Microevolutionary selection dynamics acting on immune genes of the green veined white butterfly, *Pieris napi*

Naomi L.P. Keehnen¹, Jason Hill¹, Sören Nylin¹, and Christopher W. Wheat¹

Author affiliation:

¹ Department of Zoology, Stockholm University, S-106 91 Stockholm, Sweden

Corresponding author: Naomi L. P. Keehnen Department of Zoology, Stockholm University, S-106 91 Stockholm, Sweden, naomi.keehnen@gmail.com

Keywords: selection dynamics, microevolutionary dynamics, innate immunity, genomics, Lepidoptera

Abstract

Insects rely on their innate immune system to successfully mediate complex interactions with their microbiota, as well as the microbes present in the environment. Previous work has shown that components of the canonical immune gene repertoire evolve rapidly and have evolutionary characteristics originating from interactions with fast-evolving microorganisms. Although these interactions are likely to vary among populations, there is a poor understanding of the microevolutionary dynamics of immune genes, especially in non-Dipteran insects. Here we use the full set of canonical insect immune genes to investigate microevolutionary dynamics acting on these genes between and among populations by comparing three allopatric populations of the Green Veined White butterfly, *Pieris napi* (Linné; Lepidoptera, Pieridae). Immune genes showed increased genetic diversity compared to genes from the rest of the genome and various functional categories exhibited different types of signatures of selection, at different evolutionary scales, presenting a complex pattern of selection dynamics. Signatures of balancing selection were identified in 10 genes, and 17 genes appear to be under positive selection. Genes involved with the cellular arm of the immune response as well as the Toll pathway appear to be enriched among our outlier loci, regardless of functional category. This suggests that the targets of selection might focus upon an entire pathway, more than on functional subsets across pathways. Our microevolutionary results are similar to previously observed macroevolutionary patterns from diverse taxa, suggesting that either the immune system is robust to dramatic differences in life history and microbial communities, or that diverse microbes exert similar selection pressures.

Introduction

As one of the most ubiquitous and abundant taxa on Earth, insects occur in a wide variety of environments that contain diverse microbial communities (Costello et al. 2012; Hansen & Moran 2014). Insects must protect themselves from harmful microbes in their environment (Horrocks *et al.* 2011), as well as maintain and regulate their own microbiome for the efficient exploitation of resources (Douglas 2009; Engel & Moran 2013; Broderick 2016). The ability to effectively do both depends upon the host's immune system (Hooper & Macpherson 2010; Lee & Brey 2013). Insects must rely on their innate immune system to successfully mediate these complex interactions and these interactions are likely to vary over time and among populations. However, currently we have a poor understanding of the microevolutionary dynamics acting upon the innate immune system outside a few model species of a single order of Insecta (Diptera e.g. Christophides *et al.* 2002; Lazzaro 2008). Moreover, even beyond these well-studied groups, all microevolutionary studies to date have focused upon subsets of the immune system, rather than the full set of canonical immune genes, complicating general insights into the evolution of canonical immune genes. Here, in order to gain microevolutionary insights into the canonical immune genes, we present the first population genomic study of selection dynamics on the full set of canonical immune genes in the Lepidoptera, one of the few hyper diverse orders of life on earth, containing approximately 7% of all described multicellular animal species (van Nieuwerkerken *et al.* 2011).

Insect immune system genes can be classified into four broad functional categories: pathogen-recognition genes (e.g. PGRPs, TEPs), modulators (e.g. serpins and serine proteases), the genes of the signal transduction pathways (Toll, IMD, JAK-STAT and JNK), and effector genes encoding products that directly interact with microbes (e.g. antimicrobial peptides (AMPs), or defence enzymes (e.g. pro-phenoloxidasases or those involved in the production of reactive oxygen species). Activation of the insect immune system begins with the recognition of non-self through the activation of pattern recognition receptors (PRRs), encoded by recognition genes. After recognition of a pathogen, a sequence of modulation and signalling events is initiated. The Toll and IMD pathways are directly involved with the production of AMPs whereas the JNK and JAK/STAT pathway are indirectly involved with the immune response via cellular responses such as phagocytosis and encapsulation. (Lemaitre & Hoffmann 2007)

Here we refer to the genes in these different categories as the canonical gene set, as they make up the core components of the innate immune system of invertebrates. To date the immune system genes of 35 Diptera, 4 Hymenoptera, 3 Hemiptera, 3 Lepidoptera, and 1 Coleopteran have been characterized (SM: Table 1). While a majority of canonical immune genes are shared among insects, lineage specific diversifications have been identified across several taxa, especially among recognition and effector genes (Christophides *et al.* 2002; Little & Cobbe 2005; Evans *et al.* 2007; Waterhouse *et al.* 2007; Sackton *et al.* 2007; Gerardo *et al.* 2010; Tian *et al.* 2010; Bao *et al.* 2013).

Host-pathogen interactions exert strong selection pressures upon these immune genes, and signatures of this co-evolutionary dynamic are observed in

these genes (Schlenke & Begun 2003; Sackton *et al.* 2007; Waterhouse *et al.* 2007). First, several gene families (e.g. AMPs) exhibit gene birth-death dynamics, observed in species-specific gene duplications and losses, as well as copy number variants among individuals within species (Christophides *et al.* 2002; Palmer & Jiggins 2015). These additional gene copies are thought to provide opportunities for diverse functional changes via neofunctionalization and thereby adaptations to local evolutionary pressures (Innan & Kondrashov 2010; Kondrashov 2012; Waterhouse *et al.* 2007; Sackton *et al.* 2007; Scott *et al.* 2014; Neafsey *et al.* 2015; Rottschaeffer *et al.* 2015). Second, immune genes as a whole appear to evolve faster than the genomic average (Sackton *et al.* 2007; Waterhouse *et al.* 2007), especially genes involved with the recognition of pathogens (e.g. PGRPs), a pattern likely driven by positive selection via host-pathogen interactions (Schlenke & Begun 2003; Sackton *et al.* 2007; Lazzaro & Little 2009; Bulmer *et al.* 2010). However, gaining general insights into these selection dynamics is challenging, as they vary depending upon the molecular function of the genes in question, and the taxa studied. For example, while several AMPs in *D. melanogaster* appear to experience balancing selection (Unckless *et al.* 2016; Unckless & Lazzaro 2016), the AMPs in six bumble bee species show signs of strong purifying selection (Erler *et al.* 2014). Such variation in selection across taxa could arise as a result of spatial and temporal variation in selection, or even due to differences in life history and ecological niche.

These macroevolutionary patterns, which we define as evolutionary processes at or above species level, arise via selection dynamics occurring within populations, so called microevolutionary dynamics. For example, populations of

Anopheles gambiae (Diptera, Culicidae) occupying different ecological niches exhibit signals of population specific positive selection on several immune genes, likely a result of differences in local selection on immune performance (Crawford *et al.* 2012). Population differentiation in immune genes are also seen in *Drosophila melanogaster* (Diptera, Drosophilidae) (Fabian *et al.* 2012) consistent with performance variation among natural variants within populations (Lazzaro *et al.* 2004; Unckless *et al.* 2015; Sleiman *et al.* 2015), as well as balancing selection within populations (Unckless *et al.* 2016; Unckless & Lazzaro 2016). In summary, the insect immune system has components that are highly conserved between species, with some differentiation due to species-specific expansions and losses of functional categories. In general, selection dynamics vary across genes and species, with some population level studies suggestive of local adaptation to population specific immune pressures.

While these findings provide important insights into the selection dynamics acting upon the insect immune system, there are several important biases that warrant attention (reviewed in Keehnen *et al.* 2017). First, the majority of innate immunity studies focus upon macroevolutionary dynamics, i.e. comparing the immune genes between species (e.g. Sackton *et al.* 2007; Barribeau *et al.* 2015), with only few focused upon microevolutionary timescales (SM: Table 1). Second, investigations into microevolutionary dynamics have, with the recent exception of Early and colleagues (2017), primarily focused upon subsets rather than the full canonical immune gene set. Third, the majority of evolutionary studies of the canonical insect immune system have focused on mosquitoes and fruit flies from the order Diptera. These two model organisms are exposed to very different microbes due to their extreme diets (blood vs.

fermenting fruit), which likely impacts their immune system genes. While there have been some studies in other insect taxa, e.g. in Hymenoptera that have looked at bees (pollen feeders; Harpur & Zayed, 2013; Barribeau *et al.* 2015) and ants (omnivores; Viljakainen *et al.* 2009), to date there have not been studies of exclusive herbivores, as found in the Lepidoptera clade, and the potentially different selection pressures this diet could exert. In summary, the innate immunity literature, while providing evidence that different classes of genes experience varying selection dynamics within and among primarily Dipteran species, presents a biased comparative perspective with respect to the full canonical immune gene set. Here we address this gap by conducting a microevolutionary study comparing the selection dynamics on the full canonical immune gene set between three different populations of the Green Veined White butterfly, *P. napi* (Lepidoptera).

P. napi is a common and widespread butterfly that feeds exclusively upon Brassicales plants, with limited dispersal and therefore low gene flow among populations. They, like other lepidopterans, have a female-heterogametic sex (ZW/ZZ chromosomes) system. This species produces between one to four generations per year depending on local conditions, reflecting local adaptation along latitude (Pruisscher *et al.* 2017). This variation in life history makes it an interesting species to investigate local adaptation of the immune system (Porter and Geiger, 1995). Here, in order to investigate the microevolutionary dynamics in this species, we i) performed gene annotations, ii) generated whole genome Pool-seq data for three populations of *P. napi*, and iii) used this data to investigate patterns of selection across immune genes both within and among populations of this species, and between related species species.

Methods

Gene annotation

Here we a draft genome of *P. napi*, previously only used as a database for identifying < 100 genes for a comparative analysis (Meslin *et al.* 2017). As this genome lacks an annotation, we generated a draft annotation using conducted SPALN v2.1.2 (Gotoh 2008), which is a high performance protein to genome alignment program that is exon boundary aware, with the *Danaus plexippus* protein set DPOGS2 (Zhan & Reppert 2013). In addition, the MESPA pipeline (Neethiraj *et al.* 2017), which uses SPALN to construct gene models from highly fragmented genome assemblies, was used to identify the immune genes. A set of 225 immune proteins were used for annotation, with the majority (N=205) of these being the canonical set of immune genes identified in *Bombyx mori* (Lepidoptera, Bombycidae; Tanaka *et al.* 2008), and an additional 20 candidate genes being antimicrobial peptides identified in the closely related species *Pieris rapae* (Lepidoptera, Pieridae; C. Wheat unpublished data).

