ONE GENOME, TWO SEXES: GENOMIC AND TRANSCRIPTOMIC BASES OF SEXUAL DIMORPHISM IN SPECIES WITHOUT SEXUAL CHROMOSOMES

by

ALFREDO RAGO

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

School of Biosciences

College of Life and Environmental Sciences

University of Birmingham

September 2016

Abstract

Sex in the jewel wasp *Nasonia vitripennis* is determined by whether eggs are haploid or diploid: the radically different male and female phenotypes share the same genome, showing that their sexual dimorphism is not genetic but rather a specific case of phenotypic plasticity. As a consequence, all of *Nasonia*'s genes are selected for both male and female fitness. The impact of this diverging selective pressure on the evolution of its genome and whether it is comparable to organisms with sex chromosomes are questions still largely unanswered.

In this thesis, I develop and apply a set of tools for the integrative analysis of different aspects of *Nasonia*'s biology. I characterize the improved gene set of *Nasonia* and identify several lineage-specific gene family expansions. I provide an algorithm for detection and comparison of splicing and transcription signal from transcriptomic data in non-model organisms. Finally, I identify the different regulatory processes that enable generation of disparate phenotypes using network analyses on *Nasonia*'s developmental transcriptome.

Nasonia's transcriptome shows high amounts of sex-bias not tied to linkage groups or alternative splicing. Early development shows a prevalence of sex-biased interactions between transcripts rather than single-gene upregulation, and sex-biased networks are enriched in lineage-specific regulators.

Contents

Introduction **i**
Systems Biology in the Age of 'Omics . . . i
Development as a model for Systems Biology . . . iv
Units of Study in Developmental Systems Biology . . . viii
Nasonia vitripennis as a model for Developmental Biology . . . xi
Non-Genetic Sex Determination in *Nasonia vitripennis* . . . xiii
Sexual Development and Sexual Conflict . . . xv
Summary . . . xvii

1. OGS2: GENOME RE-ANNOTATION OF THE JEWEL WASP *NASONIA VITRIPENNIS* **1**
1.1. Abstract . . . 1
1.2. Background . . . 3
1.3. Methods . . . 8
1.3.1. Gene Set Construction Process . . . 8
1.3.2. Gene set Quality and Completeness metrics . . . 8
1.3.3. Ortholog group assignments and gene family expansions . . . 10
1.3.4. Identification of fast- and slow-diverging genes in *Nasonia* relative to ants and bees . . . 13
1.3.5. Functional enrichment testing . . . 14
1.3.6. Alternative splicing analysis . . . 15
1.3.7. Additional software tools . . . 16
1.4. Results and Discussion . . . 16
1.4.1. Transcriptional and homology data complement each other . . . 16
1.4.2. Missing gene families are absent from the *Nasonia* genome . . . 18
1.4.3. Gene model quality and diversity increase . . . 20
1.4.4. *Nasonia* shows biologically relevant lineage-specific duplications . . . 22
1.4.5. Wasp lineage diversification is driven by transcriptional regulators . . . 22
1.4.6. Histone genes are enriched in lineage-specific evolution . . . 24
1.4.7. Alternative splicing and *lola* expansion . . . 26
1.4.8. Which factors promote the evolution of alternative splicing in *Nasonia*? . . . 27
1.5. Conclusions . . . 34

2. FESTA: *F*LEXIBLE *E*XON-BASED *S*PLICING AND *T*RANSCRIPTION *A*NNOTATION **37**
2.1. Abstract . . . 37
2.2. Background . . . 38
2.3. Implementation and Usage . . . 39
2.3.1. Data input and filtering . . . 39
2.3.2. Isoform detection . . . 39
2.3.3. Fine tuning parameters . . . 40
2.3.4. Caveats . . . 41
2.4. Conclusions . . . 41

3. Transcriptomic Basis of Sexual Dimorphism in *Nasonia vitripennis* 43
3.1. Abstract . . . 43
3.2. Background . . . 44
3.3. Methods . . . 50
3.3.1. Biological Materials and Data Collection . . . 50
3.3.2. Data Pre-processing . . . 51
3.3.3. Splicing Detection and Network Construction . . . 51
3.3.4. Network Topology Measurements . . . 54
3.3.5. Differential Expression of Nodes and Clusters . . . 56
3.3.6. Linkage Clusters Enriched in Sex-Biased Loci . . . 56
3.3.7. Differential Correlation Analyses . . . 57
3.3.8. Multivariate analysis of network parameters . . . 59
3.3.9. Phylostratigraphic analyses on network parameters . . . 60
3.3.10. Gene Ontology, and Protein Family Enrichment Analyses . . . 61
3.3.11. Additional software tools . . . 61
3.4. Results . . . 62
3.4.1. Stage-specificity of gene-level sex-bias . . . 62
3.4.2. Low prevalence of sex-biased splicing . . . 63
3.4.3. Genomic regions enriched in sex-biased genes . . . 63
3.4.4. Differential Cluster Expression Reveals Meiosis Genes . . . 65
3.4.5. Differential Correlation Reveals Early Sex-Biased Transcription . . . 66
3.4.6. Sex-Biased Clusters Show Different Regulatory Organizations . . . 69
3.4.7. Sex-Biased Clusters Integrate New Genes in Regulatory Positions . . . 70
3.5. Discussion . . . 74
3.5.1. Sex-bias at the Single Locus Level . . . 74
3.5.2. Sex-biased Linkage Groups . . . 75
3.5.3. Heterochrony in Gametogenesis Drives Developmental Sex-Bias Shifts 76
3.5.4. Sex-Bias in Early Development . . . 78
3.5.5. Network Structure of Sex-Biased Clusters . . . 79
3.6. Conclusions . . . 81

4. General Conclusions 84

A. Appendix 87
A.1. Additional Figures and Tables . . . 87
A.1.1. Code for the FESTA algorithm . . . 94

B. Publication 1: Function and evolution of DNA methylation in Nasonia vitripennis 96

References 117

List of Figures

1. **An example of the dangers of excessive reductionism**
   Image from Randall Munroe (xkcd.com). ii
2. **Number of genes with medium or higher support from sequence orthology, evidence of transcription, or both.**
   Medium support is defined as overlap greater than 30%. Panels show the source of evidence for genes within the ortholog and paralog subsets and the whole OGS2 17
3. **Protein divergence of OGS2 genes against orthologs in other Hymenoptera**. Every point represents a gene mapped on three coordinates originating from the corners. Each gene's distance from a corner is proportional to the average amino-acid distance of orthologs between the two clades.
   AB = ant to bee distance; AN = ant to *Nasonia* distance; BN = bee to *Nasonia* distance. Diverging genes are highlighted in orange (fast) and blue (slow) as detected by the compound ratio (A) and intersection of ratios (B). See materials and methods for full description 23
4. **Alternatively spliced introns for *lola* in Apis (blue) and Nasonia (red)** Graph shows intron spans from a common hub exon, in bases on their genomes. Blue and red bars at top of figure are short introns that join pairs of 3' end exons in *lola* gene span. 26
5. **Number of genes with alternative isoforms in OGS2** (A) split by presence of paralogs and (B) split by methylation in adult females. 27
6. **Effect of different factors on the probability of observing alternate isoforms of OGS2 gene models.** Factors are ranked by relative importance (y axis). Factors with complete support and levels of the same factor were adjusted for plotting. Effect sizes are shown as the fold change in probability from the intercept (with 95% confidence intervals). Numeric variables were log transformed prior to analysis. 29
7. **Outline of the FESTA algorithm**. Steps A-C are handled by the ClusterExons function. Step D and optional step E are handled by the AverageExons function. 39
8. **Expression Values Distribution Before Thresholding**
   Vertical red lines indicate the 50th, 66th, 90th and 99th percentiles respectively. Expression scores are reported as log ratios against the 99th percentile of random Markov probes. 51
9. **Network construction workflow**:
   I select exons based on expression within our experiment and cluster them using FESTA. Every gene is represented as a transcription node and a variable number of splicing nodes, each quantifying the inclusion ratios of a correlated set of exons. Groups of nodes with reciprocal correlations greater than 95% are collapsed into CCREs. The resulting dataset is converted into a network and clustered using WGCNA. Final figures indicate the amount of clusters, unclustered nodes and total genes in my network. See section 3.3.3 for details. 52

10. **Node Parameters in an Example Network:**
Table on the right side lists the number of connections (Con) and hub scores (Hub) of labeled nodes.
Node A is a high order co-ordinator, node B part of a 3 node complex and node C a worker. 54
11. **Number of Genes with Sex-Biased Transcription and Splicing**.
Yellow cells indicate over-representation, blue ones under-representation . . 63
12. **Sex-Bias in Expression and Correlation at the Cluster level.**
Positive values indicate male-bias, negative values indicate female-bias. . . 66
13. **Network parameters associated with sex-biased clusters**
Scaled PCA loading values indicated on the y axis. Each PC is listed with its associated level of variance explained. Within-panel percentages indicate the RI of each PC in the model set separating unbiased clusters from differentially correlated (above) or differentially expressed (below) ones. 69
14. **Proportions of genes from each taxonomic stratum in different classes of sex-biased clusters**. Proportions reported are fold-enrichment compared to the network-wide abundances of genes from each stratum. Y axis is truncated between 0.5 and 2.5 fold enrichment. 71
15. **Effect sizes of sex-bias categories on connection densities (upper) and hub scores (lower) of individual transcripts of different phylostratigraphic age.**
Asterisks indicate non-overlapping 95% intervals between sex-bias categories in the same phylogenetic stratum. All effects are calculated relative to the Metazoan stratum. For details on the modelling see section 3.3.9. 72
16. **Log counts of methylated and unmethylated genes in different classes of expression support**. 90
17. **Correlations between different cluster parameters in the *Nasonia* developmental network**
Yellow squares in the bottom left corner indicate positive correlations, blue ones negative. Lighter shades are more significant than darker ones. Numbers at the top right corner indicate the Pearson correlation score with confidence intervals in parentheses. 91

List of Tables

1. **Summary of the Official Gene Set (OGS2)** comparing all gene constructions to good constructions having expression and/or homology evidence and to the previous OGS1.2 gene models. Percentages are of the total number of genes for the set.
   * 2,935 OGS1.2 models are classified with strong homology to transposon proteins during OGS2 work, 385 models with expression and other insect homology but also transposon homology were retained in OGS2 "good" model set
   ** 5,763 additional genes of OGS2 have significant protein homology, but are not assigned as orthologs in OrthoMCL orthology analysis, 3,454 of 24,388 "good" models lack significant homology, but have expression evidence. 9
2. **The types of evidence and levels of support for Nasonia *vitripennis* gene sets.**
   Sequence-level statistics for the different types of evidence are given as proportions of the gene sets that are validated. Gene structure level statistics (ESTgene, Progene, RNAgene) are counts of the number of models that reach three structure level agreements. Homology level statistics are counts of the number of models and proportions matching proteins of reference species and paralogous (same species) proteins. See methods section for details on the evidence types and the statistics that were measured. 11
3. **Number of insect genes classified to gene families (GF) that are common among the arthropods by OrthoMCL** (ARP9, version arp11u11). Five out of nine insect species are summarized. Dupl and Singl designate the proportion duplicated and singleton genes relative to the median found among insects (Dupl:5000, Singl:10000). 18
4. **PCA Scores of individual cluster parameters**, approximated to the third digit 59
5. **Number of sex-biased genes and transcriptional events at each developmental stage.** Genes are counted as sex-biased if at least one of their transcription or splicing nodes is sex-biased. 62
6. **Linkage groups enriched in sex-biased genes.**
   Numbers indicate gene counts with their percentages compared to all genes in the linkage group. Recombination rates are expressed as centiMorgan per Mb. The last row reports median proportions and recombination rate across all linkage groups. 64
7. **Differential Expression (7a) and Differential Correlation (7b) Patterns across Development and number of Clusters and Genes which exhibit them.** Each pattern is coded as a string of five characters indicating its sex-bias status at each developmental stage from early embryo to adult: male (m), female (f), none(.). The number of genes per pattern includes all genes within all clusters that show that pattern. 65
8. Histone genes present in OGS2.0 annotated with presence or absence of lineage-specific expansions. NA entries were not assigned to orthologous groups at the level of Hymenoptera. 87

8. Histone genes present in OGS2.0 annotated with presence or absence of lineage-specific expansions. NA entries were not assigned to orthologous groups at the level of Hymenoptera. 88
9. **Consensus in the location of the OGS2 gene set on the genome assemblies of sibling species Nasonia *longicornis* and N. *giraulti*, including recent, high identity paralogs.** Almost all OGS2 genes are located on 2 sibling species draft assemblies Werren et al. (2010), using GMAP Wu and Watanabe (2005) transcript mapping. Paralog locus consensus patterns are tabulated for inparalogs (sharing orthology to other species) and uniquepar (lacking strong homology to other species). Of the total paralog families, each with several genes, most paralogs are on different scaffolds for all species. The counts of tandem paralogs with different separations are indicated. 89
10. **Gene set quality measurements**. Including deviation of protein size from the group median, and maximal bit score per species in pairwise comparisons within the arthropod orthology groups. The bit score measures both gene model artefacts of alternative gene sets within species and evolutionary divergence. Protein sizes may be more evolutionarily conserved, and may detect artefacts across and within species. See materials and methods for details on how each score is generated. 89
11. **Predictors of Connection Density (11a) and Hub Scores (11b)** with Model-Averaged Effect Sizes and Relative Importances, or probability that the factor in question is included in the best model. Stratum coefficients are relative to the *Nasonia* stratum. See section 3.3.9 for details. 92

INTRODUCTION

Systems Biology in the Age of 'Omics

2016 is an exciting time to be a biologist. Hybridization, sequencing by synthesis and mass-spectrometry can now be performed thousands of times in parallel in just a few hours. Powered by these technologies, a multitude of 'omic disciplines have been created with the goal to detect, characterize and quantify all measurable parameters of living organisms. Despite the diverse histories and applications of 'omic disciplines, most of them are based on the same fundamental assumption: Collection of large datasets brings the mechanistic basis of biological responses into the light.

This reasonable concept has unfortunately lead to the unrealistic public expectation that complete measurements of a single 'omic dimension (such as the genome) could lead to complete understanding of organismal responses (Eddy, 2013). As the last 30 years of research have shown, individual 'omics inquiries allow unprecedented insight on the mechanisms of biological responses. Yet, both responses and mechanisms vary in often unpredictable manners: epistasis, epigenetics, genotype by environment interactions, plasticity and condition-dependence are just a few of the concepts that have been borrowed or created to explain this variation in response mechanisms (Mackay and Anholt, 2006; Burggren and Crews, 2014; Olson-Manning et al., 2012; Hemani et al., 2014; Golan et al., 2014). All of them have different modes of functioning, but all are used to account for variation in responses through the effect of an additional regulatory[1] layer *via* "black box" modeling.

The pervasive presence of non-additive between-layer interactions (Huang et al., 2012; Bloom et al., 2013; Golan et al., 2014) presents a strong critique to reductionist approaches, as no explanation can be provided unless all relevant parameters are accounted for.

[1] I refer here to broad-sense epistasis, intended as the "masking" effect of the biological background on genetic changes. Strict-sense epistasis is an exception to this category, since it explains modification in a gene's action that depends on the rest of the genomic repertoire. Strict sense epistasis is conceptually more similar to a second-order term within the same level of regulation rather than an interaction term as it does not require alternative regulatory processes. Interestingly, both are generally deemed as a nuisance and neglected in traditional genetic inquiries.

In accordance, part of the scientific community is leaning towards holistic approaches, gathering data on multiple regulatory processes and integrating it to explain the final outcomes. This holistic biology or Systems Biology aims at explaining higher-level responses (organism, population or even ecosystem) using the interplay of multiple layers rather than single 'omic approaches (Civelek and Lusis, 2014; Bittleston et al., 2016).

Systems Biology can be described as an expansion of physiology that accounts for heritable differences in the regulatory mechanisms; a more inclusive genetics that accounts for physiological responses or even as the branch of cybernetics that studies how biological systems integrate internal and external information to produce adaptive outcomes. Rather than focusing on individual components of a single regulatory layer (such as causal gene mutations or key hormones) systems biology deals with the interactions between those elements (Civelek and Lusis, 2014). A genomicist might look for mutations that impede male development. A biochemist will be interested in which hormones differ between sexes. An ecologist could assess which environmental factors influence sexual development. A systems biologist will search for interactions between genes, hormones and environments to detect those that cause phenotypic changes (Bossdorf et al., 2008). Interactions within and between regulatory layers are thus integrated in a single conceptual framework that allows for the exploration of emergent properties of the whole system (i.e. Williams et al., 2011).

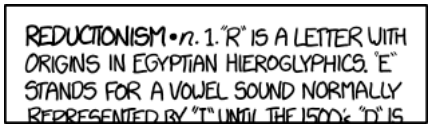


Figure 1: **An example of the dangers of excessive reductionism** Image from Randall Munroe (xkcd.com).

It is now evident that, even when focusing in a single 'omic space, biological systems display a staggering amount of complexity in the form of numerous non-linear interactions and intricate regulatory loops (i.e. Davidson, 2002; Gerstein et al., 2012; Stazic and Voß, 2016; Soshnev et al., 2016). Perhaps due to a fascination with this complexity, most systems biologists consider the description of regulatory networks as the purpose of this new discipline, with the ultimate goal of being able to generate a perfect predictive model of

any biological system as exemplified by the prevalent "blueprint", "program" and "circuit" metaphors (Stuart, 2003; Barabási and Oltvai, 2004; Marbach et al., 2012a; Rhee et al., 2014). However, this ultimate goal fails to represent the key feature that sets biological systems apart: the fact that their current structure is not the result of a goal-oriented design but rather of a whole host of factors that influenced their evolution (Knight and Pinney, 2009). Faced with the immediate usefulness of predictive modeling (i.e. Du and Elemento, 2015), the study of developmental systems' evolution might seem a purely academic pursuit. Yet, evolution is a pervasive process and even regulatory mechanisms vary in accordance to the rules of mutation, selection and random drift (Lowdon et al., 2016). Adding phylogenetic and evolutionary constraints is thus a necessary complication if we aim to predict how phenotypic responses can (or cannot) vary across species, between environments and over time (Blank et al., 2014; Botero et al., 2015). The evolutionary dynamics of biological regulatory networks ultimately underlie the key question of whether a population will be able to push the boundaries of its current physiological limits or it will be constrained by them.

It can be argued that holistic methods are needlessly complex. Predator-prey dynamics can be explained by elegant modeling equations (Abrams, 2000), local adaptation by mutation-selection models (Hendry, 2013) and gene regulation by direct molecular interactions (Roy et al., 2010; Marbach et al., 2012b). Guided by the corollary of Occam's razor[2] we should choose the simplest alternative explanation, refusing the overly complex system inquiries and instead refining current reductionist methods. There is however at least a caveat to this argument: despite the aesthetic appeal of simple and elegant explanations the true answer may still lie in more complex models. This is especially true in biological systems where the observed higher level dynamics are likely to emerge from lower-level

[2]Occam's razor is the most popular version of the parsimony principle, a key tenet in scientific and philosophical inquiries. It states that one should avoid the needless duplication of entities. Therefore, all else being equal, we should always lean towards models which require as few parameters as possible. This heuristic method is justified by the observation that we can construct a limitless number of arbitrarily complex models that fit our system equally well. However, complex models will also be harder to falsify, and should therefore be considered only if the simpler ones have already been proven incorrect.

interactions (Minelli and Fusco, 2012; Burggren and Crews, 2014, and see Hunt et al., 2013; Payne and Wagner, 2014b for an examples) or are at least sensitive to the environment of the organisms involved and the molecular context of their cells. In the following section, I will use the specific case of the development of alternative phenotypes to demonstrate how biology is rich in systems that present non-reducible multi-layer interactions and how we now have the tools to investigate them.

Development as a model for Systems Biology

Development is a fitting example of a non-reducible process. In 1958, Gurdon et al. gave conclusive proof of the genome's prime importance in determining the phenotype of an organism through the use of nuclear transfer of *Xenopus laevis* (see Gurdon, 1986 for a retrospective). This technique proves that despite transfer in a different cytological context and development in a foster mother (in the case of *placentalia*), the genome of a terminally differentiated cell is still sufficient to drive the ontogenesis of a near perfect replica of its donor. Even more strikingly, the nucleus of a species can trigger the correct developmental program when transplanted in a closely related organism's oocyte (De Robertis and Gurdon, 1977). Yet, just as cloning proves the power of DNA, it also highlights its limitations as even clones show significant differences. A beautiful proof of how stochastic effects produce differences between clones is given by calico cats. A single X-linked allele is responsible for determining whether a cat's coat will be either red (recessive) or black (dominant). Males and homozygous females display uniform coloration, but heterozygous females will display a spotted mosaic of both. This mosaicism arises from the stochastic inactivation of either copy of the X chromosome and all of its genes. Since the choice between which X chromosome to inactivate is random and happens independently in different cell lineages even clones display different color patterns (Shin et al., 2002). It is also worth noting that the susceptibility of this trait to chance effects is genetically determined. If the locus responsible for color were to be relocated on an autosome it would no longer be influenced by X-inactivation and would instead be subject to standard dominant-recessive dynamics

(Brown and Greally, 2003). It follows that genome structure can influence the degree of randomness that affects development.

As we will see in the next section, the ability of evolution to adapt to stochastic events can also be turned towards more recurrent cues, such as epigenetic factors and environmental signals. Even more interestingly, adaptations to developmental variation can act as the process that leads from a single starting point towards the highly divergent and specialized outcomes we classify as polyphenisms. Reconstructing the processes that lead from non-adaptive developmental variation to the highly constrained one observed in extant polyphenisms is a challenging problem but one that can lead to significant insights for evolutionary biology. Environmentally induced variability is widely recognized in the special case of physical and chemical factors that disrupt "normal" developmental pathways (teratogens). Biological systems can evolve mechanisms to block or neutralize environmental interferences. This ability of developmental pathways to produce the same phenotype regardless of perturbations is named robustness (Payne and Wagner, 2014b) or canalization (Waddington, 1942). The most well studied mechanisms that induces robustness are molecular buffers[3]. This broad category includes the proteins and complexes that counteract environmental perturbations on a molecular scale. They can either shield the embryo (through impermeable barriers), neutralize the damaging components (export channels and sequestering molecules) or buffer and undo damage itself (chaperones and proteasome) (Gilbert and Epel, 2009). These molecules constitute but a small selection of countermeasures that animals have evolved to protect the delicate ontogenetic dynamics. The importance of buffering mechanisms in development is underscored by experiments that ablate them. Heat shock proteins (HSPs) are a family of chaperones whose role is to assist the folding of proteins (Lindquist and Craig, 1988; Pirkkala et al., 2001). In both

[3]Robust developmental pathways can theoretically arise without dedicated buffering systems but rather due to intrinsically robust regulatory architectures (Payne et al., 2014). Feedback loops are one of such cases. Negative feedback loops protect development against temporary fluctuations in signaling molecules. Positive feedback loops ensure the irreversibility of fate determination events. Regulatory robustness presents an efficient alternative to dedicated buffering systems, yet it is difficult to prove whether it is the result of selective pressure towards robustness or a by-product of emergent biological network properties.

Drosophila and *Arabidopsis*, overloading of the buffering capacity of HSPs through extreme environmental stress or deletion leads to a slight increase in variation both between and within individuals (Rutherford and Lindquist, 1998; Queitsch et al., 2002; Takahashi et al., 2011). Similar effects can be exerted by a wide range of deletions in other genes, suggesting that robustness might not just be restricted to direct molecular buffering systems but also mediated by regulatory structures (Takahashi et al., 2012).

