Characterization of insect immune systems from genomic data

Robert M. Waterhouse, Brian P. Lazzaro, and Timothy B. Sackton

Corresponding author: RMW, robert.waterhouse@unil.ch

Abstract

Insects face a multitude of threats from the pathogens and parasites they encounter over their life cycles, and they use robust immune systems to defend themselves. This chapter provides a tutorial for the identification and annotation of genes that comprise the immune system from newly sequenced insect genomes. Insect immune responses are orchestrated by the products of a suite of genes responsible for pathogen recognition, signal transduction, and pathogen killing. Many of the genes and proteins underlying these processes can be identified based on sequence homology with related species that have been immunologically characterized. Additional components of the immune response can be identified by transcriptomic analyses to detect genes whose expression changes in response to infection stimulus. Application of our step-by-step protocols for these complementary approaches enables the characterization of insect immune systems from genomic data.

Key words immunity, infection, genome annotation, gene families, comparative genomics, transcriptomics

Running head Insect immunogenomics

1 Introduction

A major element of genome sequencing projects is the identification and annotation of the genes expected to underlie key physiological processes. The initial identification of these genes from genomic data enables subsequent functional experimentation and comparative genomic analyses to understand the evolutionary forces that drive establishment, maintenance, and diversification of these processes. In this chapter, we describe (*i*) a general framework for using sequence homology searches, and (*ii*) a detailed infection protocol for transcriptomic analyses, to identify and annotate candidate immune system genes in newly sequenced insect genomes.

The identification of genes in newly sequenced genomes is typically initiated with computational searches for homologs of genes that have been characterized in other species. This approach works well for genes that make up an evolutionarily conserved, canonical immune repertoire, such as those established over two decades of functional genetic research on the model insect *Drosophila melanogaster* [1–6] and more recent work in non-model insects [7–16]. The identification of novel genes or those with no prior ascribed functional role in immunity, however, requires experimental data to be coupled with the computational analyses. Identifying these infection-responsive genes is facilitated by the fact that the expression of many immune genes is induced by infectious challenge. This means that transcriptomic analysis of changes in gene expression after infection can be used to support inferences from homology searches and to suggest additional, sometimes novel, components of the immune system.

Homology searches are excellent for identifying conserved genes and protein domains that comprise various components of the innate immune system. This includes most immune gene families and signaling pathway members. The presence of core recognition, signaling and modulation, and effector components of the immune system indicates functional conservation across taxa, while notable absences such as the apparent degradation of the Imd pathway in pea aphids [10] can suggest possible

rewiring of the system. Computational searches will identify candidate immune-related genes from the full set of genes predicted by whole genome annotation pipelines. Manual curation may be required to validate some candidates or to confirm cases of apparent losses of otherwise widely-conserved genes. Homology searches also help to detect and quantify expansions and contractions of multi-gene families that vary in copy number across insects, such as genes encoding peptidoglycan recognition proteins (PGRPs) and members of the phenoloxidase cascade (PPOs). Unlike for the generally single-copy signaling pathway genes, defining clear orthologous relationships can be difficult for such multi-gene families, depending on the age of the gene duplications and the phylogenetic distance between the species being compared. Nevertheless, the variable numbers of such immunity genes can sometimes be interpreted as indicative of the natural selective and epidemiological pressures on the insect being studied [7, 17, 18].

Homology searches are invaluable for identifying most canonical immune genes. However, genes that have newly acquired immune functions, or evolutionarily novel genes with roles in immunity, will not be identified through homology searches using known immune gene sequences. Thus homology searches can be complemented with transcriptomic analyses to identify sets of genes whose expression levels are responsive to infection, but that are not normally considered part of the canonical immune system. In such analyses, the insect in question is challenged with a relevant infection stimulus and RNA is extracted either from the whole insect or from immunologically relevant tissues. The gene expression profiles of challenged insects can then be compared to the expression profiles of naïve insects, enabling identification of genes whose expression is induced or repressed by infection (e.g. [19, 20]). Transcriptomic analysis is especially powerful for identifying effector genes such as those encoding antimicrobial peptides (AMPs). These may be unique to specific groups of insects and the genes are often so short that they fail to be detected by computational gene-finding algorithms. However, they are often massively transcriptionally induced upon infection. Thus, transcriptomic analysis can be a powerful approach to identify effectors that would be missed by other methods (reviewed in [21]). While AMPs and other effectors have direct roles in immunity, many other differentially expressed genes may play indirect roles and as such they do not form part of the “immune

system" by any canonical definition. For example, infection often causes activation of generic stress response genes [22, 23] and a transcriptional signature of repression of basal metabolism [24, 25]. In some cases, these transcriptional responses may promote host survival, but in other cases they may even represent deleterious consequences of infection. Therefore caution must be taken and it should not be assumed that a gene is part of the immune system solely because its expression level changes after challenge.

Homology searches and transcriptomic analyses are complementary approaches to characterize genes that play a role in the insect immune system from newly sequenced genomes (henceforth referred to as the "target" or "focal" species). Sequence homology searching is powerful and allows for the identification of genes with conserved immune-related protein domains, including genes whose expression patterns do not change substantially in response to infections. Transcriptomic analyses have the advantage that they can identify novel infection-responsive genes that have not been previously characterized in other species. In this chapter, we detail a practical workflow for applying these two approaches in parallel to characterize the immune system of an insect with a newly sequenced genome.

2 Methods

2.1 Identification of canonical innate immunity genes

Characterizing the canonical innate immune gene repertoire in newly sequenced genomes follows four main steps, presented in **Figure 1**. The first is to compile a comprehensive list of immune-related genes and their protein sequences from species that have already been characterized (henceforth referred to as the "reference" species). These sequences are then used to search the genomes and gene sets (the complete set of predicted genes for a given genome) for putative homologs and characteristic protein domains. The candidate gene models can then be inspected and manually curated to ensure that they are correct and complete. Finally, phylogenetic analyses to trace the evolutionary histories of each gene

family allow for the delineation of orthologs and paralogs, and the confident characterization of a new set of canonical immune genes.

2.1.1 Compiling sets of reference sequences

1. The comparative approach to identifying immune-related genes in newly sequenced genomes relies on comparisons with previously characterized sets of immunity genes in other species. While newer investigations of immune systems across diverse insect taxa have begun to reveal novelties in different species, a great deal of the collective knowledge of the canonical insect innate immune gene repertoire nevertheless still derives from studies conducted on *D. melanogaster* (*see* **Note 1**). To start compiling sets of reference immune gene sequences, you will first need to (*i*) define the scope of your study by deciding which immune-related pathways and gene families to include, and (*ii*) select appropriate species from which to source the reference immune protein sequences.
2. Defining the scope of the immune gene repertoire to be examined requires an overview of the current understanding of the canonical insect innate immune system. The principal components of an immune response must include proteins responsible for recognition of pathogens, signal transduction once a pathogen has been recognized, and effector proteins and biomolecules that eliminate the pathogen (**Table 1**). A core set of key genes and pathways has been characterized through experimental research in different insect systems and shown to be widely conserved across divergent insect species (*see* **Note 2**). These can serve as the initial basis for homology searches, although novel genes should also be expected to emerge from each new study system. A streamlined scope would normally first focus on (*i*) canonical families of pathogen recognition receptors such as peptidoglycan recognition proteins (PGRPs) and gram-negative bacteria-binding proteins (GNBPs, also known as beta-1,3-glucan-binding proteins); (*ii*) the core members of the three main immune signaling cascades, the Toll, Imd, and JAK/STAT pathways; and (*iii*) effectors such as antimicrobial peptides (AMPs) and lysozymes (LYSs) whose expression is generally upregulated upon stimulation of these pathways. Additional core processes include immune responses such as RNA interference (RNAi), phagocytosis, apoptosis and autophagy, the defensive production of reactive oxygen species (ROS), and melanization reactions [26, 27]. Broadening the scope of the study further would

normally include (*i*) additional gene families with members implicated in pathogen recognition and/or immune response activation such as C-type lectins (CTLs), thioester-containing proteins (TEPs), or scavenger receptors (SCRs); (*ii*) genes responsible for the positive or negative regulation of core members of the main signaling pathways and cascade modulation. Ultimately, the scope of the study will be determined by size of the research team working on the project and the questions of particular biological interest for the target species.

2. The selection of appropriate reference species should be guided by published comparative characterizations of other insect genomes such as those listed in **Table 2**. Selecting several reference species will allow for better consistency checks; i.e. do searches using one reference species produce similar results as using another reference species? Comparisons between insects from the same order are the most useful, as the lower sequence divergence between more closely related species improves the success of sequence homology searches. Additionally, gene family composition will generally be more similar between closely related species, with fewer gene gains or losses since their last common ancestor. Data from the reference species should public, versioned, and recognized by their respective communities as the official assemblies and gene sets, to facilitate both repeatability of the analysis and ease of data acquisition. Data retrieval and querying will be further facilitated if the selected reference species are already hosted by an online genome browser resource such as the Bioinformatics Platform for Agroecosystem Arthropods [28], Ensembl Metazoa [29], FlyBase [30], Hymenoptera Genome Database [31], i5k at the National Agricultural Library [32], the National Center for Biotechnology Information [33], or VectorBase [34].