In order to assess the accuracy of the identified immune genes, we implemented a number of validation steps. First, we used tBLASTn (Altschul *et al.* 1990) with the 225 immune proteins to search the genome assembly, extracting hits with E-values < 0.0001 and a bitscore higher than 60. Secondly, we used MESPA to conduct protein searches of the genome for the 225 candidate genes and generate gene models (n=148 were predicted). We then compared the blast hits with the SPALN gene models, to assess gene model accuracy, filter out potential errors, and identify potential duplicates. In addition, protein sequences from the identified gene models were functionally annotated against Uniref90

(Suzek *et al.* 2007) (downloaded 18 December 2013) and NCBI-nr databases (downloaded 29 October 2015) using BLASTP (thresholds: bitscore > 60 and E-value < 0.0001). As a last quality control step, each immune gene model was manually curated for exon boundaries, reading frame and read depth using IGV (Robinson *et al.* 2011). In order to be confident in our gene models, average read depth per exon was calculated using coverageBed from Bedtools2 (Quinlan & Hall 2010), and compared to the coverage of the genome-wide genes. Immune genes that exhibited a read depth lower than were excluded from further analysis. These approaches resulted in a total of 136 immune-related genes identified in the genome, of which 25 were partially reconstructed (SM: Table 2,3).

Sample collection and DNA extraction

Adult *P. napi* butterflies were collected from three different locations during the summer of 2013. Two of the locations were in the north and south of Sweden, respectively: Abisko: 68°19'12"N: 18°41'40"E and Kullaberg: 55°59'00"N: 13°30'00"E. The third location was in north-eastern Spain (Aiguamolls: 42°13'10"N: 3°6'38E, SM:Table 4). DNA was extracted from thorax material using either a modified version of a salt-extraction protocol (Aljanabi & Martinez 1997) or E-Z 96 Tissue DNA Kit (OMEGA Bio-tek) (SM: Table 4). In both extraction methods RNase A (Qiagen) was added to digest any RNA in the sample. After extraction the DNA samples were quantified and assessed using both spectrophotometric (NanoDrop 8000, Thermo Scientific) and fluorometric quantitation assays (Qubit 2.0, Thermo Scientific). The fragmentation of DNA was assessed on a agarose gel stained with 2% GelRed. Samples with an

absorbance ratio at 260/280 > 1.7 and < 2.0 and minimal fragmentation were selected for pooling.

Library construction and sequencing

All three sources of the DNA for Illumina TruSeq library constructions were composed of 24 individuals, pooled at equimolar levels to 5 µg total. The library from Kullaberg was sex-specific, containing 24 males. The Aiguamolls and Abisko Pool-seq libraries had both equal sex ratios (12 females, 12 males). The libraries had either a 350 or 450 basepair (bp) inserts and were sequenced with paired-end (PE) reads of either 100 or 125 bp length (SM: Table 4).

Population mapping and statistics

The Pool-seq data of the three populations were mapped to the genome using Next-Gen Mapper v0.4.10 with an identity cut-off of 90% (Sedlazeck *et al.* 2013). SAMtools v1.2 was used to filter the mapped data for only correctly mapped paired-end reads, after which a mpileup file was created for further analysis (Li *et al.* 2009). Popoolation v1.2.2 was used to mask indels using a 5 bp window, centered on the indel (identify-genomic-indel-regions.pl, Kofler *et al.* 2011a). After this a uniform 30x coverage mpileup file was created for each population by randomly drawing base calls from the mpileup file without replacement (subsample-pileup.pl, Kofler *et al.* 2011a). Then, Tajima's D, nucleotide diversity (π), and the number of non-synonymous and synonymous sites per gene were calculated (Kofler *et al.* 2011a). The Tajima's D values provide an indication of the selection or demographic events by comparing two different estimates of theta (Tajima, 1989). Under idealized conditions, when a gene is evolving

neutrally, $TajD = 0$. A Tajima's D value smaller than 0 indicates that alleles are present at a lower frequency than expected under the null model, indicative of either a recent selective sweep, linkage to a swept gene, or population expansion after a recent bottleneck. A Tajima's D value larger than 0 is indicative of an excess of alleles present at intermediate frequencies, a signal of either balancing selection or population contraction. Nucleotide diversity quantifies the mean divergence between sequences, and is used to measure the degree of polymorphism within a population at a locus. Between population comparisons were conducted using the fixation index (F_{ST}), a measure of population differentiation, wherein a value of 0 indicates no differentiation and a value of 1 implies complete differentiation, i.e. populations have fixed alternative allelic states (Holsinger & Weir, 2009). F_{ST} were calculated for every Single Nucleotide Polymorphism (SNP), as well as for genes using PoPoolation2 v1.201 (Kofler *et al.* 2011b). F_{ST} of SNPs located within an exon of a gene were selected using the intersectBed function of Bedtools2 (Quinlan & Hall 2010). Two methods were used to assess if the immune genes differed from the genome-wide patterns. First, non-parametric pairwise comparisons, both between categories of genes as well as populations were performed using JMP 12 (SAS Institute Inc., 2014). Second, a permutation subset analysis was performed in R (R core team, 2016) for Tajima's D and nucleotide diversity measures, to control for differences in subset size across functional categories. we estimated the average nucleotide diversity of 10,000 sets of 136 randomly sampled genome-wide genes, and compared this distribution of means to the mean of the 136 immune genes.

Outlier analysis

To distinguish from demographic effects and to identify SNPs likely to have been affected by selection, either as the direct target or via linkage, we used the empirical null distribution of values from the 11,240 predicted gene models. Outlier values for Tajima's D and nucleotide diversity measures were identified using the 2.5 percentiles of the genome-wide distribution as thresholds, consistent with previous publications (Carlson *et al.* 2005; Sterken *et al.* 2009). Outlier values for F_{ST} had a threshold that corresponded to the genome-wide 99.5th percentile (e.g. Fabian *et al.* 2012).

Sex vs autosomal

We investigated whether the significant outliers found in the outlier analysis were due to the lower average nucleotide diversity present on sex chromosomes as these differ in effective population size and inheritance (Charlesworth *et al.* 1987). We used three steps to identify which genes were located on the sex-chromosome. First, we used the nucleotide sequences of the assembled *P. napi* scaffolds and aligned those against the *B. mori* chromosome level assembly, using BLASTn. Secondly, as there is extensive conservation of gene order in Lepidoptera (d'Alencon *et al.* 2010; Ahola *et al.* 2014), we used the *B. mori* immune gene set (Tanaka *et al.* 2008), and aligned those against the *B. mori* chromosomal assembled genome (Shimomura *et al.* 2009) using tBLASTn. Both alignments were filtered on a bitscore >60, and E-value <0.0001, and we only considered the location of the genes to be valid when both the *P. napi* scaffolds and the *B. mori* protein set were in concordance. Finally, we performed a permutation test using R to control for differences in subset size, since only 10 immune genes were found to be on the sex chromosome. We estimated the

average nucleotide diversity of 10,000 sets of 10 randomly sampled autosomal immune genes, and compared this distribution of means to the mean of the 10 sex-chromosomal immune genes.

Test of selection

To test for positive selection, polarized McDonald-Kreitman tests were performed (McDonald & Kreitman 1991), assessing polymorphism and divergence in *P. napi* from the congeneric species *P. rapae* and *P. brassicae* (Lepidoptera, Pieridae). The *P. napi* and *P. rapae* genome sequences used for the MK test were obtained from Meslin *et al.* (2017). The *P. brassicae* genome was constructed using a Pool-seq library from the thoraxes of 24 Spanish butterflies (41.818225°N, 2.302625°E; 12 males and 12 females). The DNA was extracted E-Z 96® Tissue DNA Kit (Omega Bio-tek, cat no: D1196-01) and sequenced by BGI (Beijing Genomics Institute). PCR duplicates were removed using Stacks 1.21 clonefilter (Catchen *et al.* 2013), after which Illumina contaminants were removed and the reads were filtered on a Q10 quality score using BBMap (BBMap v34.86 - Bushnell B. - <http://sourceforge.net/projects/bbmap/>). The cleaned reads were used in CLC Genomics Workbench 5.5.1 (CLC bio) to construct a de novo genome with a kmer size of 55 and bubble size 2000. The resulting, highly fragmented, genome contained 41 956 scaffolds and had a N50 of 31.399 Kb (SM: Table 5). Gene-level completeness of all three genomes was assessed using CEGMA (Parra *et al.* 2007) and (Benchmarking Sets of Universal Single-copy orthologs) using the arthropod gene set (N= 2675, Simão *et al.* 2015). Orthologues of the immune genes were identified in the *P. rapae* and *P. brassicae* genomes using SPALN v2.1.2. Multiple consensus sequences (N=20) of

the coding sequences were extracted from the *P. napi* Abisko mpileup, after which they were aligned with the orthologs of *P. rapae* and *P. brassicae*, using scripts kindly provided by V. Nolte (Nolte *et al.* 2013) that we subsequently modified. CDS alignments used a codon aware, protein based approach as implemented in MACSE (Ranwez *et al.* 2011). Genes were only included in the MK analysis if they were found with a read depth of at least 10x and no more than 250x in all three species. Afterwards, the average read depth of each gene was assessed using BCFTOOLS 1.2 (Li *et al.* 2009) and seqtk (<https://github.com/lh3/seqtk.git>). To further minimize false positives, genes with high proportions of gaps and unknown nucleotides (>5%) were excluded. Genes under positive selection are expected to have elevated rates of amino acid evolution, which can be measured by comparing the ratio of synonymous (P_S) and non-synonymous polymorphism (P_N) to the ratio of synonymous (D_S) and non-synonymous divergence (D_N). The neutrality index ($NI = D_S P_N / D_N P_S$) measures the direction and degree of departure from neutral evolution (Stoletzki & Eyre-Walker 2010). Genes that depart from neutrality (NI), and therefore are under selection, will have an unequal ratio. A neutrality index greater than 1 (i.e. $NI > 1$) a sign of negative selection, resulting in an excess of amino acid polymorphism. Because synonymous mutations are neutral, a neutrality index lower than 1 (i.e. $NI < 1$) indicates an excess of non-synonymous divergence, which is a sign of positive selection (Stoletzki & Eyre-Walker 2010).