It is especially important to underscore that the increase in variation under environmental stress is not exclusively due to stochastic failures but is mediated by a genetic component (Badyaev, 2005): a considerable portion of the variants revealed by removal of HSP buffering can be selected for and is therefore due to otherwise cryptic genetic variation (Takahashi, 2013). Once this latent diversity (Payne and Wagner, 2014a) is revealed it can be shaped by natural selection just like constitutively expressed alleles, with two evolutionary outcomes depending on the net fitness of the revealed phenotypes: variants that cause a loss of fitness will be selected against while those that increase the animal's fitness will be selected for. This process will eventually result in a more environmentally robust developmental pathway which minimizes the chances of induction of maladaptive phenotypes and maximizes those of generating the adaptive ones[4](Badyaev, 2009; Standen et al., 2014). A third outcome is possible in case the revealed phenotypes are advantageous in the environment that causes their induction but otherwise deleterious. In this scenario condition-dependent expression presents already the optimal evolutionary strategy. Natural selection can further refine the adaptiveness of the induced phenotype by placing additional traits under the control of the same regulatory mechanisms, a process called genetic accommodation (i.e. Suzuki, 2006). The long term effect of selection for environmentally sensitive expression in genes

[4]The shift from induced to constitutive phenotypes has been known for a long time by developmental biologists. The concept was first introduced by James Baldwin (1896a; 1896b). Waddington (1953) was able to select for constitutive expression of environmentally induced phenotypes in *Drosophila* and strongly advocated the term genetic assimilation to describe it, ironically leading to the more widespread adoption of the term "Waddington Effect". In more recent times, Gilbert and Epel (2009) proposed the more generic term "heterocyberny" (shift in mode of control), which has the advantage of including the opposite phenomenon of genetically encoded phenotypes shifting to environmental control and is congruent with the other three major categories of innovation of evolutionary developmental biology (heterochrony, heterometry and heterotopy).

that confer environmentally-dependent advantages is alternative phenotypes displaying a correlated suite of traits that increase the organisms' fitness in the environment that induces them while avoiding the cost of those adaptations in environments that do not require them.

It might seem that the evolution of inducible phenotypes requires an unlikely combination of factors. First, the induced phenotype must provide a context-dependent advantage from the start to be selected for. Second, selection must be strong enough to counteract the recombination forces that would break apart alleles required for generating a coordinated suite of traits. Finally, inducible and non-inducible individuals will not be discriminated by selection in a non-inducing context, further diminishing the pressure to maintain this ability[5]. Despite the theoretical difficulties in both originating and maintaining inducible phenotypes, organisms with alternative phenotypes are widespread in nature and thrive due to their ability to integrate environmental information into developmental pathways to generate adaptive phenotypes.

The paradox of alternative phenotypes (and their continuous counterpart, reaction norms) gathers interest from several areas of biology. Ecologists are intrigued by how they enable a single species to fill multiple mutually exclusive niches (Nijhout, 2003; Shine, 1989). Genomicists and evolutionary biologists are puzzled by their ability to store and quickly retrieve multiple adaptive phenotypes in a single genome (Chen et al., 2010; Simon et al., 2011). Taxonomists and developmental biologists are fascinated by how animals with near identical genomes, such as different sexes, can differ more than sister species and evolve independently of each other (Jousselin et al., 2004; Hunt et al., 2013). The sheer diversity of those questions demonstrates that alternative phenotypes are at a fortuitous crossroads of interests between different sciences, one that requires

[5]There is currently debate on whether the inducible individuals are truly selectively identical to non-inducible ones in a non inducing environment. Current theories postulate the existence of a cost of plasticity (Snell-Rood et al., 2010), which would lead to negative selection towards inducible individuals under non-inducing circumstances. A simple example is provided by the observation that a minor proportion of inducible organisms will be subject to random activation of the alternative pathway even in a non-inducing environment, with maladaptive outcomes. For the purpose of this argument the hypothesis of the two phenotypes being identical is therefore conservative.

all of these areas to be integrated to provide a satisfactory explanation. Reductionist pursuit of each individual component would incur the risk of promoting compartmentalized, incompatible and ultimately incorrect theories. Mechanistic studies without appreciation for the rules of change will lead to mere descriptions of observed patterns, unusable for generalized inferences on the evolution of plasticity. Purely evolutionary inquiries will instead model fictional constructs such as independent small-effect loci with isotropic and constant potential for gradual generation of continuous change. Only by approaching developmental mechanisms in an evolutionary framework and molecular evolution in a developmentally informed fashion we can achieve theories which adequately represent the complex reality we can observe with the lens of high-throughput data.

Units of Study in Developmental Systems Biology

A fundamental difference between Systems Biology studies and individual 'omics is that the latter often include stringent definitions of the units of interest. By contrast, System Biology studies deals with heterogeneous types of entities ranging from RNAs to histone modifications. Further to that, relationships between entities can also be altered by evolution ("re-wired", Villar et al., 2014; Cotton et al., 2015) and constitute a possible subject for inquiries by themselves (Bittleston et al., 2016). Innovation can occur by multiple modes even in a simple toy system comprised by a single transcription factor regulating a set of genes that contain a single binding motif. Regulation of the transcription factor's expression, gain or loss of the target motif by target genes or changes in the specificity of the TF to motif binding can all alter the final outcome. An accurate choice on how to represent data acquired in Systems Biology investigations is thus critical to address evolution at the level or levels of interest, as well as several other ideal properties. Data representations need to convert highly complex, redundant and noisy datasets to more manageable ones (Berger et al., 2013). This conversion implies the ability to partition between relevant variation and noise, which itself depends on formulating realistic models on the overall expected behavior of entities in our dataset. Systems Biology studies need

to integrate seamlessly diverse types of data, accounting for the idiosyncratic properties of the instruments used for data acquisition as well as the known differences in layer-specific dynamics (Lowdon et al., 2016). Perhaps most important of all, the resulting non-redundant integrated data representation needs to allow the researchers to easily trace back to the biological processes they arise from in order to facilitate interpretation and validation. It is therefore crucial to design data representations not only based on computational data-handling necessities or generic properties of the type of data, but also with a clear biological framework in mind centered around the level or levels of primary interest.

The specific problems addressed in this thesis integrate genomic and phylogenomic data but focus primarily on transcriptomic measurements in order to test hypotheses. My focus on this specific regulatory level can be justified by both practical and theoretical motivations. From a practical point of view the techniques required for transcript identification and quantification are mature in throughput and accuracy but still lacking a consistent theoretical framework from which to draw null hypotheses and expectations. While obtaining quantitatively accurate measurements of transcripts is rapidly becoming less challenging, data interpretation is still reliant on either differential expression (plagued by multiple-testing penalization) or machine learning algorithms,which aim to improve performance scores rather than test predictions. The compresence of a streamlined data collection pipeline and relative lack of mechanistic models to explain observed patterns through the underlying processes makes transcriptomics a promising field for the application of biologically-minded analysis methods. From the theoretical side, transcripts provide an obligatory step from genetic material to phenotypes. Transcriptome analysis enables inferring upstream causes of gene regulation without selecting a single regulatory mechanism (i.e. DNA methylation, transcription factor binding or chromatin remodeling). As for the downstream effects of genes, while most phenotypic effects are not carried out by RNAs themselves but rather by the proteins they code for, transcripts remain a necessary transition step between DNA and phenotypes since all genes need to be transcribed in

order to exert a function, making transcriptomic analyses at least qualitatively appropriate for most applications (Roy et al., 2010).

Transcriptomic status is determined by multiple regulatory modes which can be partially disentangled by appropriate data representations but is instead often interpreted by collapsing all signal at the gene level. This approach has its roots in a long-standing tradition of evolutionary modeling and molecular experimenting which helps define clear expectations and consistently classify deviations from the norm. However, gene-centric approaches effectively discard all information other than whole-gene expression. Condition-specific transcripts are either ignored, scored as differentially expressed genes or add up to transcriptional noise further hindering the discovery of differentially expressed genes. Using transcripts-specific expression as the unit of study preserves this additional information, but loses track of whole-gene regulation and is thus unable to address splicing dynamics. Integration of genomic data is also fundamental for an appropriate characterization of the regulatory basis of differential expression by analyzing their regulatory sequences or checking for spatially clustered groups of differentially expressed genomic regions. Phylogenomic data in the form of reliable orthology assignments and dating of genomic events are also necessary if we aim to understand the evolutionary processes that lead to the observed differentiation in gene expression and distinguish between co-evolution and co-expression. Finally, the effect of numerous transcripts is highly dependent on which other transcripts are also present in the same cellular context, an interdependence which frequently results in effects qualitatively different from the sum of their parts. Groups of coexpressed transcripts can therefore be considered interesting units for selection since their effect cannot be reduced to their individual components.

Over the course of chapter section §2 and section §3, I will show how the appropriate use of biologically-informed data representations can help disentangle otherwise inaccessible forms of gene regulation and unveil how their reciprocal contributions and interactions generate the diversity required for a single genotype to generate two sexes in our model system.

Nasonia vitripennis as a model for Developmental Biology

While alternative phenotypes are widespread in nature, several other properties are necessary to enable the exploration of interactions between genotype and environment. Alterations in the environment can be easily induced in a laboratory setup, but the same is not generally true for genotypes. Traditional genetics employs crosses or molecular techniques to selectively activate and inactivate genes, whose development in novel organisms remains a challenging and active area of research (Huang et al., 2016). Testing a series of defined and stable genotypes between different environmental conditions is also a key requirement for statistical tractability of environmental plasticity, but is only possible in species with clonal reproduction. Lastly, most 'omics explorations require a mature knowledge base. This includes, but is not limited to, a reference genome assembly for QTLs/eQTLs, a metabolomic reference database for chemical identification, a complete gene set for transcriptomics (van den Berg et al., 2010), a reference methylome for DNA modification studies and a catalog of protein modifications for molecular interaction studies.

This thesis deals with the development of *Nasonia vitripennis*, which displays sexual dimorphism in spite of its lack of sex-specific chromosomes. I focus on the specific case of sexual dimorphism as it is both widespread and already extensively modeled in pre-genomic studies. The term development is here accurate both in the biological and engineering sense, as this work serves the double purpose of describing the embryonic progression of this organisms and, at the same time, generating the knowledge base to enable further systems biology investigations into it. While the individual pieces of work each focus towards the characterization of the genome, transcriptome or methylome of *Nasonia vitripennis*, I included a systems perspective linking gene regulation mechanisms through a phylogenetic and evolutionary framework.

Nasonia vitripennis is a member of Hymenoptera (ants, bees and wasps) and to date remains the only wasp with a fully assembled and annotated genome (Werren et al., 2010). The *Nasonia* genome project also provided draft genome assemblies of the sister

species *Nasonia giraulti* and *Nasonia longicornis* by mapping them to the *vitripennis* genome. The availability of assemblies for three species within the same genus provides a solid foundation for phylogenetic inquiries. This advantage is strengthened by the weak reproductive barriers present within the *Nasonia* genus, which are for the most part enforced by bacteria-induced cytoplasmic incompatibilities (Bordenstein et al., 2003). Species cured from those parasites can be crossed in the laboratory thus allowing the study of the genetic bases of speciation through direct experimentation (Desjardins et al., 2013; Niehuis et al., 2013). *Nasonia* is also a parasitoid of dipteran larvae, and is therefore of high value for modeling predator-prey evolution with potential future applications as a natural remedy to pests (see Kaufman et al., 2001 for a case study). *Nasonia*'s venom does not kill its prey but causes its developmental arrest, converting dipteran larvae into a more suitable host for its own offspring (Rivers and Denlinger, 1995). This precise regulation of a prey's development by a predator's venom offers a fascinating window on molecular co-evolution.

The main asset of *Nasonia* as a genetic model lies in its reproductive cycle. Like other Hymenoptera *Nasonia* has haplodiploid sex determination: males are produced by unfertilized haploid eggs and females by fertilized diploid eggs. Unlike other model Hymenoptera, *Nasonia*'s life-cycle is brief, asocial and allows for repeated cycles of inbreeding. The combination of haplodiploid genetics and inbreeding allows fast and accurate analyses of its genome via crosses of homozygotic lines (Pultz et al., 2000; Pultz and Leaf, 2003). It is also important to point out that *Nasonia*'s molecular toolbox already includes targeted gene knock-outs which can be directed to either the zygotic or the parentally inherited supply of RNAs (Lynch and Desplan, 2006), an opportunity that has already been used to discover a developmental path much less reliant on maternal inheritance than that of *Drosophila* (Pultz et al., 2005). *Nasonia's* haplodiploid sex-determination is made even more intriguing by its plastic reactions to environmental conditions. So far, the list of factors with proven effects on the ratio of males per brood includes female choice (Werren, 1980), selectable alleles (Pannebakker et al., 2011), bacterial

infection (Darby et al., 2010) and selfish genetic elements both paternally (Beukeboom and Werren, 2000; Werren, 1991) and maternally (Skinner, 1982) transmitted. *Nasonia*'s sex-determination system is therefore rich with ecologically relevant and naturally occurring interferences at several regulatory layers.

For the purpose of this work I will focus on the dimorphism present between males (small, with vestigial wings, pheromone producing and short lived) and females (large, flying, venomous and long-lived). However I wish to point out that *Nasonia*, as other holometabolous insects, possesses a larval stage that is radically different in physiology and ecology from the adult form and may as well be considered an alternative phenotype occurring within the same organism at different times.

Non-Genetic Sex Determination in *Nasonia vitripennis*

From a traditional genetic standpoint sex determination might seem as an unusual place in which to look for developmental plasticity. Our focus on models with genetic sex determination has led to a tendency to consider sex ratios as an unresponsive trait fixed on the 1:1 ratio. The evolutionary argument in favor of this ratio was first postulated by Fisher in 1930 as follows: polygamous sons confer higher chances of transferring their parents' genetic inheritance in populations where females are readily available, as they will be able to sire more than a single brood. This will lead to a male-biased population where less than a female per male is available. Female producing alleles will be favored in a male-biased population, reversing the trend towards male production. This dynamic equilibrium ensures that the only evolutionarily stable sex-ratio will be 1:1 even if adaptive optima depend on the population's current sex ratio. Already in 1967 Hamilton pointed out that this model is correct only for loci that have the same number of copies in each sex (Hamilton, 1967). Y-linked loci will favor males as they are the only ones that propagate them and vice-versa for X-linked ones, which have double copies in females. The same conflict will be even more widespread in species with haplodiploid sexes, as all loci are duplicated in females and single copy in males. He also included a list of the several cases

in which production of sexes is biased among natural populations. This is indeed the case for our model organism. As we will see, *Nasonia vitripennis* presents not only a naturally female-biased life-cycle, but also the ability to plastically adapt the sex-ratios to its environment with beneficial effects for its fitness. Because of this ability, a vast amount of effort has been invested in dissecting its sex determination mechanisms.

Cytologically, *Nasonia*'s sex is determined by the number of copies of its genome like in other Hymenoptera (Heimpel and de Boer, 2008; Beukeboom and Van De Zande, 2010). However, we know that its mechanism of primary sex determination is fundamentally different than that of the main model organism of its order: *Apis mellifera*. Primary sex determination in *Apis* is controlled by a single gene (*csd* or *complimentary sex determiner*), a duplicate of the arthropod *transformer* (*tra*) . If *csd* is present as a heterozygote in the organism's genome, it will initiate the female-specific splicing of *doublesex* (*dsx*). If present as either an homozygote or an hemizygote (as in unfertilized eggs), it will instead lead to the male-specific splicing of *doublesex*. This locus thus exerts a double function as both a sex determinant and a control against inbreeding, lowering the amount of homozygosity in queens (see Gempe and Beye, 2011 for a comparative review). *Nasonia* lacks the *csd* locus, and can be inbred for several generations without leading to increased male counts (Verhulst et al., 2010a). This suggests either an independent evolution or a drift of the upstream sex determination mechanisms (Verhulst et al., 2010b). The identity of the *Nasonia* primary sex determinant remains unknown. Current consensus tends towards a gene epigenetically silenced in the maternal copy of the zygotic genome (Trent et al., 2006). If that is the case only fertilized eggs will inherit the paternal active copy of the female sex-determining locus, leading to differential *transformer* splicing and a female phenotype. However, to date such gene remains to be found and we cannot exclude several competing hypotheses (Verhulst et al., 2013). As with other arthropods, splicing appears to play a key role both in the induction and the establishment of sex. Maternal inheritance of the female-specific splicing isoform of *transformer* is necessary to induce the female developmental pathway as failure to provide sufficient amounts of maternal

transformer RNA results in diploid males (Verhulst et al., 2010a). Despite some evidence that methylation might be a regulator of splicing in *Apis mellifera*, knock-out of maternally provided methyltransferases in *Nasonia* embryos does not impede the correct processing of *transformer* (Zwier et al., 2012). I have already pointed out the numerous factors that can influence sex determination of *Nasonia* in nature. It does not seem too far-fetched to hypothesize that the complex picture emerging from the molecular level might be a consequence of the evolutionary conflict that is enacted at the ecological scale.

Sexual Development and Sexual Conflict

Compared to the numerous investigations into sex determination, sexual development remains a relatively neglected area.

It is well known that different sexes of the same species can exhibit a staggering amount of differences in phenotype and ecology. It seems reasonable to assume that within a species' genome the same genes will have different expression optima in a female or male context. In spite of that, animals of both sexes within a species share an almost identical genome. Genetic differences are limited to non-recombining regions of the genome in species with genetic sex determination, ploidy level in species with haplodiploid sex determination or none at all in species with environmental sex determination. Genes will thus tend towards the same expression pattern in both sexes. Numerous taxa indeed show a high correlation between male and female gene expression levels(Poissant et al., 2010).

Intersexual genomic constraint generates an evolutionary conflict over the optimal expression pattern of genes between sexes. Mutations that affect the expression of genes with sexually conflicting optima will increase fitness in males while decreasing the fitness of females. Sexual genetic conflict has been verified for a wide variety of species, from mammals to birds and insects (Bonduriansky and Chenoweth, 2009). In a study from Chippindale et al. (2001), *Drosophila* genotypes selected for male or female reproductive success resulted in a decrease of reproductive success in the other sex. Interestingly, larval fitness remained positively correlated in the two sexes. A similar pattern of unmasking of

differential fitness during sexual maturation is also underscored by introgression dynamics in an hybrid population of *Formica* ants (Kulmuni and Pamilo, 2014). In this population, introgressed alleles are present only in the female (heterozygous) background while they cause complete mortality in (hemyzigous) males during the larva to adult transition. Females with introgressed alleles by contrast show higher survival rate when compared to their non-introgressed siblings and are responsible for maintaining the otherwise deleterious introgressed genetic variants. Conflicts in gene expression patterns seem thus to reflect the different ecological and physiological needs that arise only after sexual maturation, emphasized in the case of holometabolous insects due to the abrupt remodeling that happens during pupation.

The tendency of genes to homogenize expression patterns between sexes is also shown by a study from Hollis et al. (2014) in which males of *Drosophila* were released from sexual-selection pressures. After 65 generations male biased genes showed a marked decrease in expression in both males and females, while male testes showed a significant decrease of expression of male-specific genes. As suggested by the previous case, when sufficient selective pressures are present intersexual genomic constraint can be solved through the evolution of sex-specific expression patterns. To date, a great deal of studies on genetic sexual conflict have been focused on the role of sexual chromosomes (Ellegren and Parsch, 2007; Parsch and Ellegren, 2013), which provide a suitable location for genes whose expression is deleterious to the homogametic sex. Although sex chromosomes have been shown to be enriched in sex-biased genes (Innocenti and Morrow, 2010) a considerable proportion of sex-biased genes are found on autosomes. Alternative solutions include the duplication and sex-specific specialization of genes (Gallach and Betrán, 2011; Baker et al., 2012; Wyman et al., 2012), or differential epigenetic silencing in the paternal or maternal genome (genomic imprinting).

Nasonia lacks a sex-specific portion of the genome. This implies that all genetic changes present in one sex will be reflected by the other, and exacerbates the genetic conflict over genes with different optima in different sexes. The co-occurrence of haplodiploidy

and absence of sexual chromosomes makes *Nasonia* an interesting model organism for the study of genomic sexual conflict. Haplodiploidy implies that the whole genome of males is derived from their mothers and will be selected in an hemizigous background. By converse only females inherit a copy of the paternal genome and are subject to the conventional dominant-regressive allele dynamics. Finally, being an holometabolous insect, *Nasonia* possesses a larval stage with a vastly decreased amount of sexual dimorphism and an ecological niche fundamentally distinct from both adult forms. What the proportion, identity and function of sex-biased genes is in this stage poses an interesting evolutionary question.

Thesis Outline

In the previous sections I have outlined how my project fits within the overarching developments in Systems Biology and Evolutionary Theory by using sexual development and transcriptomic sexual conflict as a specific case for the evolution of alternative phenotypes and multi-layered solutions of regulatory constraints. I have also clarified the reasons behind our choice of the wasp *Nasonia vitripennis* as a model system and the most relevant traits of its life-cycle.

While *Nasonia* has ecologically relevant plastic traits and provides the tools to facilitate their investigation at multiple levels, I must also underscore how little is yet known about its development. At the time of writing this thesis, searching Web of Knowledge for articles that include *Nasonia* in their topic include 878 entries of which only 41 belong to the developmental biology category. This figure is even more generous than the reality if we consider that a sizable portion of those articles deals with either its sex determination or the effects of *Nasonia* poison on its host species' development. Basic research on the unperturbed development of *Nasonia* is required in order to generate an empirically supported null-model for the role of gene expression in development, which are in turn necessary to draw testable hypotheses on the behaviour of genes under perturbed states.

The first chapter of this thesis deals explicitly with the improvement of the *Nasonia vit-*

ripennis gene set, a necessary step towards a more complete understanding of this species' gene regulation. Within this chapter I describe the improved Official Gene Set (OGS2.0), which raises the number of gene models from 18,850 to 24,388, includes non-coding genic sequences and improves intron-exon definitions. I take advantage of this more comprehensive characterization of the *Nasonia* gene set to detect gene families with lineage-specific increases in gene copy number, devise a method for the identification of genes with lineage-specific sequence conservation or innovation that does not rely on accurate reconstruction of phylogenetic tree, and characterize the traits associated with alternatively spliced genes, explicitly addressing the evidence for different models of evolution in alternative splicing.

The second chapter describes in detail the FESTA algorithm, which I developed for the analysis of non-transcriptional gene regulation processes. FESTA provides an intuitive recursive process for splicing detection and quantification based only on exon annotation and gene expression data. This method also disentangles transcription and splicing, enabling a comparative analysis of both components of gene regulation as statistically independent processes.

The third and final chapter builds upon the previous two by using the genome annotation and the gene expression analysis tool to characterize how gene expression regulation enables sexual dimorphism in the development of *Nasonia vitripennis*. This chapter includes an in-depth overview of regulation from the sub-gene level (splicing) to higher-order sex-specific coregulation, unveiling cryptic sex-bias in the early development and providing a first characterization of the network evolution of sex-biased transcriptional clusters.

I also include one additional paper whose publication I contributed to in the appendix. This paper defines DNA methylation in *Nasonia vitripennis* from a structural and functional perspective. It provides evidence that wasp DNA methylation is primarily intergenic and localized at the 5' portion of constitutively expressed genes. I contributed to this paper by adding evolutionary comparisons between gene pairs with recent putative gains of methylation against their unmethylated paralogs, which shows an overall increase in expression and decrease in expression variance and sequence evolution. These findings

justify the use of adult methylation as a coarse evaluation of genes that can be methylated both in the charachterization of the gene set (section §1) and of developmental dynamics (section §3).

1. OGS2: GENOME RE-ANNOTATION OF THE JEWEL WASP *NASONIA VITRIPENNIS*

1.1. Abstract

Background: *Nasonia vitripennis* is an emerging insect model system with haplodiploid genetics. It holds a key position within the insect phylogeny for comparative, evolutionary and behavioral genetic studies. The draft genomes for *Nasonia vitripennis* and two sibling species were published in 2010, yet a considerable amount of transcriptiome data have since been produced thereby enabling improvements to the original (OGS1.2) annotated gene set. I carry out comparative analyses showcasing the usefulness of the revised annotated gene set.

Results: The revised annotation (OGS2) now consists of 24,388 genes with supporting evidence, compared to 18,850 for OGS1.2. Improvements include the nearly complete annotation of untranslated regions (UTR) for 97% of the genes compared to 28% of genes for OGS1.2. The fraction of RNA-Seq validated introns also grow from 85% to 98% in this latest gene set. The EST and RNA-Seq expression data provide support for several non-protein coding loci and 7712 alternative transcripts for 4146 genes.

