Characterization of insect immune systems from genomic data

Identification of genes involved in physiological processes can be conducted by homology search or transcriptomic analysis. Homology search works well for evolutionarily conserved and well-studied canonical gene repertoires, but lose evolutionarily novel or not well-studied genes, which can be complemented by transcriptomic analysis.

1 Homology search

The first step of characterizing canonical gene repertoires in a newly sequenced genome is to compile reference sequences, *i.e.* protein sequences of gene repertoires from reference species that have been characterized. It requires define a scope of gene families to be included and select appropriate species from which reference sequences are drawn. For immune gene identification, principle components of immune responses should be included, *i.e.* recognition of antigene, signaling transduction and effectors (3). As for reference species, characterized species of the same order are the most useful, as the lower sequence divergence between more closely related species improves the success of sequence homology searches. Besides, closely related species share similar gene family components with less gene gain/loss events.

2 Supplementary

3 Canonical immune gene families in insects

Gram-negative binding proteins: Gram-negative binding proteins (GNBPs) or beta-1,3-glucan-binding proteins (BGBPs) are a family of carbohydrate-binding pattern recognition receptors.

Peptidoglycan binding proteins: PGRPs are pattern recognition receptors capable of recognizing the peptidoglycan from bacterial cell walls.

Fibrinogen-related proteins: FREPs (also known as FBNs) are a family of pattern recognition receptors with homology to the C terminus of the fibrinogen beta- and gamma-chains.

Galectins: GALEs bind specifically to beta-galactoside sugars and can function as pattern recognition receptors in innate immunity.

MD-2-like proteins: MLs, also known as Niemann-pick type C-2 proteins, possess myeloid-differentiation-2-related lipid-recognition domains involved in recognizing lipopolysaccharide.

Nimrods: NIMs have been shown to bind bacteria leading to their phagocytosis by hemocytes.

Scavenger receptors: SCRs are made up of different classes that function as pattern recognition receptors for a broad range of ligands including from pathogens.

Spaetzle-like proteins: The cleavage of Spaetzle results in binding of the product to the toll receptor and subsequent activation of the toll pathway; SPZs contain a cystine knot domain.

IMD pathway: Immune deficiency pathway is characterized by peptidoglycan recognition protein receptors, intracellular signal transducers and modulators, and the NF-B transcription factor relish.

Toll pathway: The intracellular components of Toll pathway signaling are homologous to the Toll-like receptor innate immune pathway in mammals, culminating in activation of the NF-B transcription factors dorsal and DIF in Drosophila.

JAK/STAT pathway: The Janus kinase protein (JAK) and the signal transducer and activator of transcription (STAT) are two core components of the JAK/STAT pathway, which is involved in cellular responses to stress or injury.

RNAi pathway: RNA interference protects against viral infections employing dicer and Argonaute proteins as well as helicases to identify and destroy exogenous double-stranded RNAs.

Caspase: Cysteine-aspartic proteases are involved in immune signaling cascades and apoptosis.

CLIP-domain serine protease: Several CLIP proteases have roles as activators or modulators of immune signaling cascades.

Inhibitor of apoptosis: IAPs are important in antiviral responses and are involved in regulating immune signaling and suppressing apoptotic cell death.

Serine protease inhibitors: Protease inhibition by serpins, or SRPNs, modulates many signaling cascades; they act as suicide substrates to inhibit their target proteases.

Thioester-containing proteins: TEPs are related to vertebrate complement factors and alpha2-macroglobulin protease inhibitors; their activation through proteolytic cleavage leads to phagocytosis or killing of pathogens.

Antimicrobial peptide: Antimicrobial peptides (AMPs) are the classical effector molecules of innate immunity; they include defensins, cecropins, and attacins that are involved in bacterial killing by

disrupting their membranes.

Lysozymes: LYSs are key effector enzymes that hydrolyze peptidoglycans present in the cell walls of many bacteria, causing cell lysis.

C-type lectins: C-type lections (CTL) are carbohydrate-binding proteins with roles in pathogen opsonization, encapsulation, and melanization, as well as immune signaling cascades.

Prophenoloxidases: PPOs are key enzymes in the melanization cascade that helps to kill invading pathogens and is important for wound healing.

Peroxidases: PRDXs are enzymes involved in the metabolism of reactive oxygen species (ROS) that are toxic to pathogens.

Superoxide dismutases: SODs are antioxidant enzymes involved in the metabolism of toxic superoxide into oxygen or hydrogen peroxide.