We searched for possible selective sweep regions, i.e. the elimination of variation in regions linked to a recently fixed beneficial mutation, on the scaffolds containing immune genes using a Hidden Markov Model (HMM) program specifically designed to handle pooled-sequencing data, Pool-HMM

(Boitard *et al.* 2012, 2013). The Pool-HMM parameter of “--theta” was set to be the genome-wide average θ for the coding regions estimated for each population. The sensitivity parameter (-k) was run from the conservative k=0.001 to the lenient k=0.0000000001, to enable us to find potential selective sweep in well supported regions.

Results

Genome assessment and annotation

The *P. napi* genome assembly (7,829 scaffolds, N50 = 300,294) was reported to contain a conserved single-copy ortholog content of 78% using CEGMA (Meslin *et al.* 2017). Here this was further assessed by quantifying the content of complete BUSCO arthropod genes, which was 74% (SM: Table 5) and higher than most other Lepidopteran genomes (Talla *et al.* 2017). The high complete BUSCO score suggests the assembly has captured the majority of core arthropod genes in full length. A total of 11,240 gene models were predicted in the *P. napi* genome, with 136 immune genes identified (SM: Table 2) and classified into five functional groups: recognition (N = 52), modulation (N = 22), signalling (N = 37), effectors (N = 20), and others (N = 5). These results are consistent with other studies in Lepidoptera, with the exception that we found fewer modulator genes (Tanaka *et al.* 2008; Xia *et al.* 2015).

Population level patterns

Nucleotide diversity (π), averaged across all gene models for each of the three populations, ranged from 0.034 to 0.035 at synonymous sites (π_{SS}) and 0.002 to 0.003 at non-synonymous sites (π_{NS}) (SM: Table 6 and 7). Nucleotide diversity in

immune genes was significantly higher than the genome-wide average for all three populations (Figure 1), showing a 20% excess in π_{SS} and up to 50% in π_{NS} (Table 1). This was further supported by the permutation analysis, where the π_{NS} and π_{SS} averages of the immune genes were located in the outermost 97.5th percentile of the distribution of the 10,000 sampled means, indicating a significantly higher nucleotide diversity for the immune genes (Figure 3). In addition, immune gene categories differed significantly from each other, driven by significantly lower π_{NS} and π_{SS} in signalling genes and higher π_{NS} in effector genes (Table 3). Populations did not differ significantly from each other in their immune gene nucleotide diversity (Table 2). In summary, nucleotide diversity differed between the functional classes of immune genes but not between populations, with effector genes exhibiting the highest diversity.

The genome-wide average of Tajima's D, calculated per gene for each of the three populations, was moderately negative for both $TajD_{SS}$ and $TajD_{NS}$ in all populations (Figure 2; SM: Table 6). Tajima's D values for the immune genes (SM: Table 7) did not differ significantly from the genome-wide means (Table 1). This was further supported by the permutation analysis, where for all three populations the $TajD_{SS}$ and $TajD_{NS}$ averages of the immune genes were located outside the 2.5th percentile borders of the distribution of the sampled means, indicating that these did not significantly differ from the genome-wide averages (Figure 3). Comparing the populations did not reveal any significant differences (Table 2). While there were no significant $TajD_{SS}$ differences between the gene categories, $TajD_{NS}$ values did differ between the different immune gene categories, with signalling genes exhibiting significantly more negative values compared to the other categories (Table 4). In summary, $TajD_{NS}$ values did differ

between the immune gene categories, but not between populations, and overall immune genes showed no significant difference from the genome-wide TajD.

Within population analyses

Using the empirical null distribution of the values from the 11,240 gene models, we identified outliers for π and TajD. A total of 22 immune genes were π -outliers, of which seven were recognition genes, five modulation genes, three signalling genes, and seven effector genes (Table 5). Nine genes exhibited excess in either π_{NS} and/or π_{SS} in all three populations, whereas six genes were outliers in only one of the populations.

The majority of TajD outlier genes appear to be experiencing population specific selection pressures (Abisko N=12, Kullaberg N=12, Aiguamolls N=8; Table 6). In the positive tail of both the TajD_{NS} and TajD_{SS} distribution, lysozyme was identified in Aiguamolls. In Abisko three TajD_{SS} positive outlier genes were identified; two recognition genes (FREP3 and Galectin3) and a signalling gene (Jun). The Kullaberg population contained the most TajD_{SS} outlier genes (N=5), three signalling genes (Ubc13, SOCS, Jun), a recognition gene (SCRIB2), and an effector gene (LLP2). Jun, part of the JNK pathway, exhibits a shift in TajD_{SS} between Abisko (TajD_{SS}=1.32) and Kullaberg (TajD_{SS}=-2.01).

Sex linked selection

The 10 immune genes found on the sex chromosome showed lower nucleotide diversity, but this pattern was only significant for Abisko & Aiguamolls π_{SS} and Aiguamolls π_{NS} (SM: Figure 1). However, none of the previously discussed nucleotide diversity outlier genes were on the sex chromosome.

For all three populations, the $TajD_{NS}$ and $TajD_{SS}$ averages of the 10 sex-linked immune genes were located outside the 2.5th percentile borders of the distribution of the sampled means, indicating that these 10 sex-linked genes did not significantly differ from autosomal immune genes.

Pairwise population differentiation: F_{ST} analyses

Genome-wide a total of 1,110,127 SNPs were identified within exon boundaries of the 11,240 predicted genes. Based on the genome-wide average F_{ST} Abisko and Kullaberg differed the most (0.035), Abisko to Aiguamolls were similar (0.031), whereas F_{ST} was slightly lower between Kullaberg and Aiguamolls (0.0276; SM: Table 8). The average F_{ST} for immune genes was similar between the three comparisons (Abisko:Aiguamolls 0.050, 0.029 (mean, std. dev); Abisko:Kullaberg 0.055, 0.033; Kullaberg:Aiguamolls 0.043, 0.024). F_{ST} values were also calculated for each SNP to represent the overall differentiation pattern between the populations. Within the exons of immune genes a total of 23,551 SNPs were identified across all populations, and average F_{ST} per population comparison in the immune genes were almost all identical to those of the whole genome, with the exception of the Abisko:Kullaberg comparison, where the mean for immune genes was lower ($M = 0.033$, $SD = 0.045$; SM: Table 8). In addition, while F_{ST} distribution of the immune genes was similar to the whole genome, no SNP in any immune gene had an F_{ST} of 1 (i.e. none were differentially fixed between any populations; SM: Figure 3 & SM: Table 8).

SNPs in immune genes were selected as candidate SNPs when occurring above the 99.5th percentile of the F_{ST} distribution ($n=11,240$ predicted genes) in at least one of the comparisons (Abisko:Aiguamolls $F_{ST} > 0.252$, Abisko:Kullaberg

$F_{ST} > 0.250$, Kullaberg:Aiguamolls $F_{ST} > 0.242$). This resulted in 267 candidate SNPs across 58 genes. Significantly more candidate SNPs were found in recognition and effector genes, while fewer SNPs were found in modulation and signalling genes than expected ($\chi^2=106.769$, $df=4$, $p<0.0001$, SM: Table 9). When comparing the different signalling pathways, the genes of the JAK/STAT pathway contained significantly more SNPs than expected ($\chi^2=115.661$, $df=7$, $p<0.0001$, SM: Table 10), indicating that the pattern of underrepresentation of signalling genes is driven by genes encoding for Toll, IMD and JNK.

Species level selection dynamics

Using the closely related species *P. rapae* and *P. brassicae* as outgroups, and variation in the Abisko population of *P. napi* as the polymorphism data, we performed McDonald-Kreitman tests on 83 immune genes. We found 17 genes ($FDR \leq 0.05$) that significantly deviate from neutral expectations, all with an NI < 1, indicating that these genes are under selection (Table 7,8). Of the 17 genes, 6 are recognition genes, 4 are modulators, another 6 are signalling genes, and 1 is an effector gene. Out of the 6 signalling genes, 4 are part of the Toll pathway.

To identify recently fixed beneficial mutations, i.e. selective sweeps, (Nielsen *et al.* 2005), Pool-HMM was performed on scaffolds containing immune genes across all three populations. We found no evidence for selective sweeps on any of these scaffolds, despite assessing a wide range of parameters.

Discussion

Here we report the first population genomic analysis of the canonical immune genes in Lepidoptera, investigating evolutionary dynamics both within and

among populations, and between species. We found that all populations of *P. napi* exhibit consistent patterns of increased genetic variation in immune genes compared to the rest of the genome, which is driven primarily by recognition, modulation, and effector genes. Furthermore, we revealed population specific patterns of both directional and balancing selection dynamics, in genes of various categories and at different evolutionary scales, presenting a complex pattern of selection dynamics acting across functional categories. However, genes involved with the cellular arm of the immune response as well as the Toll pathway appear to be enriched among these outlier loci.

Nucleotide diversity across immune genes

We observed an increased level of nucleotide diversity, especially in immune genes involved with recognition, modulation & effectors of *P. napi* compared to the genome-wide average, consistent with previous findings in Diptera (Begun & Whitley 2000; Lazzaro & Clark 2001; Schlenke & Begun 2003; Sackton *et al.* 2007). Recognition and effector genes also showed significantly higher level of differentiation (F_{ST}) between the populations compared to the other immune gene categories, suggesting that these functional categories could more often be targets of local adaptation. Genes involved with the recognition of non-self as well as effector molecules, have been suggested to be targets of co-evolutionary arms races, wherein pathogens try to inhibit their function (Sackton *et al.* 2007).

Positive selection acting on immune genes

Despite exhibiting higher genetic variation within, and differentiation between, populations, recognition and effector genes were not more likely to be under

positive selection. Instead our results revealed evidence for positive selection across a range of functional categories, predominantly genes linked to the cellular response, as well as the Toll pathway. Similar observations have been made in *Drosophila*, where genes involved with the cellular response of phagocytosis, as well as the signalling cascades often experience positive selection and appear to evolve faster, perhaps due to host-pathogen interactions (Begun & Whitley, 2000; Schlenke & Begun, 2003; Jiggins & Kim, 2007; Sackton *et al.* 2007). The cellular response is often the first line of defence after a pathogen crosses the epithelial layer, wherein pathogens likely evolve to subdue the subsequent cellular immune response, which in turn leads to selection on variants in these genes that prevent this suppression (Graham, 2013). These results suggest natural selection fixes advantageous amino acid changes across genes of an entire pathway, more than on functional subsets across pathways.