Nasonia now has among the most complete insect gene set; only 27 conserved single copy orthologs in arthropods are missing from OGS2. Its genome also contains 2.1-fold more duplicated genes and 1.4-fold more single copy genes than the *Drosophila melanogaster*

This chapter has been published as part of Rago et al. (2016). While I include here only the portion of the project I have directly worked on, I also include the authors' contributions as stated on the paper to facilitate the evaluation of my independent contribution.
I performed the statistical analyses on the gene set and wrote the manuscript.
DG conceived, designed and developed gene construction methods, and provided public web access genome database of *Nasonia*.
JHC modeled, evaluated and annotated gene constructions, and performed summary analyses.
TS provided the sequencing data and assisted in drafting the manuscript.
YK provided the comparisons between OGS2 and NCBI Annotation Release 101.
JHW and JKC conceived the study, provided scientific guidance and participated in the writing of the manuscript.
I am also grateful to the associate editor and two referees, who have critically evaluated this work during the peer review process.

genome. The *Nasonia* gene count is larger than those of other sequenced hymenopteran species, owing both to improvements in the genome annotation and to unique genes in the wasp lineage.

I identify 1008 genes and 171 gene families that deviate significantly from other hymenopterans in their rates of protein evolution and duplication history, respectively. I also provide an analysis of alternative splicing that reveals that genes with no annotated isoforms are characterized by shorter transcripts, fewer introns, faster protein evolution and higher probabilities of duplication than genes having alternative transcripts.

Conclusions: Genome-wide expression data greatly improves the annotation of the *Nasonia vitripennis* genome, by increasing the gene count, reducing the number of missing genes and providing more comprehensive data on splicing and gene structure. The improved gene set identifies lineage-specific genomic features tied to *Nasonia*'s biology, as well as numerous novel genes.

1.2. Background

The jewel wasp *Nasonia vitripennis* belongs to the superfamily Chalcidoidea, which is a vast group of hymenopterans that consists mostly of parasitoids that deposit their eggs in or on other arthropods. Parasitoids play an important role at controlling insect populations and are used extensively as an alternative to pesticides (Quicke and Others, 1997). *Nasonia* is the genetic model system for parasitoids and a model for evolutionary and developmental genetic studies (Werren and Loehlin, 2009; Lynch, 2015). As an hymenopteran, it provides a study system with naturally occurring haploid stages (males) and is a non-social relative to the ant and bee lineages, having diverged from them approximately 170-180 MYA (Werren et al., 2010; Misof et al., 2014). The *Nasonia* genus includes at least four species (Raychoudhury et al., 2010) that are partially to completely reproductively isolated by the bacterial parasite *Wolbachia*, yet can be crossed after its removal (Breeuwer and Werren, 1990; Bordenstein et al., 2003), allowing the study of speciation from both a genetic (Werren et al., 2015; Gibson et al., 2013; Niehuis et al., 2013; Loehlin and Werren, 2012) and non-genetic (Brucker and Bordenstein, 2013) perspective. The draft genome assembly of *Nasonia vitripennis* was published in 2010 (Werren et al., 2010). At that time, it provided a first comparative study of hymenopteran genomes with reference to the honeybee, *Apis mellifera*. The *Nasonia vitripennis* genome project also included genome sequences for the cross-fertile species *Nasonia giraulti* and *Nasonia longicornis*, which were aligned to the *Nasonia vitripennis* reference genome assembly. Utilizing information from these genomes, advancements have been made in areas as diverse as behavioural ecology (Pannebakker et al., 2013), speciation (Gibson et al., 2013; Niehuis et al., 2013), immune responses (Sackton et al., 2013) and DNA methylation (Wang et al., 2013).

In the coming years, projects such as the i5K and 1KITE (Misof et al., 2014) will continue to deliver new insect genomes and transcriptomes to the research community, with the goal of improving genomic knowledge for this most speciose animal clade (Barribeau and Gerardo, 2012). Expanding the taxonomic breadth and number of well annotated genomes is important to develop new research avenues, and several quality measures

are necessary for the accurate interpretation of comparative genomic, transcriptomic and epigenomic data (Waterhouse, 2015). Completeness (the number of reported genes compared to the actual number of genes in the organisms' gene set) is one such measure; an incomplete gene set may exclude the true causal genes responsible for trait variation in quantitative genetic analyses and confound the interpretation of genome-wide association studies. The accuracy and reliability of gene models are equally important for genetic and genomic studies. Erroneous models can arise either from the fragmentation of true genes or by falsely joining neighboring genes (also termed fused or chimeric models, not to be confounded with their biological counterparts) because of mismatched splice sites, missing exons, or the addition of spurious exons. False models are especially problematic for the functional study of genes by misrepresenting their true expression levels. Finally, an accurate annotation of untranslated regions is required to investigate post-transcriptional regulation. Untranslated regions (UTRs) consist of 5' and 3' terminal portions of the mRNAs, as well as introns that are removed from the final mRNA via splicing. UTRs are functionally relevant since they are often targets for regulatory mechanisms such as microRNAs mediated regulation (Pauli et al., 2011; Carthew and Sontheimer, 2009), ribosomal binding affinity (Xue et al., 2014) and transcript localization (Olesnicky and Desplan, 2007).

The quality of genome annotations is improved by using more sequence data of gene transcripts. These data often expand the initially reported gene repertoires, indicating that (except for a few model species) current gene inventories are still far from completion. The gene numbers and accuracy of annotations for model species have generally increased over decades of work (e.g. 10% more genes and 200% more alternates for Arabidopsis over 15 years, Sterck et al., 2007). Species specific, targeted strategies are employed to refine the annotated gene sets. For example, by applying specific targeted solutions to the technical challenges of annotating the *Apis mellifera* genome (largely because of its unusual base composition), its initial count of ca 10,000 genes (Weinstock et al., 2006) increased to a more acceptable gene count of 15,314 (Elsik et al., 2014). Improving a

gene set's quality however does not necessarily require targeted strategies. Integrating multiple gene-model construction algorithms and incorporating novel expression data can often provide sufficient evidence to improve existing models while also uncovering new loci and their variants. This is especially true if the source data are tissue-specific or include novel environmental conditions and developmental stages, which are likely to reveal the expression of specialized genes or transcripts (Brown et al., 2014; Gerstein et al., 2014). For example, the *Anolis carolinensis* gene set was updated in 2013 by adding tissue and embryonic specific RNA-Seq datasets, which provided sufficient new data to increase the overall gene count from 17,792 to 22,962 genes and from 18,939 to 59,373 transcripts – an increase of 29% and 210% respectively (Eckalbar et al., 2013)! These case studies indicate that we are still far from reaching the point of diminishing returns on investments at improving the annotation of eukaryote genomes. As such, the genomics community is aware that updates to integrate novel expression and sequence data must remain a priority in order to provide a more accurate representation of the real biological background of animals.

I report on a more comprehensive Official Gene Set for *Nasonia vitripennis* (OGS2), which vastly improves our understanding of its genome biology. Since its public release in 2012 (Gilbert et al., 2012), OGS2 has been used in a number of studies (Niehuis et al., 2013; Pannebakker et al., 2013; Sackton et al., 2013; Wang et al., 2013, 2015) and as a resource for comparative genomics (e.g., through databases such as OrthoDB Waterhouse et al., 2013; Kriventseva et al., 2015). Several information resource projects support the use of *Nasonia* for genomics investigations, reviewed by Lynch (Lynch, 2015). Gene set improvements of OGS2 are available at the Hymenoptera Genome Database (HGD) (Munoz-Torres et al., 2011) and more recently at WaspAtlas (Davies and Tauber, 2015). The HGD provides genome map views and BLAST sequence searches for *Nasonia*, including this OGS2 gene set, and 8 other Hymenoptera species. WaspAtlas offers gene annotation and functional information searches of *Nasonia* gene sets including OGS2, integrating expression and DNA methylation annotations. This OGS2 gene set along with

associated gene evidence and alternate gene sets are also available with genome map views and BLAST sequence homology searches through the EvidentialGene project of euGenes genome database (Gilbert, 2002). NCBI provides genome map views, sequence and gene annotation searches (Thibaud-Nissen et al., 2013) for their annotations of *Nasonia*.

Here I describe *Nasonia vitripennis* OGS2 in detail and compare it to the earlier annotation set using several quality measures. I use OGS2 for a comparative analysis of gene family expansion and sequence evolution with reference to other hymenopteran genomes. Finally, I reveal the usefulness of the novel gene set by presenting a multi-factorial analysis of the features that characterize alternatively spliced genes, demonstrating that genes with annotated isoforms are characterized by longer transcripts, greater number of introns, slower rate of protein evolution and lower probabilities of duplication when compared to genes with no alternate transcripts.

List of abbreviations

OGS2: *Nasonia* Official Gene Set 2

OGS1.2: *Nasonia* Official Gene Set 1.2

EST: Expressed Sequence Tags

RNA-Seq: RNA-sequencing

UTR: UnTranslated Region

mRNA: messenger RNA

lncRNA: long noncoding RNA

OG: Orthologous Group

BUSCO: Benchmarking Universal Single Copy Orthologs

CDS: Protein CoDing Sequences

Nvit_1.0: *Nasonia vitripennis* genome assembly 1.0

Nvit_2.1: *Nasonia vitripennis* genome assembly 2.1

NCBI-101: NCBI *Nasonia vitripennis* annotation release 101

GLMM: Generalized Linear Mixed Model

1.3. Methods

1.3.1. Gene Set Construction Process

As the gene set construction process was developed and carried out entirely by DG, it is not included in this thesis. The complete methods for the generation of the OGS2 gene set are presented in full in Rago et al. (2016).

All selected gene models are supported by some kind of evidence; *ab-initio* predictions without gene evidence are not included in OGS2. A small set of problem genes were manually curated and corrected by expert examination of evidence. A final set of 36,327 distinct loci, selected by EvidentialGene methods was compared to other available and draft *Nasonia* gene sets (table 1 and table 2). The predicted models include UTRs based on expression data and genome gene signals. Putative long non-coding genes (lncRNA) from the transcript assemblies (those with weak coding potential and no homology to reference proteins) were retained in the full gene set. The models and EST evidence were assessed with PASA for valid alternate transcripts. Gene proteins were annotated with Uniprot descriptions, and classified by evidence scores, including transposable elements.

Finally, 24,388 constructions were chosen to be “good models” (table 1), having the best match to EST and protein homology evidence. Models excluded from the "good" set include: (1) those with expressed RNA assemblies but with weak or no coding potential, (2) most of those with significant homology to known transposon proteins, and (3) those with minor or no expression and protein evidence from the quality assessment. However, 385 genes having homology to putative transposon proteins but also with expression and homology to other insect species genes were retained as an indeterminate subset annotated as "expressTE". I used the “good models” set for all downstream analyses, but note instances where the remainders include some genes of biological value.

1.3.2. Gene set Quality and Completeness metrics

The quality scores per model are calculated using the following types of evidence: (a) the level of RNA sequence coverage and tiling array signal over the gene model coordinates

Summary Statistics	OGS2 All Models	OGS2 Good Models	OGS1.2 Final Models
Genes	36,327	24,388	18,850
Protein coding genes	25,725 (71%)	24,388	15,566*
Non-coding genes	3,997 (11%)	0	0
Transposon protein genes	6,605 (20%)	385*	2,935*
Single transcript genes	32,079 (88%)	20,243 (83%)	18,759 (99.5%)
Genes assigned to ortholog**	15,176 (42%)	15,173 (62%)	–
Transcripts	44,164	32,101	18,941
Alternative transcripts	7837	7712	91
Mean isoforms per gene	1.22	1.32	1
Complete proteins	41,256 (93%)	30,521 (95%)	18,941 (100%)
Median transcript length	1571 bp	1603 bp	1176 bp
Median CDS length	777 bp	981 bp	1032 bp
Transcripts with UTR	41,313 (94%)	30,512 (95%)	5264 (28%)

Table 1: **Summary of the Official Gene Set (OGS2)** comparing all gene constructions to good constructions having expression and/or homology evidence and to the previous OGS1.2 gene models. Percentages are of the total number of genes for the set.

* 2,935 OGS1.2 models are classified with strong homology to transposon proteins during OGS2 work, 385 models with expression and other insect homology but also transposon homology were retained in OGS2 "good" model set

** 5,763 additional genes of OGS2 have significant protein homology, but are not assigned as orthologs in OrthoMCL orthology analysis, 3,454 of 24,388 "good" models lack significant homology, but have expression evidence.

on the genome assembly; (b) the number of EST and RNA sequence reads spanning the intron splice sites that matched to annotated exon ends; (c) gene structure agreement, as end-to-end match of exons in the model with the evidence in support of gene structure, summarized in table 2 for evidence structure from EST/RNA assemblies and reference proteins; (d) sequence homology to proteins from eleven species-specific reference databases using BLASTp scores of all significant matches to the reference set of genes including the number of reference protein matches, bitscore per protein match, and the similarity scores for alignments to same species paralog proteins. These quality scores are summarized for several *Nasonia* gene sets (table 2) and partitioned according to the source of evidence (EST, RNA sequences, tiled expression spans, reference sequences (*Nasonia* RefSeq), and reference species proteins. Each gene model for each locus is therefore scored by weighted evidence. Finally, the maximal evidence scored, non-overlapping model set is determined, with respect to inter-locus effects of gene joins and other factors.

Quality scores per orthologous group (10 on page 89) are calculated in the following way: for each orthology group, the median protein size of all genes among the species within the group is determined. Then for each species gene set, the maximal BLASTp bit score of a gene within that group is recorded as metric #1, and the protein size difference from the group median of that maximal match is recorded as metric #2. These metrics are averaged for all groups per species, and reported as average bit score, as average size deviation, and as percentage of size outliers (2 standard deviations below median sizes). These gene set quality measurements are provided by the Evigene scripts: "eval_orthogroup_genesets.pl" and "orthomcl_tabulate.pl". Partial gene models are a common artefact of draft gene sets, indicated by both a negative deviation from group median sizes, and larger percentage of outliers. A similar calculation is part of the OrthoDB methodology (Simão et al., 2015).

1.3.3. Ortholog group assignments and gene family expansions

Orthology of *Nasonia* protein coding genes was assigned using two methods: OrthoMCL (Li, 2003) and OrthoDB (Waterhouse et al., 2013). OrthoMCL was used during gene

Evidence	Available Evidence	Statistic	OGS1.2	Evidence prediction Set	OGS2	OGS2 Good Genes	NCBI RefSeq	RNA-Seq Assembly
EST	18 Mb	Seq. Overlap	0.506	0.814	0.768	0.715	0.672	0.724
Protein	26 Mb	Seq. Overlap	0.674	0.696	0.729	0.693	0.616	0.612
RNA	46 Mb	Seq. Overlap	0.381	0.551	0.599	0.540	0.468	0.571
RefSeq	17 Mb	Seq. Overlap	1.000	0.934	0.958	0.908	0.857	0.839
Intron	66,593	Splices Hit	0.846	0.965	0.981	0.969	0.903	0.975
TAR	75 Mb	Seq. Overlap	0.292	0.850	0.533	0.443	0.370	0.386
Transposon	28 Mb	Seq. Overlap	0.168	0.282	0.406	0.099	0.009	0.039
ESTgene	10,194	Perfect	2,737	3,996	4,952	4,900	3,631	4,293
ESTgene	10,194	Equal 66%	3,491	5,059	6,283	6,198	4,284	5,187
ESTgene	10,194	Some	6,263	9,940	11,313	11,157	7,123	8,373
Progene	44,040	Perfect	4,808	6,713	8,048	8,010	6,215	4,935
Progene	44,040	Equal 66%	7,759	12,217	14,046	13,837	9,003	8,567
Progene	44,040	Some	11,563	18,173	21,759	19,718	10,861	18,457
RNAgene	28,016	Perfect	6,004	9,531	14,899	13,804	8,502	28,016
RNAgene	28,016	Equal 66%	8,173	13,552	18,829	17,608	10,202	28,016
RNAgene	28,016	Some	11,933	19,602	24,936	22,179	12,258	28,016
Homolog	11,683	Matches	16,174	16,669	23,994	17,341	11,950	13,187
Homolog	11,683	Found	10,426	10,593	11,683	11,683	9,323	9,650
Homolog	11,683	Bits/Amino Acid	0.449	0.424	0.416	0.455	0.562	0.558
Paralog		Matches	12,843	14,503	19,423	12,576	7,904	10,520
Paralog		Bits/Amino Acid	0.459	0.450	0.564	0.517	0.554	0.635
Genome		Coding Seq.	28 Mb	31 Mb	36 Mb	29 Mb	10 Mb	16 Mb
Genome		Exon Seq.	29 Mb	52 Mb	70 Mb	45 Mb	24 Mb	24 Mb
Genome		Gene count	18,941	23,605	36,327	24,388	12,989	20,926

Table 2: **The types of evidence and levels of support for Nasonia *vitripennis* gene sets.**
Sequence-level statistics for the different types of evidence are given as proportions of the gene sets that are validated. Gene structure level statistics (ESTgene, Progene, RNAgene) are counts of the number of models that reach three structure level agreements. Homology level statistics are counts of the number of models and proportions matching proteins of reference species and paralogous (same species) proteins. See methods section for details on the evidence types and the statistics that were measured.

construction as an essential measure of gene quality, for refining gene model classifications. For OrthoMCL, related species proteomes with *Nasonia* gene models were aligned using all-by-all reciprocal best BLASTp (Altschul et al., 1990, 1997) of 11 species' proteomes (wasp plus those listed above). Alternate transcripts were removed after BLASTp matching, in order to use the most similar gene variants. Clustering of these blast alignments into gene families was also done using OrthoMCL. The resulting gene families are narrow or broad, depending on the chosen alignment options, especially the distance at which to break groups. Resulting groups are rather like the leaves at the tips of a phylogenetic tree. Further MCL clustering of these groups showed relations between many of the narrowly clustered groups. Significance criteria were applied using recommended options: a similarity p-value $<$ 1e-05, protein percent identity $>$ 40%, and MCL inflation of 1.5 (this affects the granularity of clustering). Reciprocal best similarity pairs between species, and reciprocal better similarity pairs within species (i.e., recently arisen paralogs, or in-paralogs, proteins that are more similar to each other within one species than to any protein in the other species) were added to a similarity matrix. The protein similarity matrix was normalized by species and subjected to Markov clustering (MCL; Enright et al., 2002; van Dongen, 2000) to generate ortholog groups including recent in-paralogs. An additional round of MCL clustering was applied to identify between-group relations.

After producing the *Nasonia* OGS2 genes, its protein sequences were incorporated into release-6 of the OrthoDB database (Waterhouse et al., 2013). Ortholog groups are here defined as groups of genes related by descent from a single common ancestor at the base of the taxonomic level of interest. All genes within a single ortholog group evolved from a series of speciation and/or gene duplication events from a unique ancestor. Their amino acid sequences can thus be aligned and compared with each other. Ortholog groups provide efficient units of analysis for genes over long timescales as they enable partitioning in evolutionarily relevant categories without the need to resolve precise 1 to 1 relationships. From the total 24,388 OGS2 genes, 15,173 (62%) could be assigned to an ortholog group among the Arthropoda in OrthoDB version 6.

I assessed which ortholog groups are characterized by evolutionary expansions in the *Nasonia* lineage. I selected 9601 ortholog groups that have paralogs in *Nasonia* and over 80% of the other sequenced Arthropoda. To further increase the stringency of the selection criteria, I removed all genes from this set that have any duplicates in other hymenopteran species. Of the total 9601 ortholog groups, 411 (0.05%) have duplicates specific to the *Nasonia* lineage among the Hymenoptera. I used sequence similarity searches to cross-validate the absence of ultra-conserved ortholog groups of the BUSCO dataset (OrthoDB) from the *Nasonia* genome. I retrieved protein sequences for all genes within those ortholog groups from all sequenced arthropods.

1.3.4. Identification of fast- and slow-diverging genes in *Nasonia* relative to ants and bees

I retrieved amino-acid alignments for ortholog groups among the Hymenoptera from OrthoDB version 6 and selected those that contained at least one gene in the *Nasonia* genome and at least one gene in one ant and one bee genome (8696 OGs). I generated a pairwise sequence divergence matrix, comparing all genes versus all genes within each of those ortholog groups by applying a JTT protein evolution model as implemented in the R package phangorn Schliep 2011. I then estimated the proportion of between-genus sequence divergence due to the *Nasonia* genes using the following ratio

$$\frac{AN + BN}{AN + BN + AB}$$

where AN and BN are the median pairwise amino-acid distances between the *Nasonia* gene and Ant or Bee orthologs respectively, and AB is the median pairwise distance between the ant and bee orthologs in the genes' ortholog group. I analyzed this ratio with a generalized linear mixed model (GLMM) with logit link function, using overall median sequence divergence of the ortholog group, presence of *Nasonia* paralogs and transposon-associated expression as predictors to account for the role of those factors in protein evolution. I also used the ortholog group ID as a random blocking factor to

account for individual differences in evolutionary rates between ortholog groups. I then extracted the GLMM's residuals to evaluate the remaining unexplained levels of sequence evolution. I selected genes that exceeded the 95th percentile of the distribution of residuals as highly diverging, and those below the 5th percentile as slowly diverging. I did not include relative non-synonymous to synonymous substitution rates in the GLMM because the analysis is based on protein sequence alignments scored by a weighted matrix of amino acid substitutions.

To avoid false positives due to exceedingly fast or slow protein sequence evolution in either the ant or bee clade, I also computed separately the rates of divergence between *Nasonia* and the ant or bee lineages ($\frac{AN}{AN+BN+AB}$ and $\frac{BN}{AN+BN+AB}$). I then generated two independent GLMMs for these ratios with the same factors used for the compound ratio and reported the genes that scored as significantly faster or slower (above 80th percentile or below 20th percentile) in both cases. This second set provides a high confidence list of genes that are differentially diverging in the *Nasonia* lineage but show limited differentiation between the ant and bee lineages. I point out that this is a tool to identify proteins that may be evolving more quickly at the amino acid level in the *Nasonia* clade. Because the analysis is unrooted, the method does not identify proteins that are specifically evolving more quickly since divergence of *Nasonia* from its common ancestor with ants and bees, but also includes changes from that common ancestor to the split between ants and bees. More precise evolutionary analyses will require phylogenetic reconstruction for all the genes, but the current set is useful for identifying likely candidates for divergence among these taxa. Given the very long branches involved in such analyses, use of dN/dS ratios as an index of adaptive evolution would be inappropriate due to total saturation of synonymous substitutions.

1.3.5. Functional enrichment testing

I tested all gene sets for functional enrichment of Gene Ontology (GO) terms obtained by Blast2GO (Conesa et al., 2005), using the two-tailed Fisher's exact test with a False

Discovery Rate (FDR) of 5% against the complete gene complement of *Nasonia vitripennis*. The *Nasonia* GO annotation for OGS2 was provided by the *Nasonia* community (Munoz-Torres et al., 2011). Of the 24,388 OGS2 genes with supporting evidence, 24,373 are present in the community-provided Blast2Go annotation files and 6446 of these (26,4%) have GO assignments.

1.3.6. Alternative splicing analysis

I used GLMMs to test for factors correlated with the presence or absence of alternative transcripts in OGS2. Our test factors include presence of strict sense paralogs (defined as reciprocal best sequence similarity match within the same genome versus reciprocal best match within other genomes), number of broad sense paralogs (genes within the same genome belonging to the same arthropod OrthoDB ortholog group plus one, log and z transformed), number of predicted introns (log and z transformed), transcript length (log and z transformed, using the longest transcript per gene), proportion of coding sequence over total transcript length (CDS/Transcript length, log transformed and normalized), ratio of *Nasonia*-specific protein evolution (see section 1.3.4, log and z transformed), methylation status in adult females (Wang et al., 2013) and phylostratigraphic age (Sackton et al., 2013).

I selected only genes with a complete record for all tested factors. Since the detection of isoforms is proportional to the coverage of that gene, I further restricted our analyses only to genes with both strong expression support and strong intron support, which have comparable levels of transcriptional data available. Therefore, our final dataset was comprised of 5447 genes. To estimate over-dispersion, I fitted a GLM with quasi-binomial error distribution including all analysis parameters. This model did not show over-dispersion, with a c-hat of 1. I therefore fitted subsequent models to a binomial distribution with logit link function. All subsequent models also included a random intercept error structure for each ortholog group among arthropods, to account for different selective pressure on different gene families.