3. Having defined the scope and selected the reference species, you can now proceed with compiling your sets of reference immune-related protein sequences. Published studies such as those presented in **Table 2** usually include lists of gene and/or protein identifiers of the immune genes that were identified. Use these to extract the corresponding sequences from the complete gene sets for each species. As these studies are effectively snapshots of the available data at the time of publication, they should be treated as starting points for compiling your own sets of reference sequences. By subsequently curating these initial sets, you will be able to match them with the most up-to-date information, both with respect to the latest genome assembly versions and their corresponding gene

sets, as well as to incorporate new discoveries or refinements described in the current literature. One advantage of having selected reference species with publicly browsable genomic resources is that it allows you to perform online queries with gene identifiers or names from the literature in addition to the sequence homology searches described below. Typically, the collected reference sequences will be the translated protein products of each transcript comprising each gene (*see* **Note 3**), stored in plain-text files in FASTA format. When alternative splicing produces protein products that differ substantially (e.g. a single PGRP gene that can encode one, two, or three distinct PGRP domains), it is important to collect all predicted transcripts. This will allow you to assess whether the target species genome also encodes equivalent transcripts and whether gains or losses of alternative transcripts have occurred.

Table 1. The principal components of the canonical insect innate immune gene repertoire.

Gene family or signaling pathway	Brief description
Imd pathway	The immune deficiency pathway is characterized by peptidoglycan recognition protein receptors, intracellular signal transducers and modulators, and the NF-κB transcription factor Relish.
Toll pathway	The intracellular components of the Toll signaling are homologous to the toll-like receptor innate immune pathway in mammals, culminating in activation of the NF-κB 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.
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.
Caspases	Cysteine-aspartic proteases are involved in immune signaling cascades and apoptosis.
CLIP-domain serine proteases	Several CLIP proteases have roles as activators or modulators of immune signaling cascades.
C-type lectins	CTLs are carbohydrate-binding proteins with roles in pathogen opsonization, encapsulation, and melanization, as well as immune signaling cascades.
Fibrinogen-related proteins	FREPs (also known as FBNs) are a family of pattern recognition receptors with homology to the C terminus of the fibrinogen β and γ chains.
Galectins	GALEs bind specifically to β-galactoside sugars and can function as pattern recognition receptors in innate immunity.
Gram-negative binding proteins	GNBPs (or β-1,3-glucan-binding proteins, BGBPs) are a family of carbohydrate-binding pattern recognition receptors.
Inhibitors of apoptosis	IAPs are important in antiviral responses and are involved in regulating immune signaling and suppressing apoptotic cell death.
Lysozymes	LYSs are key effector enzymes that hydrolyze peptidoglycans present in the cell walls of many bacteria, causing cell lysis.
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.
Peptidoglycan recognition proteins	PGRPs are pattern recognition receptors capable of recognizing the peptidoglycan from bacterial cell walls.
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.

Scavenger receptors	SCRs are made up of different classes that function as pattern recognition receptors for a broad range of ligands including from pathogens.
Superoxide dismutases	SODs are antioxidant enzymes involved in the metabolism of toxic superoxide into oxygen or hydrogen peroxide.
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.
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 α2-macroglobulin protease inhibitors, their activation through proteolytic cleavage leads to phagocytosis or killing of pathogens.

Table 2. Examples of comparative studies of the canonical insect innate immune repertoire.

Gene categories: Rec, recognition; Sig, signaling; Mod, modulation; Eff, effectors.

Focal species	Comparison species	Breadth of study	Reference
6 *Glossina*	*Musca domestica* *Drosophila melanogaster*	Rec, Sig, Mod, Eff	Attardo *et al*, 2019 [35]
Manduca sexta	*Bombyx mori*	Serine protease inhibitors (SRPNs)	Li *et al*, 2018 [36]
Aedes aegypti	*Aedes albopictus* *Anopheles gambiae* *Culex quinquefasciatus*	C-type lectins (CTLs)	Adelman & Myles, 2018 [37]
6 *Glossina*	Several other dipterans Outgroup blood-feeding hemipterans	Thioester-containing proteins (TEPs)	Matetovici & Van Den Abbeele, 2018 [38]
Musca domestica	*Glossina morsitans* 5 mosquitoes 7 *Drosophila*	Rec, Sig, Mod, Eff	Sackton *et al*, 2017 [7]
Pteromalus puparum	*Aedes aegypti* *Anopheles gambiae* *Apis mellifera* *Bombyx mori* *Drosophila melanogaster* *Manduca sexta*	Serine protease inhibitors (SRPNs)	Yang *et al*, 2017 [39]
Bombus impatiens *Bombus terrestris*	*Apis florea* *Apis mellifera* *Megachile rotunda* *Nasonia vitripennis* *Tribolium castaneum* *Drosophila melanogaster* *Anopheles gambiae*	Rec, Sig, Mod, Eff	Barribeau *et al*, 2015 [8]
Anopheles gambiae	20 other mosquitoes *Drosophila melanogaster*	Rec, Sig, Mod, Eff	Neafsey *et al*, 2015 [40]
Zootermopsis nevadensis	Diptera Lepidoptera Coleoptera	Rec, Sig, Mod, Eff	Terrapon *et al*, 2014 [41]
Nasonia vitripennis	*Drosophila melanogaster* *Anopheles gambiae* *Apis mellifera* *Acrythosiphon pisum*	Rec, Sig, Mod, Eff	Brucker *et al*, 2012 [42]
Aedes aegypti	*Anopheles gambiae* *Culex quinquefasciatus* 12 *Drosophila*	Caspases (CASPs)	Bryant *et al*, 2010 [43]
Culex quinquefasciatus	*Anopheles gambiae* *Aedes aegypti* *Drosophila melanogaster*	Rec, Sig, Mod, Eff	Bartholomay *et al*, 2010 [9]
Acyrthosiphon pisum	*Drosophila melanogaster* *Anopheles gambiae*	Rec, Sig, Mod, Eff	Gerardo *et al*, 2010 [10]

	Tribolium castaneum *Apis mellifera* *Pediculus humanus*		
Anopheles gambiae	*Culex quinquefasciatus* *Aedes aegypti*	Mosquito leucine-rich repeat immune proteins (LRIMs)	Waterhouse *et al*, 2010 [44]
Bombyx mori	*Drosophila melanogaster* *Anopheles gambiae* *Aedes aegypti* *Apis mellifera* *Tribolium castaneum*	Serine protease inhibitors (SRPNs)	Zou *et al*, 2009 [45]
Bombyx mori	*Drosophila melanogaster* *Anopheles gambiae* *Apis mellifera* *Tribolium castaneum*	Rec, Sig, Mod, Eff	Tanaka *et al*, 2008 [11]
Drosophila melanogaster	11 other *Drosophila*	Rec, Sig, Mod, Eff	Sackton *et al*, 2007 [12]
Aedes aegypti	*Anopheles gambiae* *Culex quinquefasciatus* *Drosophila melanogaster*	Rec, Sig, Mod, Eff	Waterhouse *et al*, 2007 [13]
Tribolium castaneum	*Drosophila melanogaster* *Anopheles gambiae* *Apis mellifera*	Rec, Sig, Mod, Eff	Zou *et al*, 2007 [14]
Apis mellifera	*Drosophila melanogaster* *Anopheles gambiae*	Rec, Sig, Mod, Eff	Evans *et al*, 2006 [15]
Anopheles gambiae	*Drosophila melanogaster*	Rec, Sig, Mod, Eff	Christophides *et al*, 2002 [16]

2.1.2 Searching gene sets for candidate immunity genes

1. The purpose of compiling a comprehensive and up-to-date set of reference sequences is to then use these as query sequences to search the gene set of the target species being investigated. Your searches should start with a global comparison (*see* **Note 4**) of the compiled sets of reference sequences against the target species' gene set. Use the BLASTp option of the Basic Local Alignment Search Tool (BLAST) suite [46] to identify the most significant matches (i.e. the highest bit scores and the lowest expectation values) to the reference protein sequences in the predicted target proteome (the translations of the predicted gene set). The National Center for Biotechnology Information (NCBI) BLAST+ user manual (https://www.ncbi.nlm.nih.gov/books/NBK279690) provides detailed installation and usage instructions, and example commands (in `monospace` type following `$` symbols) for the required steps are provided here with default parameters:

Format the protein sequences from your gene set into a searchable database:

```
$ makeblastdb -in geneset_proteins.fasta -dbtype prot -out proteinsDB
```

Search your compiled reference protein sequences against the gene set:

```
$ blastp -query reference_proteins.fasta -db proteinsDB -out referencesVSgeneset.txt
```

Produce tabular results of searching your compiled reference protein sequences against the gene set:

```
$ blastp -query reference_proteins.fasta -db proteinsDB -outfmt 6 -out
   referencesVSgenesetTAB.txt
```

The BLASTp search will provide ranked lists of putative homologs of each query sequence from the reference proteins, thereby identifying the predicted proteins encoded in the target genome that most closely resemble the reference sets of immunity proteins. You should next run reciprocal BLASTp searches using the top-scoring proteins from the target species as queries against the complete protein set from the reference species. Your reciprocal searches should return the original query protein as the top-scoring match, especially in the case of proteins encoded by immunity genes that are generally maintained across most species as single-copy orthologs (but *see* **Note 3**). In contrast, for multi-copy gene families, several proteins encoded by members of the gene family in the reference genome may be among the best-scoring matches. These reciprocal sequence homology searches will provide support for the lists of putative immunity genes, but you will need to perform downstream phylogenetic analyses (see **Section 2.1.3** step 6 below) in order to confirm single-copy orthologs and resolve the relationships among members of multi-copy gene families.

