounds, including methoxyfenozide and tebufenozide, showed that this Hemiptera insect is insensitive to these types of chemicals (Amor et al., 2012). Exposure methoxyfenozide or tebufenozide did not cause decreased survival or sublethal effects on reproductive parameters (Amor et al., 2012). Follow-up homology modeling of the flower bug EcR LBD and docking studies with both chemicals revealed steric hindrance caused by isoleucine 55 (I55), essentially preventing interaction with tebufenozide (Amor et al., 2012). Further, upon comparing the ligand binding domains between flower bug (insensitive) and select Lepidopterans, lysine 48 (K48), phenylalanine 222 (F222), and glutamine 227 (Q227) were identified in the Hemiptera as structurally different, restricting the LBD cavity leading to a steric clash upon docking methoxyfenozide or tebufenozide (Amor et al., 2012). These individual residue positions linked to species insensitivity were aligned across taxa using SeqAPASS (Table 1).

Level 3 evaluation of the ten residues identified from the available literature across species indicate that V384 was conserved only in species of the Lepidoptera order, including beneficial Lepidopterans, such as butterflies including the common yellow swallowtail (*Papilio machaon*), Asian swallowtail (*Papilio xuthus*), monarch (*Danaus plexippus*), squinting bush brown (*Bicyclus anynana*), and buckeye (*Junonia coenia*). Those residues associated with the insensitivity of species in the Hemiptera order were consistently not found in species of the Lepidoptera order, though inconsistently present among Hymenopterans (bees) and representative species from the Arachnida, Branchiopoda, Actinopteri, Mammalia, Amphibia, and Aves taxonomic lineages.

Overall the evaluation of conservation of the primary amino acid sequence, ligand binding domain, and all 10 individual residues indicate that EcR is conserved among Lepidopterans, including butterflies. These data provide a line-of-evidence to suggest that further toxicity testing should be conducted to better understand the possible adverse effects of molt-accelerating compounds on beneficial Lepidopterans. Further, with several butterfly species currently listed as threatened or endangered (U.S. Fish & Wildlife Service, March 28, 2016; http://www.fws.gov/endangered/), the SeqAPASS evaluation can provide valuable insights as to the potential chemical susceptibility of such species lacking empirical toxicity test data.

Case example 2: Predicting pollinator susceptibility to insecticides

Background

Neonicotinoids are neuroactive insecticides, similar in structure to nicotine, that act as nicotinic acetylcholine receptor (nAChR) agonists, effectively blocking normal cholinergic synaptic transmission, leading to death in numerous plant-sucking pests (Figure 3a) (Dani and Bertrand, 2007; Tomizawa and Casida, 2001; Tomizawa and Casida, 2003). Examples of these

insecticides include imidacloprid, clothianidin, and thiamethoxam, which are registered worldwide, making up approximately 17% of the global insecticide market (Jeschke and Nauen, 2008). Neonicotinoids are commonly used to protect important crop plants such as cereals, cotton, grain, legumes, fruits, and vegetables (Elbert *et al.*, 2008). These chemicals also are used in veterinary medicine as insecticides (Genchi *et al.*, 2000; Rust, 2005). One reason for the popularity of neonicotinoids is their selective toxicity toward insects as opposed to vertebrates. There are a number of causes for their selective nature, including cross-species differences in physicochemical and steric interactions at the nAChR (Ihara *et al.*, 2007).

Query formulation: Protein characterization and targeted species

The molecular target of neonicotinoids, the nAChR, is a highly complex pentameric protein consisting of non-α and/or α heteromers and α homomer subunits, each with four transmembrane domains and an extracellular domain containing a LBD (Jones and Sattelle, 2010). From affinity labeling studies it is known that α subunits contain a pair of adjacent cysteines that form a disulphide bond and a series of conserved aromatic residues important for binding the natural ligand, acetylcholine (Jones and Sattelle, 2010). A number of key amino acid residues have been identified that partially explain the selective nature of neonicotinoids and their preferential binding affinity to the insect nAChR compared to mammalian nAChRs (Matsuda *et al.*, 2000; Tomizawa and Casida, 2003). Although full crystal structures are not available for the insect nAChR, there has been valuable structural characterization of the extracellular portion of the nAChR from crystal structures of acetylcholine-binding proteins from the mollusk, *Aplysia californica* (high neonicotinoid sensitivity), and the snail, *Lymnaea stagnalis* (exhibiting lower neonicotinoid susceptibility), which share homology (Tomizawa *et al.*, 2008). Further structural insights, including identification of residues important for

interaction with common neonicotinoids such as imidacloprid, have been gained from chicken or rat (vertebrates; relatively insensitive to neonicotinoids) and Drosophila (insect; highly sensitive to neonicotinoids) nAChR subunit hybrids and site directed mutagenesis studies (Dederer *et al.*, 2013; Shimomura *et al.*, 2004).

Neonicotinoids target aphids, flies, beetles, weevils, termites, stink bugs, and leafhoppers, but it is suspected that they also could affect non-target pollinator species such as bees (Elbert et al., 2008). As described with the previous example in the context of the SeqAPASS analysis, querying all possible targeted insects, and examining all subunits for sequence similarity could be performed. However the data would likely be daunting to interpret. To better focus the analysis, due to interest in understanding possible pollinator susceptibility to neonicotinoids, we formulated and conducted our SeqAPASS query using the honey bee (*Apis mellifera*) nAChR subunits for query sequences. From the NCBI protein database, seven α subunits (α1-4, α6-8), some with alternative splice variants (from α4 and α6 subunits), and two β subunits (β1 and β2) were identified for the honey bee nAChR and subsequently utilized in SeqAPASS.

Level 1: Primary amino acid sequence comparisons

Level 1 evaluation of primary amino acid sequences across species showed that, overall, there was substantial similarity across invertebrate species to most honey bee nAChR subunits. Specifically, SeqAPASS data indicated that Insecta, Malacostraca, Arachnida, Merostomata, and Branchiopoda classes shared the greatest sequence similarity for the majority of the honey bee nAChR subunits including, α1, α2, α3, α4, α4', α8, and β1 (Supplemental Data, Figure S1); however, from examination of the data, acetylcholinesterase receptor and cholinergic receptor subunits usually had the greatest percent similarity across non-insect species (Supplemental Data, Table S1). To address our specific question pertaining to potential honey bee susceptibility

to neonicotinoids, we focused on the comparison of primary amino acid sequence similarity of honey bee nAChR subunits within the class Insecta (Figure 4a). These data illustrate a relatively large degree of conservation of the honey bee nAChR subunits with those of other insects, including several pests targeted by neonicotinoid chemicals (e.g., aphids, flies, cockroaches). The mean percent similarity for α1-4 subunits, α6 variants, α7-8 subunits and β1 subunit ranged between 34.7 – 47.2%, with the majority of ortholog candidates identified with similarity of >28.1%. The primary amino acid sequence data also illustrate that the honey bee α9 and β2 subunits are rather unique, sharing limited similarity with other insects, consistent with the findings of Jones et al. (2006), whom indicated that honey bee is only the second insect known to have more than one non-α (β2) subunit. Most insects possess at least one divergent subunit, having limited sequence similarity to all other nAChR subunits. This type of cross-species knowledge could be relevant in terms of designing insecticides that have enhanced effectiveness for targeted pests, yet are safe for beneficial insects.

Level 2: Functional Domain Analysis

Level 2 evaluation of nAChR subunits for prediction of potential honey bee susceptibility to neonicotinoids focused on comparisons of sequence similarity of LBDs. The NCBI conserved domain database identified two domains for each nAChR subunit, including the neurotransmitter-gated ion-channel LBD (extracellular region; pfam02931) and the neurotransmitter-gated ion-channel transmembrane region (forms the ion channel; pfam02932). A number of studies describe the extracellular LBD as the key region of neonicotinoid binding (Jones and Sattelle, 2010; Shimomura *et al.*, 2003; Tomizawa et al., 2008; Toshima *et al.*, 2009), hence the focus on this domain for the susceptibility predictions. As anticipated, the SeqAPASS Level 2 data indicated that Insecta, Malacostraca, Arachnida, Merostomata, and

Branchiopoda classes shared the greatest sequence similarity for the majority of the honey bee nAChR subunits including, α1, α2, α3, α4, α4', α8, and β1 (Supplemental Data, Figure S2). In specifically comparing the honey bee LBDs for each nAChR subunit for sequence similarity across the class Insecta, there was an overall increase in percent similarity relative to that observed through evaluation of the primary amino acid sequence (Figure 4b). For all nAChR subunits, excluding α9 and β2, the mean percent similarity amongst insects was greater than 49.3%. Consistent with the primary amino acid sequence analysis, several insects targeted by neonicotinoid insecticides shared high sequence similarity at the LBD with the honey bee, providing further evidence of potential susceptibility.

Level 3: Individual amino acid residue comparison

Level 3 analysis of sequence similarity requires knowledge of the chemical/protein interaction, preferably for otherwise closely-related species. A number of studies have demonstrated key amino acid residues in the extracellular LBD seemingly important for interactions with neonicotinoids (Matsuda et al., 2000; Shimomura *et al.*, 2005; Shimomura *et al.*, 2006; Shimomura et al., 2004; Shimomura et al., 2003; Toshima et al., 2009). The present case study focuses on an example demonstrating how this information can be used to inform susceptibility predictions using SeqAPASS.