I estimated the support of individual factors by fitting a full model incorporating all parameters, then compared this model to others incorporating all factor combinations by applying the Akaike Information Criterion, corrected for finite sample size (AICc). I calculated the relative importance of factors as the sum of weights of all models containing that factor over the total weight of all models within the set. Since the final model set contained several models with similar AICc values (additional file 1, see attached disk), I choose to present the results as model-averaged estimates rather than to choose a single best model.

1.3.7. Additional software tools

Most statistical analyses were performed in R version 3.0.0 (R Core Team, 2013) using the following packages: plyr (Wickham, 2011) and reshape2 (Wickham, 2007) for data handling, phangorn for sequence analyses (Schliep, 2011), lme4 (Bates et al., 2013) for GLMMs, MuMIn (Barton, 2011) for multi-model comparisons and model-averaging, vcd (Meyer et al., 2014) and ggplot2 (Wickham, 2009) for plotting.

1.4. Results and Discussion

1.4.1. Transcriptional and homology data complement each other

I compared the relative contribution of both expression and homology to the construction of gene models in OGS2. Expression data supports 17,925 genes (74% of OGS2) at strong or medium ($>\frac{2}{3}$ and $>\frac{1}{3}$ expression overlap, respectively) levels of evidence. Strong or medium homology support ($>\frac{1}{3}$ sequence overlap) is present for 17,238 genes (71% of OGS2). The intersection of strong and medium support from both lines of evidence contains 12,912 genes (53% of OGS2, figure 2), suggesting a high degree of convergence (p-value = 2E-14, Fisher's exact test).

While still significant (p-value = 1E-8, Fisher's exact test, N=13,861), the level of convergence between expression and orthology support decreases to 44% for the subset of duplicated genes, likely due to a reduced relative support of expression data (figure 2).

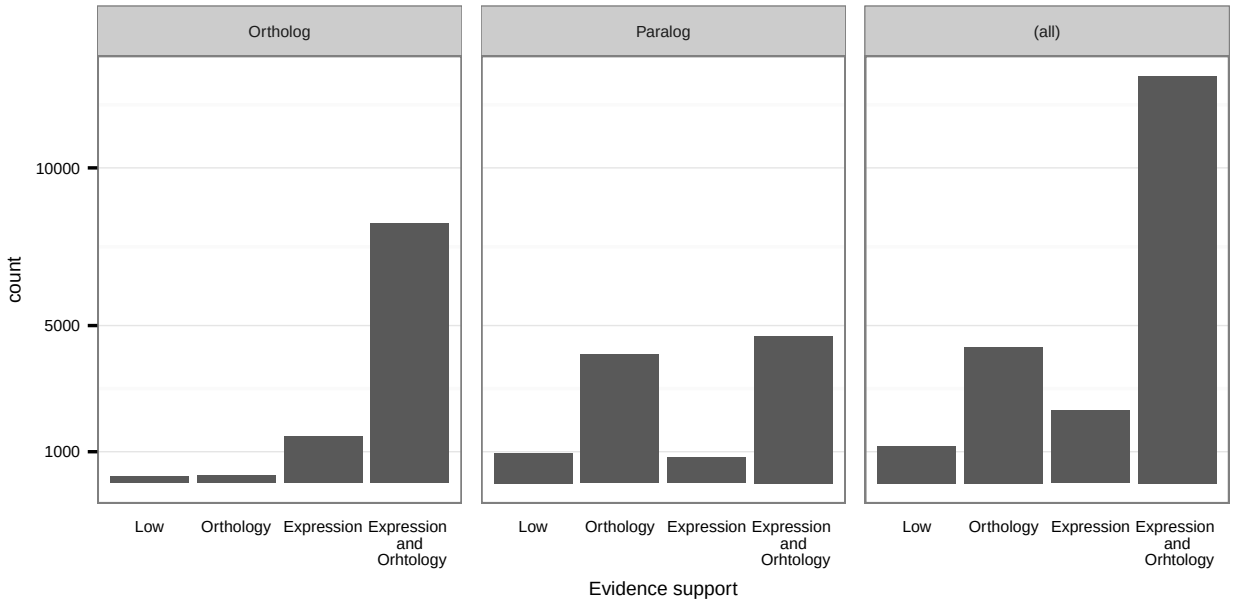


Figure 2: **Number of genes with medium or higher support from sequence orthology, evidence of transcription, or both.**
Medium support is defined as overlap greater than 30%. Panels show the source of evidence for genes within the ortholog and paralog subsets and the whole OGS2

The decrease in expression support can be explained by a more restricted expression profile for paralogs, which often arises after gene duplication events (Van de Peer et al., 2009). Therefore, further transcriptomic data from different tissue types and conditions should increase the level of convergence between the orthology and expression sets. Conversely, genes without duplicates show greater convergence between orthology and expression support (81% of 24,388 genes, figure 2).

Most of the 24,388 OGS2 genes that map to the *Nasonia vitripennis* genome assembly also map to the genome assemblies of sibling species *Nasonia longicornis* and *Nasonia giraulti* (Werren et al., 2010) using GMAP (Wu and Watanabe, 2005); 664 do not map to *Nasonia longicornis*, and 735 do not map to *Nasonia giraulti* (391 are missing in both, yet 50 of these have non-wasp orthologs). All 4,141 high identity paralog loci from *Nasonia vitripennis* map to assemblies of both siblings, though some are overlapping loci (table 9). The majority of paralog mapping patterns are the same for all 3 species (i.e., their relative positions are shared for all three species): 83% (3442/4141) of the paralogs for

Gene Sets	Gene Families (GF)			Gene Counts					Proportions	
	GF	Ortholog GF	GF Missing Genes	Genes	Species Specific Genes	Species Specific Paralogs	Single Ortholog Genes	Duplicated Ortholog Genes	Dupl	Singl
OGS2	10,293	8,983	92	24,296	5,446	6,686	8,239	3,925	2.1	1.4
Apis	8,591	8,560	170	10,145	987	88	8,182	888	0.2	0.9
Harpegnathos	9,633	9,291	107	15,029	2,943	1,567	8,710	1,809	0.7	1.2
Tribolium	8,893	8,388	116	16,985	4,586	2,163	7,608	2,628	1.0	1.2
Drosophila	8,464	7,636	187	14,289	2,824	2,556	6,994	1,915	0.9	1.0

Table 3: **Number of insect genes classified to gene families (GF) that are common among the arthropods by OrthoMCL** (ARP9, version arp11u11). Five out of nine insect species are summarized. Dupl and Singl designate the proportion duplicated and singleton genes relative to the median found among insects (Dupl:5000, Singl:10000).

all species, 99% (4098/4141) of the paralogs for 2 or 3 species. The differences include both real biological differences and assembly errors. Of the 2481 paralogs on separate scaffolds of the *Nasonia vitripennis* genome, 328 overlap first paralog spans in other species, therefore may be missing or mis-assembled. Of 239 tandemly arrayed paralogs in *Nasonia vitripennis*, 128 are also tandem in other species, 101 are on separate scaffolds in other species, and 69 overlap first paralog spans in other species (ie. missing or mis-assembled).

I also report that 3558 genes (15% of OGS2) have no homology support and are therefore annotated only by means of expression data, and that 1818 genes (7.5% of OGS2) have no expression support and are therefore annotated only by means of orthology matching. Eight hundred and thirty-three (833) genes in OGS2 are expert-curated including 38 that span different scaffolds, odorant genes, and other cases that could not be annotated automatically. Finally, 374 transcripts have complete proteins from transcript assemblies that do not match genome sequence due to genome gaps and frame-shifts.

1.4.2. Missing gene families are absent from the *Nasonia* genome

I assessed the level of completeness of the OGS2 gene set using OrthoMCL to classify genes into orthologous gene families that are common to arthropods (table 3 and table 10). The comparison of genes among nine species indicates that OGS2 is equally or more complete than the other insect gene sets, having fewer missing gene families, and similar numbers of orthologous gene groups and single copy orthologs. Additionally, OGS2 reveals

that *Nasonia* has twice the number of duplicated genes than *Drosophila melanogaster* or *Tribolium castaneum*, both with homology (in-paralogs) and without (unique duplicates), plus a greater number of unique singletons. Measures of protein sizes and alignment score (table 10) indicate that OGS2 genes are larger on average than genes from other versions of the *Nasonia* annotated gene sets, yet near to the *Apis mellifera* ortholog gene sizes. The transcript assemblies contain 62 orthologous gene groups that are not included within OGS2 because these transcripts are only poorly positioned onto the *Nasonia* genome assembly. These may be included in a more complete gene set as transcript assemblies, but are not yet part of this genome-mapped OGS2 gene set. A total of 75 orthologous gene groups are missing in *Nasonia* but present in 9 other insect genomes.

I also used the OrthoDB method to independently assess completeness. I counted the number of missing conserved single-copy genes that are otherwise present among the sequenced Arthropoda (Benchmarking Sets of Universal Single-Copy Orthologs [BUSCO] in OrthoDB Release-6), as well as the multi-copy *Nasonia* genes that are otherwise classified as single copy in other Arthropoda. For the majority of gene families, there were no discrepancies between the results obtained from OrthoDB and OrthoMCL. Although the BUSCO results suggest that OGS2 lacks 67 of the 3377 (2%) conserved ortholog groups, further analyses found all but 27. Conserved families missing in *Nasonia* OGS2 according to OrthoDB can be attributed to (i) genome artifacts (10 missing genes were found split across assembly scaffolds, or lost in gaps but found in transcript assembly), (ii) gene model artifacts (9 loci were apparent join errors appended to a second gene protein), (iii) OrthoDB discrepancies at classifying proteins to families (25 loci were assigned to different gene families by OrthoMCL and by OrthoDB family). Twenty-seven conserved single copy genes are either truly missing or sufficiently diverged to avoid detection. This number is comparable to those in other Arthropoda, which lack a number of BUSCO genes ranging from 3 (*Drosophila erecta*) to 708 (*Strigamia maritima*), with a median of 42.

Experimental evidence supports the lineage-specific gene loss for the three BUSCO genes involved in developmental regulation: *short gastrulation* (*sog*, OG EOG6S4MX5), *spaetzle 3*

(OG EOG61C5BT) and *daughters against dpp* (*Dad* or *smad6*, OG EOG69CNQ7). Despite their ultra-conserved status across currently sequenced arthropods, detailed investigations of *Nasonia* development suggest that those genes are truly absent from its genome due to modifications in the BMP signaling pathway (Buchta et al., 2013) rather than because of omissions in the current annotation. Since genes in the BUSCO set are defined as single-copy in 90% of 30 arthropod species, I compared the number of duplicated BUSCO genes in OGS2 to estimate the fraction of potential false gene duplications. I counted 141 (4%) multiple-copy OGS2 of the total 3377 BUSCO single-copy gene families (additional file 2, see attached disk). Of those, 62 (44%) are reported as duplicates uniquely for *Nasonia*, 61 for *Nasonia* plus one additional species, and 18 for *Nasonia* plus two other species. Other species have similar rates of duplicated single-copy genes: 78 for *Apis mellifera* and *Harpegnathos saltator*, 96 for *Pogonomyrmex barbatus*, 119 for *Atta cephalotes* (all Hymenoptera), 107 for *Anopheles*, and 437 for *Aedes* mosquitos. *Nasonia* OGS2 is therefore well within the observed range of duplications of BUSCO genes.

To further assess whether the reported duplicates are likely to be false models, I removed the best supported gene from each orthologous group and measured the expression support of the remaining models. One hundred and fifty-three (153) out of 175 genes (87%) show medium or strong support for expression and only 2 have no expression support. Lineage-specific duplications are supported by the observation that the majority of genes belonging to ultra-conserved ortholog groups display moderate to strong expression, even after removing the most supported duplicate and map to different genomic locations (data not shown).

1.4.3. Gene model quality and diversity increase

OGS2 improves our knowledge of the *Nasonia* genome in several ways (table 1). First, the number of annotated genes climbs from 18,850 to 24,388 (an increase of 29%). This greater completeness of the *Nasonia* gene set is corroborated by the sharp decrease in Arthropod ortholog groups missing from the *Nasonia* genome. OGS1.2 lacked 609 ortholog

groups that are present in all other Arthropoda (OrthoDB Release-5). Only 331 conserved OGs are now missing from OGS2 when compared to the same subset of species (OrthoDB Release-6) and 253 when considering all currently available arthropod species.

The spans of coding exons are very similar between OGS2 and OGS1.2 for 10,583 loci, which have a median percent equivalence of 92% between both sets. Changes in coding sequences are mostly attributable to error correction such as splitting and merging of models: 1617 original gene models (10% of OGS1.2) have been split into separate genes in OGS2, while 3555 OGS2 genes (15% of OGS2) contain a portion of an OGS1.2 split gene, and 494 OGS2 genes result from the joining of two or more OGS1.2 fragment genes (30 from three or more). Moreover, the proportion of genes with UTR extensions is now near complete: 23,069 (95%) of OGS2 gene models have annotated UTRs compared to only 5,264 genes (28%) within OGS1.2. These gene models match 98% of 66,593 intron locations on the genome assembly, identified by multiple reads of expressed RNA (>3; table 2), compared to 85% within OGS1.2 and 90% within NCBI-11 RefSeq. Intron splice sites are strong indicators of genes, including species-specific genes. This measure therefore indicates a high level of gene set completeness, independent of protein homology. Finally, OGS2 dramatically increased the number of annotated transcripts from 91 alternate transcripts in 91 genes (0.5% of OGS1.2, Table S4 in Werren et al., 2010) to 7712 transcripts among 4146 genes (17% of OGS2). Therefore, OGS2 increases the completeness of the reported *Nasonia* gene repertoire and the quality of gene models as well as allowing a first overview of *Nasonia* transcriptional diversity.

The current release also increases the diversity of annotated wasp genes. Of all OGS2 gene models, 12,296 (50%) could not be assigned a putative function via orthology with other annotated genes. Four thousand, six hundred and fifty-six (4656) genes from this subset (38%) could be assigned to 2334 arthropod orthologous groups, 490 of which (21%) are present as multiple copy in *Nasonia*. The remaining 7640 genes with no known function are found exclusively in OGS2 and could not be assigned to orthologous groups shared with other arthropods (OrthoDB, release 6). This subset is likely to include both incorrect

models and innovations along the wasp lineage. Three thousand, nine hundred and eighty-three (3983) of those *Nasonia*-only genes (52%) are present as duplicates in OGS2, a proportion that is significantly greater than that reported for the whole genome (fisher's exact test, p-value $<$ 2.2E-16). Of the 7640 lineage-specific genes with no annotated function, 4498 (59%) have been newly annotated in OGS2.

1.4.4. *Nasonia* shows biologically relevant lineage-specific duplications

Our examination of the updated gene families of OGS2 identified 411 Arthropoda ortholog groups that have duplicated exclusively in the *Nasonia* lineage (4% of all ortholog groups within OGS2). These groups consist of 1230 genes, of which 599 loci (49%) have no assigned homolog (additional file 3, see attached disk). The most frequent category among annotated expanded genes within the "good models" set is that of transposon associated proteins (102 genes, 30 ortholog groups), followed by kinases/phospatases (38 genes, 16 ortholog groups) and odorant receptors (23 genes, 7 ortholog groups). The enzyme 5-hydroxyprostaglandin dehydrogenase (6 paralogs, 2 ortholog groups) also shows an evolutionarily interesting lineage-specific expansion. This protein is essential for male pheromone processing, and is a prime candidate for driving mate selection and speciation, based on positional cloning of genes involved in pheromone differences between *Nasonia* species (Niehuis et al., 2013).

1.4.5. Wasp lineage diversification is driven by transcriptional regulators

I calculated the sequence divergence of each *Nasonia* gene from its orthologs in both ants and bees. I then selected *Nasonia* genes that have a significantly higher or lower proportion of sequence divergence to ant and bee orthologs when compared to the rest of the *Nasonia* gene set (see section 1.3.4 for details). This method identified 504 genes (the most extreme 5% of the frequency distribution) for both the rapidly and the slowly evolving gene categories (figure 3 A and additional file 4, see attached disk).

I also adopted a more stringent approach by measuring the divergence scores of *Nasonia*

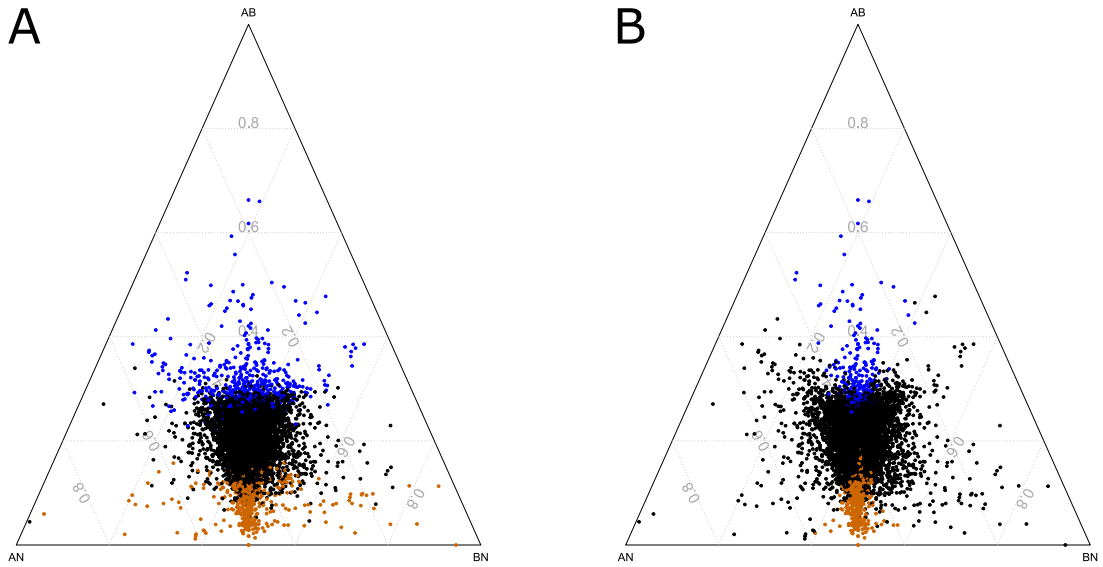


Figure 3: **Protein divergence of OGS2 genes against orthologs in other Hymenoptera**. Every point represents a gene mapped on three coordinates originating from the corners. Each gene's distance from a corner is proportional to the average amino-acid distance of orthologs between the two clades. AB = ant to bee distance; AN = ant to *Nasonia* distance; BN = bee to *Nasonia* distance. Diverging genes are highlighted in orange (fast) and blue (slow) as detected by the compound ratio (A) and intersection of ratios (B). See materials and methods for full description

genes against genes of the ant and bee lineages separately, then selecting only those genes that scored as rapidly or slowly diverging in both. This intersection method identified 596 and 394 genes that have differentially accelerated or slowed evolutionary rates in the *Nasonia* clade, respectively (figure 3 B and additional file 4, see attached disk). I note that both methods are unrooted, which therefore identify genes with greater divergence in *Nasonia* relative to bees and to ants, not to the common ancestor of these three lineages.

In all subsets, the most significantly enriched Gene Ontology terms are "nuclear location" for the cellular component category, "DNA/chromatin binding" for the molecular function category and "transcriptional regulation" for the biological process category. These data are consistent with the view that evolution of unique metazoan traits occurs more by changes in transcriptional regulators rather than in structural proteins (Knoll, 1999; Chen and Rajewsky, 2007).

1.4.6. Histone genes are enriched in lineage-specific evolution

Although histone genes are generally highly conserved, I identified several members of the histone complex with sequences that evolved relatively rapidly in the *Nasonia* lineage. Specifically, I observe a greater rate of sequence divergence for the histone proteins H2A when compared to ant and bee variants. Histone H2A proteins package DNA into chromatin and are implicated in epigenetically mediated gene expression regulation in vertebrates (Pauls et al., 2001; Hardy et al., 2009; Talbert and Henikoff, 2010). Regulatory variants of H2A histones are also present in the *Apis mellifera* genome (Lyko et al., 2010). There are currently twenty-four (24) H2A genes within OGS2, 22 of which are assigned to a single ortholog group (OG) (Arthropoda OG EOG6VT4F0) and 18 of which are assigned to a single Hymenoptera group (OG EOG65QGR3). Compared to other Hymenoptera, this ortholog group is more rapidly evolving in *Nasonia* and has a greater number of paralogs: four times greater than *Linepithema humile* (the 2nd highest number with only five copies). However, I cannot rule out that the number of H2A genes in other hymenopterans is underestimated, especially considering the comparable number of H2A genes that are found in other arthropods (e.g. 21 in *Daphnia pulex*, 22 in the *Culex quinquefasciatus*, 22 in *Drosophila melanogaster*). As of now, only two *Nasonia* H2A genes have strong homology with genes within Hymenoptera, while most others have higher scoring BLAST sequence similarity matches among vertebrate histones. This pattern can be explained by a lineage specific increase in protein sequence evolution, which would decrease the similarity between histones of *Nasonia* and of other Hymenoptera, and therefore increase their relative similarity to those of more distantly related species by a phenomenon called long-branch attraction. Thus, even though this result is most likely an artifact, it is still indicative of a faster evolutionary rate of *Nasonia* histones compared to those of other Hymenoptera.

Histone H3 is known to exhibit a wide range of modifications, many of which have known effects on the transcriptional status of the underlying genes (Gerstein et al., 2014; Müller et al., 2002). Several *Nasonia* H3 proteins (Hymenoptera OG EOG6R4ZDK)

appear to significantly evolve less rapidly when compared to ant and bee orthologs. I find that this apparently slower evolutionary rate of this orthologous group is due to a mis-identification of this OG, which is comprised of at least two different paralogs at the base of the Hymenoptera lineage (additional file 5, see attached disk). One of these putative sub-groups is retained in two copies across all Hymenoptera. The other sub-group is present in 2-4 copies in most Hymenoptera; yet *Nasonia* has 14 copies. The combination of an artefactual fusion of two OGs and unequal representation of *Nasonia* duplicates between the two groups is therefore the cause for an apparent slower relative evolutionary rate; the the correct interpretation consists of a lineage-specific expansion. *Nasonia* also retains an H3 gene of the OrthoDB group EOG62V6ZW, which is shared with other arthropods but not with other Hymenoptera, and and H3 gene of the OrthoDB group EOG6ZCRM6, which is seemingly lost in the bee lineage.

The *Nasonia* H2B histone proteins are encoded by 21 genes; only four are assigned to an ortholog group containing other Hymenoptera genes (EOG6Z8X7C of OrthoDB, whereas 8 are assigned to an OrthoMCL group). All genes are diverging at comparable rates while *Nasonia*'s copy number within this orthology group is similar to that of other hymenopterans (5 in *Pogonomyrmex barbatus* and *Atta cephalotes*). The remaining seventeen H2B histones could not be analyzed by our method, as they are not assigned to other Hymenoptera H2B histone gene families (OrthoDB, release 6). Although these genes may be mis-identified by the annotation pipelines the NCBI-101 gene set independently annotates 18 of these 21 loci as H2B histone proteins, suggesting that this annotation may indeed comprise a *Nasonia*-specific expanded histone gene cluster(s).

I found that families of histone modification enzymes have specifically expanded in the *Nasonia* genome: 4 of 38 histone-related gene families (10%) meet our criteria for lineage-specific expansion (see methods section). By comparison, expansions are found in only 0.013% of gene families for the rest of the genome. Our data therefore suggests that the *Nasonia* genome is enriched for histone modification enzymes due of lineage-specific gene expansions (table 8; p-value = 0.024, Fisher's Exact test). The finding suggests

that histone modification, rather than DNA methylation, may play an important role in the lineage-specific features of epigenetic modulation in *Nasonia*, consistent with findings that DNA methylation does not differ between the sexes in *Nasonia*, nor correlate with epigenetic changes in gene expression (Wang et al., 2015).