2. The next step is to complement the global protein-protein homology searches of gene set with protein-domain-level searches. Run InterProScan [47] on the proteins from the target species' gene set and the reference protein sequences to obtain detailed domain-level annotations of all protein sequences with significant matches to profiles from the InterPro member databases [48]. Next, use the InterPro domains that characterize each of the different immune gene families or pathway members (**Table 2**) to identify genes from the target species that encode proteins with significant matches to these domains (*see* **Note 5**). For example, serine protease inhibitors (serpins, or SRPNs) are recognized by the 'Serpin superfamily' (IPR036186) or 'Serpin family' (IPR000215) profiles, or related profiles such as 'Serpin, conserved site' (IPR023795) or 'Serpin domain' (IPR023796). Exercise caution when the characteristic domains are promiscuous, meaning when they are also present in gene families unrelated to immunity, or when two or more distinct domains characterize a particular immune gene family. For example, Toll-like receptors (TLRs, or TOLLs) contain 'Leucine-rich repeat' domains, but these are also found in many other types of proteins so their

presence is not, on its own, diagnostic of TOLLs. Instead, TOLLs are more specifically characterized by several 'Leucine-rich repeat' domains followed by a 'Toll/interleukin-1 receptor homology (TIR) domain'. The European Bioinformatics Institute provides detailed InterProScan installation and usage instructions (https://www.ebi.ac.uk/interpro/interproscan.html); the example here uses profiles from the Pfam database:

Scan the gene set protein sequences and compiled sets of reference sequences for matches to InterPro domains:

```
$ ./interproscan.sh -appl Pfam -i geneset_proteins.fasta -f tsv -iprlookup
$ ./interproscan.sh -appl Pfam -i reference_proteins.fasta -f tsv -iprlookup
```

3. A third approach to searching the target species' gene set for candidate immunity genes is to use profiles built from the reference sequences. First, align each set of orthologous or homologous reference immunity protein sequences collected from several reference species using tools such as PRANK [49] or MAFFT [50]. Next, convert the resulting multiple protein sequence alignments into sequence profiles using HMMER [51]. The HMMER suite of tools can then be used to search the profiles against the target species' gene set. Here we present some examples of the commands that need to be run, but please see the user guides and installation instructions for the alignment tools and HMMER for full details. The input proteins in FASTA format should consist of orthologs or homologs from each of the reference species. Specifically, each FASTA file should contain only proteins encoded by homologs of a single gene or conserved gene family and the entire analysis should be repeated for each gene or gene family in the study.

Multiple protein sequence alignment example using PRANK:

```
$ prank -d input_proteinset1.fasta -o aligned_proteinset1.aln
```

Multiple protein sequence alignment example using MAFFT:

```
$ mafft input_proteinset1.fasta > aligned_proteinset1.aln
```

Convert a multiple protein sequence alignment to a profile using HMMER:

```
$ hmmbuild proteinset1.hmm aligned_proteinset1.aln
```

Combine all your profiles into a single profile library (here just three sets shown):

```
$ cat proteinset1.hmm proteinset2.hmm proteinset3.hmm > profile_library
```

Compress and index the library of profiles:

```
$ hmmpress profile_library
```

Search the library of profiles against the target species' gene set using HMMER:

```
$ hmmscan profile_library geneset_proteins.fasta
```

2.1.3 Curating candidate immune-related genes

1. Your global protein sequence and profile searches and protein domain searches will result in lists of candidate immune-related genes from the target species. With good supporting data, especially from transcriptomics (as described below in **Section 2.2**), automated prediction pipelines applied to well-assembled genomes generally produce gene sets with a high coverage of the true gene content [52–54]. The task nevertheless remains challenging, and accurate predictions at the detailed level of gene intron/exon structures can be difficult to achieve even with extensive supporting data. Manual curation aims to verify that the automatically predicted gene models identified through your sequence and domain searches are in agreement with the available supporting evidence. You may undertake the curation process with a small team or you may bring together several groups of researchers and/or students (e.g. [55–57]) to examine your lists of candidate immunity genes. For a small team, the curation process may focus on quality control and targeted appraisal of specific genes of interest. For example, quality control of seemingly anomalous results can confirm true novelties, such as the multi-PGRP-domain PGRP proteins encoded in the banded demoiselle genome [58]. For a larger research community the aims may be broader and may include taking advantage of researchers' expertise to build a rich knowledge base for the target species. The tools and approaches described here are useful for both small- and large-scale curation efforts.
2. Several computational resources need to be set up so that the genomic data from the target species can be easily queried by users with little or no bioinformatics expertise. You can achieve a local setup of the necessary resources with relatively modest computational equipment and the installation of several freely available bioinformatics packages and software. The key components should include a genome browser and a sequence search interface. A particularly useful platform that allows for sequence-based database searching is the combination of the JBrowse genome viewer [59] with the Apollo annotation feature editor plug-in [60], and SequenceServer [61]. Software installation is beyond the scope of this chapter but is described in detail in the respective setup and user guides. These resources will provide you with a user-friendly environment to interrogate the genomics data

without requiring experience with running command-line bioinformatics tools. They also offer the flexibility to search gene-by-gene for specific genes of interest, to search using sequences from species or genes that were not included in the compiled sets of reference sequences, or to use sequences from the target species to search for within-species homologs.

3. A tBLASTn search of the reference immunity sequences against the target species' genome assembly will enable visualization of genomic loci with homology to the reference proteins. tBLASTn uses the provided reference protein sequences to search the six-frame translations of the genome assembly nucleotides and is more sensitive than nucleotide-nucleotide searches. The tBLASTn results are useful because the automated pipeline used to predict gene models in the target species may have missed or misannotated some genes or exons, meaning that they would be impossible or difficult to identify from searching only the predicted gene set. You should produce tabular format outputs of the tBLASTn searches because these can be loaded as data tracks for visualization within a genome browser after converting them into General Feature Format (GFF) output files (*see* **Note 4**). The following commands illustrate how this can be achieved:

Format your genome assembly into a searchable database:

```
$ makeblastdb -in genome_assembly.fasta -dbtype nucl -out assemblyDB
```

Produce tabular results of searching your compiled reference protein sequences against the genome assembly:

```
$ tblastn -query reference_proteins.fasta -db assemblyDB -outfmt 6 -out
   referencesVSassemblyTAB.txt
```

4. The locations of the best hits define genomic loci that likely encode orthologs or homologs of the reference sequences. Visualizing these using a genome browser enables you to assess how much of the reference sequence aligns to the target assembly and how well these alignments match up to the predicted gene model (*see* **Note 6**). Complementary supporting evidence comes from transcriptomics data in the form of RNA sequencing (RNA-seq) reads from samples prepared from your target species. The RNA-seq reads may derive from your own infection experiments (see **Section 2.2** below), but if other datasets are available then it is advisable to also include these as additional supporting data. You will need to align the reads to the genome assembly in order to visualize them in a genome browser, typically as both stacked individual read alignments and read coverage plots (*see* **Note 4**). Several bioinformatics tools are able to align reads to an assembly (e.g.

HISAT2 [62] or STAR [63]) and coverage plots can be built using bamCoverage from the DeepTools suite [64]. Here we present some examples of the commands that need to be run, but please see the user guides for full details.

Build an index of your genome assembly then align fastq format RNA-seq reads using HISAT2:

```
$ hisat2-build genome_assembly.fasta index_name
$ hisat2 -x index_name -1 sample_1.fastq -2 sample_2.fastq -S hisat2-mapped.sam
```

Build an index of your genome assembly then align fastq RNA-seq reads to your assembly using STAR:

```
$ STAR --runMode genomeGenerate --genomeDir star-index --genomeFastaFiles
   genome_assembly.fasta
$ STAR --genomeDir star-index --readFilesIn sample_1.fastq sample_2.fastq --outSAMtype BAM
   SortedByCoordinate
```

Produce an RNA-seq read coverage file using bamCoverage:

```
$ bamCoverage -b Aligned.sortedByCoord.out.bam -o rnaseq-coverage.bw
```

5. With the necessary resources in place, the next step is to examine the genomic locus encoding each candidate immunity gene in order to establish whether the predicted model is well supported (*see* **Note 7**). Well-supported models generally show RNA-seq coverage and spliced RNA-seq read alignments that match the intron-exon structure of the entire model and tBLASTn alignments for most of the model. Typical minor edits to improve the models include altering the intron-exon boundaries to match the aligned RNA-seq reads, removing non-supported exons (i.e. predicted exons that have no tBLASTn alignments and no aligned RNA-seq reads), or adding exons missed by the automated prediction pipeline (i.e. regions with tBLASTn alignments and/or aligned RNA-seq reads where no exon was predicted). For example, **Figure 2** shows how editing an incorrectly predicted intron-exon boundary to match the supporting RNA-seq read alignments produces a full length gene model for *Dicer-2*. More substantial edits include the merging of two or more neighboring predicted gene models that in fact encode a single gene, or the splitting of gene models where the automated gene prediction has incorrectly fused neighboring genes. Automated gene predictors are prone to such erroneous fusing of neighboring genes when the genes are homologous or have arisen from tandem gene duplication events. Thus it is worth paying particular attention to the gene model predictions of members of multi-copy gene families. In addition, it is often challenging for automated pipelines to correctly predict two or more alternative transcripts from the same gene, so

manual editing may be required to distinguish the individual transcripts based on the available supporting data.