Evidence suggests that ticks, as well as related arachnids are largely insensitive to neonicotinoids (McCall *et al.*, 2004). Erdmanis et al. (2012) conducted a sequence analysis focused on the β1 nAChR subunit for selected insect species including the black-legged tick (*Ixodes scapularis*). The β1 subunit had previously been reported to be associated with the development of adaptive resistance to a common neonicotinoid, imidacloprid, in aphid (*Myzus persicae*), and seemingly had resulted from an arginine (positively charged) to threonine

(uncharged) substitution (Bass *et al.*, 2011). Erdmanis et al. (2012) found that within the loop D regions (loops A-C and D-F are conserved regions of the extracellular LBD of adjacent subunits) of nAChR β1 subunits of select insect species, a conserved arginine residue exists; however, at that position for two arachnid species, the black-legged tick and wolf spider (*Pardosa pseudoannulata*), an uncharged glutamine residue is instead present. Further, alignment of loop D nAChR β1 subunits across multiple tick species demonstrated conservation of the identified uncharged glutamine residue (Erdmanis et al., 2012). And, when structures of nAChRs complexed with imidacloprid were modeled, it was found that the arginine residue could optimally interact electrostatically with the negatively charged areas of the neonicotinoid, while the uncharged glutamine could not (Erdmanis et al., 2012).

The study by Erdmanis et al. provides the necessary data for a Level 3 SeqAPASS analysis. We specifically compared Arachnida and Insecta classes at each Level (i.e., 1, 2, and 3) of the protein analysis (Figure 5 and Table 2). For the Level 3 protein comparison, Insecta sequences that shared sequence similarity to the honey bee β1 subunit were simultaneously aligned with the black-legged tick (NCBI accession XP_002406474.1) glutamine residue at position 81 (Table 2). All Insecta species selected for the alignment had an arginine residue that aligned with the tick glycine residue present in the nAChR β1 subunit. The insecta species included the honey bee and other beneficial insects (e.g., common eastern bumble bee (*Bombus impatiens*), domestic silkworm (*Bombyx mori*), monarch butterfly), and insecticide-targeted pests such as the red flour beetle (*Tribolium castaneum,*), flies (e.g., *Drosophila melanogaster*); desert locust (*Schistocerca gregaria*), brown planthopper (*Nilaparvata lugens*), aphids (e.g., *Myzus persicase*), budworm (*Heliothis virescens*), armyworm (*Spodoptera exigua*), and American cockroach (*Periplaneta americana*) (Table 2).

Although there are other examples of variations in key individual residues in nAChR LBDs across subunits (Matsuda et al., 2000; Shimomura et al., 2005; Shimomura et al., 2006; Shimomura et al., 2004; Shimomura et al., 2003; Toshima et al., 2009), this case study demonstrates the type of analysis that can be conducted using SeqAPASS to explore relative intrinsic susceptibility using available literature. As new studies are published and more sequence information becomes available, greater evidence could be collected to evaluate predictions based on SeqAPASS results. Overall, after analyzing various levels of protein sequence/structural similarity, SeqAPASS data predict that honey bees have the potential to be susceptible to neonicotinoid insecticides. Specifically, Level 1 analysis revealed broad taxonomic conservation of nAChR, with invertebrate sequences most similar to the honey bee; Level 2 data showed high sequence similarity of the LBD across insects, including those targeted by neonicotinoids, compared to the honey bee; and, finally, the Level 3 analysis indicated that key residues involved in the chemical/protein interaction of target insects were also conserved in bee species. Taken together these data can be used as a line-of-evidence for extrapolation of AOP information (Figure 3b) or in a risk assessment, particularly when species-specific (e.g., honey bee) empirical toxicity data are lacking.

Case example 3: Predicting cross-species susceptibility to strobilurin fungicides

Query formulation: Protein characterization and targeted species

Strobilurins, originally discovered in mushrooms, have become an important class of agricultural fungicides widely used to treat a number of plant pathogens in crops (Balaba, 2007). For example, strobilurins protect cereals, nuts, fruit and vegetable crops, grapevines, and turfgrass against soil-borne and foliar pathogens including mildew, leaf mold, leaf spot, rusts and anthracnose (Bartlett et al., 2002). Azoxystrobin has been identified as the world's leading

compound for controlling plant diseases (www.syngenta.com, accessed April 2016) and is registered in over 70 countries (Rodrigues *et al.*, 2013).

The molecular target of strobilurins is cytochrome b, a mitochondrial enzyme involved in cellular respiration. Specifically, these chemicals bind the quinol-oxidizing site (Q_o), blocking electron transport between cytochrome b and cytochrome c1, thereby inhibiting ATP synthesis, halting energy production, and killing the fungal pathogen (Balba, 2007; Bartlett *et al.*, 2002). Due to the central role of cytochrome b in cellular respiration for eukaryotes, this mitochondrial gene has been sequenced for many taxa and, in fact, has commonly been used to determine phylogenetic relationships between organisms (Castresana, 2001). Consequently, cytochrome b is one of the well-represented proteins, across species, in the NCBI database.

Level 1 and Level 2 sequence comparisons

To focus the SeqAPASS query for the identification of potential non-target species that may be susceptible to strobilurin fungicides we used the pathogen, common corn rust (*Puccinia sorghi*), as a query species and cytochrome b as the query protein (NCBI protein accession ABB54707.1). This protein is identified as a partial sequence in the NCBI database and, therefore, is not an ideal query protein in this sense. However, the coverage of the corn rust sequence includes the ubiquinol-binding site of importance to the action of strobilurins, which is the most critical aspect for our analysis. The NCBI conserved domain database identifies two specific domain hits for the cytochrome b enzyme, including the N-terminus/b6/petB (cd00284) and the C-terminus/b6/petD (cd00290) domains. Due to the role of the N-terminus in forming the ubiquinol/ubiquinone and heme binding sites, this domain was selected for the SeqAPASS Level 2 functional domain analysis.

As anticipated, due to the ubiquitous role of cytochrome b in eukaryotic cellular respiration, data from the Level 1 (primary amino acid sequence) and Level 2 (conserved domain) sequence comparisons across species revealed high conservation of the mitochondrial enzyme (Supplemental Data, Figure S3a and b). All protein sequences sharing sequence similarity with common corn rust cytochrome b were identified as ortholog candidates. Further, greater sequence similarity at the N-terminus domain, which includes the ubiquinol binding site, was identified across species compared to the primary amino acid sequences as a whole. From these analyses, SeqAPASS data indicate that all species, representing 110 taxonomic groups, would be predicted as potentially susceptible to strobilurin fungicides. However, strobilurin fungicides have been reported to have low toxicity to a variety of non-target plants, birds, mammals, earthworms, and certain insects (Bartlett et al., 2002), so there are clearly factors not reflected in the Level 1 and 2 analyses that contribute to the differences in measured toxicity for these species.

Level 3: Individual amino acid residue comparison

To conduct a Level 3 SeqAPASS protein comparison across species, we examined instances where mutations in the cytochrome b gene led to field resistance in target species. Specifically, it has been reported that a single point mutation in several agriculturally important plant pathogens (GGT or GGA [sensitive, wild-type] to GCT or GCA [resistant, mutant]) in the cytochrome b gene, resulting in a glycine to alanine substitution at position 143, confers resistance to strobilurin fungicides (Grasso *et al.*, 2006; Ishii *et al.*, 2001; Ishii *et al.*, 2007; Kim *et al.*, 2003). Another mutation, phenylalanine to leucine (F129L), at position 129 has been implicated in field resistance to strobilurins as well (Kim et al., 2003). Therefore, positions 129 and 143 were assessed using SeqAPASS methods, aligning all cytochrome b sequences from the

fungal classes Pucciniomycetes (which includes the query species *P. sorghi*), Exobasidiomycetes, Ustilaginomycetes, Eurotiomycetes, and several organism classes reported to have low sensitivity (Diaz-Espejo *et al.*, 2012; Tomizawa and Casida, 2003; Zhang *et al.*, 2010) to strobilurins, including two plant classes, Rosids and Liliopsida, Mammalia (top 300 species sharing sequence similarity in Level 1 SeqAPASS analysis), Actinopterygii (bony fish; top 300 species sharing sequence similarity in Level 1 SeqAPASS analysis), Aves (top 300 species sharing sequence similarity in Level 1 SeqAPASS analysis), Amphibia, and Insecta (Supplemental Data, Tables S2-14). All cytochrome b sequences from both the fungal and non-sensitive organism classes contained the wild-type (strobilurin sensitive) glycine aligned with position 143 and phenylalanine aligned with position 129, (with the exceptions of the fungus, *Histoplasma capsulatum G186AR*, which did not align at position 129 and tassel rope-rush (*Baloskion tetraphyllum*), having alanine aligning at position 143). Overall, all the taxa examined shared the two residues, again supporting the prediction of likely susceptibility for multiple species known to be relatively insensitive to strobilurins.