Signatures of balancing selection

In addition to positive selection, genes involved in the immune system are considered likely targets of balancing selection (Leffler *et al.* 2012), and while such dynamics are well known for vertebrate MHC loci (Edwards & Hedrick 1998), these have only recently been documented in several AMPs of *Drosophila* species (Unckless & Lazzaro 2016), and it is currently unknown how common this is across Insecta. Two genes, the AMP Attacin-1 and the homolog to BmSerpín-6 (SPN6), showed an excess π_{SS} and π_{NS} in all three populations, a possible signature of balancing selection. Attacin AMPs are effective against Gram negative bacteria, by attacking the outer membrane of the bacterial cell (Carlsson *et al.* 1991). Interestingly, Attacin genes exhibit high levels of

synonymous nucleotide diversity in *D. melanogaster* (Lazzaro & Clark 2001), and show trans-species polymorphisms across the *Drosophila* phylogeny (Unckless & Lazzaro 2016). Whether additional members of Insecta show signatures of balancing selection in this gene family is an interesting prospect, and could suggest fluctuating selection across the phylogeny.

The second gene, SPN6, is a member of the Serine protease inhibitors (serpins) that mediate cellular immune responses to wounding and infection in insects (Jiang & Kanost 2000). In Lepidoptera, the *B. mori* ortholog BmSPN6 is strongly up-regulated after bacterial infection (Tanaka *et al.* 2008), and in *Manduca sexta* (Lepidoptera, Sphingidae), the ortholog to this gene, MsSerp6, is involved in the production of melanin via the PPO cascade (Zou & Jiang 2005). Melanin immobilizes, agglutinates, or suffocates microbes (Whitten & Coates, 2017). Our results are in agreement with previous work in *Anopheles* that found Serpins to show excess levels of amino acid diversity, consistent with either reduced constraint or balancing selection (White *et al.* 2007; Obbard *et al.* 2009). Although an excess of functional non-synonymous polymorphism and synonymous polymorphism at nearby neutral sites is consistent with long-term balancing selection, validations using trans-species polymorphisms (Charlesworth 2006) and functional assays are needed to identify if this increased nucleotide diversity is predictive for increased resistance to bacteria. (Storz & Wheat 2010).

Population specific dynamics

Besides species level selection dynamics, our analysis also revealed population differences in their patterns of variation, consistent with population specific

selection dynamics, perhaps as a result of variation in local pathogen pressures. Two immune genes, Gloverin3 (an AMP) and SOCS (part of JAK/STAT signalling), revealed patterns of either very strong constraint or the recent fixation of a novel mutation in Aiguamolls. In contrast, the majority of population specific outlier genes revealed an excess of intermediate frequency variation, consistent with the adaptive maintenance of a polymorphism i.e. balancing selection. In *Drosophila*, the most common immune genes reported to experience balancing selection are AMPs (Unckless *et al.* 2016; Unckless & Lazzaro 2016). Our positive Tajima's D outliers did not contain AMPs, although one Tajima's D outlier was an effector gene, lysozyme in Aiguamolls. Furthermore, in the Kullaberg population the effector LLP2, a Lysozyme-like protein, showed a negative TajD_{SS} but a significantly positive TajD_{NS}. A possible explanation for this observation could be the target of (balancing) selection is a NS mutation that has recently increased in frequency, without time for SS difference to be accumulated between the maintained variant alleles (SM: Figure 2). Although this pattern was not identified in *Drosophila*, lysozyme appears to be under balancing selection in *Apis mellifera* (Harpur and Zayed, 2013), highlighting the need for more data from more species to assess whether this gene family could frequently be a target of selection among insects.

Outlier analyses using nucleotide diversity revealed an unexpected pattern of high π_{NS} in the AMP Gallerimycin within the Kullaberg population. Closer investigation revealed excessive linkage among four closely linked NS variants, which appear at roughly 50% frequency in the sample, along with very low frequency SS variation within one of the haplotype groups. Although these patterns suggest strong balancing selection at this locus, there is a potential for

two recently diverged loci to have their alleles mapping to this one locus. However, this appears to be unlikely. First, this is not observed in the other populations and the read depth was consistent with a single locus. Second, we would expect two such loci that differ by four NS sites to also differ by several SS sites, given the NS divergence. This suggests that our observation is indeed an outlier locus of excess NS diversity, and a target of balancing selection. Gallerimycin is an antifungal defensin-like peptide, similar to *Drosophila* Drosomycin (Schuhmann *et al.* 2003). This antifungal AMP was found to be ineffective in *D. melanogaster* against the generalist insect pathogens *Metarhizium*, potentially as a result of strong selective pressure for resistance against this AMP (Lu *et al.* 2015). While it is difficult to interpret our results in relation to the as-of-yet unknown natural pathogens and their interaction virulence proteins, it is possible that the pattern of excess NS diversity is a consequence of selection acting between fungal pathogens and AMPs.

In addition to the effector molecules discussed above, three recognition genes linked to the cellular arm of the immune response, FREP and Galectin3 in Abisko, and SCR-B2 in Kullaberg (Wang *et al.* 2005; Hanington *et al.* 2010), showed signatures of balancing selection. However, the recognition genes identified in our study, to our knowledge, have not been previously found to be under balancing selection in other insects, although cellular recognition genes appear to be a frequent target of positive selection in *Drosophila*, suggested to be a result of molecular interactions between host and pathogens (Sackton *et al.* 2007). Overall, the results of the outlier analysis suggest that population specific selection dynamics exist, and short-term balancing selection might be a more common mode of selection in immune genes than previously thought.

594

595 *Sex-linked bias*

596 Kullaberg shows the most population specific outlier genes. However, this
597 library was constructed from only males, which is potential confounding factor.
598 Since sex chromosomes have an overall lower nucleotide diversity due to their
599 lower effective population size and limited recombination, there was a potential
600 for a sex-linked bias when identifying outliers of low nucleotide diversity or
601 skewed Tajima's D, where patterns of selection could be driven by chromosomal
602 properties, rather than potential selection on immunity. However, only ten of the
603 studied immune genes were located on the sex chromosome, and none of these
604 were identified as significant outliers for reduced nucleotide diversity. Thus, a
605 sex-linked bias on our observed patterns is unlikely and the outliers found in the
606 Kullaberg population, which consisted of only males, are driven by something
607 else than sex-linked bias.

608

609 *Study limitations*

610 Our study had several limitations due to the high heterozygosity within our
611 species. First, the assembly used was fragmented. However, it had a high BUSCO
612 gene content and were able to identify and annotate coding regions for a large
613 fraction of the genome and the canonical immune genes in particular. However,
614 our annotation relied upon the use of SPALN2 (Iwata & Gotoh 2012). Although
615 SPALN2 exhibits excellent performance in generating genomic annotations using
616 proteins from divergent species, our annotations have room for improvement
617 using RNA-Seq data from *Pieris* and traditional annotation pipelines such as
618 Maker (Holt & Yandell 2011). Nevertheless, we have confidence in the over

10,000 gene models generated and all our analysis were performed only within their coding regions to ensure high confidence, with the trade-off that non-coding genomic regions were not studied, nor was the potential role of immune genes being clustered assessed. Second, our Pool-seq data does not allow us to look at recombination and linkage disequilibrium aspects of the site frequency spectrum. Given that chromosomal rearrangements and recombination rates butterflies are more than an order of magnitude higher compared to *Drosophila* (Bures *et al.* 2012; Martin *et al.* 2016), the information of such linkage dynamics are likely to be limited especially in the absence of hard sweeps, as is the cases in our dataset. Third, the stringent filtering that enabled us to be confident in the single copy status of gene models has the potential consequence of preventing our study of copy number variation (CNV). Future work using improved assemblies, individual level genomic data as well as RNA-seq data should enable us to properly address CNV dynamics and their selection dynamics in the immune genes.

Furthermore, all three populations showed an overall negative genome-wide Tajima's D distribution, suggestive of population bottleneck events in the past, some of which could have been shared. While we have tried to minimize the confounding effect of demography in our tests of selection (e.g. outlier analysis and MK tests), such bottleneck effects could interact with the hitchhiking of deleterious alleles to inflate our identified candidates for selection (e.g. Fay 2011). Thus, as in all molecular tests of selection, demographic effects could account for some fraction of these patterns. In such light, the overlap of our candidate genes with previous findings could be indicative of wide-spread

deleterious mutational effects on, or the functional importance of, these classes of genes.

A final consideration is the general lack of good annotations for these phenotypic level responses, which is common in non-model organisms. This impedes attempts at understanding the selection dynamics at work. Here, the gene function of proteins was determined based on homology inference and *P. napi* lacks experimental studies confirming their direct role in its immune system, a limitation that is common for many genomic studies. Nevertheless, by using several tests to infer evolutionary forces acting on immune genes, our population genomic approach provides a new level of understanding for the scale at which selection dynamics interact with the insect immune system.

Conclusions

In summary, immune genes exhibited a general pattern of increased genetic diversity compared to the rest of the genome, primarily driven by recognition, modulation and effector genes. While evidence of directional and balancing selection was found across functional categories, findings could be grouped as being involved primarily with the cellular response and Toll signalling. These findings support the following conclusions. First, although the pathogens Lepidoptera encounter compared to Diptera likely differ due to their respective diets and niches, the overall findings for *P. napi* are very consistent with the previous literature. Second, this suggests that either the immune system is robust to the level of microbial differences between these Orders, or the generalist microbes encountered are simply similar from the immune system perspective. Third, the microevolutionary dynamics we observe here are

consistent with the previously observed macroevolutionary patterns, suggesting that variable microevolutionary dynamics give rise to a stable macroevolutionary pattern across Insecta.

Acknowledgements

The authors would like to thank Constantí Stephanescu for sample collection in Spain, Maria de la Paz Celorio Mancera for her help during lab work, Ramprasad Neethiraj for general bioinformatic help, Viola Nolte for providing her scripts for the MK analysis, and Peter Pruisscher for help with the genome assembly of *P. brassicae*. PP and Uli Theopold are also thanked for their helpful comments during writing. We thank three anonymous reviewers for helpful comments on the earlier draft of the manuscript. The work was funded by the Knut and Alice Wallenberg Foundation (KAW2012.0058) and the Swedish Research Council (2012-3715).