6. One reason for checking and correcting the candidate immune-related gene models is to facilitate subsequent phylogenetic analysis of immune genes or gene families of particular interest, including where putative duplications/expansions have been noted from the initial searches. Molecular phylogenetic analysis aims to reconstruct the evolutionary histories of sets of homologous sequences. Conceptually, this is achieved by contrasting the species phylogeny with the inferred gene trees to enable the confident assignment of orthologous relations [65]. In practice there are many different methodological approaches and bioinformatics tools designed for preparing and analyzing the sequence data required for phylogenetic tree construction, the discussion of which is beyond the scope of this chapter. One suite of such tools that is particularly user-friendly for novices is the Molecular Evolutionary Genetics Analysis (MEGA) software [66]. In the context of characterizing your sets of newly identified putative immune-related genes, the phylogenetic analyses will allow you to (*i*) confirm or refine orthologous relations suggested by your reciprocal sequence homology searches and (*ii*) place putative gene duplications or losses in their appropriate evolutionary contexts.

2.2 Identification of infection-responsive genes

While searching based on sequence homology is a valuable approach to identify canonical immune genes in new species, some immunologically important genes may be novel to the target species or otherwise difficult to identify from sequence data. In many cases, however, expression of these genes is responsive to infection [21]. These can include both genes that are directly involved in immune defense, and also genes that are regulated as a consequence of infection. Using RNA-sequencing (RNA-seq), it is possible to obtain a direct readout of the transcriptional response to infection.

There are a number of important experimental design issues to consider before embarking on RNA-seq based identification of immune-responsive genes [67]. Two key requirements must be met for a successful experiment. First, in order for the protocol outlined below to be successful, a mostly complete draft genome with a high-quality gene set must exist for the target insect. While it is possible to use RNA-seq data to build a *de novo* transcriptome [68, 69] (**and see Chapter 2 of this book**) or to aid gene prediction for an draft genome without a gene set [62, 70], this is beyond the scope of this chapter and we do not recommend it unless there is no alternative. Second, it must be possible to experimentally infect the target insect in the laboratory. Ideally, the insect can be maintained for several generations under controlled conditions to eliminate effects of previous exposure to pathogenic challenges or other stimuli that could modulate the immune response.

The simplest experimental design to identify genes that are transcriptionally responsive to infection would include just a single control condition (either naive, untreated insects or sterilely wounded insects), and a single experimental condition at some time post infection with the desired infectious challenge. More complex designs could include multiple controls, multiple pathogenic agents, and/or multiple time points. As a general rule of thumb, a minimum of three biological replicates should be included for each experimental treatment and control, although additional replicates will increase statistical power [71–74]. If the target insect is so small that sufficient RNA is hard to obtain from a single insect, pools of genetically similar (or ideally identical) individuals can be used, but this does not eliminate the need for multiple biological replicates of the experiment.

2.2.1 Artificial infections for RNA-seq analysis

Insects mount different immune responses to different types of infectious challenge (e.g., bacterial, fungal, viral, protozoan, nematode, etc.), and different challenges will therefore elicit different transcriptional responses. Injection with bacteria or bacterial cell wall and membrane components is often used as a generic immune stimulus for identification of genes that are transcriptionally responsive to infection [19, 20]. Here, we detail a protocol for infection of a small insect like *Drosophila* or a mosquito with a live bacterium. The protocol is demonstrated visually in [75] and can be modified for

larger insects or for other infectious agents. The experimenter should choose the most appropriate challenge for the system being queried and modify delivery of the challenge accordingly.

1. In order to minimize experimental noise, all insects should be reared in the laboratory without exposure to pathogens prior to the experiment. This will allow optimal comparison of the expression profiles of infected insects to unchallenged controls. Biological replicates should be collected for both challenged and unchallenged insects (*see* **Note 8**). For small insects or small tissue samples taken from larger insects, the material from multiple individuals can be pooled within each biological replicate. Using co-reared insects that are the same age and sex will minimize experimental noise, although in some cases it may be of interest to make comparisons across life stages, sexes, or rearing conditions (*see* **Note 9**).
2. Culture the infectious agent and prepare it for infection. In the case of bacterial challenge, infection may be delivered with a single bacterium or a mixture of different bacteria, and the bacteria may be either alive or killed by incubation at 60ºC for 30 minutes (*see* **Note 10**).
3. Challenge the insects in the infection treatment. Bacteria, planktonic fungi, and viruses can be injected into insects with a microcapillary needle. Live bacteria may also be introduced with a septic pinprick (demonstrated in detail in [75]) (*see* **Note 11**). Other challenges, such as infection with filamentous fungi (e.g. [76]) or eukaryotic parasites (e.g. [77]), require different methods.
4. Collect the insects at the prescribed time point post-infection (s*ee* **Note 12**). RNA may be isolated immediately or the insects may be flash-frozen in liquid nitrogen and stored at -80ºC until RNA extraction is to be performed. If RNA will be performed using a TriZOL (Invitrogen) extraction, the insects or insect tissue may be stored at -80ºC in TriZOL.
5. Isolate high-quality RNA from the infected and control insects. There are a variety of protocols and commercial kits available for RNA isolation, and any of these should be work well for RNA sequencing. Isolations using TriZOL reagent (Invitrogen) are reliable and inexpensive. A thorough protocol for RNA isolation using TriZOL is outlined in **Chapter 2** of this volume. Consult with the facility that will perform your RNA sequencing to see whether they have preferences or recommendations as to which RNA isolation procedure should be employed.

6. Perform the RNA sequencing (RNA-seq) on your infected and control insect material. In most circumstances, we recommend that inexperienced practitioners outsource library preparation and sequencing to a core facility or commercial provider. The library preparation is highly technical and labor intensive, and the technology changes quickly. Unless a very large number of libraries are going to be generated, the cost savings associated with doing the preparation yourself are generally not worth the effort or the risk of failed reactions. Therefore, if possible, use a facility that will accept RNA shipped on dry ice and that prepares their libraries and performs sequencing in-house. The optimal read length and depth of sequencing will depend on project budget and a variety of other factors that will vary among projects. For the analysis described below, we recommend a minimum of 10 million fragments sequenced per replicate, using at least 40 bp paired-end reads. Increasing read depth to 20-30 million fragments per replicate can be beneficial if project scope and funding allow (*see* **Note 13**), and increasing read length to 75 bp will decrease the number of reads that map ambiguously to multiple locations in the genome (e.g., reads from members of closely related gene families).

2.2.2 Performing differential expression analysis

1. The first step in differential expression analysis is using a read alignment or pseudoalignment (*see* **Note 14**) to estimate expression of each transcript or gene (*see* **Note 15**). Here we present one option for this, but there are many alternative choices (*see* **Note 16**). The protocol here assumes you have paired-end sequencing reads from your core facility or commercial provider, in fastq format. We describe optional quality control and trimming steps in **Note 17**. A workflow of the steps required to perform differential expression analysis is presented in **Figure 3**. In the following steps, command lines are given with variables (file names, species, and sample identifiers) that will need to be changed for each experiment in curly braces {}. Commands are given in `monospace type`.
2. This protocol uses commands from the kallisto program [78] (https://pachterlab.github.io/kallisto/) and should run in less than an hour per sample on a typical laptop computer. Software installation is beyond the scope of this chapter but is described in detail here: https://pachterlab.github.io/kallisto/download. The first step in using kallisto is to prepare the index.

Indexing takes a plain-text FASTA file containing the nucleotide sequences of all transcripts from the gene set of a given genome and converts it into a format that allows for subsequent rapid pseudoalignment of the RNA-seq reads to the transcripts. The complete set of transcripts from the gene set to be analyzed is referred to in the kallisto documentation as the 'reference transcriptome' to which the RNA-seq reads will be mapped. For your target species you should obtain the FASTA file of transcripts from the official gene set provided by public databases (e.g. Ensembl, FlyBase, NCBI, VectorBase). If only available in-house then use the FASTA file of transcripts resulting from the full genome annotation pipeline.

3. Prepare a reference transcriptome index for kallisto. First, make a working directory and copy the transcriptome FASTA file to it. You can then index this file and proceed to quantify transcript abundances. You will obtain a {SAMP}_out directory for each sample/replicate you generated, which can be used with sleuth (or other tools) as described below to estimate differentially expressed transcripts and genes per condition.

In the working directory and with kallisto installed:

```
$ kallisto index -i {INDEX_NAME}.idx {TRANSCRIPTOME}.fasta
```

Quantify abundance of transcripts in each sample, where {SAMP} is the fastq base name for a particular replicate/condition:

```
$ kallisto quant -i {INDEX_NAME}.idx -o {SAMP}_out -b 100 {SAMP}_R1.fastq.gz {SAMP}_R2.fastq.gz
```

5. There are many toolkits for detecting genes with differential expression between conditions. Here we present protocols for using sleuth [79], but discuss alternatives in **Note 18**. Note that sleuth requires the technical bootstraps generated by kallisto for full functionality, and thus we only recommend this protocol to be used with data analyzed first by kallisto.