1.4.7. Alternative splicing and *lola* expansion

OGS2 includes alternate transcripts assembled from available expressed sequence using genome-mapped assembly and de-novo assembly methods. A total of 7712 alternate forms are identified for 4145 genes (17% of the total reported genes). One thousand, seven hundred and twenty-five (1725) genes (42%) have at least 3 isoforms, 219 genes (5%) have at least 6 isoforms and 26 genes have at least 10 isoforms. One gene (*longitudinals lacking* or *lola*) has a notable expansion of over 180 alternate forms, of which 89 are included in the OGS2 gene set. The remaining alternative transcripts are identified by read splice introns. Named for its observable wing phenotype in *Drosophila*, *lola* is also expressed in many tissues and developmental stages, and has a putative role in neuronal development (Giniger et al., 1994). *Lola* alternate transcripts all share a common 5' set of six exons, with one hub exon that branches to alternate 3' coding sequences of 500-900 bp, spanning 350 kb of the genome, with a new alternate each 1400 bases (median). *Apis mellifera* shares this *lola* alternate expansion, with 58 annotated alternates branching over 200 kb from the single hub exon, as shown in figure 4. In both species, additional alternates may be discovered with further expression evidence, as the regular spacing in *Nasonia* suggests up to 250 may fit into this region of the genome. Examination

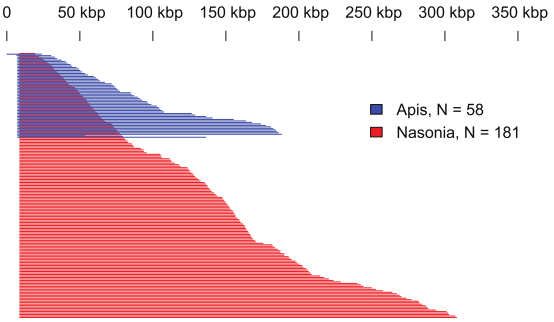


Figure 4: **Alternatively spliced introns for *lola* in Apis (blue) and Nasonia (red)** Graph shows intron spans from a common hub exon, in bases on their genomes. Blue and red bars at top of figure are short introns that join pairs of 3' end exons in *lola* gene span.

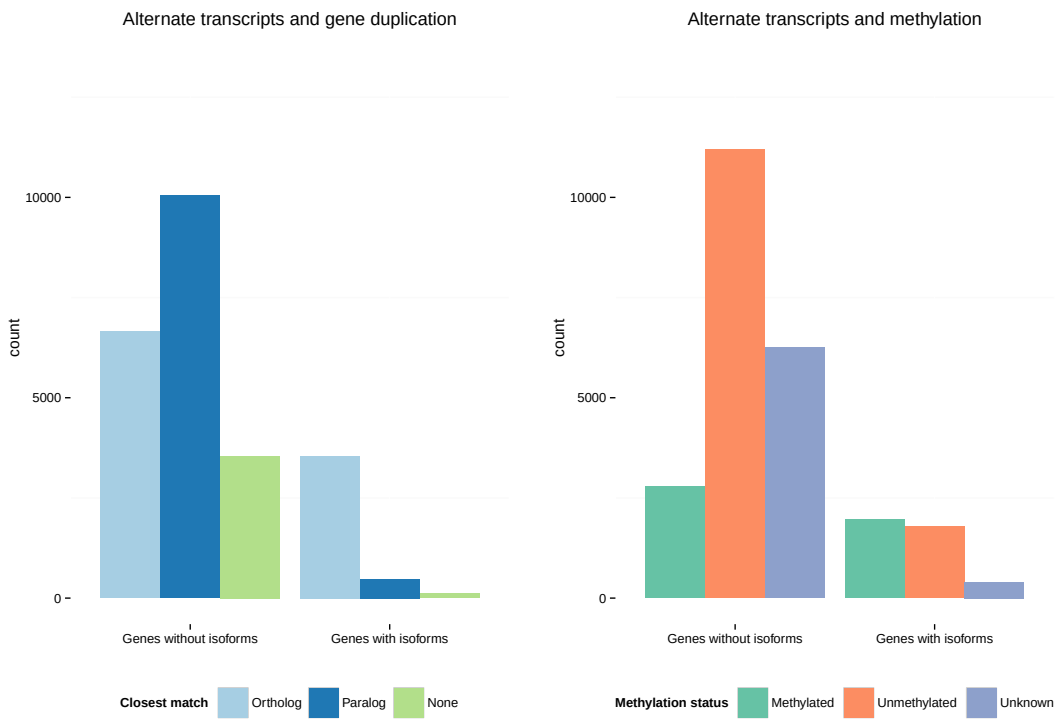


Figure 5: **Number of genes with alternative isoforms in OGS2** (A) split by presence of paralogs and (B) split by methylation in adult females.

of non-hymenopteran insects shows no similarly large expansion for *lola*.

The *Nasonia* gene with the second largest number of isoforms is the neuronal developmental transcription factor *fruitless*, with 17 alternative isoforms. *Fruitless* was already characterized as having an unique gene structure in *Nasonia* compared to Diptera, and its differential splicing is involved in both development and sexual differentiation (Bertossa et al., 2009). Two other *fruitless* paralogs are also reported within OGS2, while no other insect genome shows paralogs for this gene. Other genes with a high number of reported isoforms include mostly transcription factors and various kinases/phosphatases (additional file 6, see attached disk).

1.4.8. Which factors promote the evolution of alternative splicing in *Nasonia*?

The augmented number or genes with reported isoforms in OGS2 allowed an examination of factors that contribute to the evolution of this regulatory mechanism. From a total of 4146 genes with reported isoforms, only 476 (11% of all genes with isoforms, 2% of

OGS2) have annotated paralogs (figure 5 A). This proportion is significantly less (p-value <2.2xE-16, Fisher's Exact Test) than the product of proportions of genes with alternative transcripts and that of genes with duplicates (17% x 43% = 7.3%). In addition, genes without paralogs also have a greater number of introns than those with duplicate copies in the genome (Kruskal-Wallis rank sum test, p-value <2.2E-16 for both strict and broad sense paralogs). Possible interpretations of these patterns are considered in the discussion section below.

Methylation has been proposed as a molecular mechanism for the regulation of alternative splicing in humans (Shukla et al., 2011). In Hymenoptera, studies of both bees and ants consistently locate methylation target sites at the intron-exon junctions (Lyko et al., 2010; Bonasio et al., 2012; Flores et al., 2012). However, a study on the *Nasonia* methylome (Wang et al., 2013) reports alternative transcripts in non-methylated genes and no correlation between presence of alternate splicing and methylation status. I re-tested for the overrepresentation of alternative splicing with OGS2 sets of known methylated and known non-methylated genes (reported in Wang et al., 2013) (figure 5 B). Results indicate a significant overrepresentation of isoforms among methylated genes (p-value = 2.2e-16, Fisher's exact test), with alternative transcripts reported for 41% of methylated genes, while only 14% of non-methylated genes have transcript isoforms.

To exclude spurious results due to correlation with unaccounted variables, I fitted a generalized linear mixed model (GLMM) to estimate the probability of observing alternative transcripts in OGS2 genes according to a variety of factors (see methods section for details). The final statistical model (figure 6) is composed of the following co-factors: strict sense paralogy (presence of a reciprocal best match within the genome), number of broad-sense paralogs (OGS2 genes within the same arthropod ortholog group), ratio of *Nasonia*-specific protein evolution within Hymenoptera (see section 1.3.4), number of introns, methylation status in adult female and furthest matching ortholog. I also fitted a random error structure to account for individual differences between ortholog groups.

Expression level and intron support are also expected to be main predictors of observed

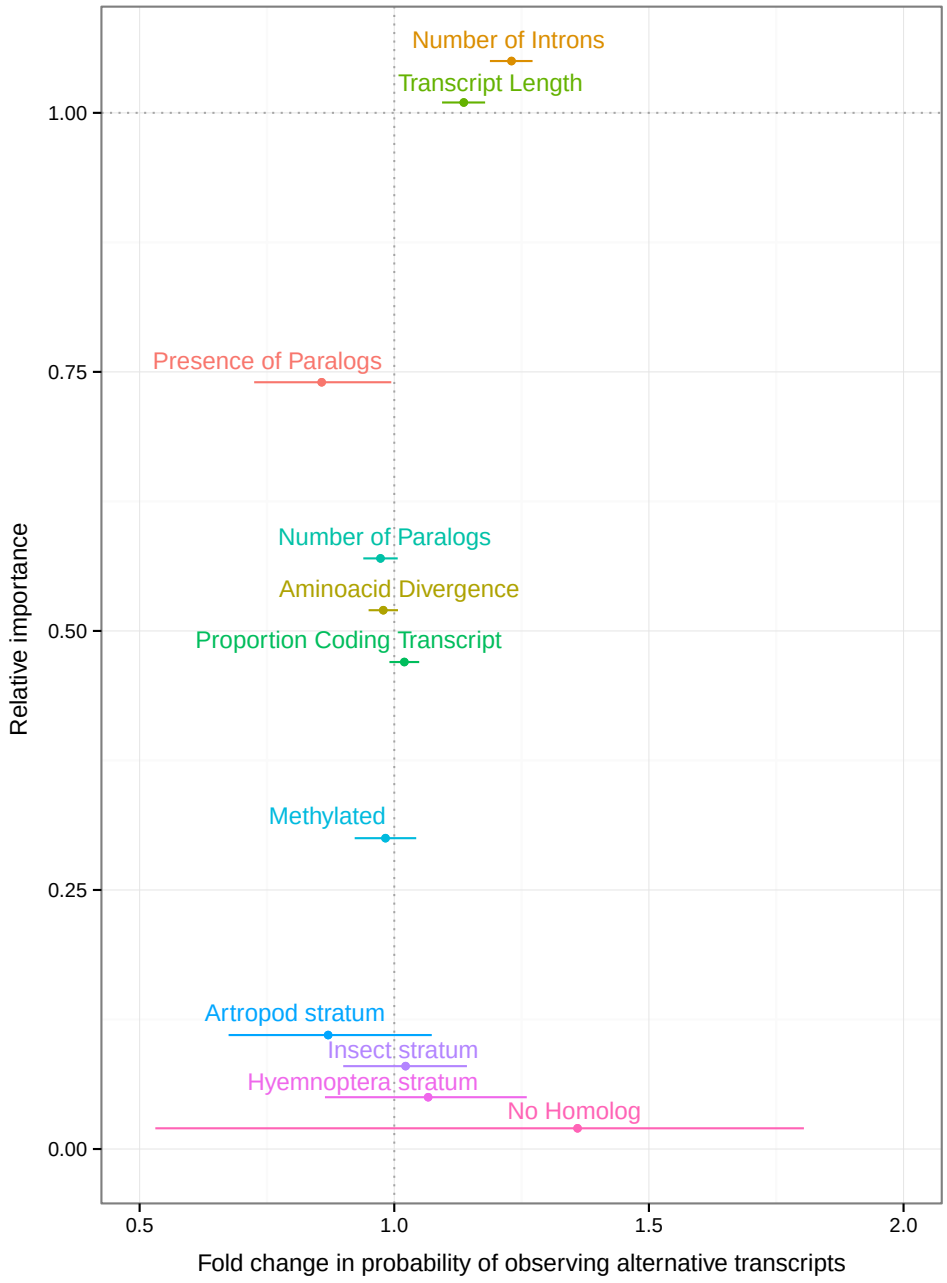


Figure 6: **Effect of different factors on the probability of observing alternate isoforms of OGS2 gene models.** Factors are ranked by relative importance (y axis). Factors with complete support and levels of the same factor were adjusted for plotting. Effect sizes are shown as the fold change in probability from the intercept (with 95% confidence intervals). Numeric variables were log transformed prior to analysis.

alternative isoforms, since isoforms of genes with greater transcript abundances will be easier to detect via RNA-Seq. I could not include expression and intron support as factors in our analyses due to their high correlation with methylation status (see figure 16). I therefore restricted our analyses to the subset of genes that have both strong expression and strong intron support (N=5447, figure 6).

Results indicate that the number of predicted introns and transcript length are positive predictors of alternative isoforms. Both findings are consistent with recent studies on the *Apis* transcriptome (Flores et al., 2012). The presence of introns enables the evolution of alternative splicing, since the latter requires differential inclusion of exons. The role of transcript length is more difficult to interpret. It is possible that genes with longer transcripts simply reflect better annotation quality. Alternatively, longer transcripts may allow for longer intronic sequences, which may facilitate the emergence of alternative splicing by providing a greater number of targets for the generation of novel splice sites or by switching from the intron signaling mechanism to the more error prone exon signaling mechanism (Roy et al., 2008). I explicitly included coding sequence to transcript length ratios among factors of interest to study these effects. I found that the proportion of coding transcript sequence (CDS/transcript length) is less well supported than transcript length itself (47% relative importance versus 100%). Furthermore, genes with higher proportions of non-coding sequence have a lower probability of displaying alternative transcripts. Even by assuming a role for intronic to exonic sequence length proportions, I find that shorter exons are prevalent among spliced genes, contrary to both the novel splice site and exon definition modes of new isoform generation. I should however note that the prevalence of long introns flanking alternative exons appears to be primarily driven by isoforms that comprise a minor proportion of all splice variants of a gene (Roy et al., 2008). It is therefore possible that the slight skew towards genes with low proportions of intronic sequences might be driven by issues in annotating low-abundance isoforms rather than by biological constraints.

Our initial genome-wide analyses detected a correlation between methylation and

alternative splicing. However, I observe alternative transcripts for non-methylated genes as well as methylated genes. This finding indicates that methylation is not necessary for alternative splicing in *Nasonia*. Furthermore, after focusing on the subset of genes with strong expression and intron support, methylation status in adult females is only weakly correlated with presence of isoforms (relative importance 30%).

I find low support for a negative correlation between *Nasonia*-specific sequence divergence and probability of observing alternative splicing. Methylated genes are known to have a slower rate of protein sequence evolution in *Nasonia* (Wang et al., 2013), while the presence of paralogs often increase protein evolutionary rates by releasing pleiotropic constraints on individual gene copies. Yet, rate of sequence evolution and lack of isoforms remained correlated, even after controlling for the effect of methylation and paralogy (relative importance 52%). This finding suggests that, despite the relatively low level of support, the inverse correlation between protein sequence evolution and alternative splicing may be direct result, rather than being derived from indirect correlations, and is consistent with studies of the *Apis* genome (Flores et al., 2012).

Both measures of paralogy (by reciprocal best hits or number of genes within the same arthropod ortholog group) retained a moderate level of support (74% and 57% respectively) when compared to other factors. Presence and number of paralogs are correlated with a lower probability of observing alternative transcripts. Since I performed all our analyses on the subset of genes with strong expression support, I can dismiss an effect due to the relatively lower expression support available for duplicated genes (see figure 2). The relatively large confidence intervals of the estimated effect of this factor on the probability of observing splicing of a given gene may either indicate a weak effect or result from the under-representation of paralogs in our subset (6% of the "good expression" gene set versus 43% of OGS2).

Finally, I tested whether isoforms are observed more or less frequently amongst genes which emerged at a specific taxonomic level by using furthest phylostratigraphic match as a proxy for gene age (Domazet-Lošo and Tautz, 2010). While average probabilities

decrease with gene age, this trend was not validated as statistically significant (data not shown). Furthermore, no single gene age category significantly alters the probability of observing alternative splicing in its assigned genes (relative importance: 0.07).

The inverse relationship between alternative splicing and gene duplication in particular is consistent with observations on the evolution of mammalian model species' genomes (Kopelman et al., 2005). There are currently several competing models that explain the negative correlation between gene family size and number of isoforms.

The "function sharing" model hypothesizes that duplication events reduce the selective pressure to maintain alternative transcripts in both gene copies (Roux and Robinson-Rechavi, 2011). This model is based on the assumption that both paralogs and isoforms provide equal opportunities for functional diversification. The reduced selective constraint would lead to the reciprocal loss of isoforms and subfunctionalization of the gene copies (Su, 2005). Such a scenario had been proposed for the *Dscam* genes in Arthropoda (Brites et al., 2013). The function-sharing model predicts that genes will gradually accumulate isoforms that are lost shortly after duplication events.

By contrast, Roux and Robinson-Rechavi (2011) proposed an "age-dependent" model, in which the inverse correlation between duplication and gain of isoforms is not direct but rather arises independently because of structural properties. Short gene length could be advantageous for whole gene duplication, while genes with an already high number of exons will have a higher propensity towards single exon duplication due to replication and recombination errors (Roux and Robinson-Rechavi, 2011). The lower numbers of isoforms for genes with duplicates would thus result from the different rates of accumulation of isoforms and duplicates rather than loss of redundant transcripts. This hypothesis has been criticized in depth (Su and Gu, 2012).

Finally, the underlying equivalence between the diversification potential of duplication and alternative splicing assumed by both the function-sharing and the age-dependent models is refuted by (Talavera et al., 2007). This finding suggests that a gene's probability of having isoforms rather than duplicates might be less dependent on its structural

properties and more dependent on the different adaptive potential of the novel proteins generated by two diversification modes, or functional constraint. Our analyses support longer transcripts and high numbers of exons as predictors of the presence of isoforms. While this is in agreement with the age-dependent model, I do not find a significant correlation between age of a gene family and the presence of isoforms. This could be either be caused by an actual lack of correlation, inaccurate dating (Moyers and Zhang, 2014) or by the fact that the divergence from the most recent outgroup (~180 MYA) is sufficiently great that every new family gains at least one detectable isoform.

Absence of duplicates has moderate support as a predictor of splicing, even after controlling for the structural properties of genes. Together with the lack of support for gene family age, this observation is congruent with the predictions of the function-sharing model. However, I must point out that a true test to falsify the function-sharing model would require testing the significance of the date from last duplication event, which I could not measure with our dataset. Comparisons between the sibling species *Nasonia giraulti* and *Nasonia longicornis* are especially suited to this task, as they provide a sufficiently short timescale to assess transcriptome changes lead by duplication when compared to more basal Hymenoptera.

Since I lack estimates on the potential functional overlap of duplicates and isoforms in the genes I analyzed, I could not explicitly test the independent model. However, the fact that I observe a strong effect of structural gene properties runs contrary to the expectation of a process driven by their different potential to generate adaptive variants.

In conclusion, while I find no evidence for age itself being a determinant of the presence of isoforms, I do find strong support for structural gene properties. This might be explained by an hybrid model in which the final outcome is determined both by the propensity of a gene to produce isoforms (or duplicates), and by their differential fixation because of their adaptive potential (independent model) or overlap (function-sharing model).

I must point that our study assesses the presence or absence of isoforms, rather than their number, and only considers the subset of highly expressed genes, which might have

different selective pressures than restricted ones. Our choices are necessary to provide a fair comparison, since lowly expressed genes have intrinsically lower probabilities of having observable isoforms and the number of isoforms is likely to increase as more diverse RNA samples are sequenced. However, they also skew our analysis towards a non-random subset of genes, which might be subject to different selective pressures. As such, tackling a truly comprehensive analysis of splicing and duplication in the *Nasonia* genome will require more sequencing efforts.

1.5. Conclusions

OGS2 provides a major quantitative and qualitative update to the toolbox for *Nasonia*'s genomics research. Better-defined UTRs enable the study of post-transcriptional regulation via targeting of small RNAs. Novel reported isoforms provide a more accurate representation of gene expression. I also highlight interesting areas for future molecular biology research using this organism, such as histone modification. Furthermore, I provide an estimate of the most unique traits of the *Nasonia* genome when compared with other Hymenoptera, which can assist the discovery of genetic mechanisms underlying the typical features of this lineage.

The advances in gene annotation for OGS2 are notable today, however as gene evidence accumulates in the future, new and improved gene sets will need to be constructed until a verifiably complete and biologically accurate gene set is produced. Transcriptomic data in the form of high quality and inexpensive RNA-Seq is now the leading form of gene evidence for most genome projects, surpassing gene prediction and mapping of reference gene proteins. Along with abundant high quality RNA-Seq for the model *Drosophila*, Tribolium, and other insects, the Apis *mellifera* gene set has recently been improved by addition of several billion paired reads, sufficient for the assembly of all but the weakly expressed genes. This approach has been employed at NCBI for updated genome-based models, and at EvidentialGene with RNA-only assemblies. The RNA assemblies may well surpass genome-modeled genes for orthology completeness as well as species-unique

completeness (Gilbert, 2013).

As a proof of concept, all of the novel data that enabled the annotation improvements made by OGS2 are derived from functional genomics methods (RNA-Seq, tiling arrays and ESTs). Transcriptomic data can thus improve genome annotation, even when the underlying genome assembly is frozen. As shown by the publication of results from the modENCODE *Drosophila* project (Roy et al., 2010), new genes and transcripts are discovered, even for a genome that has been intensively investigated for over half a century. Our modeling estimated that 50% of all *Nasonia* loci may possess alternative transcripts, comparable to the 57% observed from the *Drosophila* transcriptome (Brown et al., 2014), whereas I recovered alternates from RNA assemblies at only 17% of all loci. Therefore, even though it is unlikely that the addition of novel data will drastically increase the gene count for the *Nasonia* genome, I expect an increase in the number of reported isoforms with the addition of stage, tissue and condition specific transcriptomes. Perhaps more importantly, new data will increase the quality of gene models, where RNA transcript assemblies will validate and improve gene structures, an unresolved subset of which I believe are fragments or gene joins, and will provide further evidence for intron/exon patterning.

Our phylogenetic analyses were restricted in scope to the portion of the genome that could be assigned to an ortholog group, and its interpretation hindered by the large number of genes of unknown function. In order for the genomics of this organism to be better linked to its biology, there is a pressing need for more functional studies tailored to *Nasonia*'s unique features. Genome wide association studies and quantitative trait loci are especially complimentary for this purpose, as they provide a first connection between the well-defined transcriptionally active regions and biologically relevant traits (Mackay et al., 2009; Ayroles et al., 2009). As a final note, OGS2 is currently rich in models that have little support. These lowly supported models might prove to be a valuable resource for future studies on the unique features of the wasp lineage, as their current status as low-level support loci could either be indicative of a restricted expression pattern or of a

recent evolution or emergence in the hymenopteran phylogeny.

2. FESTA: *F*LEXIBLE *E*XON-BASED *S*PLICING AND *T*RANSCRIPTION *A*NNOTATION

2.1. Abstract

I introduce FESTA, an R based algorithm that allows detection of alternative splicing based on experiment-specific exon expression data. FESTA disentangles alternative splicing signal from whole-gene transcription, facilitating the discovery and characterization of novel regulatory events even in the absence of transcript annotations or paired-end data. I also include customization options to increase its applicability on different platforms and experimental designs as well as a tool for the conversion from transcript expression to inclusion ratios.

Availability and implementation: The scripts described are presented in the supplementary materials of this thesis and as additional file 7 in the attached disk.

2.2. Background

Alternative splicing is a widespread feature of eukaryotic gene regulation which can be represented as a two step process. Transcription generates the total amount pre-mRNA per gene locus, whereas splicing determines the proportions of each alternative transcript that is produced. Based on this model I can distinguish between constitutive exons, which are present across all isoforms and facultative exons, which are present in only a subset of alternative transcripts.

Commonly used methods discard information contained in constitutive exons or average it to match the proportions provided by transcript-specific exons (Trapnell et al., 2010), effectively conflating transcription and splicing dependent signal. Furthermore if reads are mapped to pre-annotated transcripts novel transcriptional events might be missed entirely. Dataset-specific estimation of constitutive and transcript-specific exons is therefore advisable for the discovery of novel alternative splicing events relevant to the design of interest (i.e. Dai et al., 2012; McManus et al., 2014).

Correlation based exon clustering is a simple implementation of such a method (Patrick et al., 2013). Since strong correlations among exons arise from their coexpression as part of a single transcript, every cluster represents either an alternative transcript or the subset of exons that are present across all isoforms (constitutive exons, Chen, 2013). Constitutive exon clusters will be present in all isoforms and can therefore be identified by having an absolute expression value either higher or equal to any other exon group. Despite being intuitive and effective, correlation based hierarchical clustering is limited by its choice of an a priori threshold.

In this chapter I define a simple algorithm that solves this issue by setting gene-specific thresholds based on highly customizable biological expectations. I also provide a function to calculate inclusion ratios of alternative exon groups in order to allow analysis of transcription independent effects of splicing.