To further probe the available cytrochrome b sequence information and generate testable hypotheses as to possible sequence differences that could indicate strobilurin susceptibility, the NCBI COBALT was used to compare sequences and SeqAPASS Level 3 was used to align residues of interest across species. Using COBALT, the strobilurin sensitive query species, *P. sorghi*, cytochrome b was aligned with all cytochrome b sequences from the top ten most similar (from SeqAPASS Level 1 analysis) fungi classes (Pucciniomycetes, Exobasidiomycetes, Microbotryomycetes, Ustilaginomycetes, Malasseziomycetes, Agaricomycetes, Eurotiomycetes, Pneumocystidomycetes, and Tremellomycetes, and Dothideomycetes) (Supplemental Data, Table S15). From this alignment, 25 residues were completely conserved amongst all fungal

species assessed. These 25 conserved amino acid residues were then aligned with all cytochrome b sequences from insensitive plants (i.e., members of the Rosids and Liliopsida classes). This alignment yielded two amino acid residues that were unique to fungi species, a leucine and a threonine residue, found at positions 193 and 232, respectively, when using the *P. sorghi* cytochrome b sequence as the template (Supplemental Data, Table S16). Empirical data indicate that daphnids vary in sensitivity to strobilurin fungicides due to clonal variation (Warming *et al.*, 2009), with 48-hr median lethal concentrations (LC50) generally between 16 -14000 µg/L, depending on the fungicide analyzed (Bartlett et al., 2002). Therefore, since daphnids have relatively low sensitivity to strobilurins, the two residue positions that were unique upon comparison of fungi to plants were aligned with Branchiopoda species (including *D. pulex*) cytochrome b sequences, which resulted in the identification of one residue, threonine (T) at position 232 as unique to fungi. Species from the class Branchiopoda maintain a conserved valine (V) or methionine (M) aligning with leucine at position 193. Because valine (having a similar side chain to leucine) aligns with the Branchiopoda species that show limited sensitivity, it is unlikely that position L193 found in fungi is linked to strobilurin susceptibility. Threonine on the other hand was unique to fungi species when compared to sequences from Branchiopoda. Therefore, SeqAPASS was used to align threonine at position 232 with other species known to have variable or low sensitivity to strobilurin fungicides (such as the fish, amphibians, birds, and mammals) (Table 3). Interestingly for all other taxonomic groups evaluated in Level 3, a glycine residue aligned with threonine at position 232. Overall, this analysis leads to the hypothesis that T232 may be a key amino acid residue in determining cross-species susceptibility to strobilurin fungicides. This observation could be used to focus toxicity testing or further, guide site-directed mutagenesis experiments (G231T substitution; possibly with daphnids) to evaluate the

hypothesis that this amino acid residue may be of importance in modulating toxicity to strobilurins.

In summary, the strobilurin example demonstrates how empirical evidence can contribute to the SeqAPASS analysis, when, for example, predicted susceptibility based on Level 1 or 2 analyses does not correlate with measured toxicity across species. In these instances, it is possible that other aspects leading to susceptibility, such as differential pharmacokinetics or life-history considerations, may be stronger determinants of sensitivity across species. Alternatively, as illustrated above, SeqAPASS also could allow for exploration of sequence information at a level that would allow for hypothesis generation relative to cross-species sequence/structural differences in molecular targets.

Application of SeqAPASS data

The three case studies were presented as a means to demonstrate the process of strategically conducting SeqAPASS analyses at different levels of protein complexity, appropriate to problem formulation, and illustrate how empirical data complement the susceptibility predictions, to both guide the analyses and develop testable hypotheses for enhanced evaluations of chemical toxicity. These examples, along with those previously described, focusing on predicting susceptibility across taxa to a human pharmaceutical, veterinary drug, and common insecticide (LaLone et al., 2013a), provide the foundational work for further evaluating the utility of the SeqAPASS application in cross-species extrapolation. We anticipate that SeqAPASS capabilities and the data derived will continue to expand, for example relative to AOP development and cross-species extrapolation of HTS data. Recently, Russom et al. (2014) developed an AOP for activation of acetylcholinesterase leading to acute mortality, demonstrating how knowledge of molecular target conservation at the level of the molecular

initiating event can aid in defining the taxonomic domain of applicability for the AOP. That study also demonstrated strong correlation between protein sequence similarity and species sensitivity to organophosphates and carbamates known to act via acetylcholinesterase inhibition, providing further evidence for the utility of sequence-based methods for predicting susceptibility. Consistent with the approach described herein, McRobb et al. (2014), recently demonstrated the importance of considering conserved functional domains, such as binding pockets within ligand binding domains, which focus the comparative protein analysis on the site of the chemical/protein interaction (when such knowledge is available) for identifying conservation across species. The recognition that evaluation of functional domains provides a higher degree of resolution when considering cross-species conservation of chemical protein targets illustrates the value of the newly incorporated conserved domain analysis in SeqAPASS. Overall, as the SeqAPASS tool continues to evolve, for example with the addition of Level 3 evaluations to the automated pipeline, the tool enables rapid evaluation of protein sequence/structural information. This greatly expands the application of the ever increasing amount of such data for purposes of species extrapolation relevant to chemical toxicity. For example, from January to March 2016, the NCBI protein database increased by +4.5 million sequences (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/). Increasingly, molecular sequence/structural data are being applied to intelligently design chemicals (e.g., pesticides, drugs). Our utilization of these data are pertinent to understanding potential adverse effects across species, particularly where limited empirical data exist. However, it is essential to continually engage in research efforts that aid in defining the domain of applicability for sequence-based predictions of susceptibility. One goal in creating the SeqAPASS application and making it publically accessible, is to provide a foundation for the broader scientific community

to further assess protein sequence/structure-based predictions of chemical susceptibility. Further, an aspiration of future work is to engage researchers whom use or develop in vitro receptor binding, transcriptional activation, or enzyme activity assays across species, to develop focused comparisons of such data to susceptibility predictions, and iteratively refine and/or elaborate upon new or existing protein comparison methods to test and improve predictions. It is our hope that the SeqAPASS application will continue to evolve with the science and bioinformatic capabilities to meet the needs of researchers and regulators for use in species extrapolation.

Tables

Table 1.Tobacco budworm ecdysone receptor Level 3 analysis of individual amino acid residue positions across species