References

- Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., Rastas, P., et al. (2014). The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nature Communications*, **5**, 4737–9. <http://doi.org/10.1038/ncomms5737>
- Aljanabi, S. M., Martinez, I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research*, **25**, 4692–4693. doi: 10.1093/nar/25.22.4692
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410. [http://doi.org/10.1016/S0022-2836\(05\)80360-2](http://doi.org/10.1016/S0022-2836(05)80360-2)
- Bao, Y.-Y., Qu, L.-Y., Zhao, D., Chen, L.-B., Jin, H.-Y., Xu, L.-M., et al. (2013). The genome- and transcriptome-wide analysis of innate immunity in the brown planthopper, *Nilaparvata lugens*. *BMC Genomics*, **14**(160), 1–22. <http://doi.org/10.1186/1471-2164-14-160>
- Barribeau, S. M., Sadd, B. M., Plessis, du, L., Brown, M. J. F., Buechel, S. D., Cappelle, K., et al. (2015). A depauperate immune repertoire precedes evolution of

- sociality in bees. *Genome Biology*, 16(1). <http://doi.org/10.1186/s13059-015-0628-y>
- Bartholomay, L. C., Waterhouse, R. M., Mayhew, G. F., Campbell, C. L., Michel, K., Zou, Z., et al. (2010). Pathogenomics of *Culex quinquefasciatus* and meta-analysis of infection responses to diverse pathogens. *Science*, 330(6000), 88–90. <http://doi.org/10.1126/science.1193162>
- Begun, D. J., & Whitley, P. (2000). Adaptive evolution of relish, a *Drosophila* NF-kappaB/IkappaB protein. *Genetics*, 154(3), 1231–1238.
- Belvin, M. P., & Anderson, K. V. (1996). A conserved signalling pathway: The *Drosophila* Toll-Dorsal pathway. *Annual Review of Cell and Developmental Biology*, 12(1), 393–416. <http://doi.org/10.1146/annurev.cellbio.12.1.393>
- Boitard, S., Schlötterer, C., Nolte, V., Pandey, R. V., & Futschik, A. (2012). Detecting selective sweeps from pooled next-generation sequencing samples. *Molecular biology and evolution*, 29(9), 2177–2186. <https://doi.org/10.1093/molbev/mss090>
- Boitard, S., Kofler, R., Françoise, P., Robelin, D., Schlötterer, C., & Futschik, A. (2013). Pool - hmm: a Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. *Molecular ecology resources*, 13(2), 337–340. doi:10.1111/1755-0998.12063
- Broderick, N. A. (2016). Friend, foe or food? Recognition and the role of antimicrobial peptides in gut immunity and *Drosophila*–microbe interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1695), 20150295. <http://doi.org/10.1098/rstb.2015.0295>
- Bulmer, M. S., Lay, F., & Hamilton, C. (2010). Adaptive evolution in subterranean termite antifungal peptides. *Insect Molecular Biology*, 19(5), 669–674. <http://doi.org/10.1111/j.1365-2583.2010.01023.x>
- Bureš, P., Zedek, F., & Marková, M. (2012). Holocentric Chromosomes. In *Plant Genome Diversity Volume 2* (pp. 187–208). Vienna: Springer Vienna. http://doi.org/10.1007/978-3-7091-1160-4_12
- Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* 22(11):3124–3140. <http://doi:10.1111/mec.12354>
- Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., & Nickerson, D. A. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, 15(11), 1553–1565. <http://doi.org/10.1101/gr.4326505>
- Carlsson, A., Engström, P., Palva, E. T., & Bennich, H. (1991). Attacin, an antibacterial protein from *Hyalophora cecropia*, inhibits synthesis of outer membrane proteins in *Escherichia coli* by interfering with omp gene transcription. *Infection and Immunity*, 59(9), 3040–3045.
- Charlesworth, B., Coyne, J. A., & Barton, N. H. (1987). The Relative Rates of Evolution of Sex Chromosomes and Autosomes. *The American Naturalist*, 130(1), 113–146. <http://doi.org/10.1086/284701>
- Charlesworth, D. (2006). Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genetics*, 2(4), e64. <http://doi.org/10.1371/journal.pgen.0020064>

- Christophides, G. K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., et al. (2002). Immunity-related genes and gene families in *Anopheles gambiae*. *Science*, **298**(5591), 159–165. <http://doi.org/10.1126/science.1077136>
- Cong, Q., Shen, J., Warren, A. D., Borek, D., Otwinowski, Z., & Grishin, N. V. (2016). Speciation in Cloudless Sulphurs Gleaned from Complete Genomes. *Genome Biology and Evolution*, **8**(3), 915–931. <http://doi.org/10.1093/gbe/evw045>
- Costello, M. J., Wilson, S., & Houlding, B. (2012). Predicting Total Global Species Richness Using Rates of Species Description and Estimates of Taxonomic Effort. *Systematic Biology*, **61**(5), 871–883. <http://doi.org/10.1093/sysbio/syr080>
- Crawford, J. E., Bischoff, E., Garnier, T., Gneme, A., Eiglmeier, K., Holm, I., et al. (2012). Evidence for Population-Specific Positive Selection on Immune Genes of *Anopheles gambiae*. *G3 (Bethesda, Md.)*, **2**(12), 1505–1519. <http://doi.org/10.1534/g3.112.004473>
- d'Alençon, E., Sezutsu, H., Legeai, F., Permal, E., Bernard-Samain, S., Gimenez, S., ... & Flutre, T. (2010). Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements. *Proceedings of the National Academy of Sciences*, **107**(17), 7680–7685.
- Douglas, A. E. (2009). The Microbial Dimension in Insect Nutritional Ecology. *Functional Ecology*, **23**(1), 38–47. <http://doi.org/10.2307/40205500?ref=search-gateway:05c35b588781cb16496d37ba776f67e6>
- Early, A. M., Arguello, J. R., Cardoso-Moreira, M., Gottipati, S., Grenier, J. K., & Clark, A. G. (2017). Survey of Global Genetic Diversity Within the *Drosophila* Immune System. *Genetics*, **205**(1), 353–366. <http://doi.org/10.1534/genetics.116.195016>
- Edwards, S. V., & Hedrick, P. W. (1998). Evolution and ecology of MHC molecules: from genomics to sexual selection. *Trends in Ecology & Evolution*, **13**(8), 305–311. [http://doi.org/10.1016/S0169-5347\(98\)01416-5](http://doi.org/10.1016/S0169-5347(98)01416-5)
- Engel, P., & Moran, N. A. (2013). The gut microbiota of insects – diversity in structure and function. *FEMS Microbiology Reviews*, **37**(5), 699–735. <http://doi.org/10.1111/1574-6976.12025>
- Erler, S., Lhomme, P., Rasmont, P., & Lattorff, H. M. G. (2014). Rapid evolution of antimicrobial peptide genes in an insect host–social parasite system. *Infection, Genetics and Evolution*, **23**, 129–137. <http://doi.org/10.1016/j.meegid.2014.02.002>
- Evans, J. D., Aronstein, K., Chen, Y. P., Hetru, C., Imler, J. L., Jiang, H., et al. (2006). Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Molecular Biology*, **15**(5), 645–656. <http://doi.org/10.1111/j.1365-2583.2006.00682.x>
- Fabian, D. K., Kapun, M., Nolte, V., Kofler, R., Schmidt, P. S., Schlötterer, C., & Flatt, T. (2012). Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Molecular Ecology*, **21**(19), 4748–4769. <http://doi.org/10.1111/j.1365-294X.2012.05731.x>
- Fay, J. C. (2011). Weighing the evidence for adaptation at the molecular level. *Trends in Genetics*, **27**(9), 1–7. <http://doi.org/10.1016/j.tig.2011.06.003>

- Gerardo, N. M., Altincicek, B., Anselme, C., Atamian, H., Barribeau, S. M., de Vos, M., et al. (2010). Immunity and other defenses in pea aphids, *Acyrtosiphon pisum*. *Genome Biology*, **11**(2), R21. <http://doi.org/10.1186/gb-2010-11-2-r21>
- Gotoh, O. (2008). Direct mapping and alignment of protein sequences onto genomic sequence. *Bioinformatics*, **24**(21), 2438–2444. <http://doi.org/10.1093/bioinformatics/btn460>
- Graham, A. L. (2013). Editorial Optimal immunity meets natural variation: the evolutionary biology of host defence. *Parasite Immunology*, **35**, 315–317. <http://doi.org/10.1111/pim.12073>
- Hanington, P. C., Forsy, M. A., Dragoo, J. W., Zhang, S.-M., Adema, C. M., & Loker, E. S. (2010). Role for a somatically diversified lectin in resistance of an invertebrate to parasite infection. *Proceedings of the National Academy of Sciences*, **107**(49), 21087–21092. <http://doi.org/10.1073/pnas.1011242107>
- Hansen, A. K., & Moran, N. A. (2014). The impact of microbial symbionts on host plant utilization by herbivorous insects. *Molecular Ecology*, **23**(6), 1473–1496. <http://doi.org/10.1111/mec.12421>
- Harpur, B. A., & Zayed, A. (2013). Accelerated Evolution of Innate Immunity Proteins in Social Insects: Adaptive Evolution or Relaxed Constraint? *Molecular Biology and Evolution*, **30**(7), 1665–1674. <http://doi.org/10.1093/molbev/mst061>
- Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, **10**, 639–650. <http://doi.org/10.1038/nrg2611>
- Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *Bmc Bioinformatics*, **12**(1), 491. <http://doi.org/10.1186/1471-2105-12-491>
- Hooper, L. V., & Macpherson, A. J. (2010). Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nature Reviews Immunology*, **10**(3), 159–169. <http://doi.org/10.1038/nri2710>
- Horrocks, N. P. C., Matson, K. D., & Tieleman, B. I. (2011). Pathogen Pressure Puts Immune Defense into Perspective. *American Zoologist*, **51**(4), 563–576. <http://doi.org/10.1093/icb/icr011>
- Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, **11**(2), 97–108. <http://doi.org/10.1038/nrg2689>
- Irving, P., Troxler, L., Heuer, T. S., Belvin, M., Kopczynski, C., Reichhart, J. M., et al. (2001). A genome-wide analysis of immune responses in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(26), 15119–15124. <http://doi.org/10.1073/pnas.261573998>
- Iwata, H., & Gotoh, O. (2012). Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Research*, **40**(20), e161–e161. <http://doi.org/10.1093/nar/gks708>
- Jiang, H. B., & Kanost, M. R. (2000). The clip-domain family of serine proteinases in arthropods. *Insect Biochemistry and Molecular Biology*, **30**(2), 95–105. [http://doi.org/10.1016/S0965-1748\(99\)00113-7](http://doi.org/10.1016/S0965-1748(99)00113-7)