Open R and ensure that the sleuth package is installed, as well as tidyverse which is used for some data manipulation tasks (*see* **Note 19**):

```
$ library(sleuth)
$ library(tidyverse)
```

Set the path to your kallisto output files:

```
$ kall_path <- {PATH/TO/FILES}
```

Get sample identifiers from names of kallisto runs:

```
$ sample_id <- dir(file.path(kall_path))
```

Get the directories where the kallisto runs are saved:

```
$ kal_dirs <- data.frame(sample_id = sample_id, path = file.path(kall_path, sample_id))
```

Load the table that associates sample identifiers with treatments and add file paths. You will need to create this yourself (*see* **Note 20**):

```
$ s2c<-read_table("{PATH/TO/TABLE}")%>% full_join(kal_dirs, by=c("sample_id" = "sample_id")
```

Load gene to transcript map (*see* **Note 21**):

```
$ t2g<-read_table("{T2G_FILE"})
```

Run sleuth prep, note this aggregates transcript level counts into gene level counts:

```
$ so<-sleuth_prep(s2c, extra_bootstrap_summary=TRUE, read_bootstrap_tpm=TRUE, target_mapping =
t2g, aggregation_column = 'gene_id')
```

Fit a sleuth model (*see* **Note 22**):

```
$ so<-sleuth_fit(so, ~treatment, 'full')
$ so<-sleuth_wt(so, "inf", which_model = "full")
```

Output results:

```
de_genes <- sleuth_results(so, test="inf")
```

Note that there are many quality control and plotting options available in sleuth, which can be explored using the built-in Shiny server. To launch run:

```
$ sleuth_live(so)
```

3 Notes

Note 1. In addition to the references presented in the introduction, literature reviews that focus on different pathways or responses can provide additional details as to the expected structure and function of immune system components (e.g. on antiviral immunity [80], or the Imd [81], JAK/STAT [82], or Toll [83] pathways). While studies of the *Drosophila* immune system provide a rich knowledge base for understanding insect immunity, this model should be considered as a sample of the full spectrum immunity in insects. Experimental examination of immune responses in other insects have revealed many features that are widespread, such as melanization reactions and presence of the principal immune signaling pathways. However, they have also identified many lineage-specific features that differ greatly from observations to date in flies. For example, adult *Drosophila* have very few

circulating hemocytes (blood cells) [84] so the relative importance of cellular immunity is probably underestimated in *Drosophila* relative to other insects. With the great diversity of insect species (over 500 million years of evolution), and the variety of pathogens they encounter in their various ecological niches, such differences are to be expected.

Note 2. Immune-related genes of the canonical repertoire in fact comprise many genes that may not have direct experimental evidence supporting their roles in immunity. It is also important to note that many genes and pathways have pleiotropic functions, meaning a single gene can produce proteins that are involved in different biological processes, so being classified as a canonical immunity gene does not preclude involvement in other processes. Similarly, the sub-classification of genes into recognition, signal transduction, modulation, or defense/effector phases is a useful framework, but it does not necessarily exclude the possibility of the protein being involved in other processes.

Note 3. For gene models with alternative transcripts, it is advisable to collect the sequences for each transcript that produces a distinct protein product through alternative splicing, because (*i*) annotation prediction of alternative transcripts by automated pipelines is particularly challenging so having a reference set of possible transcripts will help to build accurate gene models during curation; and (*ii*) being able to select equivalent transcripts will make downstream phylogenetic analyses more robust and, in the case of alternatively spliced protein domains, will allow for domain-based analyses. It should also be noted that sequence homology searches with the different protein products of alternative transcripts may obscure truly reciprocal best matches at the level of the gene. These can generally be resolved by examining the genomic loci to determine equivalence at the transcript level.

Note 4. Performing global searches of all the compiled sets of reference protein sequences against the proteins from the gene set will require running some bioinformatics sequence analysis tools. Working with colleagues who have experience running such analyses will allow novice team members to learn these key skills. Installing the required software and setting up the resources to run a local genome browser and sequence search interface can be achieved with a range of freely available bioinformatics tools. Aligning RNA-seq reads to

the genome assembly and producing tracks for visualization in a genome browser will greatly facilitate the process of manually curating the candidate immune-related genes. Providing detailed instructions for installing and running these tools is beyond the scope of this chapter. Instead, team members should be able to relatively easily set up these necessary resources following instructions in the references and links provided herein. These tools will greatly facilitate both the gene identification and curation steps, e.g. being able to visualize the genomic locations of the sequences that produce significant matches to the reference protein sequences (using the tabular tBLASTn results) in order to find genes that may have been missed by the automated gene prediction pipeline as well as highlighting possible errors in the predicted gene models that need to be corrected during manual curation.

Note 5. Examining the results from running InterProScan on the compiled sets of reference proteins will provide an up-to-date summary of which proteins encoded in the target genome contain domains that are characteristic of members of the canonical immune gene repertoire. It is important to note that InterPro entry types range from general to specific: homologous superfamily, protein family, domain, repeat, or site. Thus the more general entry types may recognize a much broader set of proteins than the immune genes of interest. For example, the prophenoloxidases (PPOs) are recognized by the 'Hemocyanin/hexamerin' family (IPR013788) profile, which also recognizes insect hexamerins (storage proteins).

Note 6. The alignments that define significant matches between the reference protein sequences and the target assembly are not expected to correspond perfectly to the predicted gene model in the target species. Evolutionary divergence between the reference and target species means that only the relatively well conserved regions of most proteins will produce confident alignments. Highly diverged regions, regions of low-complexity sequence, and short exons may produce no significant hits and therefore could appear as non-supported parts of the gene model. In addition, the alignment boundaries are unlikely to match exactly the intron/exon boundaries of the gene model since tBLASTn searches do not take putative splice sites into account. Thus, the homology searches serve to identify the most likely genomic loci encoding genes of interest and they provide support for the predicted gene

model, but differences between the alignment coordinates and the gene model are to be expected.

Note 7. Detailed practical guidelines for performing manual curation of predicted gene models and assessing the supporting evidence using the Apollo online collaborative genomic annotation editor are provided in the documentation and user guide materials (http://genomearchitect.github.io). Additional training materials include several webinars available through YouTube, e.g. from the Bioinformatics Platform for Agroecosystem Arthropods https://www.youtube.com/watch?v=BMeSwdKiO_E or from the European Molecular Biology Laboratory Australia Bioinformatics Resource https://www.youtube.com/watch?v=Wec7ZlXykQc.

Note 8. The simplest possible experimental design is a single control (three replicates of either untreated insects or sterilely wounded insects) compared to three replicates of infected insects assayed at a single timepoint post-infection. More complicated experiments might include a time series after infection to capture transcriptional dynamics in response to infection. Depending on the goals and scope of the project, a variety of options are feasible. More complex designs (e.g., those with more than a single control and a single infected treatment) will require more complicated analysis.

Note 9. Exact age of insects will depend substantially on the species and goals of the project (e.g., comparisons across life stages or sexes may be of interest). In general, to minimize uncontrolled noise, ensuring that the experimental insects are of roughly the same age and the same sex is standard practice. The number of individual insects depends on size and the amount of RNA that can be obtained from single individuals. Your sequencing provider can tell you how much starting material is necessary for library preparation, which provides a starting point for the infection experimental design.

Note 10. Challenge with a single bacterial strain will give a clean measurement of the transcriptional response to that bacterium, whereas challenge with a pool of bacterial species (e.g., including both Gram-negative and Gram-positive) will reveal a broader spectrum of responses but will not allow determination of which genes are responding to which microbe.

Live bacterial infection will stimulate transcriptional responses to both the presence of bacteria (e.g., immune stimulation by peptidoglycan) as well as responses to pathogenic damage caused by infection, which can also be a strong trigger of immune responses [85]. The ideal bacterial concentration is one that is sufficient to induce a strong immune response without causing substantial mortality so that immune responses do not become conflated with transcriptional signatures of death. In most cases pilot experiments using different concentrations and measuring mortality over time will be necessary to calibrate the proper dosage. Challenge with dead bacteria or purified bacterial components eliminates concerns about host mortality and often is sufficient for stimulating a robust response [25]. It should be noted that some pathogens are capable of suppressing host responses (e.g. [86]), so heat-killing these prior to infection may yield a stronger response. Pathogens such as viruses, nematodes and protozoa generally need to be alive in order to infect so these should not be heat-killed unless required by the specific objectives of the experiment. A standard method for culturing bacteria prior to infecting *D. melanogaster* is shown visually in [75].

Note 11. Delivering infection by septic pinprick is less quantitatively controlled than performing injections with a microcapillary needle, but also requires less equipment and technical proficiency. For many experimental designs, especially those using a mixed pool of bacteria to elicit a broad spectrum immune response, precise quantification of the challenges is probably unnecessary. It should be noted, however, that septic pinprick delivers fairly low infection dose that may not be sufficient to stimulate a robust response in large insects such as large caterpillars and beetles. For these insects, microcapillary injection may be required.

Note 12. The time after infection at which to measure expression is an important decision. Bacterial infections elicit a rapid response in insects, and sampling at 8-12 hours post-infection is common and experimentally convenient (allowing infection in the morning and freezing of infected insects in the evening, or infections in the evening and freezing the following morning) [7, 87, 88]. However, transcriptional dynamics vary depending on the pathogenic agent and other experimental variables [25, 89]. Therefore it is advisable to perform preliminary experiments before collecting samples for sequencing to be able to select the

most appropriate conditions and time points. These pilot studies could involve low-coverage RNA-seq from a single sample across multiple time points or could involve quantitative PCR of candidate immune effectors, such as antimicrobial peptides, that provide reliable readouts of immune system activation.

Note 13. In general, power to detect differential expression scales more with replicate number than with reads per sample [71]. So for a fixed amount of sequencing, there is more experimental gain in sequencing a greater number of replicates to individually lower depth than sequencing fewer replicates to higher depth. However, given a fixed number of replicates, increasing depth will also increase resolution and power up to a point. Sequencing depth can be adjusted to the scope of the project and available budget.

Note 14. There are two approaches to determining which transcript a read arises from. The traditional approach uses standard read alignment metrics to map a particular read to a genome (or transcriptome) sequence, and then uses the mapping position to determine the transcript. There are many programs that can perform this alignment procedure, as recent benchmarking studies show [90]. The pseudoalignment approach instead uses representations of transcripts and reads to find a fast match; this has the benefit of greatly increased speed and computational efficiency, at no cost to accuracy [91].