Insect Classification	Scientific Name	Common Name	Order	Protein Name	[b]Q331	[c]P353	[c]M360	[a]V402	[b]V434	[b]Q521	[b]M525	[c]I527	[c]K532	[b]L536
Pest	*Heliothis virescens*	tobacco budworm	Lepidoptera	Ecdysone receptor	Q	P	M	V	V	Q	M	I	K	L
Pest	*Helicoverpa armigera*	cotton bollworm	Lepidoptera	ecdysone receptor	Q	P	M	V	V	Q	M	I	K	L
Pest	*Agrotis ipsilon*	black cutworm moth	Lepidoptera	EcR B-like protein	Q	P	M	V	V	Q	M	I	K	L
Pest	*Spodoptera littoralis*	African cotton leafworm	Lepidoptera	ecdysone receptor	Q	P	M	V	V	Q	M	I	K	L
Pest	*Spodoptera exigua*	beet armyworm	Lepidoptera	ecdysone receptor	Q	P	M	V	V	Q	M	I	K	L
Pest	*Spodoptera litura*	moths	Lepidoptera	ecdysone receptor	Q	P	M	V	V	Q	M	I	K	L
Pest	*Amyelois transitella*	moths	Lepidoptera	PREDICTED: ecdysone receptor isoform X1	Q	P	M	V	V	Q	M	I	K	L
Pest	*Scirpophaga incertulas*	moths	Lepidoptera	ecdysone receptor	Q	P	M	V	V	Q	M	I	K	L
Pest	*Chilo suppressalis*	striped riceborer	Lepidoptera	ecdysone receptor B1 isoform	Q	P	M	V	V	Q	M	I	K	L
Pest	*Omphisa fuscidentalis*	bamboo borer	Lepidoptera	ecdysone receptor B1 isoform	Q	P	M	V	V	Q	M	I	K	L
Beneficial	*Papilio machaon*	common yellow swallowtail	Lepidoptera	PREDICTED: ecdysone receptor isoform X1	Q	P	M	V	V	Q	M	I	K	L
Beneficial	*Papilio xuthus*	Asian swallowtail	Lepidoptera	PREDICTED: ecdysone receptor isoform X1	Q	P	M	V	V	Q	M	I	K	L
Pest	*Plodia interpunctella*	Indianmeal moth	Lepidoptera	ecdysone receptor	Q	P	M	V	V	Q	M	I	K	L
Beneficial	*Danaus plexippus*	monarch butterfly	Lepidoptera	ecdysteroid receptor EcR-B	Q	P	M	V	V	Q	M	I	K	L
Pest	*Plutella xylostella*	diamondback moth	Lepidoptera	ecdysteroid receptor	Q	P	M	V	V	Q	M	I	K	L
Pest	*Choristoneura fumiferana*	eastern spruce budworm	Lepidoptera	ecdysteroid receptor EcR-B	Q	P	M	V	V	Q	M	I	K	L
Pest	*Ectropis obliqua*	moths	Lepidoptera	ecdysteroid receptor	Q	P	M	V	V	Q	M	I	K	L
Pest	*Grapholitha molesta*	moths	Lepidoptera	ecdysone receptor	Q	P	M	V	V	Q	M	I	K	L
Pest	*Manduca sexta*	tobacco hornworm	Lepidoptera	Ecdysone receptor	Q	P	M	V	V	Q	M	I	K	L
Beneficial	*Bombyx mori*	domestic silkworm	Lepidoptera	EcRB1	Q	P	M	V	V	Q	M	I	K	L
Pest	*Operophtera brumata*	winter moth	Lepidoptera	Ecdysone receptor A	Q	P	M	V	V	Q	M	I	K	L
Pest	*Sesamia nonagrioides*	Mediterranean corn borer	Lepidoptera	ecdysone receptor	Q	P	M	V	V	Q	M	I	K	L
Beneficial	*Bicyclus anynana*	squinting bush brown	Lepidoptera	ecdysteroid receptor	Q	P	M	V	V	Q	M	I	K	L
Beneficial	*Junonia coenia*	buckeye	Lepidoptera	ecdysteroid receptor	Q	P	M	V	V	Q	M	I	K	L
Beneficial	*Apis mellifera*	honey bee	Hymenoptera	ecdysteroid receptor A isoform	Q	K	I	M	T	Q	M	F	K	L
Beneficial	*Apis dorsata*	giant honeybee	Hymenoptera	PREDICTED: ecdysone receptor-like	Q	K	I	M	T	Q	M	F	K	L
Beneficial	*Apis florea*	little honeybee	Hymenoptera	PREDICTED: ecdysone receptor isoform X2	Q	K	I	M	T	Q	M	F	K	L
Beneficial	*Bombus terrestris*	buff-tailed bumblebee	Hymenoptera	PREDICTED: ecdysone receptor isoform X2	Q	K	I	M	T	Q	M	F	K	L
Beneficial	*Bombus impatiens*	common eastern bumble bee	Hymenoptera	PREDICTED: ecdysone receptor isoform X2	Q	K	I	M	T	Q	M	F	K	L
Beneficial	*Melipona quadrifasciata*	bees	Hymenoptera	Ecdysone receptor	Q	K	I	M	T	Q	M	F	K	L
Pest	*Polistes dominula*	European paper wasp	Hymenoptera	EcR	Q	R	I	M	T	Q	M	Y	K	L
Beneficial	*Habropoda laboriosa*	bees	Hymenoptera	Ecdysone receptor	Q	K	I	M	T	Q	M	F	K	L
Beneficial	*Orius laevigatus*	Insidious flower bug	Hemiptera	Amar et al. 2012	H	K	I	M	V	Q	M	F	Q	L
Pest	*Nezara viridula*	southern green stink bug	Hemiptera	ecdysone receptor isoform	H	R	I	M	V	Q	M	F	Q	L
Pest	*Nilaparvata lugens*	brown planthopper	Hemiptera	ecdysteroid receptor	H	R	I	T	V	Q	M	F	K	L
Pest	*Bemisia tabaci*	sweet potato white fly	Hemiptera	-	H	R	I	M	T	M	T	F	K	V
Beneficial	*Hydropsyche incognita*	caddisflies	Trichoptera	EcR	Q	P	I	M	I	R	M	I	K	L
Non-insect	*Daphnia magna*	crustaceans	Branchiopoda	ecdysone receptor A1	Q	K	M	C	T	L	M	F	K	L
Non-insect	*Ixodes scapularis*	black-legged tick	Arachnida	AamEcRA1, putative	S	R	I	G	S	M	M	F	Q	L
Non-insect	*Oreochromis niloticus*	Nile tilapia	Actinopteri	PREDICTED: oxysterols receptor LXR-alpha	-	R	L	T	F	V	Q	F	Q	L
Non-insect	*Xenopus (Silurana) tropicalis*	western clawed frog	Amphibia	oxysterols receptor LXR-beta	K	R	L	T	F	V	Q	F	Q	L
Non-insect	*Homo sapiens*	human	Mammalia	oxysterols receptor LXR-alpha isoform 4	R	R	L	T	F	V	Q	F	Q	L
Non-insect	*Picoides pubescens*	downy woodpecker	Aves	Oxysterols receptor LXR-alpha	Q	E	L	T	F	V	Q	F	Q	L

[a]Tobacco budworm (O18473.1) V402 (green text) aligned with V384 from Billas et al.

[b]Tobacco budworm (O18473.1) Q331, V434, Q521, M525, and L536 (blue text) aligned with *B. tabaci* H200, T304, M389, T393, and V404 from Carmichael et al., respectively

[c]Tobacco budworm (O18473.1) P353, M360, I527 and K532 (red text) aligned with *O. laevigatus* K48, I55, F222, and Q227 from Amar et al., respectively

Amino acid residues are represented by single letter acronyms: Q, Glutamine; P, Proline; M, Methionine; V, Valine; I, Isoleucine; K, Lysine; L, Leucine; T, Threonine; F, Phenylalanine; H, Histidine; S, Serine; R, Arginine; E, Glutamic Acid; C, Cysteine; G, Glycine.

Dashed line separates insect species from non-insect species

Table 2. Honey bee nAChR Level 3 SeqAPASS analysis of individual amino acid residue position across species

NCBI protein accession	Taxonomic group	Scientific Name	Common Name	Amino Acid Residue	Position
XP_002406474.1	Arachnida	[a]*Ixodes scapularis*	black-legged tick	Q	81
NP_001073028.1	Insecta	*Apis mellifera*	honey bee	R	80
ADH15831.1	Insecta	*Apis cerana cerana*	bees	R	79
XP_012347218.1	Insecta	*Apis florea*	little honeybee	R	114
XP_006611530.1	Insecta	*Apis dorsata*	giant honeybee	R	115
XP_003486051.1	Insecta	*Bombus impatiens*	common eastern bumble bee	R	79
XP_003393394.1	Insecta	*Bombus terrestris*	buff-tailed bumblebee	R	79
XP_003701924.1	Insecta	*Megachile rotundata*	alfalfa leafcutting bee	R	79
XP_011157846.1	Insecta	*Solenopsis invicta*	red fire ant	R	79
NP_001234897.1	Insecta	[b]*Nasonia vitripennis*	jewel wasp	R	80
NP_001103418.1	Insecta	*Tribolium castaneum*	red flour beetle	R	77
ETN60728.1	Insecta	*Anopheles darlingi*	American malaria mosquito	R	80
NP_523927.2	Insecta	[b]*Drosophila melanogaster*	fruit fly	R	81
XP_001660921.1	Insecta	*Aedes aegypti*	yellow fever mosquito	R	80
XP_001866461.1	Insecta	*Culex quinquefasciatus*	southern house mosquito	R	80
ABA39253.1	Insecta	[b]*Schistocerca gregaria*	desert locust	R	85
ACJ07013.1	Insecta	[b]*Nilaparvata lugens*	brown planthopper	R	81
NP_001166819.1	Insecta	*Bombyx mori*	domestic silkworm	R	89
XP_014356082.1	Insecta	*Papilio machaon*	common yellow swallowtail	R	80
AKV94624.1	Insecta	*Periplaneta americana*	American cockroach	R	87
EHJ77616.1	Insecta	*Danaus plexippus*	monarch butterfly	R	139
CAB87995.1	Insecta	[b]*Myzus persicae*	green peach aphid	R	81
AHJ11206.1	Insecta	*Locusta migratoria*	migratory locust	R	84
AAD09810.1	Insecta	[b]*Heliothis virescens*	tobacco budworm	R	80
ABU41521.1	Insecta	[b]*Spodoptera exigua*	beet armyworm	R	80
KDR22508.1	Insecta	*Zootermopsis nevadensis*	termites	R	48

[a]Ixodes scapularis selected as template sequence for comparison to sequences from Insecta (Erdmanis et al., 2012)

[b]Species targeted by neonicotinoid insecticides

Amino acid residues are represented by single letter acronyms: Q, Glutamine and R, Arginine

Table 3. Common corn rust cytochrome b Level 3 SeqAPASS analysis of individual amino acid residues across species: hypothesis generation