- Jiggins, F. M., & Kim, K. W. (2007). A screen for immunity genes evolving under positive selection in *Drosophila*. *Journal of Evolutionary Biology*, **20**(3), 965–970. <http://doi.org/10.1111/j.1420-9101.2007.01305.x>
- Keehnen, N. L. P., Rolff, J., Theopold, U., & Wheat, C. W. (2017). Insect Antimicrobial Defences: A Brief History, Recent Findings, Biases, and a Way Forward in Evolutionary Studies. In *Advances in Insect Physiology* (Vol. 52, pp. 1–33). <http://doi.org/10.1016/bs.aiip.2017.02.003>
- Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., et al. (2011a). PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS ONE*, **6**(1). <http://doi.org/10.1371/journal.pone.0015925>
- Kofler, R., Pandey, R. V., & Schlötterer, C. (2011b). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, **27**(24), 3435–3436. <http://doi.org/10.1093/bioinformatics/btr589>
- Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings. Biological Sciences / the Royal Society*, **279**(1749), 5048–5057. <http://doi.org/10.1098/rspb.2012.1108>
- Lazzaro, B. P. (2008). Natural selection on the *Drosophila* antimicrobial immune system. *Current Opinion in Microbiology*, **11**(3), 284–289. <http://doi.org/10.1016/j.mib.2008.05.001>
- Lazzaro, B. P., & Clark, A. G. (2001). Evidence for Recurrent Paralogous Gene Conversion and Exceptional Allelic Divergence in the Attacin Genes of *Drosophila melanogaster*. *Genetics*, **159**(2), 659–671.
- Lazzaro, B. P., & Little, T. J. (2009). Immunity in a variable world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364**(1513), 15–26. <http://doi.org/10.1098/rstb.2008.0141>
- Lazzaro, B. P., Scurman, B. K., & Clark, A. G. (2004). Genetic basis of natural variation in *D. melanogaster* antibacterial immunity. *Science*, **303**(5665), 1873–1876. <http://doi.org/10.1126/science.1092447>
- Lee, W.-J., & Brey, P. T. (2013). How microbiomes influence metazoan development: insights from history and *Drosophila* modeling of gut-microbe interactions. *Annual Review of Cell and Developmental Biology*, **29**(1), 571–592. <http://doi.org/10.1146/annurev-cellbio-101512-122333>
- Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., et al. (2012). Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species? *PLoS Biology*, **10**(9), e1001388–9. <http://doi.org/10.1371/journal.pbio.1001388>
- Lemaitre, B., & Hoffmann, J. (2007). The Host Defense of *Drosophila melanogaster*. *Annu Rev Immunol*, **25**(1), 697–743. <http://doi.org/10.1146/annurev.immunol.25.022106.141615>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079. <http://doi.org/10.1093/bioinformatics/btp352>
- Little, T. J., & Cobbe, N. (2005). The evolution of immune-related genes from disease carrying mosquitoes: diversity in a peptidoglycan- and a thioester-recognizing protein. *Insect Molecular Biology*, **14**(6), 599–605. <http://doi.org/10.1111/j.1365-2583.2005.00588.x>

- Lu, H.-L., Wang, J. B., Brown, M. A., Euerle, C., & Leger, R. J. S. (2015). Identification of *Drosophila* Mutants Affecting Defense to an Entomopathogenic Fungus. *Nature Publishing Group*, 1–12. <http://doi.org/10.1038/srep12350>
- Martin, S. H., Möst, M., Palmer, W. J., Salazar, C., McMillan, W. O., Jiggins, F. M., & Jiggins, C. D. (2016). Natural Selection and Genetic Diversity in the Butterfly *Heliconius melpomene*. *Genetics*, 203(1), 525–541. <http://doi.org/10.1534/genetics.115.183285>
- McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, 351(6328), 652–654. <http://doi.org/10.1038/351652a0>
- Meslin, C., Cherwin, T. S., Plakke, M. S., Hill, J., Small, B. S., Goetz, B. J., et al. (2017). Structural complexity and molecular heterogeneity of a butterfly ejaculate reflect a complex history of selection. *Proceedings of the National Academy of Sciences of the United States of America*, 114(27), E5406–E5413. <http://doi.org/10.1073/pnas.1707680114>
- Mesquita, R. D., Vionette-Amaral, R. J., Lowenberger, C., Rivera-Pomar, R., Monteiro, F. A., Minx, P., et al. (2015). Genome of *Rhodnius prolixus*, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. *Proceedings of the National Academy of Sciences of the United States of America*, 112(48), 14936–14941. <http://doi.org/10.1073/pnas.1506226112>
- Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., et al. (2015). Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. *Science*, 347(6217), 1258522.1–1258522.8. <http://doi.org/10.1126/science.1258522>
- Neethiraj, R., Hornett, E. A., Hill, J., & Wheat, C. W. (2017). Investigating the genomic basis of discrete phenotypes using a Pool-Seq only approach: new insights into the genetics underlying color variation in diverse taxa. *Molecular Ecology*, 26(19), 4990–5002. <http://doi.org/10.1111/mec.14205>
- van Nieuwerkerken EJ, Kaila L, Kitching IJ *et al.* (2011) Order Lepidoptera Linnaeus, 1758. In Z-Q. Zhang (Ed.), *Animal Biodiversity: An outline of higher-level classification and survey of taxonomic richness* (Vol. 3148, pp. 212–221). Auckland, New Zealand: Magnolia Press. (Zootaxa, Vol. 3148).
- Nolte, V., Pandey, R. V., Kofler, R., & Schlotterer, C. (2013). Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana*. *Genome Research*, 23(1), 99–110. <http://doi.org/10.1101/gr.139873.112>
- Obbard, D. J., Welch, J. J., & Little, T. J. (2009). Inferring selection in the *Anopheles gambiae* species complex: an example from immune-related serine protease inhibitors. *Malaria Journal*, 8(1), 117–15. <http://doi.org/10.1186/1475-2875-8-117>
- Palmer, W. J., & Jiggins, F. M. (2015). Comparative Genomics Reveals the Origins and Diversity of Arthropod Immune Systems. *Molecular Biology and Evolution*, 32(8), 2111–2129. <http://doi.org/10.1093/molbev/msv093>
- Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9), 1061–1067. <http://doi.org/10.1093/bioinformatics/btm071>

- Porter, A. H., & Geiger, H. (1995). Limitations to the inference of gene flow at regional geographic scales—an example from the *Pieris napi* group (Lepidoptera: Pieridae) in Europe. *Biological Journal of the Linnean Society*, **54**, 329–348. <http://doi.org/10.1111/j.1095-8312.1995.tb01041.x>
- Pruisscher, P., Larsdotter-Mellström, H., Stefanescu, C., Nylin, S., Wheat, C. W., & Gotthard, K. (2017). Sex-linked inheritance of diapause induction in the butterfly *Pieris napi*. *Physiological Entomology*, **42**(3), 257–265. <http://doi.org/10.1111/phen.12194>
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842. <http://doi.org/10.1093/bioinformatics/btq033>
- R Core Team (2016) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLoS ONE*, **6**(9), e22594. <http://doi.org/10.1371/journal.pone.0022594>
- Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, **29**(1), 24–26. <http://doi.org/10.1038/nbt.1754>
- Rottschaefer, S. M., Crawford, J. E., & Riehle, M. M. (2015). Population Genetics of *Anopheles coluzzii* Immune Pathways and Genes. *G3: Genes/ Genomes/ Genetics*, **5**, 329–339. <http://doi.org/10.1534/g3.114.014845/-/DC1>
- Sackton, T. B., Lazzaro, B. P., Schlenke, T. A., Evans, J. D., Hultmark, D., & Clark, A. G. (2007). Dynamic evolution of the innate immune system in *Drosophila*. *Nature Genetics*, **39**(12), 1461–1468. <http://doi.org/10.1038/ng.2007.60>
- Schlenke, T. A., & Begun, D. J. (2003). Natural selection drives *Drosophila* immune system evolution. *Genetics*, **164**(4), 1471–1480. <http://doi.org/10.1093/oxfordjournals.molbev.a003804>
- Schuhmann, B., Seitz, V., Vilcinskas, A., & Podsiadlowski, L. (2003). Cloning and expression of gallerimycin, an antifungal peptide expressed in immune response of greater wax moth larvae, *Galleria mellonella*. *Archives of Insect Biochemistry and Physiology*, **53**(3), 125–133. <http://doi.org/10.1002/arch.10091>
- Scott, J. G., Warren, W. C., Beukeboom, L. W., Bopp, D., Clark, A. G., Giers, S. D., et al. (2014). Genome of the house fly, *Musca domestica* L., a global vector of diseases with adaptations to a septic environment. *Genome Biology*, **15**(10), 466. <http://doi.org/10.1186/s13059-014-0466-3>
- Sedlazeck, F. J., Rescheneder, P., & Haeseler, von, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, **29**(21), 2790–2791. <http://doi.org/10.1093/bioinformatics/btt468>
- Shimomura, M., Minami, H., Suetsugu, Y., Ohyanagi, H., Satoh, C., Antonio, B., et al. (2009). KAIKObase: an integrated silkworm genome database and data mining tool. *BMC Genomics*, **10**(1), 486. <http://doi.org/10.1186/1471-2164-10-486>
- Sleiman, M. S. B., Osman, D., Massouras, A., Hoffmann, A. A., Lemaitre, B., & Deplancke, B. (2015). Genetic, molecular and physiological basis of variation in *Drosophila* gut immunocompetence. *Nature Communications*, **6**, 7829. <http://doi.org/10.1038/ncomms8829>

989 Stoletzki, N., & Eyre-Walker, A. (2010). Estimation of the Neutrality Index.
990 *Molecular Biology and Evolution*, **28**(1), 63–70.
991 <http://doi.org/10.1093/molbev/msq249>