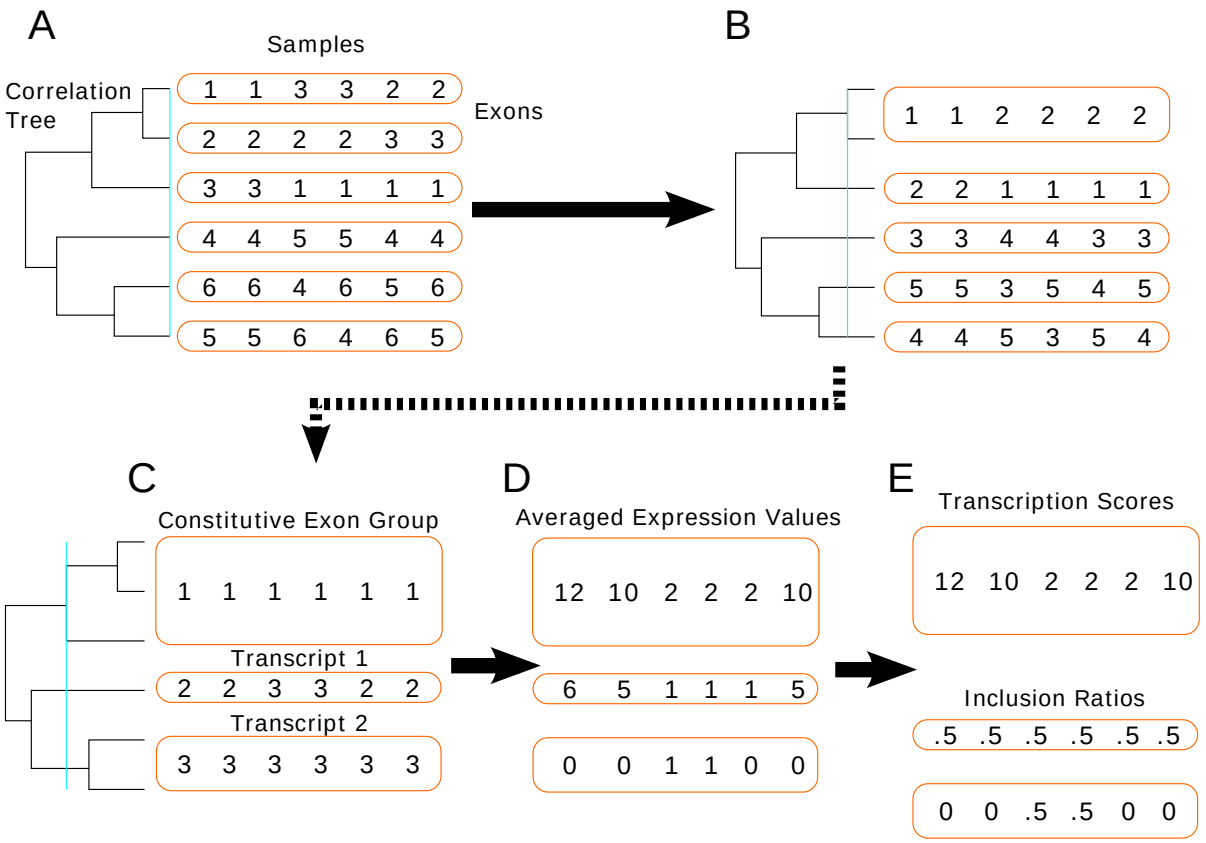


Figure 7: **Outline of the FESTA algorithm**. Steps A-C are handled by the ClusterExons function. Step D and optional step E are handled by the AverageExons function.

2.3. Implementation and Usage

2.3.1. Data input and filtering

FESTA requires two input files: an exon by sample expression table and an exon to gene assignment table. In order to avoid spurious grouping resulting from correlations in the noise component I advise thresholding raw expression data, removing all values that score below minimal signal and excluding all exons that lack expression in a sufficient number of biological replicates for at least one of the dataset's conditions.

2.3.2. Isoform detection

Figure 7 shows an outline of the FESTA algorithm, which is applied to iteratively to each gene. If a gene has more than one expressed exon, FESTA calculates a clustering

tree based on the correlation matrix of expressed exons. FESTA then cuts the tree at the lowest level (one exon per cluster) and ranks each group's expression in each sample (figure 7, A).

If any exon group is ranked first or tied for first across all samples I consider it to be the constitutive part of the gene, record the cluster assignments at the tree cut level and proceed on to the next gene. If no exon group ranks as first or tied across all samples, FESTA moves up a level in the hierarchical clustering tree, averages expression scores in exon groups with more than one exon and re-calculates the exon group rankings across the dataset (figure 7, B).

FESTA iteratively calculates the expression rankings of exon groups at each level until a single exon group shows the highest expression across all samples (figure 7, C). If no exon group meets constitutive exon criteria at the highest level, the algorithm converges on single-group clustering: all exons are annotated as constitutive and the gene is reported as lacking significant splicing events.

FESTA generates a single expression score for each group by averaging the expression scores of all its exons (figure 7, D). These raw expression scores can be directly used for analyses on individual transcript abundance. Alternative splicing events can also be converted to inclusion ratios by dividing them by the transcription score of their gene (figure 7, E). Inclusion ratios range between zero (if the isoform is absent) and one (if all transcripts produced by the gene include those exons) and can be used to analyze the effects of alternative splicing independently of the main gene's overall expression.

2.3.3. Fine tuning parameters

I include two main parameters can be changed to affect the sensitivity and power of the main clustering algorithm: significant digits and number of exceptions.

Significant digits allows the user to define numerical accuracy of expression measurements. Setting a high number of significant digits will result in less ties between exon groups but might cause over sensitivity to minor fluctuations in expression values between biologically

co-expressed exons. Fewer significant digits increase the number of ties in rankings, decreasing the ability to differentiate constitutive exons from highly expressed alternative exons.

Number of exceptions allows to increase the permissiveness of constitutive exon group definitions. If this number is greater than zero, constitutive groups are re-defined as being first or tied with any other group in all samples except the exceptions. For instance, in case the dataset includes 25 samples, exon groups will be identified as constitutive if they are first or tied in at least 24 samples if exception number is set to 1, at least 23 samples if it is set to 2 and so on. This parameter enables setting tree-cut height based on experimental design considerations, with more stringent values resulting in less isoforms and larger constitutive exon groups and more permissive values resulting in more isoforms and smaller constitutive exon groups.

2.3.4. Caveats

There are three caveats regarding FESTA's current implementation. Firstly, the algorithm depends on the number of biological replicates to generate accurate exon rankings. Secondly, it does not currently make use of paired-end data. Lastly, as the algorithm attempts to identify isoform-specific exon groups it will not be able to detect isoforms characterized by different combinations of the same exons such as in the case of hypervariable combinatorial genes.

2.4. Conclusions

I present an intuitive method for the detection of transcription and splicing in transcriptomic data which requires only an exon by sample expression table. FESTA allows the end user to customize sensitivity using easily interpreted parameters which can be tuned to the experimental design and the instrument's sensitivity. FESTA's output is a reduced transcript by sample table, which retains only the splice variants observed in the experiments and can be directly used in downstream analyses. The optional conversion from

transcript abundances to splicing ratios allows the investigation of the effects of increasing the proportions of specific isoforms rather than their absolute abundances, allowing for a comparative study of the impact of transcriptional and splicing regulation.

3. Transcriptomic Basis of Sexual Dimorphism in *Nasonia vitripennis*

3.1. Abstract

Background The generation of sexually dimorphic phenotypes requires a series of sex-specific regulatory processes to occur throughout development. A more detailed description of earlier sex-bias patterns is required for understanding the true extent of sex-specific selection on the genome and how such sex-specific patterns are achieved. In order to expand our understanding of developmental sex-bias dynamics, I apply a series of network-based methods to disentangle and characterize the impact of differential expression, splicing, linkage, gene duplication and whole cluster co-regulation in the sexual development of *Nasonia vitripennis*.

Results Sex-lethal and several other sex-biased genes show clustering on the genome. Sex-biased transcription appears to be more prevalent than sex-biased splicing. Few transcripts shift from female to male biased expression (or *vice versa*) during development. Sex-biased interactions reveal several regulatory events in early development. Compared to unbiased clusters, sex-biased clusters show enrichment for novel or fast evolving genes which occupy potentially regulatory positions.

Conclusions *Nasonia* shows significant amounts of transcriptional sex-bias across all of its development, often in a stage-specific fashion. Early sex-bias appears to be driven by transcript-transcript interactions rather than single-gene differential expression. Clustering of sex-biased genes is present for several regions, despite the lack of sex-determining loci. Sex-biased clusters appear to have rapidly integrated new and fast-evolving genes in potentially regulatory positions, suggesting a dynamic evolutionary history of sexual development.

3.2. Background

While sex-determination cascades have been explored in a wide array of organisms (Bull and Others, 1983; Cho et al., 2007; Verhulst and van de Zande, 2014), we still lack knowledge on how sex-biased expression evolves and how it affects phenotypic evolution. Most studies on sex-bias have so far focused on the specific case of genes with differences in their mean expression level between adult males and females, primarily in organisms with genetic sex determination (Innocenti and Morrow, 2010; Chang et al., 2011). This focus has led to several interesting findings such as a tendency of sex-biased genes to arise from gene duplications, evolve more rapidly than non-biased ones and to accumulate in sex-specific portions of the genome (Vibranovski et al., 2009; Gallach and Betrán, 2011; Jaquiéry et al., 2013; Dean and Mank, 2014). Some of these processes have been linked to the uneven action of selective pressure on sex-biased loci, which leads to tug-of-war dynamics between male and female-specific optimization in gene function, or intragenomic sexual conflict (Ellegren and Parsch, 2007; Dean et al., 2012; Mank et al., 2013; Parsch and Ellegren, 2013).

At the same time, the focus on adult differential expression and sex chromosomes embraces only a small subset of ways by which transcriptomic bias can achieve between-sex differences. Specific transcripts can display transient sex-bias in earlier developmental stages, which is not maintained in the adult (Perry et al., 2014; Mank et al., 2010). Such cases are especially likely for genes involved the establishment and development of sex-specific cell fates, which need to act before the adult forms are completely functional (Badyaev, 2002; Sun et al., 2015). Analyses of earlier developmental stages can reveal genes that are male biased in some stages and female biased in others (Mank et al., 2010; Zhao et al., 2011). Such changes in sex-bias within the same gene are likely to create sexual conflict, since the same locus will be under selection for female-specific and male-specific functions in different developmental stages: a scenario that I call developmental sexual conflict. Differential splicing has also the potential to generate sex-biased transcripts without affecting overall gene expression (Telonis-Scott et al., 2008; Hartmann et al., 2011),

causing exon-specific evolution even in conserved genes (Parker et al., 2013) and potentially mediating sexual conflict similarly to sex-specific duplicates. A role for splicing regulation in sex determination has been characterized for the auto regulatory *transformer* (*csd* in *Apis* and *fem* in *Nasonia*) loop in several insect species (Verhulst et al., 2010b, 2013), but measurements of transcriptome-wide prevalence and role of sex-biased isoforms remain confined to either standard model organisms such as *Drosophila* (Telonis-Scott et al., 2008; Hartmann et al., 2011; Chang et al., 2011; Brown et al., 2014) or specific genes of interest (i.e. Bertossa et al., 2009).

Characterizing sex-biased gene expression is of even greater importance for the wasp lineage represented by *Nasonia vitripennis*, which differs from most of the aforementioned models on several accounts. As a member of the Hymenoptera, *Nasonia* shares haplodiploid sex-determination with ants and bees. Male and female *Nasonia* lack sex-specific genomic regions and must share all genes between sexes, allowing for analyses of sex-bias that avoid the complications of between-sex genetic differences (Heimpel and de Boer, 2008; Godfray, 2010). We know that *Nasonia*'s primary sex-determination mechanism is different from that of other Hymenoptera, as it does not rely on the *csd* locus, but we lack a clearer identification for which mechanism may have replaced it (Kamping et al., 2007; Heimpel and de Boer, 2008). Studies on *Nasonia* gene expression have so far focused either on adults, individual tissues or specific pathways (i.e. Pers et al., 2016) so that a transcriptome-wide description of its developmental gene expression is currently lacking. Despite the lack of sex-specific genomic regions, a recent study estimated that adult *Nasonia* shows expression bias in over 75% of its genes (Wang et al., 2015). This finding begets the question of how such differences are established over the course of development and how early on we can detect significant differences in the gene expression of the two sexes. Phenotypic descriptions of *Nasonia*'s embryonic development show no clear divergence in their morphologies (Bull, 1982; Pultz and Leaf, 2003), which become sexually dimorphic only during pupation. Conversely, molecular studies show evidence of sex-biased transcription for the sex-determining genes *transformer* and *doublesex* as early as 7 and

12 hours from oviposition, respectively (Verhulst et al., 2013; Zwier et al., 2012). The lack of morphological dimorphism between sexes before pupation allows us to putatively assign early sex-biased expression to sex-determination, as opposed to development of dimorphic adults which occurs during pupation. Thus, using time-series data allows us to detect the onset of sex-bias for different categories of genes and distinguish their different roles in sex determination (early development) and sexual development (late development) from the adult functions of reproduction, flight, poison and pheromone production. From a Systems Biology perspective, developmental time series data are especially valuable as they provide the required complexity to distinguish between stable associations and transient interactions as well as allowing for detection of directional effects thanks to the explicit presence of a time component.

I choose to analyze *Nasonia*'s sexual development by reconstructing its coexpression network. A major advantage of network-based frameworks is that they can detect groups of regulatory events acting in concert (transcriptional modules) and estimate relationships between their members based on their connections (Langfelder and Horvath, 2008; Zampieri et al., 2008; Mozhui et al., 2012). By measuring connections, I was able to identify potential regulators and assess whether interactions between groups of nodes are sex-biased themselves (Hudson et al., 2009; Hsu et al., 2015). Such contingent interactions are especially interesting in the context of gene regulation, since different transcription factors can combine non-additively, making it crucial to know their interacting partners in order to predict their effect (Ament et al., 2012; Spitz and Furlong, 2012; Boyle et al., 2014). In the specific case of sexual development, it is conceivable that the same genes may cause a sex-specific effect only when coexpressed (Arnold et al., 2009; Van Nas et al., 2009), giving rise to transcriptional modules whose effect is elicited by sex-biased changes in the correlation between their members rather than in their overall expression values (de la Fuente, 2010). Combinatorial effects of this nature are impossible to detect by independently testing transcripts for changes in mean expression but can be identified *via* differential correlation analyses on transcriptional modules (Tesson et al., 2010; Yang

et al., 2013).

Compared to standard enrichment testing, network methods enable us to ask not only if specific types of genes are over-represented in a cluster (Conesa et al., 2005; Subramanian et al., 2005) but also to evaluate whether they occupy preferential positions within their topology (Khatri et al., 2012). For instance, network-based methods can be used to separate condition-specific genes between molecular workers which carry out molecular functions (i.e. structural proteins and enzymes) and hubs which regulate their behaviour (i.e. Pierson et al., 2015). The ability to assign putative functions independently of homology assignments is especially valuable for non-model species, since it allows both to identify new study targets among lineage-specific genes and to estimate if genes with known homologs have evolved new functions (i.e. Nawaz et al., 2012). At a larger scale, the structure of entire transcriptional modules can be assessed via several network parameters such as centralization and density, allowing systematic comparisons of their regulatory structures (Jeong et al., 2000; Horvath and Dong, 2008). Development-spanning network reconstructions provide a necessary comparison for testing the generality of more targeted pathway analysis studies (i.e. Pers et al., 2016) and validating models of evolutionary change via network remodelling. In the case of *Nasonia*'s sexual dimorphism, I focus on the hypothesis that the organization of sex-biased clusters may facilitate the rapid evolution of sex-biased genes. Two main traits of network structure have been predicted to influence the evolutionary rates of individual genes: module density and hierarchical organization. Modules with high density are predicted to show decreased rates of regulatory evolution for two reasons. First, altering the behaviour of a gene with several molecular interactions is likely to result in several epistatic effects with random fitness consequences, decreasing the chances of achieving a net increase in fitness (Kauffman, 1987; Papakostas et al., 2014). Second, individual regulators in a highly interconnected network are unlikely to produce new phenotypes due to the counterbalancing effects of other regulators in the same module. Since dense networks are predicted to hinder rapid regulatory evolution, I expect to find lower average densities among sex-biased clusters. Hierarchical network have instead been

shown to strongly facilitate the evolution of new adaptive regulatory interactions under several simulated circumstances (Mengistu et al., 2016), but have received little attention from the empirical community. I thus set out to test whether hierarchical organization may be involved in the rapid evolution of sex-biased clusters using the *Nasonia* system.

In this chapter, I generate a hybrid transcription and splicing network (developed in section §2) to detect how different regulatory processes shape sex-bias at the transcriptome level in *Nasonia vitripennis*. I a prevalence of sex-biased transcription over sex-baised splicing, and outline three linkage groups enriched for sex-biased genes. I develop and utilize differential correlation analyses to identify cryptic sex-bias in early stages and highlight a previously unrecognized potential role for histone modification in inducing sex-bias. Lastly, I find that sex-biased clusters show higher hierarchical organization, and enrichment for recently evolved genes in potentially regulatory positions.

Definition of Terms

CCRE Constitutively Coexpressed Regulatory Events. Sets of nodes (both splicing and transcription) with more than 95% correlation among themselves. CCREs are represented by a single node in all downstream analyses. See Splicing Detection and Network Construction for details.

Developmental Sexual Conflict Sexual conflict arising from selection on the same gene for male-biased expression in specific developmental stages and female-biased selection in others.

Differential Expression (DE) Differences in the mean expression of a node or cluster between different types of samples. In our case refers to differences in mean expression values between sexes, or sex-biased differential expression.

Differential Correlation (DC) Differences in the within-cluster connection density between different types of samples. In our case refers to differences in within-cluster correlations between sexes, or sex-biased differential correlation.

Sex-Bias Generic term indicating sex-specific bias in at least one parameter of an element. Can refer to sex-biased expression of nodes, sex-biased expression of clusters or sex-biased correlation of clusters.

Splicing Node Node representing the expression of a specific gene's isoform relative to the expression of all isoforms. Each gene has a number of splicing nodes equal to the number of splicing events detected, which varies from zero to the number of exons. See section §2 for details.

Transcription Node Node representing the total production of RNAs from a single gene locus, irrespective of their final isoform. Each gene has only one transcription node. See See section §2 for details.

Density Proportion (0-1 bound). Indicates the number of observed connections compared to the maximum possible connections. In the case of cluster density it refers to connections between nodes within the cluster, with 0 indicating that no connection is observed and 1 that all nodes are connected between each other. In the case of node density it indicates the number of connections with other nodes of the same clusters, with 0 indicating that the node has no within-cluster connections and 1 indicating that the node is connected to all other nodes in the same cluster.

Hub Score Product of proportions (0-1 bound). High values indicate that a specific node is highly connected to other nodes which are not otherwise connected among themselves (hub). Low values indicate that a specific node is lowly connected to nodes which are already interconnected (worker).

3.3. Methods

3.3.1. Biological Materials and Data Collection

The data used in this study consists of a developmental time series of transcriptional activity of whole animals in males and females of the jewel wasp *Nasonia vitripennis*. The experimental design comprises five distinct developmental stages: early embryo (0-10 hr old), late embryo (18-30 hr old), 1st instar larvae (~51 hr old), yellow pupa stage (~14 days) and sexually mature virgin adults. More specifically, the first embryo stage comprises the development from a single zygote to the late blastoderm, just before the beginning of gastrulation. The late embryo stage starts after the end of gastrulation and comprises most of the remaining pre-hatching development, including segmentation and organogenesis (for reference timings see Bull, 1982).

All animals used for data collection come from the highly inbred strain AsymCX (Werren et al., 2010). Each of these conditions was sampled in triplicate for each sex. Due to the different number of cells at different stages, different numbers of individuals were sampled pooled for each biological replicate as follows: 300-900 individuals for early embryos, 140-500 for late embryos, 245-520 for 1st instar larvae, 20 for pupae and adults. Pupae and adults were produced by mated females and sexed by visual examination prior to extraction and sequencing. Since sexing by visual examination is not possible before the pupal stage, male embryonic and larval samples were collected from virgin females, which produce only males. Female embryonic and larval samples were collected from mated females, which produced ~83% female offspring.

Expression values were measured via single-channel whole-genome tiling path microarrays using custom NimbleGen high-density 2 (HD2) arrays (Lopez and Colbourne, 2011), consisting of 8.4 million probes with a 50-60 nt length spanning the *Nasonia* genome at 33 bp intervals, as well as 27,000 Markov probes which are absent from the genome for noise detection (see below). Further details on animal breeding, RNA extraction and microarray processing are available in the supplementary materials of Werren et al. (2010).

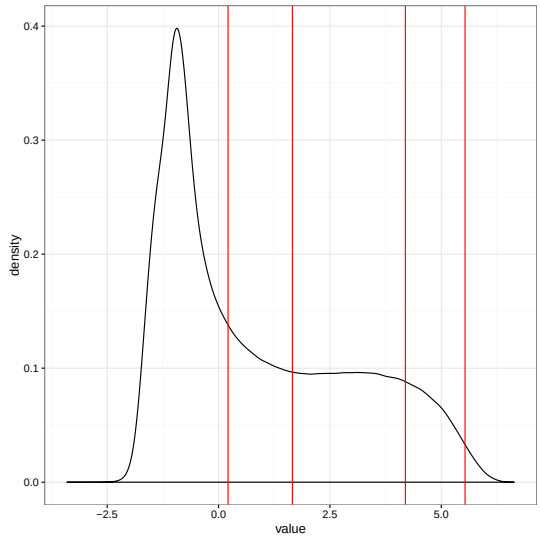


Figure 8: **Expression Values Distribution Before Thresholding**
Vertical red lines indicate the 50th, 66th, 90th and 99th percentiles respectively. Expression scores are reported as log ratios against the 99th percentile of random Markov probes.

3.3.2. Data Pre-processing

Individual probes were assigned to exons according to the latest release of the Official *Nasonia* Gene Set (OGS2.0, see Chapter 1 and Rago et al., 2016). Expression for each exon was measured as the log_2 ratio of the 99th quantile of the random Markov probes on their arrays. I determined a sensible expression cut-off by examining the distribution of exon expression values across the whole experiment (figure 8). Based on this assessment, I collapsed all values below the 66th expression percentile to zero in order to avoid spurious signal from random noise variation among non-expressed exons. Lastly, I retained only exons which showed expression above our threshold in at least two out of three replicates for at least one biological condition.