NCBI protein accession	Taxonomic group	Scientific Name	Common Name	Amino Acid Residue	Position
ABB54707.1	Pucciniomycetes	[a]*Puccinia sorghi*	basidiomycetes	T	232
ABB54704.1	Pucciniomycetes	*Puccinia horiana*	basidiomycetes	T	232
AAY67655.1	Pucciniomycetes	*Puccinia triticina*	wheat leaf rust	T	232
ABB54702.1	Pucciniomycetes	*Puccinia graminis f. sp. tritici*	basidiomycetes	T	232
ABB54701.1	Pucciniomycetes	*Puccinia coronata var. avenae f. sp. avenae*	basidiomycetes	T	232
AKJ25533.1	rosids	*Geranium macrorrhizum*	eudicots	G	238
AKJ25530.1	rosids	*Geranium brycei*	eudicots	G	238
AKJ25531.1	rosids	*Geranium endressii*	eudicots	G	238
AKJ25549.1	rosids	*Monsonia emarginata*	eudicots	G	238
YP_009137027.1	rosids	*Geranium maderense*	Madiera cranesbill	G	238
AFA35868.1	Liliopsida	*Haworthia retusa*	monocots	G	227
AFA35863.1	Liliopsida	*Aloe bulbillifera*	monocots	G	227
AFA35869.1	Liliopsida	*Kniphofia caulescens*	monocots	G	227
AFA35870.1	Liliopsida	*Trachyandra smalliana*	monocots	G	225
AFA35866.1	Liliopsida	*Bulbine alooides*	monocots	G	226
YP_004347612.1	Actinopteri	*Auriglobus modestus*	bronze puffer	G	231
YP_005970.1	Actinopteri	*Mola mola*	ocean sunfish	G	231
AHH25205.1	Actinopteri	*Lagocephalus inermis*	smooth blaasop	G	231
BAD17866.1	Actinopteri	*Dicologlossa cuneata*	wedge sole	G	231
AKO90264.1	Actinopteri	*Caesio cuning*	redbelly yellowtail fusilier	G	231
YP_009003495.1	Amphibia	*Caecilia gracilis*	Surinam caecilian	G	232
YP_009155367.1	Amphibia	*Rhacophorus dennysi*	Blanford's whipping frog	G	232
YP_007374585.1	Amphibia	*Caecilia volcani*	Cocle caecilian	G	232
AFV95394.1	Amphibia	*Paramesotriton chinensis*	Chinese warty newt	G	232
AKP94994.1	Amphibia	*Kaloula rugifera*	Szechuan narrowmouth toad	G	232
ADD25510.1	Aves	*Caprimulgus saturatus*	Dusky nightjar	G	232
ADD25500.1	Aves	*Caprimulgus nigriscapularis*	Black-shouldered nightjar	G	232
ABW72646.1	Aves	*Lanius collurio*	red-backed shrike	G	232
ADD25486.1	Aves	*Caprimulgus anthonyi*	Scrub nightjar	G	232
ADD25501.1	Aves	*Caprimulgus parvulus*	birds	G	232
NP_008632.1	Branchiopoda	*Daphnia pulex*	common water flea	G	231
NP_775076.1	Branchiopoda	*Triops cancriformis*	crustaceans	G	232
YP_009128474.1	Branchiopoda	*Streptocephalus sirindhornae*	crustaceans	G	231
YP_001054762.1	Mammalia	*Anomalurus sp. GP-2005*	rodents	G	231
AFY10058.1	Mammalia	*Petaurista elegans*	spotted giant flying squirrel	G	231
AAG24443.1	Mammalia	*Neotoma mexicana*	Mexican woodrat	G	231
ABA81947.1	Mammalia	*Neotoma isthmica*	rodents	G	231
NP_542242.1	Mammalia	*Tachyglossus aculeatus*	Australian echidna	G	231

[a]*Puccinia sorghi* selected as template sequence for comparison to selected sequences across taxonomic groups
Amino acid residues are represented by single letter acronyms: T, Threonine and G, Glycine

Figure Legends

Figure 1. Schematic of strategic approach for predicting cross-species susceptibility using the SeqAPASS tool. Beginning with the query formulation stage (a) and moving through each level of protein sequence/structural comparison. (b) Level 1 describes analysis for predictions based on primary amino acid sequence comparisons; (c) Level 2 provides a means to compare functional domain sequences; and (d) Level 3 methodology allows for evaluation of key individual amino acid residue positions across species. Portions of the diagram surrounded by the grey oval represent areas where information collected from the query formulation stage are essential to inform the analysis.

Figure 2. Boxplots depicting SeqAPASS data illustrating the percent similarity across species compared to tobacco budworm (*H. virescens*) ecdysone receptor (EcR) examining the primary amino acid sequences (a) and the ligand binding domain (b) ○ represents the tobacco budworm EcR and • represents the species with the highest percent similarity within the specified taxonomic group. The top and bottom of each box represent the 75^{th} and 25^{th} percentiles, respectively. The top and bottom whiskers extend up to 1.5 times the interquartile range. The mean and median values for each taxonomic group are represented by horizontal thick and thin black lines on the box, respectively. The dashed line indicates the cut-off for intrinsic susceptibility predictions (based on ortholog analysis).

Figure 3. Putative adverse outcome pathway for nicotinic acetylcholine receptor activation leading to mortality (a). For this putative AOP example, relevant to the SeqAPASS analysis, the chemical class has been defined with neonicotinoids and the taxonomic domain under consideration was limited to insect pests. SeqAPASS analyses can be used to extrapolate AOPs (b) based on conservation of the molecular initiating event (MIE). Dashed arrow represents the

weight of evidence provided by the various levels of the SeqAPASS analysis to extrapolate the MIE to honey bees.

Figure 4. Boxplots to describe SeqAPASS data illustrating the percent similarity of all sequences from species of the Insecta organism class compared to each honey bee (*Apis mellifera*) nicotinic acetylcholine receptor (nAChR) subunit examining both the primary amino acid sequences (a) and the ligand binding domain (b). The dashed line represents the honey bee nAChR subunits having 100% similarity when comparing the sequence to itself. The top and bottom of each box represent the 75th and 25th percentiles, respectively. The top and bottom whiskers extend up to 1.5 times the interquartile range. The mean and median values for each subunit percent similarity are represented by horizontal thick and thin black lines on the box, respectively.

Figure 5. Boxplots to describe SeqAPASS data illustrating the percent similarity of all sequences from the Insecta and Arachnida organism classes for the honey bee (*Apis mellifera*) nicotinic acetylcholine receptor (nAChR) beta 1 subunit examining both the primary amino acid sequences (a) and the ligand binding domain (b). ○ represents the honey bee nAChR beta 1 subunit and • represents the species with the highest percent similarity within Arachnida. The top and bottom of each box represents the 75th and 25th percentiles. The top and bottom whiskers extend up to 1.5 times the interquartile range. The mean and median values for each subunit percent similarity are represented by horizontal thick and thin black lines on the box, respectively.

Supplementary Data Description

Supplemental Data, Tables and Figures

Table S1. SeqAPASS output derived from queries of honey bee alpha 1, alpha 2, alpha 3, alpha 4, alpha 4', alpha 6 (contains exon 8a), alpha 6 (contains exon 8b), alpha 7, alpha 8, alpha 9, beta 1, and beta 2.

Tables S2-14. Level 3 SeqAPASS results aligning common corn rust cytochrome b phenylalanine (F) 129 and glycine (G) 143 to hit proteins from fungal classes (S2-7), plant classes (S8-9), Actinopteri (bony fish; S10), Amphibia (S11), Aves (S12), Mammalia (S13), and Insecta (S14).

Table S15. COBALT alignment of sequences from the top 10 (based on Level 1 percent similarity) fungi classes identifying 25 conserved amino acid residues.

Table S16. Level 3 SeqAPASS results aligning common corn rust cytochrome b phenylalanine (L) 193 and glycine (T) 232 to hit proteins from plant classes identifying 2 unique amino acid residues conserved in fungi.

Figure S1. Boxplots depicting SeqAPASS data illustrating the percent similarity across species compared to each of the honey bee (*Apis mellifera*) nicotinic acetylcholine receptor (nAChR) subunits examining the primary amino acid sequences. For each graph, ○ represents the honey bee nAChR subunit and • represents the species with the highest percent similarity within the specified taxonomic group. The top and bottom of each box represents the 75th and 25th percentiles, respectively. The top and bottom whiskers extend up to 1.5 times the interquartile range. The mean and median values for each subunit percent similarity are represented by horizontal thick and thin black lines on the box, respectively. The dashed line indicates the cut-off for intrinsic susceptibility predictions (based on ortholog analysis).

Figure S2. Boxplots depicting SeqAPASS data illustrating the percent similarity across species compared to each of the honey bee (*Apis mellifera*) ligand binding domains of the nicotinic acetylcholine receptor (nAChR) subunits examining the functional domain sequences. For each graph, ○ represents the honey bee nAChR subunit and • represents the species with the highest percent similarity within the specified taxonomic group. The top and bottom of each box represents the 75th and 25th percentiles. The top and bottom whiskers extend up to 1.5 times the interquartile range. The mean and median values for each subunit percent similarity are represented by horizontal thick and thin black lines on the box, respectively. The dashed line indicates the cut-off for intrinsic susceptibility predictions (based on ortholog analysis).

Figure S3. Boxplots depicting SeqAPASS data illustrating the percent similarity across species compared to corn rust (*P. sorghi*) cytochrome b examining the primary amino acid sequences (a) and the N-terminus quinol binding site domain (b). ○ represents the corn rust cytochrome b and • represents the species with the highest percent similarity within the specified taxonomic group. The top and bottom of each box represents the 75th and 25th percentiles, respectively. The top and bottom whiskers extend up to 1.5 times the interquartile range. The mean and median values for each subunit percent similarity are represented by horizontal thick and thin black lines on the box, respectively. The dashed line indicates the cut-off for intrinsic susceptibility predictions (based on ortholog analysis).

Supplemental Data, SeqAPASS

SeqAPASS S1. SeqAPASS user guide. Also available on 'Home' page of SeqAPASS application.