992 Storz, J. F., & Wheat, C. W. (2010). Integrating evolutionary and functional
993 approaches to infer adaptation at specific loci. *Evolution; International Journal*
994 *of Organic Evolution*, **64**(9), 2489–2509. [http://doi.org/10.1111/j.1558-
995 *5646.2010.01044.x*](http://doi.org/10.1111/j.1558-5646.2010.01044.x)

996 Suen, G., Teiling, C., Li, L., Holt, C., Abouheif, E., Bornberg-Bauer, E., et al. (2011).
997 The Genome Sequence of the Leaf-Cutter Ant *Atta cephalotes* Reveals Insights
998 into Its Obligate Symbiotic Lifestyle. *PLoS Genetics*, **7**(2).
999 <http://doi.org/10.1371/journal.pgen.1002007>

1000 Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef:
1001 comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*,
1002 **23**(10), 1282–1288. <http://doi.org/10.1093/bioinformatics/btm098>

1003 Talla, V., Suh, A., Kalsoom, F., Dincă, V., Vila, R., Friberg, M., et al. (2017). Rapid
1004 Increase in Genome Size as a Consequence of Transposable Element
1005 Hyperactivity in Wood-White (Leptidea) Butterflies. *Genome Biology and*
1006 *Evolution*, **9**(10), 2491–2505. <http://doi.org/10.1093/gbe/evx163>

1007 Tanaka, H., Ishibashi, J., Fujita, K., Nakajima, Y., Sagisaka, A., Tomimoto, K., et al.
1008 (2008). A genome-wide analysis of genes and gene families involved in innate
1009 immunity of *Bombyx mori*. *Insect Biochemistry and Molecular Biology*, **38**(12),
1010 1087–1110. <http://doi.org/10.1016/j.ibmb.2008.09.001>

1011 Tian, C., Gao, B., Fang, Q., Ye, G., & Zhu, S. (2010). Antimicrobial peptide-like genes
1012 in *Nasonia vitripennis*: a genomic perspective. *BMC Genomics*, **11**(1), 187.1–
1013 187.19. <http://doi.org/10.1186/1471-2164-11-187>

1014 Unckless, R. L., & Lazzaro, B. P. (2016). The potential for adaptive maintenance of
1015 diversity in insect antimicrobial peptides. *Philosophical Transactions of the*
1016 *Royal Society B: Biological Sciences*, **371**(1695), 20150291.
1017 <http://doi.org/10.1098/rstb.2015.0291>

1018 Unckless, R. L., Howick, V. M., & Lazzaro, B. P. (2016). Convergent Balancing
1019 Selection on an Antimicrobial Peptide in *Drosophila*. *Current Biology*, **26**(2),
1020 257–262. <http://doi.org/10.1016/j.cub.2015.11.063>

1021 Unckless, R. L., Rottschaefer, S. M., & Lazzaro, B. P. (2015). The Complex
1022 Contributions of Genetics and Nutrition to Immunity in *Drosophila*
1023 *melanogaster*. *PLoS Genetics*, **11**(3), e1005030.
1024 <http://doi.org/10.1371/journal.pgen.1005030>

1025 Viljakainen, L., Evans, J. D., Hasselmann, M., Rueppell, O., Tingek, S., & Pamilo, P.
1026 (2009). Rapid evolution of immune proteins in social insects. *Molecular*
1027 *Biology and Evolution*, **26**(8), 1791–1801.
1028 <http://doi.org/10.1093/molbev/msp086>

1029 Wang, X., Zhao, Q., & Christensen, B. M. (2005). Identification and
1030 characterization of the fibrinogen-like domain of fibrinogen-related proteins
1031 in the mosquito, *Anopheles gambiae*, and the fruitfly, *Drosophila melanogaster*,
1032 genomes. *BMC Genomics*, **6**(1), 114. [http://doi.org/10.1186/1471-2164-6-
1033 *114*](http://doi.org/10.1186/1471-2164-6-114)

1034 Waterhouse, R. M., Kriventseva, E. V., Meister, S., Xi, Z., Alvarez, K. S.,
1035 Bartholomay, L. C., et al. (2007). Evolutionary dynamics of immune-related
1036 genes and pathways in disease-vector mosquitoes. *Science (New York, N.Y.)*,
1037 **316**(5832), 1738–1743. <http://doi.org/10.1126/science.1139862>

- White, B. J., Hahn, M. W., Pombi, M., Cassone, B. J., Lobo, N. F., Simard, F., & Besansky, N. J. (2007). Localization of Candidate Regions Maintaining a Common Polymorphic Inversion (2La) in *Anopheles gambiae*. *PLoS Genetics*, **3**(12), e217–11. <http://doi.org/10.1371/journal.pgen.0030217>
- Whitten, M. M. A., & Coates, C. J. (2017). Re-evaluation of insect melanogenesis research: Views from the dark side. *Pigment Cell & Melanoma Research*, **30**(4), 386–401. <http://doi.org/10.1111/pcmr.12590>
- Xia, X., Yu, L., Xue, M., Yu, X., Vasseur, L., Gurr, G. M., et al. (n.d.). Genome-wide characterization and expression profiling of immune genes in the diamondback moth, *Plutella xylostella* (L.). *Scientific Reports*, **5**, 9877 EP –. <http://doi.org/10.1038/srep09877>
- Zhan, S., & Reppert, S. M. (2013). MonarchBase: the monarch butterfly genome database. *Nucleic Acids Research*, **41**(Database issue), D758–63. <http://doi.org/10.1093/nar/gks1057>
- Zou, Z., & Jiang, H. (2005). *Manduca sexta* serpin-6 regulates immune serine proteinases PAP-3 and HP8. cDNA cloning, protein expression, inhibition kinetics, and function elucidation. *The Journal of Biological Chemistry*, **280**(14), 14341–14348. <http://doi.org/10.1074/jbc.M500570200>
- Zou, Z., Evans, J. D., Lu, Z., Zhao, P., Williams, M., Sumathipala, N., et al. (2007). Comparative genomic analysis of the *Tribolium* immune system. *Genome Biology*, **8**(8), R177–R177. <http://doi.org/10.1186/gb-2007-8-8-r177>

Data Accessibility

The three raw Pool-seq libraries (*P. napi* Kullaberg, *P. napi* Abisko & *P. napi* Aiguamolls), *P. brassicae* draft genomes, MK gene alignments for significant genes, Tajima's D, pi & F_{ST} result files, and scripts will be made available at Dryad.

Author Contributions

N.L.P.K. and C.W.W. conceived and designed the study, with J.H. and S.N. providing additional input. N.L.P.K. performed the lab work. N.L.P.K. analysed the data. J.H. constructed the *P. napi* & *P. rapae* genomes. N.L.P.K., C.W.W. wrote the manuscript. All the authors read the paper.

Figures

Figure 1 Scaled density distribution plots of both genome-wide (grey) and immune genes specific nucleotide diversity (red). The lines denote the 2.5% most extreme values. Nucleotide diversity was analyzed separately for non-synonymous (NS) and synonymous (SS) sites. Nucleotide diversity, both NS and SS, was significantly higher in immune genes compared to the genome-wide average.

Figure 2 - Comparison of Tajima's D values genome-wide and in immune genes for each population. Tajima's D was analyzed separately for non-synonymous (NS) and synonymous (SS) sites. There was no significant difference between the immune genes and the rest of the genome.

Figure 2 – Distribution plots of the permutations for genome-wide nucleotide diversity and Tajima's D values. Abisko plots are a-d, Kullaberg e-h and Aiguamolls (i-l). The red line denotes the averages of these values for the immune genes.

Tables

Table 1 – Wilcoxon pairwise comparisons between genome-wide and immune gene distributions of nucleotide diversity and Tajima’s D for synonymous and non-synonymous sites for each population.

	Population	W	Z	P-value	
π_{NS}	Abisko	903049.5	3.72	0.0002	***
	Kullaberg	881363	3.61	0.0003	***
	Aiguamolls	920073.5	4.24	0.0001	***
π_{SS}	Abisko	885647.5	3.24	0.0012	***
	Kullaberg	844967	3.24	0.0089	**
	Aiguamolls	898702	3.66	0.0002	***
TajD_{NS}	Abisko	818811.5	1.47	0.1416	
	Kullaberg	750728.5	0.07	0.9409	
	Aiguamolls	786041.5	0.66	0.5107	
TajD_{SS}	Abisko	816500	1.40	0.1601	
	Kullaberg	812634	1.74	0.0813	
	Aiguamolls	789158.5	0.74	0.4598	

Table 2 – Wilcoxon pairwise comparison of nucleotide diversity and Tajima’s D between populations for both non-synonymous (_{NS}) and synonymous (_{SS}) sites.

	Comparison	Z	P-value
π_{NS}	Aiguamolls - Kullaberg	0.40	0.6874
	Aiguamolls - Abisko	-0.20	0.8447
	Kullaberg - Abisko	-0.61	0.5410
π_{SS}	Aiguamolls - Kullaberg	0.97	0.3303
	Aiguamolls - Abisko	-0.26	0.7962
	Kullaberg - Abisko	-1.21	0.2273
TajD_{NS}	Aiguamolls - Kullaberg	0.41	0.6794
	Aiguamolls - Abisko	-1.12	0.2635
	Kullaberg - Abisko	-1.49	0.1360
TajD_{SS}	Aiguamolls - Kullaberg	0.70	0.4845
	Aiguamolls - Abisko	-1.85	0.0644
	Kullaberg - Abisko	-0.99	0.3207

Table 3 – Wilcoxon pairwise comparison of nucleotide diversity between immune gene categories. Genes belonging to the functional category of effectors were significantly different to most other categories, with the exception of modulators. The functional gene category “others” was excluded from the analysis due to small (N=5) sample size.

		π_{SS}			π_{NS}		
Comparison		Z	P-value		Z	P-value	
Effectors	Signalling	2.27833	0.0227	*	4.48293	<0.0001	***
Effectors	Recognition	1.18726	0.2351		2.21692	0.0266	*
Modulation	Recognition	0.19270	0.8472		2.22884	0.0258	*
Effectors	Modulation	-0.76902	0.4419		0.40308	0.6869	
Signalling	Recognition	-2.82682	0.0047	**	-3.97397	<0.0001	***
Signalling	Modulation	-3.65994	0.0003	***	-4.96012	<0.0001	***

Table 4 – Wilcoxon pairwise comparison of Tajima’s D between immune gene categories. Genes belonging to the functional category of signalling were significantly different (more negative) to all other categories. The functional gene category “others” was excluded from the analysis due to small (N=5) sample size.