Note 15. For the purposes of identifying genes regulated by infection, aggregating results to gene-level summaries (in which expression values are aggregated across all alternative isoforms of a gene) is often the most desirable outcome. There is some debate about the best way to do this e.g. [92]; we have presented one option but there are alternatives such as those described in the discussion here: https://pachterlab.github.io/sleuth/walkthroughs. In addition, when evaluating alternative splicing and related questions, it is essential to estimate transcript-level differential expression instead of gene-level differential expression.

Note 16. We present a method using kallisto [78] to generate expression estimates for use in downstream pipelines, but there are several alternatives, including salmon, which also uses pseudoalignment [93], RSEM, which uses full alignment [94], and others. Kallisto has the

considerable advantage of low compute requirements, meaning a typical experiment can be analyzed on a laptop computer without the need for dedicated computing clusters.

Note 17. Trimming low quality reads generally is not necessary for RNA-seq differential expression analysis, although removing adaptors can be useful if your reads have substantial adaptor contamination. There are a number of tools for doing this, including Trimmomatic [95], and NGmerge [96].

Note 18. There are a wide variety of R packages that can fit differential expression models to RNA-seq data, including DESeq2 [97], limma voom [98], and edgeR [99]. We focus on sleuth here, as it is designed to work with the output of kallisto, but all of the listed tools perform well.

Note 19. For most packages, including tidyverse and dependencies (but not sleuth), it should be possible to install them using the `install.packages("{PACKAGE NAME}")` command. See the tidyverse documentation and the sleuth documentation for additional details.

Note 20. Sleuth requires a table that has sample_id as one column, and the treatment (e.g., infected, control) as the second column, in order to match samples to conditions. This can be prepared in Excel or similar spreadsheet software, saved as a CSV file, and loaded into R.

Note 21. To aggregate transcript-level results into gene-level counts requires a file mapping transcript identifiers to gene identifiers. This should be a text file with two columns, one with transcript identifiers matching the transcripts used in kallisto, and the other with gene_id.

Note 22. Sleuth uses two approaches to estimate significance of differential expression. A Wald test, which compares two conditions, and a likelihood ratio test, which can compare arbitrary nested models. In this case, we show how to run a simple Wald test comparing an infected sample and control sample, for a simple experiment with only two conditions. For more complex experiments, a likelihood ratio test may be more useful. See the sleuth manual for details.

Acknowledgements

RMW acknowledges the Swiss National Science Foundation (grants PP00P3_170664 and CRSII5_186397), the Department of Ecology and Evolution, University of Lausanne, and Swiss Institute of Bioinformatics, Switzerland. BPL acknowledges the Cornell Institute of Host-Microbe Interactions and Disease, Department of Entomology, Cornell University, Ithaca, New York, USA. TBS acknowledges the Informatics Group, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, USA.

References

1. Ligoxygakis P (2017) Advances in insect physiology. Volume 52, Insect immunity. Academic Press, San Diego
2. Buchon N, Silverman N, Cherry S (2014) Immunity in Drosophila melanogaster-from microbial recognition to whole-organism physiology. Nat Rev Immunol 14:796–810. doi: 10.1038/nri3763
3. Imler J-L (2014) Overview of Drosophila immunity: a historical perspective. Dev Comp Immunol 42:3–15. doi: 10.1016/j.dci.2013.08.018
4. Rolff J, Reynolds SE (2009) Insect infection and immunity: evolution, ecology, and mechanisms. Insect Infect Immun Evol Ecol Mech. doi: 10.1093/acprof:oso/9780199551354.001.0001
5. Ferrandon D, Imler J-L, Hetru C, Hoffmann JA (2007) The Drosophila systemic immune response: sensing and signalling during bacterial and fungal infections. Nat Rev Immunol 7:862–74. doi: 10.1038/nri2194
6. Lemaitre B, Hoffmann J (2007) The host defense of Drosophila melanogaster. Annu Rev Immunol 25:697–743. doi: 10.1146/annurev.immunol.25.022106.141615
7. Sackton TB, Lazzaro BP, Clark AG, Wittkopp P (2017) Rapid expansion of immune-related gene families in the house fly, Musca domestica. Mol Biol Evol 34:857–872. doi: 10.1093/molbev/msw285
8. Barribeau SM, Sadd BM, du Plessis L, et al (2015) A depauperate immune repertoire precedes evolution of sociality in bees. Genome Biol 16:83. doi: 10.1186/s13059-015-0628-y
9. Bartholomay LC, Waterhouse RM, Mayhew GF, et al (2010) Pathogenomics of Culex quinquefasciatus and meta-analysis of infection responses to diverse pathogens. Science (80-) 330:88–90. doi: 10.1126/science.1193162
10. Gerardo NM, Altincicek B, Anselme C, et al (2010) Immunity and other defenses in pea aphids, Acyrthosiphon pisum. Genome Biol 11:R21. doi: 10.1186/gb-2010-11-2-r21
11. Tanaka H, Ishibashi J, Fujita K, et al (2008) A genome-wide analysis of genes and gene families involved in innate immunity of Bombyx mori. Insect Biochem Mol Biol 38:1087–1110. doi: 10.1016/j.ibmb.2008.09.001
12. Sackton TB, Lazzaro BP, Schlenke TA, et al (2007) Dynamic evolution of the innate immune system in Drosophila. Nat Genet 39:1461–1468. doi: 10.1038/ng.2007.60
13. Waterhouse RM, Kriventseva E V., Meister S, et al (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. Science (80-) 316:1738–1743. doi: 10.1126/science.1139862
14. Zou Z, Evans JD, Lu Z, et al (2007) Comparative genomic analysis of the Tribolium immune system. Genome Biol 8:R177. doi: 10.1186/gb-2007-8-8-r177
15. Evans JD, Aronstein K, Chen YP, et al (2006) Immune pathways and defence mechanisms in honey bees Apis mellifera. Insect Mol Biol 15:645–656. doi: 10.1111/j.1365-2583.2006.00682.x
16. Christophides GK, Zdobnov E, Barillas-Mury C, et al (2002) Immunity-related genes and gene families in Anopheles gambiae. Science (80-) 298:159–165. doi: 10.1126/science.1077136
17. Waterhouse RM, Wyder S, Zdobnov EM (2008) The Aedes aegypti genome: A comparative perspective. Insect Mol Biol 17:1–8. doi: 10.1111/j.1365-2583.2008.00772.x

18. Olafson PU, Aksoy S, Attardo GM, et al (2019) Functional genomics of the stable fly, Stomoxys calcitrans, reveals mechanisms underlying reproduction, host interactions, and novel targets for pest control. bioRxiv 623009. doi: 10.1101/623009
19. Altincicek B, Vilcinskas A (2007) Analysis of the immune-inducible transcriptome from microbial stress resistant, rat-tailed maggots of the drone fly Eristalis tenax. BMC Genomics 8:326. doi: 10.1186/1471-2164-8-326
20. Sackton TB, Clark AG (2009) Comparative profiling of the transcriptional response to infection in two species of Drosophila by short-read cDNA sequencing. BMC Genomics 10:259. doi: 10.1186/1471-2164-10-259
21. Sackton TB (2019) Comparative genomics and transcriptomics of host–pathogen interactions in insects: evolutionary insights and future directions. Curr Opin Insect Sci 31:106–113. doi: 10.1016/J.COIS.2018.12.007
22. Ekengren S, Hultmark D (2001) A family of Turandot-related genes in the humoral stress response of Drosophila. Biochem Biophys Res Commun 284:998–1003. doi: 10.1006/bbrc.2001.5067
23. Brun S, Vidal S, Spellman P, et al (2006) The MAPKKK Mekk1 regulates the expression of Turandot stress genes in response to septic injury in Drosophila. Genes to Cells 11:397–407. doi: 10.1111/j.1365-2443.2006.00953.x
24. De Gregorio E, Spellman PT, Rubin GM, Lemaitre B (2001) Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. Proc Natl Acad Sci 98:12590–12595. doi: 10.1073/pnas.221458698
25. Troha K, Im JH, Revah J, et al (2018) Comparative transcriptomics reveals CrebA as a novel regulator of infection tolerance in D. melanogaster. PLOS Pathog 14:e1006847. doi: 10.1371/journal.ppat.1006847
26. Hillyer JF (2016) Insect immunology and hematopoiesis. Dev Comp Immunol 58:102–118. doi: 10.1016/j.dci.2015.12.006
27. Bartholomay LC, Michel K (2018) Mosquito immunobiology: the intersection of vector health and vector competence. Annu Rev Entomol 63:145–167. doi: 10.1146/annurev-ento-010715-023530
28. Legeai F, Shigenobu S, Gauthier JP, et al (2010) AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. Insect Mol Biol 19 Suppl 2:5–12. doi: IMB930 [pii] 10.1111/j.1365-2583.2009.00930.x
29. Kersey PJ, Allen JE, Allot A, et al (2018) Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res 46:D802–D808. doi: 10.1093/nar/gkx1011
30. Thurmond J, Goodman JL, Strelets VB, et al (2019) FlyBase 2.0: the next generation. Nucleic Acids Res 47:D759–D765. doi: 10.1093/nar/gky1003
31. Elsik CG, Tayal A, Unni DR, et al (2018) Hymenoptera Genome Database: Using HymenopteraMine to enhance genomic studies of hymenopteran insects. Methods Mol Biol 1757:513–556. doi: 10.1007/978-1-4939-7737-6_17
32. Poelchau MF, Chen MJM, Lin YY, Childers CP (2018) Navigating the i5k workspace@NAL: A resource for arthropod genomes. Methods Mol Biol 1757:557–577. doi: 10.1007/978-1-4939-7737-6_18
33. Sayers EW, Agarwala R, Bolton EE, et al (2019) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 47:D23–D28. doi: 10.1093/nar/gky1069
34. Giraldo-Calderón GI, Emrich SJ, MacCallum RM, et al (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res 43:D707-13. doi: 10.1093/nar/gku1117
35. Attardo GM, Abd-Alla AMM, Acosta-Serrano A, et al (2019) The Glossina genome cluster: comparative genomic analysis of the vectors of African trypanosomes. bioRxiv 531749. doi: 10.1101/531749
36. Li M, Christen JM, Dittmer NT, et al (2018) The Manduca sexta serpinome: Analysis of serpin genes and proteins in the tobacco hornworm. Insect Biochem Mol Biol 102:21–30. doi: 10.1016/j.ibmb.2018.09.008
37. Adelman ZN, Myles KM (2018) The C-type lectin domain gene family in Aedes aegypti and their role in arbovirus infection. Viruses 10:367. doi: 10.3390/v10070367
38. Matetovici I, Van Den Abbeele J (2018) Thioester-containing proteins in the tsetse fly (*Glossina*) and their response to trypanosome infection. Insect Mol Biol 27:414–428. doi: 10.1111/imb.12382
39. Yang L, Mei Y, Fang Q, et al (2017) Identification and characterization of serine protease inhibitors in a parasitic wasp, Pteromalus puparum. Sci Rep 7:15755. doi: 10.1038/s41598-017-16000-5
40. Neafsey DE, Waterhouse RM, Abai MR, et al (2015) Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science (80-) 347:1258522–1258522. doi: 10.1126/science.1258522
41. Terrapon N, Li C, Robertson HM, et al (2014) Molecular traces of alternative social organization in a termite genome. Nat Commun 5:3636. doi: 10.1038/ncomms4636