SeqAPASS S2. R workspace data visualization user guide. Also available on 'About' page of SeqAPASS application.

Acknowledgements – We thank B. Blackwell and J. Nichols for providing thoughtful review comments on an earlier version of the paper. We also thank T. Transue and C. Simmons from Lockheed Martin Environmental Modeling and Visualization Laboratory for creating the SeqAPASS v1.0 application. Further, we thank D. Lane, and S. Walata, for their computer programming expertise and advice in developing a prototype version of the SeqAPASS database and interface. Also, thanks to K. Nelson for her assistance on the development of an early draft of the SeqAPASS user guide. This manuscript has been reviewed in accordance with the requirements of the US EPA Office of Research and Development; however, the recommendations made herein do not represent US EPA policy. Mention of products or trade names does not indicate endorsement by the US EPA.

References

Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., and Dessimoz, C. (2012). Resolving the Ortholog Conjecture: Ortholgs Tend to Be Weakly, but Significantly, More Similar in Funtion than Paralogs. *PLoS Comput Biol* **8**(5), e1002514.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology* **215**(3), 403-10.

Amor, F., Christiaens, O., Bengochea, P., Medina, P., Rouge, P., Vinuela, E., Smagghe, G. (2012). Selectivity of diacylhydrazine insecticides to the predatory bug Orius laevigatus: in vivo and modelling/docking experiments. *Pest Manag Sci* **68**, 1586-94

Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., Mount, D. R., Nichols, J. W., Russom, C. L., Schmieder, P. K., et al. (2010). Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ Toxicol Chem* **29**(3), 730-41.

Antes, I. (2010). DynaDock: A new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. *Proteins* **78**(5), 1084-104.

Arnold, K., Bordoli, L., Kopp, J. and Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**(2), 195-201.

Balba, H. (2007). Review of strobilurin fungicide chemicals. *Journal of environmental science and health. Part. B, Pesticides, food contaminants, and agricultural wastes* **42**(4), 441-51.

Bartlett, D. W., Clough, J. M., Godwin, J. R., Hall, A. A., Hamer, M. and Parr-Dobrzanski, B. (2002). The strobilurin fungicides. *Pest management science* **58**(7), 649-62.

Bass, C., Puinean, A. M., Andrews, M., Cutler, P., Daniels, M., Elias, J., Paul, V. L., Crossthwaite, A. J., Denholm, I., Field, L. M., Foster, S. P., Lind, R., Williamson, M. S. and Slater, R. (2011). Mutation of a nicotinic acetylcholine receptor beta subunit is associated with resistance to neonicotinoid insecticides in the aphid Myzus persicae. *BMC neuroscience* **12**, 51.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**(1), 235-42.

Billas, I. M. L., Iwema, T., Garnier, J. M., Mitschler, A., Rochel, N., and Moras, D. (2003) Structural adaptability in the ligand-binding pocket of the ecdysone hormone receptor. *Nature* **426**, 91-96.

Bonneton, F., Zelus, D., Iwema, T., Robinson-Rechavi, M., and Laudet, V. (2003). Rapid Divergenc eof the Ecdysone Recptor in Diptera and Lepidoptera Suggests Coevolution Between ECR and USP_RXR. *Mol Biol Evol* **20**(4), 541-553.

Carlson, G. R. (2000) Tebufenozide: A Novel Caterpillar Control Agent with Unusually High Target Selectivity. Ch 2. In Green Chemical Sytheses and Processes. ACS Symposium Series; Americal Chemical Society : Washington, D.C.

Carlson, G. R., Dhadialla, T. S., Hunter, R., Jansson, R. K., Jany, C. S., Lidert, Z., and Slawecki, R. A. (2001). The chemical and biological properties of methoxyfenozide, a new insecticidal ecdysteroid agonist. *Pest Manag Sci* **57**, 115-119.

Carlson, H. A. (2002). Protein flexibility is an important component of structure-based drug discovery. *Current pharmaceutical design* **8**(17), 1571-8.

Carmichael, J. A., Lawrence, M. C., Graham, L. D., Pilling, P. A., Epa, V. C., Noyce, L., Lovrecz, G., Winkler, D. A., Pawlak-Skrzecz, A., Eaton, R. E., et al. (2005)The X-ray structure of a hemipteran ecdysone receptor ligand-binding domain. *J Biol Chem* **280**(23), 22258-22269

Castresana, J. (2001). Cytochrome b phylogeny and the taxonomy of great apes and mammals. *Molecular biology and evolution* **18**(4), 465-71.

Dani, J. A. and Bertrand, D. (2007). Nicotinic acetylcholine receptors and nicotinic cholinergic mechanisms of the central nervous system. *Annual review of pharmacology and toxicology* **47**, 699-729.

Dederer, H., Berger, M., Meyer, T., Werr, M. and Ilg, T. (2013). Structure-activity relationships of acetylcholine derivatives with Lucilia cuprina nicotinic acetylcholine receptor alpha1 and alpha2 subunits in chicken beta2 subunit hybrid receptors in comparison with chicken nicotinic acetylcholine receptor alpha4/beta2. *Insect molecular biology* **22**(2), 183-98.

Diaz-Espejo, A., Cuevas, M. V., Ribas-Carbo, M., Flexas, J., Martorell, S. and Fernandez, J. E. (2012). The effect of strobilurins on leaf gas exchange, water use efficiency and ABA content in grapevine under field conditions. *Journal of plant physiology* **169**(4), 379-86.

Elbert, A., Haas, M., Springer, B., Thielert, W. and Nauen, R. (2008). Applied aspects of neonicotinoid uses in crop protection. *Pest management science* **64**(11), 1099-105.

Erdmanis, L., O'Reilly, A. O., Williamson, M. S., Field, L. M., Turberg, A. and Wallace, B. A. (2012). Association of neonicotinoid insensitivity with a conserved residue in the loop d binding region of the tick nicotinic acetylcholine receptor. *Biochemistry* **51**(23), 4627-9.

Genchi, C., Traldi, P. G. and Bianciardi, P. P. (2000). Efficacy of imidacloprid on dogs and cats with natural infestations of fleas, with special emphasis on flea hypersensitivity. *Veterinary therapeutics : research in applied veterinary medicine* **1**(2), 71-80.

Grasso, V., Palermo, S., Sierotzki, H., Garibaldi, A. and Gisi, U. (2006). Cytochrome b gene structure and consequences for resistance to Qo inhibitor fungicides in plant pathogens. *Pest management science* **62**(6), 465-72.

Ihara, M., Shimomura, M., Ishida, C., Nishiwaki, H., Akamatsu, M., Sattelle, D. B. and Matsuda, K. (2007). A hypothesis to account for the selective and diverse actions of neonicotinoid insecticides at their molecular targets, nicotinic acetylcholine receptors: catch and release in hydrogen bond networks. *Invertebrate neuroscience : IN* **7**(1), 47-51.

Ishii, H., Fraaije, B. A., Sugiyama, T., Noguchi, K., Nishimura, K., Takeda, T., Amano, T. and Hollomon, D. W. (2001). Occurrence and molecular characterization of strobilurin resistance in cucumber powdery mildew and downy mildew. *Phytopathology* **91**(12), 1166-71.

Ishii, H., Yano, K., Date, H., Furuta, A., Sagehashi, Y., Yamaguchi, T., Sugiyama, T., Nishimura, K. and Hasama, W. (2007). Molecular Characterization and Diagnosis of QoI Resistance in Cucumber and Eggplant Fungal Pathogens. *Phytopathology* **97**(11), 1458-66.

Jeschke, P. and Nauen, R. (2008). Neonicotinoids-from zero to hero in insecticide chemistry. *Pest management science* **64**(11), 1084-98.

Jones, A. K., Raymond-Delpech, V., Thany, S. H., Gauthier, M. and Sattelle, D. B. (2006). The nicotinic acetylcholine receptor gene family of the honey bee, Apis mellifera. *Genome research* **16**(11), 1422-30.

Jones, A. K. and Sattelle, D. B. (2010). Diversity of Insect Nicotinic Acetylcholine Receptor Subunits. In *Insect Nicotinic Acetylcholine Receptors* (S. H. Thany, Ed.)^ Eds.), pp. 25-43. Landes Bioscience and Springer Science+Business Media.

Kavlock, R., Chandler, K., Houck, K., Hunter, S., Judson, R., Kleinstreuer, N., Knudsen, T., Martin, M., Padilla, S., Reif, D., et al. (2012). Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. *Chemical research in toxicology* **25**(7), 1287-302.

Kim, Y. S., Dixon, E. W., Vincelli, P. and Farman, M. L. (2003). Field Resistance to Strobilurin (Q(o)I) Fungicides in Pyricularia grisea Caused by Mutations in the Mitochondrial Cytochrome b Gene. *Phytopathology* **93**(7), 891-900.

LaLone, C. A., Villeneuve, D. L., Burgoon, L. D., Russom, C. L., Helgen, H. W., Berninger, J. P., Tietge, J. E., Severson, M. N., Cavallin, J. E. and Ankley, G. T. (2013a). Molecular target sequence similarity as a basis for species extrapolation to assess the ecological risk of chemicals with known modes of action. *Aquat Toxicol* **144-145**, 141-54.