		TajD _{SS}			TajD _{NS}	
Comparison		Z	P-value		Z	P-value
Effectors	Signalling	2.59545	0.0094	*	0.98531	0.3245
Effectors	Recognition	1.30500	0.1919		-0.22728	0.8202
Modulation	Recognition	0.54028	0.5890		-0.64478	0.5191
Effectors	Modulation	0.59598	0.5512		0.14589	0.8840
Signalling	Recognition	-2.24116	0.0250	*	-1.91202	0.0559
Signalling	Modulation	-2.13345	0.0329	*	-1.10269	0.2702

Table 5 – Immune genes that had synonymous nucleotide diversity (π_{SS}) or non-synonymous nucleotide diversity (π_{NS}) that corresponded to the negative or positive 2.5th percentile of the genome-wide nucleotide diversity distribution. The outlier value is indicated by *, wherein the positive 2.5% outliers are underscored, and the negative 2.5% are in italics. The corresponding values of the other populations are shown in grey.

Genename	Functional category	Abisko		Kullaberg		Aiguamolls	
		π_{NS}	π_{SS}	π_{NS}	π_{SS}	π_{NS}	π_{SS}
BGRP1	Recognition	0.003	<u>0.08*</u>	0.003	0.062	0.002	0.066
BGRP2	Recognition	0.003	0.052	<u>0.013*</u>	0.036	0.004	0.036
CTL12	Recognition	<i>0*</i>	<u>0.128*</u>	<i>0*</i>	<u>0.1*</u>	<i>0*</i>	<u>0.095*</u>
CTL13	Recognition	<u>0.014*</u>	0.014	0.009	0.007	<u>0.012*</u>	0.013
CTL15	Recognition	0.002	<u>0.089*</u>	0.004	<u>0.082*</u>	0.002	<u>0.093*</u>
SCRB2	Recognition	<u>0.012*</u>	0.04	0.002	0.039	<u>0.011*</u>	0.038
TEP2	Recognition	0.006	0.067	0.008	0.074	0.009	<u>0.094*</u>
CLIP1 (PPAE)	Modulation	0.007	0.035	<u>0.011*</u>	0.031	0.008	0.047
CLIP11	Modulation	<u>0.012*</u>	0.075	<u>0.010*</u>	<u>0.079*</u>	<u>0.011*</u>	<u>0.082*</u>
SPN1	Modulation	0.011	0.048	0.009	0.039	<u>0.013*</u>	0.041
SPN2	Modulation	0.007	0.035	0.008	0.024	<u>0.011*</u>	0.032
SPN6	Modulation	<u>0.011*</u>	<u>0.091*</u>	<u>0.02*</u>	<u>0.114*</u>	<u>0.012*</u>	<u>0.116*</u>
Pellino	Toll Signalling	0.005	<u>0.088*</u>	0.004	0.065	0.004	<u>0.081*</u>
Tollip_v	Toll Signalling	<u>0.015*</u>	0.033	<u>0.01*</u>	0.043	0.008	0.034
SOCS	JAK/STAT Signalling	<i>0*</i>	0.045	<i>0*</i>	0.021	<i>0*</i>	<i>0*</i>
Attacin1	Effectors	0.007	<u>0.078*</u>	0.01	<u>0.091*</u>	<u>0.01*</u>	<u>0.086*</u>
CecropinE	Effectors	<i>0*</i>	0.017	<u>0.013*</u>	0.057	<u>0.011*</u>	0.061
Gallerimycin	Effectors	<u>0.016*</u>	0.033	<u>0.02*</u>	<i>0*</i>	<u>0.018*</u>	0.02
Gloverin1	Effectors	<u>0.013*</u>	0.072	<i>0*</i>	<u>0.398*</u>	<u>0.018*</u>	<u>0.092*</u>
Gloverin3	Effectors	<u>0.013*</u>	0.053	-	-	<i>0*</i>	<i>0*</i>
Lebocin	Effectors	<u>0.019*</u>	0.026	<u>0.018*</u>	0.016	<u>0.016*</u>	0.017
Defensin1	Effectors	<u>0.021*</u>	0.063	<u>0.02*</u>	0.066	<u>0.018*</u>	0.066

1 **Table 6-** Immune genes that had a synonymous (TajD_{SS}) or non-synonymous
2 (TajD_{NS}) Tajima's D value that corresponded to the negative or positive 2.5th
3 percentile of the genome-wide Tajima's D distribution as indicated by *.
4

Genename	Functional Category		Population	TajD _{NS}	TajD _{SS}
CTL1	Recognition		Kullaberg	1.22*	-0.48
CTL15	Recognition		Aiguamolls	-1.96*	0.10
CTL17	Recognition		Aiguamolls	1.13*	-0.42
CTL8	Recognition		Abisko	1.22*	-0.51
Draper	Recognition		Abisko	1.35*	-0.15
FREP3	Recognition		Abisko	0.42	0.84*
Galectin3	Recognition		Abisko	-0.83	0.79*
PGRP_S6	Recognition		Kullaberg	-2.04*	-0.19
SCRB2	Recognition		Kullaberg	0.51*	0.86*
CLIP5	Modulation		Kullaberg	-2.03*	-0.50
SPN5	Modulation		Aiguamolls	1.11*	0.14
Dredd	IMD	Signalling	Abisko	1.40*	0.40
Ubc13	IMD	Signalling	Kullaberg	0.00	0.77*
PIAS	JAK/STAT	Signalling	Aiguamolls	1.88*	-0.09
SOCS	JAK/STAT	Signalling	Kullaberg	0.00	1.01*
Jun	JNK	Signalling	Abisko	0.00	1.32*
Jun	JNK	Signalling	Kullaberg	0.00	-2.01*
ECSIT	Toll	Signalling	Abisko	-1.84*	-0.21
ECSIT	Toll	Signalling	Aiguamolls	-1.97*	-1.00
Pellino	Toll	Signalling	Abisko	1.24*	0.41
Rel	Toll	Signalling	Abisko	1.11*	-0.21
Toll10_2	Toll	Signalling	Aiguamolls	-1.86*	-0.60
Toll10_2	Toll	Signalling	Kullaberg	-1.89*	-0.75
Tollip_d	Toll	Signalling	Aiguamolls	-2.04*	-0.70
Tube	Toll	Signalling	Abisko	-2.04*	-1.16
Cecropin2	Effectors		Kullaberg	-2.00*	-0.28
Gallerimycin	Effectors		Kullaberg	1.84*	0.00
LLP	Effectors		Abisko	-2.16*	-1.04
Hinnavin	Effectors		Kullaberg	-2.00*	-0.28
LLP1	Effectors		Abisko	1.72*	-0.26
LLP2	Effectors		Kullaberg	1.01	-1.66*
Lysozyme	Effectors		Aiguamolls	1.50*	1.34*

Table 7 – Significant P-values of polarized McDonald–Kreitman test of immune genes using *P. brassicae* and *P. rapae* as closely related reference. If the gene has been linked to certain arm of the immune response, e.g. cellular response, this is indicated in the Response column. C-type lectins (CTL) are often involved with the cellular response, although no functional proof that these two genes (CTL17 & CTL19) are involved with the cellular response has been found. PO stands for Phenoloxidase response.).

Genename	Function	Response	Dn	Ds	Pn	Ps	DN/DS	PN/PS	NI	p-value
BGRP1	Recognition	Humoral (Toll)	62	62	8	70	1.00	0.11	0.11	0.000
CTL17	Recognition	Cellular	7	5	9	38	1.40	0.24	0.17	0.011
CTL19	Recognition	Cellular	6	4	4	28	1.50	0.14	0.10	0.006
NimB	Recognition	Cellular	5	7	1	25	0.71	0.04	0.06	0.008
PGRP_L6	Recognition	Humoral (Toll)	9	5	1	12	1.80	0.08	0.05	0.004
SCRB13	Recognition	Cellular	55	46	16	34	1.20	0.47	0.39	0.010
CLIP2 (BAEE)	Modulation	Cellular/PO	6	10	1	22	0.60	0.05	0.08	0.013
CLIP4	Modulation		4	6	2	24	0.67	0.08	0.13	0.039
SPN10	Modulation		10	3	18	31	3.33	0.58	0.17	0.013
SPN5	Modulation		6	11	2	29	0.55	0.07	0.13	0.017
Pelle	Toll	Humoral	3	5	0	14	0.6	0	0	0.036
Rel	Toll	Humoral	4	5	0	17	0.80	0.00	0	0.008
Toll6	Toll	Humoral	10	12	11	62	0.83	0.18	0.21	0.006
Toll7_3	Toll	Humoral	13	11	34	91	1.18	0.37	0.32	0.015
Iap2	IMD	Humoral	13	14	7	44	0.93	0.16	0.17	0.002
Hem	JNK		2	12	0	113	0.17	0.00	0	0.011
PPO2	Effectors	PO	11	13	4	53	0.85	0.08	0.09	0.000

Table 8 – Immune genes found to be under either positive or balancing selection. The genes experiencing positive selection were identified by MK- tests, whereas the genes under balancing selection were identified via outlier analysis. The colour represents in which population the pattern was found, with blue representing Abisko, red Kullaberg, orange Aiguamolls, and black indicating all three populations.

	Recognition	Modulation	Signalling	Effectors
Positive selection (MK test)	BGRP1	CLIP2 (BAEE)	Pelle	PPO2
	CTL17	CLIP4	Rel	
	CTL19	SPN10	Toll6	
	NimB	SPN5	Toll7-3	
	PGRPL6		Iap2	
	SCRB13		Hem	
Balancing selection	Galectin (TajD _{SS})	SPN6(π_{SS} & π_{NS})	Jun (TajD _{SS})	Lysozyme (Taj _{SS} &Taj _{NS})
	FREP3 (TajD _{SS})		UBC13 (TajD _{SS})	Attacin1 (π_{SS} & π_{NS})
	SCRB2 (Taj _{SS} &Taj _{NS})		SOCS (TajD _{SS})	Gallerimycin (π_{SS} & π_{NS})