3.3.3. Splicing Detection and Network Construction

In order to disentangle transcription and splicing signal, I utilized the FESTA algorithm (see section §2) restricting the number of significant digits to 3 and allowing a maximum of one exception in the whole experiment. Since splicing nodes are expressed as 0-1 bound ratios, I rescaled transcription nodes to a 0-1 space by dividing them by the maximum

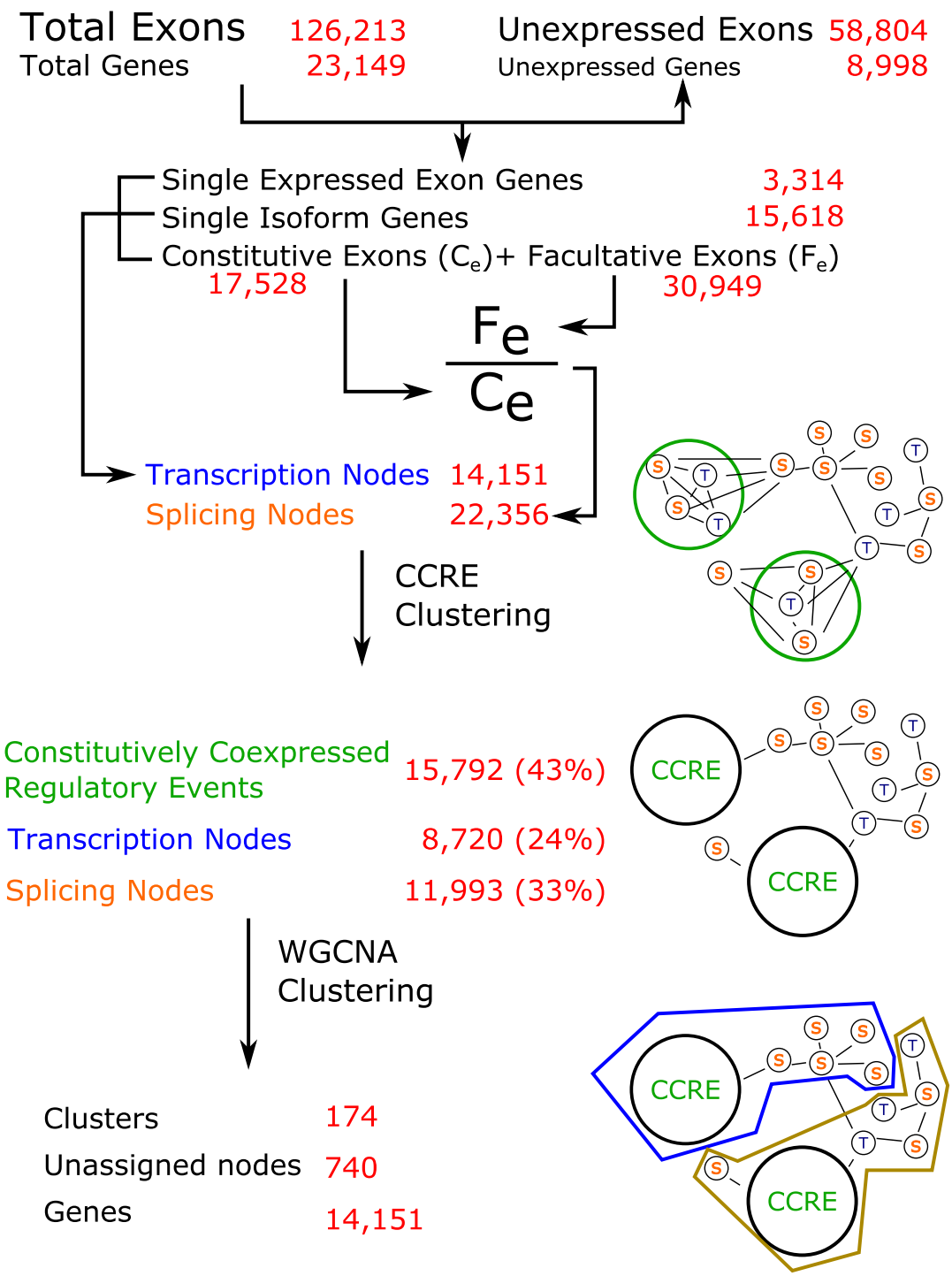


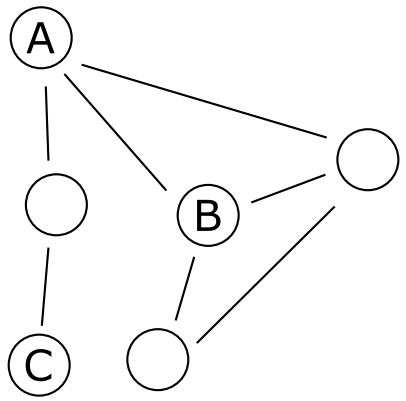
Figure 9: **Network construction workflow**:
I select exons based on expression within our experiment and cluster them using FESTA. Every gene is represented as a transcription node and a variable number of splicing nodes, each quantifying the inclusion ratios of a correlated set of exons. Groups of nodes with reciprocal correlations greater than 95% are collapsed into CCREs. The resulting dataset is converted into a network and clustered using WGCNA. Final figures indicate the amount of clusters, unclustered nodes and total genes in my network. See section 3.3.3 for details.

expression value observed. Our final dataset thus comprises one transcription node per gene and a variable number of splicing nodes, each representing a splicing event.

I collapsed all nodes with reciprocal correlation values higher than 95% into Constitutively Correlated Regulatory Events (CCREs) using the collapseRows function from the WGCNA package. This step enables us to reduce the dimensionality of our dataset by representing sets of nearly identical nodes as a single unit and indicates their possible shared role across development. Further to that, reducing highly correlated nodes to a single unit allows us to avoid the possible over-interpretation of closely tied nodes by reporting all of them as potentially significant. Our approach is conceptually similar to that of Constitutively Coexpressed Links (CCELs) in Hsu et al. (2015). I chose to represent each CCRE using expression scores of the node with the most correlation-based connections to other nodes in the same CCRE, as it is the most representative of the average behavior of other CCRE members. In the special case of CCREs with only two nodes, I used the one with the highest mean expression.

I constructed an undirected weighted interaction network using the R package WGCNA (Langfelder and Horvath, 2008). WGCNA infers between-transcript links based on power-transformed robust correlation scores. Since it does not require the input of pre-defined pathways or functional classes it is ideal for the analysis of species with high amounts of expression data but sparse functional genomic annotation. WGCNA is also able to rapidly calculate large networks, a key feature for enabling the permutation-based approaches that I implemented to monitor differentiall correlations (see 3.3.7). Finally, results obtained can be directly compared with the wealth of other studies employing the same workflow.

In order to make correlation measures tractable using graph-based approaches, WGCNA suggest power-transforming pairwise correlation scores (Zhang and Horvath, 2005), effectively increasing the gap between weak and strong links and thus the method's specificity. Most natural network studies show a power-distribution of connectivity across nodes (Jeong et al., 2000; Wagner and Fell, 2001; Barabási and Oltvai, 2004), with few highly connected nodes and many lowly connected ones, also called a scale-free degree distribution. Based



	Con	Hub
A	3	2
B	3	1
C	1	1

Figure 10: **Node Parameters in an Example Network:**
Table on the right side lists the number of connections (Con) and hub scores (Hub) of labeled nodes.
Node A is a high order co-ordinator, node B part of a 3 node complex and node C a worker.

on this "scale-free topology criterion" (Zhang and Horvath, 2005), I selected the lowest power that generated a scale-free correlation network. Hierarchical clustering applied to the topological overlap matrix (TOM) based on the power-transformed correlations identified 174 groups of co-expressed nodes; plus a group of 740 nodes that do not conform to any expression cluster (grey cluster).

An outline of the network construction process is presented in figure 9.

3.3.4. Network Topology Measurements

I decided to decided to quantify two main parameters per node: connection densities and hub-scores. My decision to focus on these two main parameters is based on their ability to classify nodes in three main categories, which are discusseed after the description of each individual parameter.

I define connection densities as the number of observed connections per node, normalized by the theoretical maximum possible number of connections. In an undirected weighted network, this parameter can be calculated with the formula $Kd_i = \frac{\sum k_i}{\frac{N-1}{2}}$, where $\sum k_i$is the sum of the weights for all connections to node i and $\frac{N-1}{2}$ is the maximum number of links in an undirected network of size N. This measure quantifies the relative importance of a node as a measure of its direct connections to other nodes within the same network

(i.e. nodes A and C in fig **10**). While useful to estimate the interactions of individual genes, connection density does not account for the different regulatory potential of different connections. For instance, in figure 10, nodes A and B have the same connection density. However, removing the connections from node A would split the network in two, whereas removing the connections from node B would only have a minor impact since its neighbours are already connected. In order to define this topological property, I calculate hub scores using the formula

$$Hub_i = Kd_i \cdot (1 - \frac{n_i}{max(n_i)})$$

Where Kd_i represents the connection density of node i, and $\frac{n_i}{max(n_i)}$ represents the clustering coefficient of node i, or the observed connectivity between nodes connected to node i divided by their maximum possible connectivity with each other. Consequently, nodes with high hub scores have a high number of connections to nodes that are not otherwise connected among themselves, and are likely to be involved in the coordination of multiple processes. Since I calculate hub scores by penalizing connection densities, their scores cannot be higher than densities themselves. This creates three potential combinations of parameters: low density and low hub score will indicate marginal nodes (i.e. node C in fig 10), high density and high hub scores will indicate regulators (node A in fig 10), and high density and low hub scores will indicate genes that are likely to be part of cooperative interactions (node B in fig 10).

I calculated both scores considering only nodes within the same transcriptional cluster, in order to provide an accurate representation of their internal regulation. I computed within-cluster network statistics for each node in a cluster using the fundamentalNetworkConcepts function from the WGCNA package, as well as weighted betweenness using the tnet package (Opsahl, 2009).

3.3.5. Differential Expression of Nodes and Clusters

I assessed the differential expression of individual nodes using generalized linear models (GLMs) as implemented in the LIMMA package (Smyth, 2005) using the formula

$$Expression \sim Stage + Stage : Sex$$

Which accounts for stage-specific differences in gene expression via the factor *Stage* and considers sex only as a second-order interaction term with stage-specific expression changes ($Stage : Sex$). The individual p-values were converted to local False Discovery Rates (lfdr), which represent the individual probability of each hypothesis to be a false positive via the R package fdrtool (Strimmer, 2008). All contrasts with a lfdr lower than 5% were considered significant.

In order to detect cluster-level bias, I extracted the first principal component (module eigengene) of each cluster and applied linear models using the same formulas as per individual nodes. I converted all p-values to lfdr scores and considered significant all contrasts with lfdr lower than 5%. Since this method assesses the differences in their mean expression between sexes, I refer to it as differential expression (DE) in the rest of the chapter.

3.3.6. Linkage Clusters Enriched in Sex-Biased Loci

I annotated each gene locus as being sex-biased if at least one of its child transcription or splicing nodes scored as differentially expressed between sexes in at least one stage. Since each node is tested for sex-bias independently at each developmental stage, it is possible for a single gene to be both male and female biased at different stages. Likewise, different transcription and splicing nodes from the same gene can show bias in either sex. Genes that fall in either category are unlikely to be subject to sex-specific selection and are thus excluded from linkage group enrichment analyses. I mapped all genes in our network to the linkage map published in Desjardins et al. (2013). I tested each individual

linkage group for enrichment in male or female biased genes via one-tailed Fisher's exact test, compared to the overall proportions of male and female-biased genes across all other linkage groups. This process generated two p-values per linkage group: one for female bias enrichment and one for male bias enrichment. Finally, I applied FDR correction to the p-values using the package fdrtool, and reported all clusters with a lfdr score lower than 5%.

3.3.7. Differential Correlation Analyses

In contrast with differential expression based methods, differential coexpression testing classifies groups of genes as biologically interesting if they show a differential increase or decrease in their correlations in the conditions of interest. Methods to analyze differential coexpression can be divided in two main categories: untargeted methods identify changes in transcript-transcript interactions (Tesson et al., 2010; Ma et al., 2011; Hsu et al., 2015; Liu et al., 2016), while targeted ones measure correlation changes in pre-defined groups of transcripts (Yang et al., 2013; Cao et al., 2014). In order to allow direct comparisons between differential correlations and differential expression data, I developed a targeted method and applied it to the coexpression clusters found via network construction, a strategy also known as semi-targeted. Developing a new method was necessary since most available ones are designed for two-sample tests or to detect individual sample deviation from a pre-defined baseline (Yu et al., 2011; Walley et al., 2012; Liu et al., 2016), and are thus unable to account for multi-level and nested experimental designs. Conversely, untargeted methods would incur in steep costs in both power and computational time as well as hinder comparisons with cluster-level differentiale expression.

I applied a sub-sampling based procedure, recording the effect that the removal of male and female samples have on specific cluster parameters. Since our main focus is the detection of sex-specific co-regulation, I employed a sub-sampling strategy that removes all possible combinations of 3 samples within each stage. This sub-sampling strategy maintains a constant number of samples used for the generation of each sub-network, while

altering the proportion of samples from each sex in a stage-specific manner.

I applied the WGCNA process of network construction to each of the sub-sampled datasets using the same power transformation and node to cluster assignments as per main network construction (see section 3.3.3), measuring the within-cluster density of each cluster in every sub-sampled network. Since WGCNA-based cluster density is effectively a power-transformed measure of correlation between nodes in a cluster, I refer to its differential change as differential correlation (DC) throughout the chapter. Within-cluster density is a proportion measure and as such it is distributed on a 0-1 scale, where 1 indicates that all possible connections between nodes are observed and 0 that none of them is. It can therefore be analyzed using GLMs with a gamma error distribution and logit link function. I fitted the following model to each cluster

$$Density \sim Stage + Stage : Sex + Network\,Density$$

which allows me to detect stage-specific sex-bias in cluster density ($Stage : Sex$ term) while controlling for stage-specific and whole-network increases in connectivity. In order to validate whether observed density bias is likely to be generated by random chance I fitted the same GLM to 1000 datasets generated by randomly permuting sex-labels. I then extracted p-values for the $Stage : Sex$ interactions for each cluster from the GLMs of both the permuted and observed datasets. I estimated the probability that each case of sex-bias is due to random chance by calculating the local fdr of observed $Stage : Sex$ p-values compared with the distribution of p-values generated by randomly permuted dataset. Finally, I corrected for multiple-hypothesis testing by calculating the lfdr score for each cluster's $Stage : Sex$ lfdr score against all other clusters' lfdr scores. I considered all $Stage : Sex$ interactions with a lfdr score lower than 10% as significant, leading to the expectation of less than 2 false discoveries.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Density	-0.45	0.10	-0.11	-0.41	0.10	-0.12	0.08	0.76
Centralization	-0.40	0.37	0.09	-0.13	-0.42	0.26	0.58	-0.31
Heterogeneity	0.28	0.53	0.20	0.41	-0.46	-0.27	-0.09	0.37
Number of Nodes	0.41	0.32	-0.03	-0.36	0.01	0.73	-0.24	0.14
Excess Splicing	-0.17	0.37	-0.67	0.48	0.35	0.16	0.05	-0.01
Excess Duplication	-0.23	0.21	0.70	0.27	0.56	0.18	0.00	0.03
Median Clustering Coefficient	-0.45	0.28	-0.01	-0.17	-0.14	-0.13	-0.74	-0.32
Diameter	0.33	0.47	-0.03	-0.43	0.38	-0.49	0.20	-0.25

Table 4: **PCA Scores of individual cluster parameters**, approximated to the third digit

3.3.8. Multivariate analysis of network parameters

Network and sub-network parameters display several non-trivial correlations (Dong and Horvath, 2007; Horvath and Dong, 2008). Consistently, I observe strong non-independence between our parameters of interest (see figure 17). I employed Principal Component Analysis (PCA) to deconvolute the latent independent components that affect network parameters. Factors included in PCA analysis are cluster size (number of nodes), density, centralization, heterogeneity and median cluster coefficient as defined in Horvath and Dong (2008), as well as cluster diameter (the longest among shortest paths within the network). I also included relative proportion of splicing nodes and relative proportion of nodes arising from duplicated genes. Both proportions were normalized by their respective network wide abundances before PCA. All variables were centered and scaled before PCA.

Each of the principal components (PCs) extracted by PCA represents a single linear combination of the factors provided that maximizes the degree of variance between clusters and minimizes the reciprocal correlation with other PCs. I determined the biological significance of each PC by comparing the relative contribution of each parameter to their score (as estimated by parameter loadings, table 4). Since my goal is to identify whether any of the latent variables can discriminate between the different classes of sex-biased clusters, I used binomial GLMs including all PCs as predictors. I fitted three separate model sets using the following dependent variables: differentially expressed cluster, differentially correlated cluster, clusters with both differential expression and correlation. I then computed model

sets containing all possible combinations of factors for each of the three main models and estimated each factor's probability of being included in the best model of its set (relative importance or RI) using AICc based rankings as implemented in the R package MuMIn, (Barton, 2011). The results for differentially correlated clusters and clusters with both differential expression and correlation are identical (data not shown), most likely because only 4 clusters with differential correlations show no differential expression in at least one stage. Due to this matching, I only report results for the model set targeting differentially correlated clusters.

In order to detect whether any PC differs significantly between differentially expressed and differentially correlated clusters, I fitted a fourth binomial model set including only clusters with either differential expression or differential correlation, using differential correlation as a dependent variable and the 8 PCs as its predictor.

3.3.9. Phylostratigraphic analyses on network parameters

I retrieved the phylostratigraphic annotation (Domazet-Lošo and Tautz, 2010) of *Nasonia* OGS2.0 from Sackton et al. (2013). I used GLMs to test for the impact of phylostratigraphic age on each node's within-cluster connection density and hub scores by fitting the following models

$$ConnectionDensity \sim ClusterSize+Stratum+DE+DC+Stratum:DE+Stratum:DC$$

$$HubScore \sim ClusterSize + Stratum + DE + DC + Stratum:DE + Stratum:DC$$

That estimates the ability of taxonomic strata to predict connection density and hub scores both independently (term $Stratum$) and while interacting with my two main sex-biased parameters (terms $Stratum:DE$ and $Stratum:DC$), after controlling for variation

in connection densities due to sex-bias parameters (terms DE and DC) and cluster size. Since connection densities and hub scores are expressed as 0-1 bound variables I used a gamma error distribution and a logit link function for GLM analyses. I subsequently fitted all possible nested models and produced model-averaged parameter estimates and RIs for each factors using AICc based rankings (as implemented in Barton, 2011).

3.3.10. Gene Ontology, and Protein Family Enrichment Analyses

For Gene Ontology (GO) and PFAM (Protein Family database) enrichment, I used the interface provided by Wasp Atlas, which returns FDR-corrected q-values for over-representation of GO and PFAM categories in the gene set of interest by using one-tailed FDR corrected hypergeometric over representation tests (Davies and Tauber, 2015). The input I used for enrichment testing was either genes (for linkage group enrichment) or transcription nodes (for transcriptional cluster enrichment). Throughout the chapter I consider significant only GO and PFAM terms with a q-value lower than 0.01.

3.3.11. Additional software tools

Most statistical analyses were performed in R version 3.2.2 (R Core Team, 2013) using the following packages: plyr (Wickham, 2011) and reshape2 (Wickham, 2007) for data handling, vcd (Meyer et al., 2014) and ggplot2 (Wickham, 2009) for plotting.

3.4. Results

3.4.1. Stage-specificity of gene-level sex-bias

A large portion of nodes display sex-biased expression or splicing when tested individually (see table 5). Male biased genes are prevalent in the pupal stage, whereas female-biased transcriptional events are most frequent in the adult stage. Larvae show the least amount of transcriptomic bias between sexes. Only one transcript (Nasvi2EG005321 or *Feminizer*) is sex-biased across the whole development, followed by *Doublesex* (Nasvi2EG010980), which is female-biased in all stages from late embryo onward (>18 hours old). The low number of transcripts consistently differentially expressed across multiple stages is most likely due to the low number of sex-biased events in pre-pupal stages. Only 751 transcripts (2% of all transcripts) show sex-bias in the embryonic or larval stages.

Stage:	Genes		Transcripts	
	Male	Female	Male	Female
Embryo, 10 hr	26	145	29	174
Embryo, 18 hr	187	185	202	220
Larva, 51 hr	17	121	17	137
Pupa, yellow	1,392	434	2,779	581
Adult	3,194	3,093	5,167	5,953

Table 5: **Number of sex-biased genes and transcriptional events at each developmental stage.** Genes are counted as sex-biased if at least one of their transcription or splicing nodes is sex-biased.

Transcripts that are both male and female biased in different developmental stages are considerably less frequent than expected by chance (Fisher's exact test, p-value ~0): only 508 transcripts, generated by 373 genes (26% of all genes in our final dataset). The majority (66%) of these transcripts display shifts from male bias in pupae to female bias in adults, and 52% of them are assigned to clusters which show the same developmental sex-bias pattern (see section section 3.4.4). Other patterns which include both male and female bias across development consist of male-biased expression in adults and female biased expression in pre-adult stages (female bias in pupa: 57 transcripts, female bias in larva: 23 transcripts, female bias in late embryo: 27 transcripts, and female bias in early embryo: 8 transcripts). Interestingly, transcripts with pre-pupal sex-bias are significantly

more likely to show both male and female-bias in different stages than transcripts with post-pupal bias only (Fisher's exact test, p-value ~0).

3.4.2. Low prevalence of sex-biased splicing

Genes with sex-biased transcription are ~50% more frequent than genes with sex-biased splicing (6041 versus 3944). Over 67% of genes with sex-biased splicing also show sex-biased transcription, whereas less than 44% of genes with sex-biased splicing are also subject to sex-biased transcription (figure 11). Only 1294 genes show sex-biased splicing alone, compared with 3391 genes with only sex-biased transcription. Taken together, these observations indicate that transcriptional bias is the main determinant of transcriptome-wide differences between sexes. My estimates on the adult proportion of the sex-biased adult *Nasonia* transcriptome are consistent with those previously reported (Wang et al., 2015). I include the full annotation of each *Nasonia* transcript included in this study as additional file 8 in the attached disk.

3.4.3. Genomic regions enriched in sex-biased genes

Non-recombining regions can provide a suitable location for multiple co-adapted alleles which need to be co-inherited to provide a fitness benefit. Such supergenes have been observed in a few polymorphic species (Joron et al., 2011; Thompson and Jiggins, 2014), and could act as pseudo sex-chromosomes. I investigated whether such regions are present in the *Nasonia* genome by testing individual linkage groups for enrichment in male or female-biased genes. Two clusters show enrichment for female-

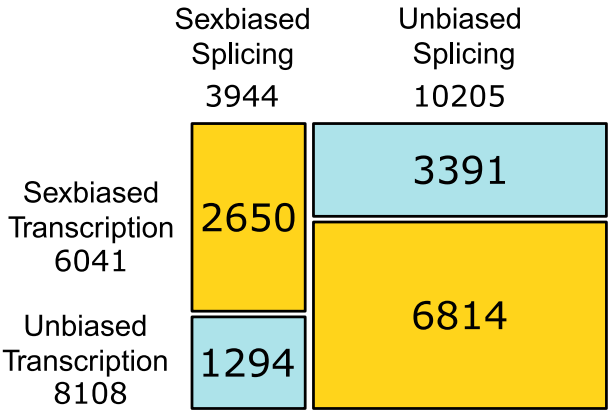


Figure 11: **Number of Genes with Sex-Biased Transcription and Splicing**. Yellow cells indicate over-representation, blue ones under-representation

Linkage Group	Enrichment	Male-biased Genes		Female-biased Genes		Total Genes	Recombination Rate
4.1	Male	49	45%	12	11%	109	$9.3 \cdot 10^{-2}$
1.065	Female	3	8.3%	18	50%	36	$3.8 \cdot 10^{-1}$
5.072	Female	12	8.0%	59	40%	147	$8.6 \cdot 10^{-2}$
Median Values	NA		17%		17%		$2.5 \cdot 10^{-1}$

Table 6: **Linkage groups enriched in sex-biased genes.**
Numbers indicate gene counts with their percentages compared to all genes in the linkage group. Recombination rates are expressed as centiMorgan per Mb. The last row reports median proportions and recombination rate across all linkage groups.

biased genes and one for male-biased ones (see table 6).

In particular, female-biased group 1.065 is the location of *Nasonia*'s sex-lethal (Nasvi2EG000104), homolog of the primary signal of *Drosophila*'s sex-determination cascade. The same linkage group also houses histone deacetylase 3, a key component of histone-mediated gene regulation. Female biased linkage group 5.072 is strongly enriched for the GO terms "apoptosis of nurse cells" (GO:0045476) and several other developmental terms related to photoreceptor and neuronal development (R3,R4 and R7 cell development and brain morphogenesis). Most genes on the male-biased linkage group 4.1 belong to cysteine-rich secretory secretory proteins (PF00188.21). While these proteins are currently annotated as venom allergens, I hypothesize that the same secretory domains may in this case be involved in sperm production, as is suggested by expression patterns of their homologs in *Drosophila* (Kovalick and Griffin, 2005).

Overall, the male enriched linkage group accounts for 1.2% of male-biased genes and the female-enriched linkage groups for 2.0% of female-biased genes. While theory predicts selection for lower recombination rates in sex-biased genomic regions, recombination rates in all three linkage groups fall within the interquartile range of recombination rates of all linkage groups.