42. Brucker RM, Funkhouser LJ, Setia S, et al (2012) Insect Innate Immunity Database (IIID): an annotation tool for identifying immune genes in insect genomes. PLoS One 7:e45125. doi: 10.1371/journal.pone.0045125
43. Bryant B, Ungerer MC, Liu Q, et al (2010) A caspase-like decoy molecule enhances the activity of a paralogous caspase in the yellow fever mosquito, Aedes aegypti. Insect Biochem Mol Biol 40:516–523. doi: 10.1016/j.ibmb.2010.04.011
44. Waterhouse RM, Povelones M, Christophides GK (2010) Sequence-structure-function relations of the mosquito leucine-rich repeat immune proteins. BMC Genomics 11:531. doi: 10.1186/1471-2164-11-531
45. Zou Z, Picheng Z, Weng H, et al (2009) A comparative analysis of serpin genes in the silkworm genome. Genomics 93:367–375. doi: 10.1016/j.ygeno.2008.12.010
46. Camacho C, Coulouris G, Avagyan V, et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. doi: 10.1186/1471-2105-10-421
47. Jones P, Binns D, Chang HY, et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. doi: 10.1093/bioinformatics/btu031
48. Mitchell AL, Attwood TK, Babbitt PC, et al (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 47:D351–D360. doi: 10.1093/nar/gky1100
49. Löytynoja A (2014) Phylogeny-aware alignment with PRANK. Methods Mol Biol 1079:155–170. doi: 10.1007/978-1-62703-646-7_10
50. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–80. doi: 10.1093/molbev/mst010
51. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195. doi: 10.1371/journal.pcbi.1002195
52. Waterhouse RM (2015) A maturing understanding of the composition of the insect gene repertoire. Curr Opin Insect Sci 7:15–23. doi: 10.1016/j.cois.2015.01.004
53. Waterhouse RM, Seppey M, Simão FA, et al (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35:543–548. doi: 10.1093/molbev/msx319
54. Waterhouse RM, Seppey M, Simão FA, Zdobnov EM (2019) Using BUSCO to assess insect genomic resources. In: Methods Mol. Biol. Humana Press, New York, NY, pp 59–74
55. Pennisi E (2000) Ideas fly at gene-finding jamboree. Science (80-) 287:2182–2184. doi: 10.1126/science.287.5461.2182
56. Saha S, Hosmani PS, Villalobos-Ayala K, et al (2017) Improved annotation of the insect vector of citrus greening disease: biocuration by a diverse genomics community. Database (Oxford). doi: 10.1093/database/bax032
57. Hosmani PS, Shippy T, Miller S, et al (2019) A quick guide for student-driven community genome annotation. PLoS Comput Biol 15:e1006682. doi: 10.1371/journal.pcbi.1006682
58. Ioannidis P, Simao FA, Waterhouse RM, et al (2017) Genomic features of the damselfly Calopteryx splendens representing a sister clade to most insect orders. Genome Biol Evol 9:415–430. doi: 10.1093/gbe/evx006
59. Buels R, Yao E, Diesh CM, et al (2016) JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol 17:66. doi: 10.1186/s13059-016-0924-1
60. Dunn NA, Unni DR, Diesh C, et al (2019) Apollo: Democratizing genome annotation. PLoS Comput Biol 15:e1006790. doi: 10.1371/journal.pcbi.1006790
61. Priyam A, Woodcroft BJ, Rai V, et al (2015) Sequenceserver: a modern graphical user interface for custom BLAST databases. bioRxiv 033142. doi: 10.1101/033142
62. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360. doi: 10.1038/nmeth.3317
63. Dobin A, Davis CA, Schlesinger F, et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. doi: 10.1093/bioinformatics/bts635
64. Ramírez F, Ryan DP, Grüning B, et al (2016) deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 44:W160–W165. doi: 10.1093/nar/gkw257
65. Altenhoff AM, Dessimoz C (2012) Inferring orthology and paralogy. Methods Mol Biol 855:259–279. doi: 10.1007/978-1-61779-582-4_9
66. Hall BG (2013) Building phylogenetic trees from molecular data with MEGA. Mol Biol Evol. doi: 10.1093/molbev/mst012
67. Conesa A, Madrigal P, Tarazona S, et al (2016) A survey of best practices for RNA-seq data analysis. Genome Biol 17:13. doi: 10.1186/s13059-016-0881-8
68. Crawford JE, Guelbeogo WM, Sanou A, et al (2010) De novo transcriptome sequencing in Anopheles funestus using Illumina RNA-seq technology. PLoS One 5:e14202. doi: 10.1371/journal.pone.0014202
69. Haas BJ, Papanicolaou A, Yassour M, et al (2013) De novo transcript sequence reconstruction from

RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512. doi: 10.1038/nprot.2013.084
70. Holt C, Yandell M (2011) MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491. doi: 10.1186/1471-2105-12-491
71. Ching T, Huang S, Garmire LX (2014) Power analysis and sample size estimation for RNA-Seq differential expression. RNA 20:1684–1696. doi: 10.1261/rna.046011.114
72. Schurch NJ, Schofield P, Gierliński M, et al (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22:839–851. doi: 10.1261/rna.053959.115
73. Robles JA, Qureshi SE, Stephen SJ, et al (2012) Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics 13:484. doi: 10.1186/1471-2164-13-484
74. Todd E V., Black MA, Gemmell NJ (2016) The power and promise of RNA-seq in ecology and evolution. Mol Ecol 25:1224–1241. doi: 10.1111/mec.13526
75. Khalil S, Jacobson E, Chambers MC, Lazzaro BP (2015) Systemic bacterial infection and immune defense phenotypes in Drosophila melanogaster. J Vis Exp e52613. doi: 10.3791/52613
76. Taylor K, Kimbrell DA (2007) Host immune response and differential survival of the sexes in Drosophila. Fly (Austin) 1:197–204. doi: 10.4161/fly.5082
77. Schlüns H, Sadd BM, Schmid-Hempel P, Crozier RH (2010) Infection with the trypanosome Crithidia bombi and expression of immune-related genes in the bumblebee Bombus terrestris. Dev Comp Immunol 34:705–709. doi: 10.1016/j.dci.2010.02.002
78. Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34:525–527. doi: 10.1038/nbt.3519
79. Pimentel H, Bray NL, Puente S, et al (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods 14:687–690. doi: 10.1038/nmeth.4324
80. Mussabekova A, Daeffler L, Imler J-L (2017) Innate and intrinsic antiviral immunity in Drosophila. Cell Mol Life Sci 74:2039–2054. doi: 10.1007/s00018-017-2453-9
81. Myllymäki H, Valanne S, Rämet M (2014) The Drosophila imd signaling pathway. J Immunol 192:3455–62. doi: 10.4049/jimmunol.1303309
82. Myllymäki H, Rämet M (2014) JAK/STAT pathway in Drosophila immunity. Scand J Immunol 79:377–85. doi: 10.1111/sji.12170
83. Valanne S, Wang J-H, Rämet M (2011) The Drosophila Toll signaling pathway. J Immunol 186:649–56. doi: 10.4049/jimmunol.1002302
84. Bosch PS, Makhijani K, Herboso L, et al (2019) Blood cells of adult Drosophila do not expand, but control survival after bacterial infection by induction of Drosocin around their reservoir at the respiratory epithelia. bioRxiv 578864. doi: 10.1101/578864
85. Buchon N, Poidevin M, Kwon H-M, et al (2009) A single modular serine protease integrates signals from pattern-recognition receptors upstream of the Drosophila Toll pathway. Proc Natl Acad Sci 106:12442–12447. doi: 10.1073/pnas.0901924106
86. Apidianakis Y, Mindrinos MN, Xiao W, et al (2005) Profiling early infection responses: Pseudomonas aeruginosa eludes host defenses by suppressing antimicrobial peptide gene expression. Proc Natl Acad Sci 102:2573–2578. doi: 10.1073/pnas.0409588102
87. Sackton TB, Werren JH, Clark AG (2013) Characterizing the infection-induced transcriptome of Nasonia vitripennis reveals a preponderance of taxonomically-restricted immune genes. PLoS One 8:e83984. doi: 10.1371/journal.pone.0083984
88. Gupta SK, Kupper M, Ratzka C, et al (2015) Scrutinizing the immune defence inventory of Camponotus floridanus applying total transcriptome sequencing. BMC Genomics 16:540. doi: 10.1186/s12864-015-1748-1
89. Doublet V, Poeschl Y, Gogol-Döring A, et al (2017) Unity in defence: Honeybee workers exhibit conserved molecular responses to diverse pathogens. BMC Genomics 18:207. doi: 10.1186/s12864-017-3597-6
90. Baruzzo G, Hayer KE, Kim EJ, et al (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods 14:135–139. doi: 10.1038/nmeth.4106
91. Yi L, Pimentel H, Bray NL, Pachter L (2018) Gene-level differential analysis at transcript-level resolution. Genome Biol 19:53. doi: 10.1186/s13059-018-1419-z
92. Soneson C, Love MI, Robinson MD (2016) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 4:1521. doi: 10.12688/f1000research.7563.2
93. Patro R, Duggal G, Love MI, et al (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14:417–419. doi: 10.1038/nmeth.4197
94. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a