LaLone, C. A., Villeneuve, D. L., Cavallin, J. E., Kahl, M. D., Durhan, E. J., Makynen, E. A., Jensen, K. M., Stevens, K. E., Severson, M. N., Blanksma, C. A., et al. (2013b). Cross-species sensitivity to a novel androgen receptor agonist of potential environmental concern, spironolactone. *Environ Toxicol Chem* **32**(11), 2528-41.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**(21), 2947-8.

Li, L., Stoeckert, C. J., Jr. and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13**(9), 2178-89.

Magrane, M. and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database : the journal of biological databases and curation* **2011**, bar009.

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* **39**(Database issue), D225-9.

Matsuda, K., Shimomura, M., Kondo, Y., Ihara, M., Hashigami, K., Yoshida, N., Raymond, V., Mongan, N. P., Freeman, J. C., Komai, K. and Sattelle, D. B. (2000). Role of loop D of the alpha7 nicotinic acetylcholine receptor in its interaction with the insecticide imidacloprid and related neonicotinoids. *Br J Pharmacol* **130**(5), 981-6.

McCall, J. W., Alva, R., Irwin, J. P., Carithers, D. and Boeckh, A. (2004). Comparative Efficacy of a Combination of Fipronil/(S)-Methoprene, a Combination of Imidacloprid/Permethrin, and Imidacloprid Against Fleas and Ticks When Administered Topically to Dogs. *J Appl Res Vet Med* **2**, 74-77.

McRobb, F. M., Sahagun, V., Kufareva, I. and Abagyan, R. (2014). In silico analysis of the conservation of human toxicity and endocrine disruption targets in aquatic species. *Environ Sci Technol* **48**(3), 1964-72.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. and Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry* **30**(16), 2785-91.

National Research Council. (2007). *Toxicity Testing in the 21st Century: A Vision and Strategy*. The National Academies Press, Washington, DC.

Organisation of Economic Cooperation and Development (2013). Guidance Document of Developing and Assessing Adverse Outcome Pathways. In, Ed.)^ Eds.), Vol. No. 184, pp. 45 pages. Organisation for Economic Cooperation and Development, Paris, France.

Papadopoulos, J. S. and Agarwala, R. (2007). COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics* **23**(9), 1073-9.

R Core Team. (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* Available at: http://www.R-project.org/. Accessed July 18 2014.

Remm, M., Storm, C. E. and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of molecular biology* **314**(5), 1041-52.

Rodrigues, E. T., Lopes, I. and Pardal, M. A. (2013). Occurrence, fate and effects of azoxystrobin in aquatic ecosystems: a review. *Environment international* **53**, 18-28.

Russom, C. L., LaLone, C. A., Villeneuve, D. and Ankley, G. T. (2014). Development of an Adverse Outcome Pathway for Acetylcholinesterase Inhibition Leading to Acute Mortality. *Environmental Toxicology and Chemistry*.

Rust, M. K. (2005). Advances in the control of Ctenocephalides felis (cat flea) on cats and dogs. *Trends in parasitology* **21**(5), 232-6.

Shimomura, M., Satoh, H., Yokota, M., Ihara, M., Matsuda, K. and Sattelle, D. B. (2005). Insect-vertebrate chimeric nicotinic acetylcholine receptors identify a region, loop B to the N-terminus

of the Drosophila Dalpha2 subunit, which contributes to neonicotinoid sensitivity. *Neuroscience letters* **385**(2), 168-72.

Shimomura, M., Yokota, M., Ihara, M., Akamatsu, M., Sattelle, D. B. and Matsuda, K. (2006). Role in the selectivity of neonicotinoids of insect-specific basic residues in loop D of the nicotinic acetylcholine receptor agonist binding site. *Molecular pharmacology* **70**(4), 1255-63.

Shimomura, M., Yokota, M., Matsuda, K., Sattelle, D. B. and Komai, K. (2004). Roles of loop C and the loop B-C interval of the nicotinic receptor alpha subunit in its selective interactions with imidacloprid in insects. *Neuroscience letters* **363**(3), 195-8.

Shimomura, M., Yokota, M., Okumura, M., Matsuda, K., Akamatsu, M., Sattelle, D. B. and Komai, K. (2003). Combinatorial mutations in loops D and F strongly influence responses of the alpha7 nicotinic acetylcholine receptor to imidacloprid. *Brain research* **991**(1-2), 71-7.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. A Genomic Perspective on Protein Families. *Science* **278**(5338), 631-37.

The Dow Chemical Company. (2014). Product Safety Assessment Methoxyfenozide. Form No. 233-00380-MM-0814X, 1-6.

Tomizawa, M. and Casida, J. E. (2003). Selective toxicity of neonicotinoids attributable to specificity of insect and mammalian nicotinic receptors. *Annual review of entomology* **48**, 339-64.

Tomizawa, M. and Casida, J. E. (2001). Structure and diversity of insect nicotinic acetylcholine receptors. *Pest management science* **57**(10), 914-22.

Tomizawa, M., Maltby, D., Talley, T. T., Durkin, K. A., Medzihradszky, K. F., Burlingame, A. L., Taylor, P. and Casida, J. E. (2008). Atypical nicotinic agonist bound conformations conferring subtype selectivity. *Proc Natl Acad Sci U S A* **105**(5), 1728-32.

Toshima, K., Kanaoka, S., Yamada, A., Tarumoto, K., Akamatsu, M., Sattelle, D. B. and Matsuda, K. (2009). Combined roles of loops C and D in the interactions of a neonicotinoid insecticide imidacloprid with the alpha4beta2 nicotinic acetylcholine receptor. *Neuropharmacology* **56**(1), 264-72.

Walker, S. D. and McEldowney, S. (2013). Molecular docking: a potential tool to aid ecotoxicity testing in environmental risk assessment of pharmaceuticals. *Chemosphere* **93**(10), 2568-77.

Warming, T. P., Mulderij, G. and Christoffersen, K. S. (2009). Clonal variation in physiological responses of Daphnia magna to the strobilurin fungicide azoxystrobin. *Environ Toxicol Chem* **28**(2), 374-80.

Wickham, H. (2009) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wieman, H., Tondel, K., Anderssen, E. and Drablos, F. (2004). Homology-based modelling of targets for rational drug design. *Mini reviews in medicinal chemistry* **4**(7), 793-804.

Zhang, Y., Zhang, X., Chen, C., Zhou, M. and Wang, H. (2010). Effects of fungicides JS399-19, azoxystrobin, tebuconazloe, and carbendazim on the physiological and biochemical indices and grain yield of winter wheat. *Pest Biochem Physiol* **98**, 151-157.

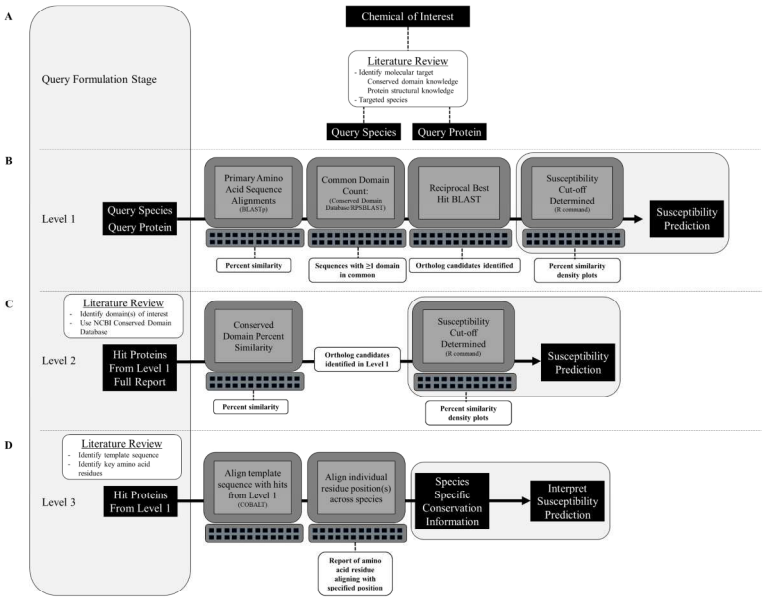


Figure 1. Schematic of strategic approach for predicting cross-species susceptibility using the SeqAPASS tool. Beginning with the query formulation stage (a) and moving through each level of protein sequence/structural comparison. (b) Level 1 describes analysis for predictions based on primary amino acid sequence comparisons; (c) Level 2 provides a means to compare functional domain sequences; and (d) Level 3 methodology allows for evaluation of key individual amino acid residue positions across species. Portions of the diagram surrounded by the grey oval represent areas where information collected from the query formulation stage are essential to inform the analysis.
279x215mm (300 x 300 DPI)