reference genome. BMC Bioinformatics 12:323. doi: 10.1186/1471-2105-12-323
95. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–20. doi: 10.1093/bioinformatics/btu170
96. Gaspar JM (2018) NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. BMC Bioinformatics 19:536. doi: 10.1186/s12859-018-2579-2
97. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. doi: 10.1186/s13059-014-0550-8
98. Law CW, Chen Y, Shi W, Smyth GK (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 15:R29. doi: 10.1186/gb-2014-15-2-r29
99. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140. doi: 10.1093/bioinformatics/btp616

Figures

Figure 1. Workflow of steps required for canonical immune gene identification.

Protein sequences of immune-related genes from selected reference species are first collected based on the current knowledge of insect innate immunity. These are then used as reference query sequences and sequence hidden Markov model (HMM) profiles for homology searches of the gene set (protein sequences) of the target species to be investigated. Complementary protein domain searches are used to identify genes that contain domains in common with the reference immunity genes. Results from the sequence and domain searches are then used to prioritize the inspection of the candidate immunity genes and curate their predicted gene models to ensure they are as complete and accurate as possible. This will benefit from the results from homology searches of the reference query sequences against the genome assembly as well aligned RNA sequencing (RNAseq) reads from the target species. Combined phylogenetic analysis of homologous reference and target candidate sequences to build gene trees then allows for the confirmation or rejection of the candidate immune-related genes and the characterization of their orthologous or paralogous relationships.

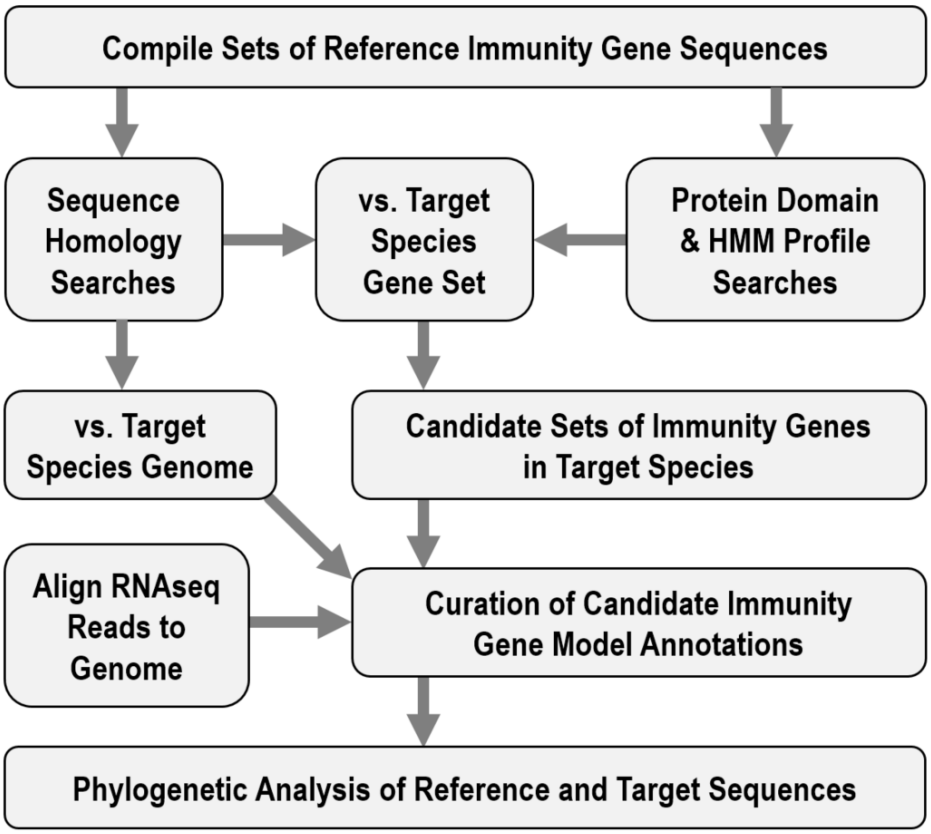

Figure 2. Example of how manual curation can improve automatically predicted gene models.

The top panel shows the curated gene model and the original prediction of the *Dicer-2* gene on the reverse strand (i.e. the 5′ start is on the right and the 3′ end is on the left of the figure) from a mosquito genome. Exons are shown as rectangles connected with lines indicating introns, with predicted coding sequence (CDS) regions in light blue and predicted untranslated regions (UTRs) shown in white. RNAseq read coverage is presented below the gene models in dark blue, clearly showing where reads from the mature messenger RNA align to the genome. Below that are alignments from tBLASTn searches with the Dicer-2 protein (AGAP012289) and the Dicer-1 protein (AGAP002836) from *Anopheles gambiae* (the reference immune protein sequences). The lower panel shows the alignments of individual RNAseq reads to this locus (in dark grey, with colors indicating mismatches between the reads and the reference genome assembly), with reads that map across potential splice junctions connected with black lines. Editing just one intron-exon boundary to match the supporting RNAseq and tBLASTn evidence (shown with the red arrow) corrects the gene model. The first six exons were incorrectly predicted to form a multi-exon 5′ UTR (all white rectangles) in the original gene model. In the curated gene model all six exons now form part of the CDS (i.e. the regions that will be translated into protein), with just a short 5′ UTR at the start of the first exon. The translation of the curated gene model now encodes a full-length Dicer-2 protein.

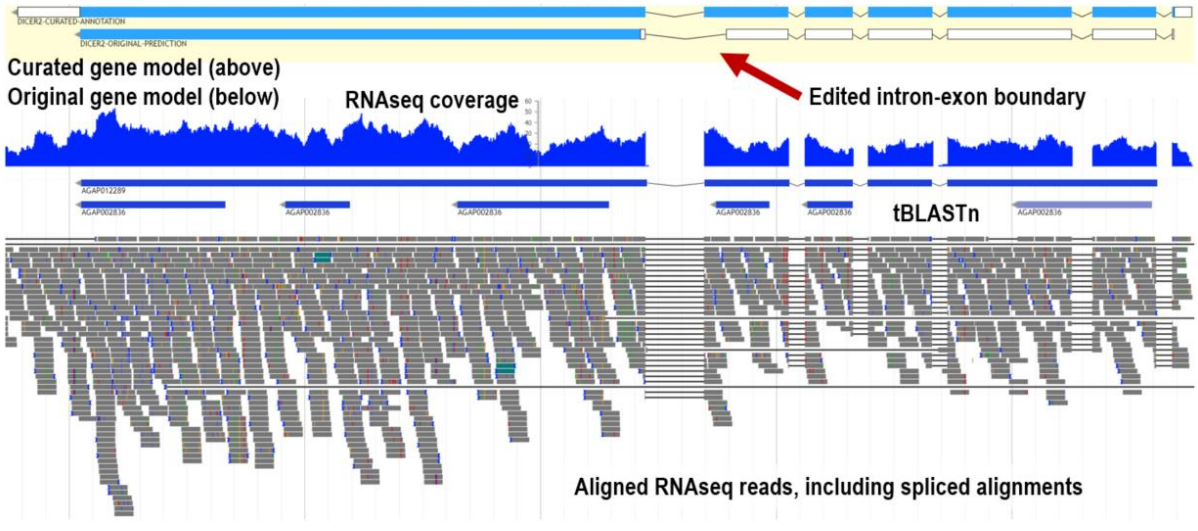

Figure 3. Workflow of steps required for immune transcriptome analysis.

Immune transcriptome analysis can proceed once the RNAseq reads (in fastq format) from all the infection and control samples have been obtained. The analysis also requires the complete set of transcripts from the gene set annotation of the target species, which may also contain updated gene model annotations based on manual curation described in **Section 2.1.3** of this chapter. In the kallisto documentation, this complete set of transcripts is referred to as the 'reference transcriptome' to which the RNA-seq reads will be mapped. RNA-seq reads (possibly after pre-processing; *see* **Note 17**) are mapped to transcripts by kallisto using a pseudoalignment step that then allows for the quantification of transcript abundances from each condition to determine expression levels of each gene and isoform. Finally, differential expression of genes and isoforms among conditions is modeled using sleuth/R to define sets of infection-responsive genes.