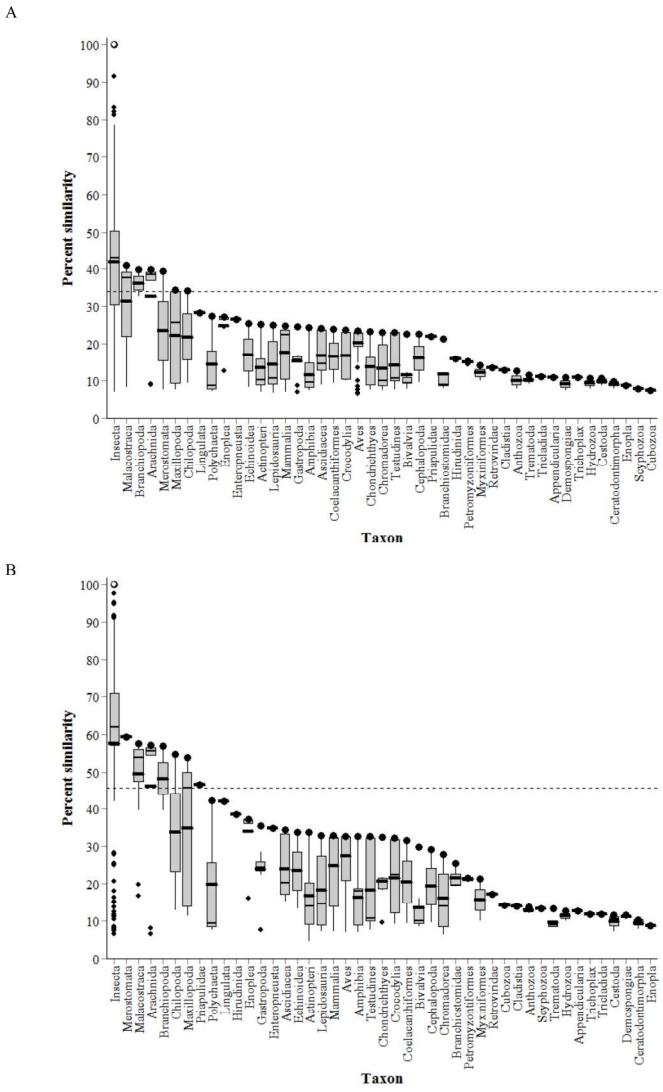


Figure 2. Boxplots depicting SeqAPASS data illustrating the percent similarity across species compared to tobacco budworm (H. virescens) ecdysone receptor (EcR) examining the primary amino acid sequences (a) and the ligand binding domain (b) o represents the tobacco budworm EcR and • represents the species with the highest percent similarity within the specified taxonomic group. The top and bottom of each box represent the 75th and 25th percentiles, respectively. The top and bottom whiskers extend up to 1.5 times the interquartile range. The mean and median values for each taxonomic group are represented by horizontal thick and thin black lines on the box, respectively. The dashed line indicates the cut-off for intrinsic susceptibility predictions (based on ortholog analysis).
215x279mm (300 x 300 DPI)

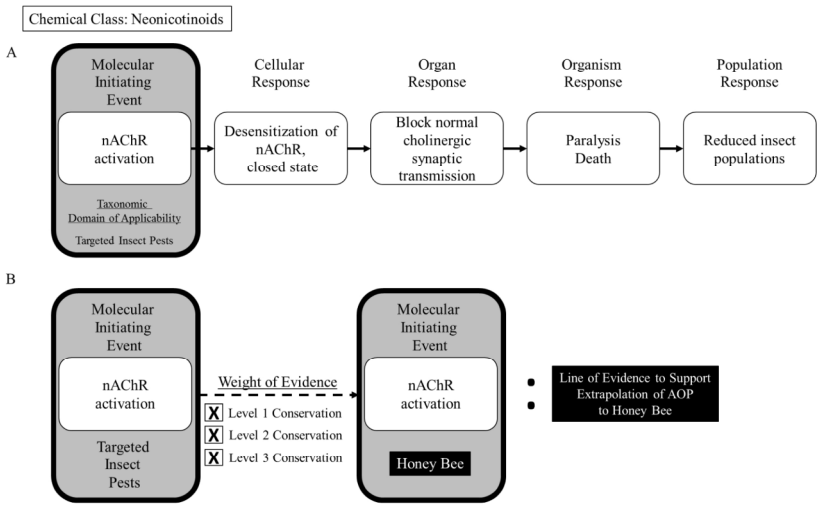
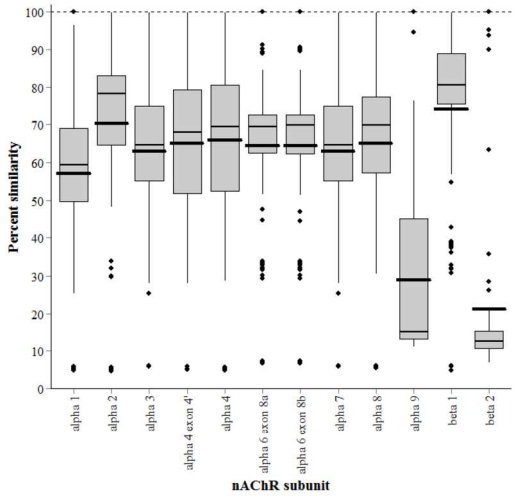


Figure 3. Putative adverse outcome pathway for nicotinic acetylcholine receptor activation leading to mortality (a). For this putative AOP example, relevant to the SeqAPASS analysis, the chemical class has been defined with neonicotinoids and the taxonomic domain under consideration was limited to insect pests. SeqAPASS analyses can be used to extrapolate AOPs (b) based on conservation of the molecular initiating event (MIE). Dashed arrow represents the weight of evidence provided by the various levels of the SeqAPASS analysis to extrapolate the MIE to honey bees.
279x215mm (300 x 300 DPI)

A



B

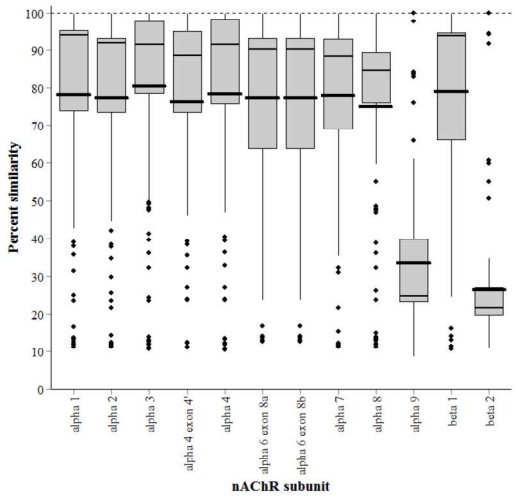


Figure 4. Boxplots to describe SeqAPASS data illustrating the percent similarity of all sequences from species of the Insecta organism class compared to each honey bee (Apis mellifera) nicotinic acetylcholine receptor (nAChR) subunit examining both the primary amino acid sequences (a) and the ligand binding domain (b). The dashed line represents the honey bee nAChR subunits having 100% similarity when comparing the sequence to itself. The top and bottom of each box represent the 75th and 25th percentiles, respectively. The top and bottom whiskers extend up to 1.5 times the interquartile range. The mean and median values for each subunit percent similarity are represented by horizontal thick and thin black lines on the box, respectively.
215x279mm (300 x 300 DPI)

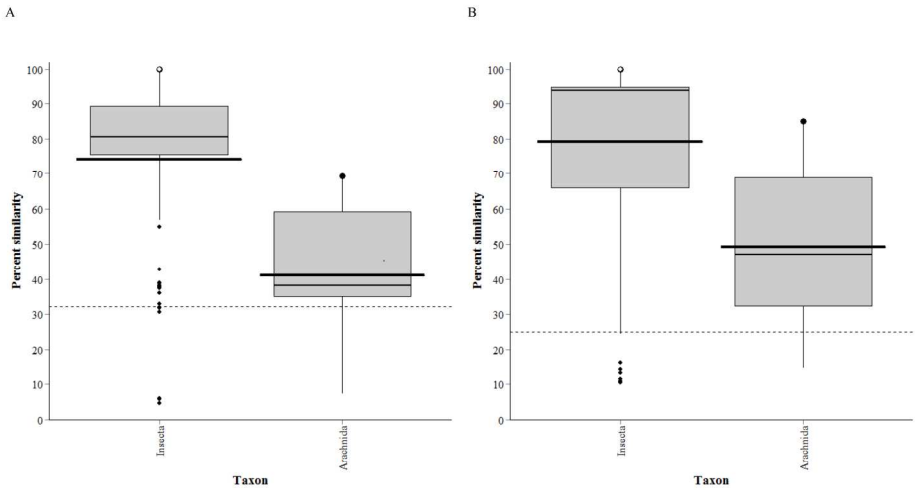


Figure 5. Boxplots to describe SeqAPASS data illustrating the percent similarity of all sequences from the Insecta and Arachnida organism classes for the honey bee (Apis mellifera) nicotinic acetylcholine receptor (nAChR) beta 1 subunit examining both the primary amino acid sequences (a) and the ligand binding domain (b). o represents the honey bee nAChR beta 1 subunit and • represents the species with the highest percent similarity within Arachnida. The top and bottom of each box represents the 75th and 25th percentiles. The top and bottom whiskers extend up to 1.5 times the interquartile range. The mean and median values for each subunit percent similarity are represented by horizontal thick and thin black lines on the box, respectively.
279x215mm (300 x 300 DPI)