Sex-Bias Pattern	Number of Clusters	Number of Genes
Unbiased	91	15,418 (52%)
....f	32	6,418 (22%)
....m	29	4,929 (17%)
...f.	4	452 (1.5%)
...m.	4	1,120 (3.8%)
...mf	3	324 (1.0%)
...mm	1	343 (1.2%)
..f..	1	64 (0.2%)
.f...	3	431 (1.4%)
.m...	2	97 (0.3%)
.mf..	1	59 (0.2%)
f....	1	79 (0.3%)

(a)

Sex-Bias Pattern	Number of Clusters	Number of Genes
Unbiased	144	25,137 (85%)
....f	6	1,145 (3.9%)
....m	8	967 (3.3%)
...m.	4	1,312 (4.4%)
..f..	1	64 (0.2%)
.m...	5	661 (2.2%)
f....	3	373 (1.3%)
m....	1	75 (0.3%)

(b)

Table 7: **Differential Expression (7a) and Differential Correlation (7b) Patterns across Development and number of Clusters and Genes which exhibit them.** Each pattern is coded as a string of five characters indicating its sex-bias status at each developmental stage from early embryo to adult: male (m), female (f), none(.). The number of genes per pattern includes all genes within all clusters that show that pattern.

3.4.4. Differential Cluster Expression Reveals Meiosis Genes

Differential expression testing at the cluster level shows quantitatively similar results to single-node testing (table 7a). Almost half of all transcriptional clusters (81 out of 172) are differentially expressed at some point in development. Most differential-expression based sex-bias occurs in pupal and adult stages (73 differentially expressed clusters), whereas only 8 clusters shows differential expression in pre-pupal stages. The complete annotation of all clusters is included in the attached disk as additional file 9.

Four clusters alternate between male and female sex bias in different developmental stages. Cluster green3 shifts from male bias in late embryos to female bias in larvae, and is primarily constituted by retrotranscriptases and unannotated multi-copy genes. It can therefore be attributed to transposon-related activity rather than developmentally related processes. The remaining three clusters (antiquewhite4, lightpink2 and yellow4) shift from male bias in pupae to female bias in adults. Antiquewhite4 and yellow4 comprise multiple isoforms of the *Nasonia* homologs of SAK (Nasvi2EG010310) involved in the formation of

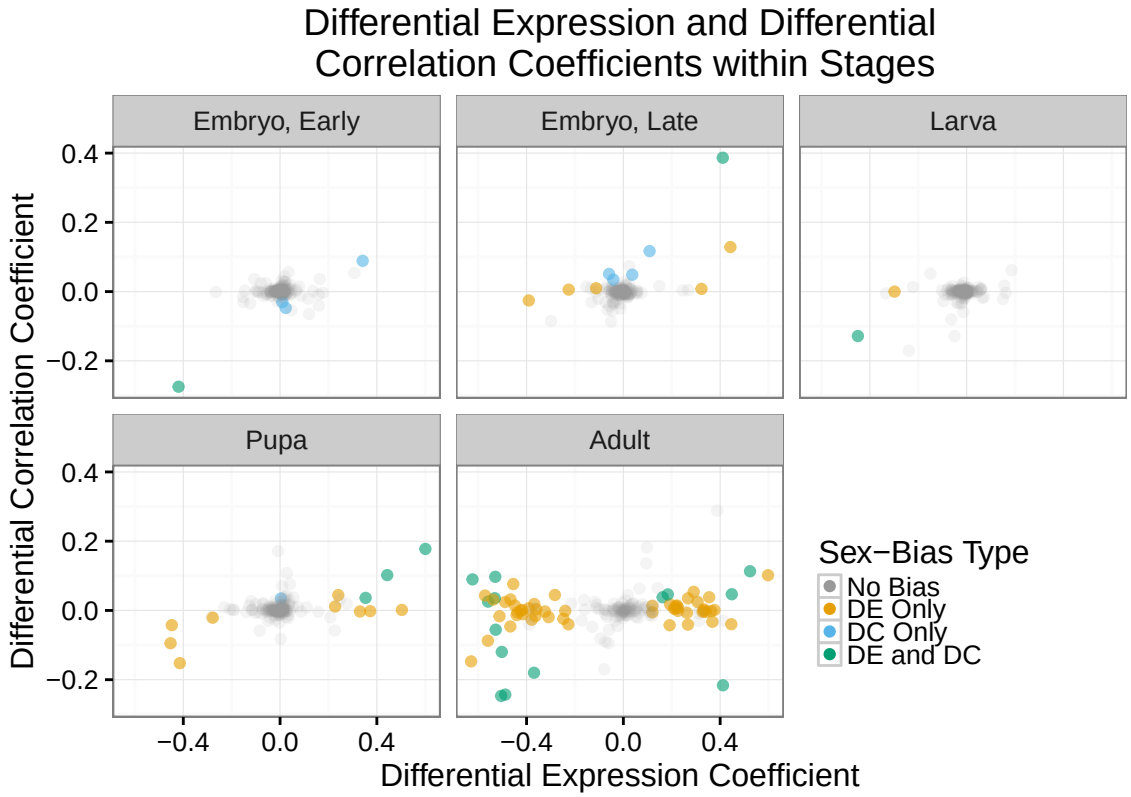


Figure 12: **Sex-Bias in Expression and Correlation at the Cluster level.** Positive values indicate male-bias, negative values indicate female-bias.

sperm anoxeme (Bettencourt-Dias et al., 2005) and Cyclin B (Nasvi2EG014042) which triggers mitotic division, and are accordingly enriched in meiosis and gametogenesis related GO terms. Cluster lightpink2 contains several gens conding for amino acid binding proteins (i.e. condensin, Nasvi2EG004100). Since male gametogenesis occurs during pupation and female gametogenesis during adulthood, the shift in sex-bias observed in these clusters is likely caused by a sex-related heterochronic shift of gametogenesis. I also note that cross-referencing the top-ranking hubs in each of those clusters (CCRE 226, 345 and 3023 respectively) with their Wasp Atlas entries reveals that other studies have found them to be moderately to extremely testes-biased (Akbari et al., 2013).

3.4.5. Differential Correlation Reveals Early Sex-Biased Transcription

Differential correlation based analyses present several discrepancies from differential expression in both timing and direction (table 7b and figure 12). No single cluster shows

significant differential sex-biased correlation in more than one stage; several clusters show differential correlation in the earliest stages, and 5 out of the 65 clusters with differential expression towards one sex in adults show opposite bias in their differential correlation.

Among the clusters that show differential correlation in early embryos, only cluster navajowhite3 is also differentially expressed. This cluster is enriched by several GO terms related to nucleosome assembly, comprising primarily modified histone genes and possibly including the histone acetyltransferase complex H4/H2A HAT (genes Nasvi2EG008990, Nasvi2EG008772, Nasvi2EG024702, and Nasvi2EG008770). One of those genes is assigned to histone H1, one to histone H2A and two to histone H2B. These histone H2A and H2B nodes currently lack sufficient homology to be assigned to an orthologous group and are likely to be modified according to a lineage-specific expansion (Rago et al., 2016, see also section 1.4.6). Histone H1 is part of the most likely hub of this cluster (CCRE108), alongside an isoform of sex-lethal interactor (Nasvi2EG016490), and bällchen (Nasvi2EG003614) whose *Drosophila* ortholog is involved in the maintenance of neuronal and germline stem-cells via histone phosphorilation (Herzig et al., 2014; Yakulov et al., 2014).

Two more clusters (lavenderblush3 and palevioletred2) show female-biased correlation during early embryogenesis. Both are also differentially over-expressed in adult males. Neither shows enrichment in informative GO terms. CCRE 493 is the most hub-like node in cluster lavenderblush3 and is comprised by the transcriptional nodes of gene Nasvi2EG018256 (a CDK inhibitor enriched in *Nasonia* testes Akbari et al., 2013), *Nasonia*'s Yellow-f protein (Nasvi2EG033442) and Nasvi2EG003903 or Inositol-trisphosphate 3-kinase A, whose *Drosophila* homolog is necessary for correct wing formation (Dean et al., 2016). The primary hub of palevioletred2 is CCRE 180, which groups two poorly characterized transcription nodes: the putative chitinase Nasvi2EG007678 and the SMYD-2 like N-lysine methyltransferase Nasvi2EG001109, both of which are enriched in *Nasonia* testes (Akbari et al., 2013). The same cluster also includes two fatty acyl-CoA reductases (Nasvi2EG017071 and Nasvi2EG025693) homologous to *Drosophila* and *Culex* male

sterility proteins. Only the cluster darkseagreen2 shows significant male-biased correlation in early embryos. Darkseagreen2 is strongly enriched in GO terms related to stem-cell fate determination, neurogenesis and down-regulation of RNAs. Its hub node CCRE 1500 contains several poorly annotated genes alongside a splicing event for Nasvi2EG022761, a homeobox-like transcription factor and Nasvi2EG006781, isoform of a testis-biased putative telomerase.

While the direction of sex-bias is generally consistent between the two regulatory modes, I find that 5 of the 20 clusters with simultaneous differential expression and correlation show different bias between expression and correlation. All of those exceptions are observed in adults. Four of these clusters (antiquewhite4, plum, plum3 and thistle3) are more expressed in females but more strongly correlated in males, whereas cluster antiquewhite2 is more expressed in males but more correlated in females. These discrepancies could be caused by differential tissue representation between the adult phenotypes, since females possess much larger gonads than males in proportion to their body. The increased proportion of gonadal tissue could increase representation of non-sex specific and male-biased gonadal transcripts in females, as well as their average expression compared to male gonadal transcripts. An increase in mean representation would affect differential expression analyses but not correlation-based ones, which rely on the relative change of node expression. This seems to be the case for cluster antiquewhite4, which as mentioned earlier is likely to be involved in gametogenesis. I also observe an enrichment for gametogenesis, neurogenesis, and histone modification associated terms in the cluster plum3, while the cluster thistle3 is enriched in GO terms related to germ cell development and splicing regulation. All genes contained in the hub nodes of those clusters show moderate testes-bias in adults (Akbari et al., 2013). Cluster plum does not show significant enrichment in gametogenesis related terms but rather is enriched in ribosomal biogenesis and RNA-processing related terms. Both genes within its hub (CCRE106) are testes-biased (Akbari et al., 2013), suggesting it may also be involved in either spermatogenesis or testicular functioning. By contrast, cluster antiquewhite2 is enriched mostly in generic OG terms related to signal transduction and

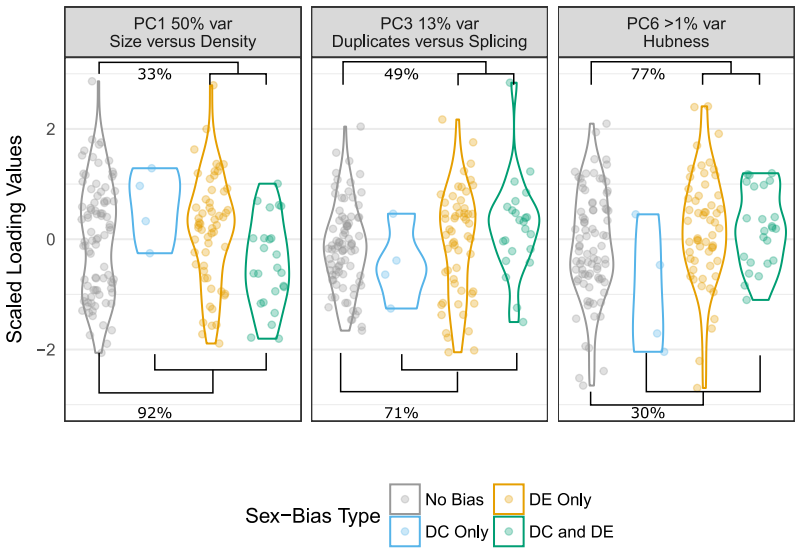


Figure 13: **Network parameters associated with sex-biased clusters**
Scaled PCA loading values indicated on the y axis. Each PC is listed with its associated level of variance explained. Within-panel percentages indicate the RI of each PC in the model set separating unbiased clusters from differentially correlated (above) or differentially expressed (below) ones.

its hub contains several isoforms of Nasvi2EG010141, a calcitonin receptor enriched in female heads (Hoedjes et al., 2015). I find it likely that this cluster may be involved in female-specific neuronal functioning and its apparent over-expression in males may be due to the relative smaller size of female brains compared to their gonads.

3.4.6. Sex-Biased Clusters Show Different Regulatory Organizations

I identify three components of cluster architechture which are significantly different (relative importance >70%) between sex-biased and non sex-biased clusters, shown in figure 13. The strongest association (RI 92%) is between differentially correlated clusters and PC 1, which is also the only factor with a significant ability to discriminate between clusters with sex-biased expression and clusters with sex-biased correlation (RI 93%). The lower scores of differentially correlated clusters on PC 1 indicate that they tend to have smaller sizes but higher density and a less centralized structure. According to theoretical models, the higher density of differentially correlated clusters would predict lower evolutionary potential via network re-wiring compared to both differentially expressed and non sex-biased clusters.

Differentially correlated clusters are also moderately associated (RI 71%) with PC 3, which is positively correlated with enrichment in duplicated genes and negatively correlated with enrichment for splicing nodes. This finding is in accordance with theories on how gene duplication can solve sexual conflict at the gene level, but are supported only for differentially correlated clusters. Taken together with the low potential of evolution by re-wiring, the enrichment in duplicates could indicate that these clusters evolve primarily by adding new genes to the existing network.

Despite the fact that PC 7 explains less than 1% of between-cluster variance, it is the only PC significantly associated with differentially expressed clusters (RI 77%). PC 7 is positively correlated with cluster centralization and negatively correlated with median clustering coefficient. The highest PC 7 scores of differentially expressed clusters indicate a more hierarchical structure, with a stronger divide between hyperconnected hubs and peripheral worker nodes. Thus, while differentially expressed clusters have average distribution of densities, their structure could still be promoting rapid turnover of regulatory interactions.

3.4.7. Sex-Biased Clusters Integrate New Genes in Regulatory Positions

In order to validate whether sex-biased clusters show faster evolution compared to non sex-biased ones, I compared the proportions of gene ages present in each category (figure 14). All types of sex-biased clusters are more frequently comprised by genes whose most ancient match is at the *Nasonia* (or wasp) taxonomic level, although the effect is more pronounced in clusters with differential correlations. Compared to clusters that show only differential expression, clusters with differential correlations appear depleted of genes from more ancient strata, such as Hymenoptera, Insecta, and Metazoa. I further combined data from the gene's age with their network properties to address whether new genes present in sex-biased clusters are more likely to be in regulatory positions than new genes in non sex-biased clusters. I tested whether gene age can predict the number of interactions with other genes using within cluster connection density and its regulatory potential using hub scores (figure 15, see materials and methods for details).

Figure 14: **Proportions of genes from each taxonomic stratum in different classes of sex-biased clusters**. Proportions reported are fold-enrichment compared to the network-wide abundances of genes from each stratum. Y axis is truncated between 0.5 and 2.5 fold enrichment.

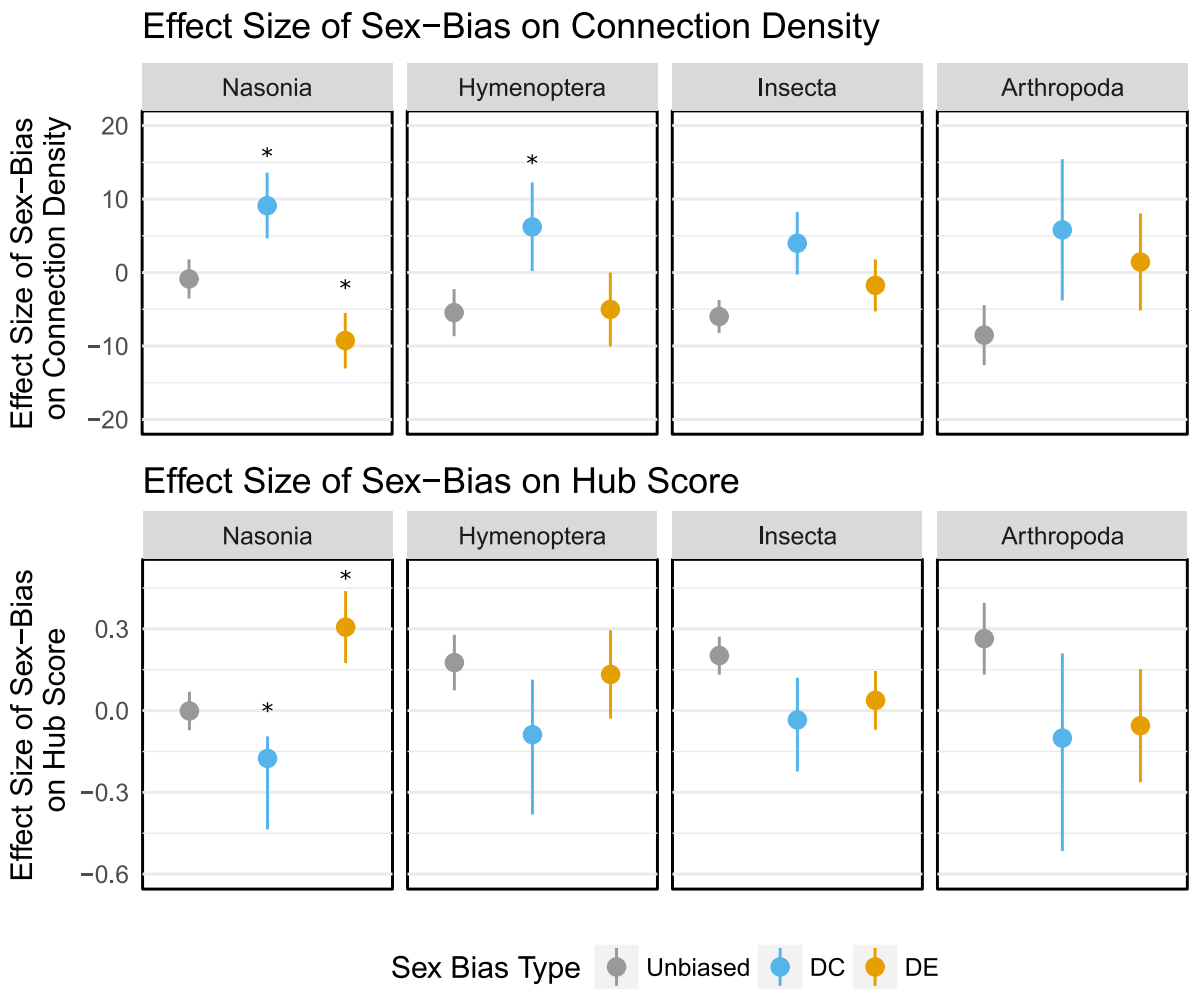


Figure 15: **Effect sizes of sex-bias categories on connection densities (upper) and hub scores (lower) of individual transcripts of different phylostratigraphic age.**
Asterisks indicate non-overlapping 95% intervals between sex-bias categories in the same phylogenetic stratum. All effects are calculated relative to the Metazoan stratum. For details on the modelling see section 3.3.9.

Nodes from different phylogenetic strata show wide variation in both density and hub-score, as indicated by the high relative importance (RI 100%) of the term stratum for both models. The density of *Nasonia* stratum nodes increases even further in differentially correlated clusters (fig figure 15), making *Nasonia*-stratum genes the strongest interactors of all phylogenetic layers. Conversely, the hub scores of transcripts in differentially correlated clusters remains constant across the different strata. Based on these two parameters, new nodes in differentially correlated clusters appear to have large number of connections but low regulatory potential, consistently with co-worker type nodes.

Nasonia and Hymenoptera-stratum nodes in differentially expressed clusters are instead characterized by high hub scores and low connectivity (100% and 95% RI, respectively). This distribution indicates that nodes from the *Nasonia* and Hymenoptera strata conform to the expectation of hubs, which bridge connections between otherwise independent group of genes and enable co-ordinated regulation. Conversely, nodes from the older Insecta, Arthropoda and Metazoa strata show high connectivities and low hub scores within differentially expressed clusters, suggesting low regulatory potential but high co-operation at the molecular level. Accordingly, Metazoan stratum nodes are enriched in protein complexes (GO:0043234, q-value ~3.1 e^{-96}), such as flagellar proteins in cluster tomato and skyblue3, and spindle formation in cluster thistle. Although using homology to assign functions to *Nasonia*-specific genes is impossible, their topological properties are highly indicative of their preferential role as regulators of sex-biased transcriptional clusters.

The different position of new genes in differentially expressed and differentially correlated clusters is consistent with their general topological properties. High density differentially correlated clusters show low capacity to evolve new regulators. Low density and high hierarchy differentially expressed clusters instead seem to allow the rapid integration of new regulators.

3.5. Discussion

My assessment of sex-bias across the development of *Nasonia vitripennis* leads to several discoveries. At the gene level, I observe a prevalence of sex-biased transcription over splicing. I find an extremely limited number of genes which shift between male and female bias across different developmental stages, suggesting developmental constraint of sex-bias. I also identify several genomic regions enriched in male and female-biased genes, one of which contains the key sex-determining gene sex-lethal. At the cluster level, I report two different types of sex-biased clusters with specific temporal expression patterns and topological properties. Differentially correlated clusters show a surprising amount of cryptic sexual dimorphism in the earliest developmental stages. Differentially expressed clusters instead have a more hierarchical structure, with new or fast-evolving genes in key regulatory positions.

3.5.1. Sex-bias at the Single Locus Level

My analyses at the gene-level suggest that regulation of whole-gene transcriptional levels may be the most frequent means to induce transcriptome-wide differentiation between sexs. I find far more loci with evidence of sex-biased gene expression than sex-biased splicing. More importantly, the majority of genes with sex-biased transcription do not show sex-biased splicing, whereas most genes with sex-biased splicing also show sex-biased transcription. This inequality suggests that transcriptional regulation might be the prevalent method of solving within-locus sexual conflict, while sex-biased splicing may in most cases be a byproduct a gene's transcriptional bias. This finding is consistent with studies in *Drosophila* development (Brown et al., 2014), which show that the majority of splice variation is observed either between tissues or between stages and that the few consistently sex-specific splice variants in adults can be attributed to sex-specific tissues. Nonetheless, I describe 1,294 (~10%) genes that show sex-biased splicing and lack sex-biased transcription in the parent gene. This proportion is more than double the frequency reported for adult *Drosophila* by Brown et al. (2014) but consistent with

earlier *Drosophila* estimates from studies aiming at the specific detection of sex-biased alternative splicing (Telonis-Scott et al., 2008; Hartmann et al., 2011). It is also worth noting that Brown et al. (2014) measured transcript expression via RNAseq technologies whereas both earlier *Drosophila* studies and this manuscript rely on microarrays. As such, more molecular data from a range of methods is required in order to validate my findings on the scope of sex-biased alternative splicing.

Despite the abundance of sex-biased transcripts, only one gene (*Feminizer*) shows consistent sex-bias across all developmental stages and the majority of sex-bias is observed in either pupal or adult stages. This pattern sets *Nasonia* apart from *Drosophila,* in which 50 to 60% of sex-biased genes retain their expression bias across all stages (Perry et al., 2014), but is closer to estimates from vertebrates (Mank et al., 2010) and potentially in accord with the pattern observed for caste-specific genes in ants (Ometto et al., 2011). Sexual conflict solution via gene duplication is moderately supported by our analyses: Differentially correlated clusters are the only clusters that show some enrichment for duplicates and the result is confounded by the high negative correlation with enrichment for alternative splicing. As such, I cannot currently determine whether this higher proportion of duplicates has arisen from duplication and subfunctionalization of conflicting genes or as a by-product of their lack of single-copy spliced genes.

3.5.2. Sex-biased Linkage Groups

I find three main genomic regions enriched in sex-biased genes, one of which contains the *Nasonioa* ortholog to the key *Drosophila* sex-determiner sex-lethal (Gempe and Beye, 2011). A possible causal explanation of genomic co-localization is that short genetic distance will lower chances of recombination between each gene. As such, co-localization allows reliable co-inheritance of different genes and enables them to evolve as a single supergene (Thompson and Jiggins, 2014). The concept of supergene has already been invoked to explain genomic clustering of several traits which provide fitness advantages only when co-expressed (Joron et al., 2011; Kunte et al., 2014). The selective advantage of

co-inheritance is also considered to promote the formation of sex chromosomes via linkage of sex-biased genes to the sex-determining locus (Charlesworth and Mank, 2010).

It would be tempting to attribute the female-bias enrichment of the linkage group containing *Nasonia*'s sex-lethal to the formation of a pseudo sex-chromosome. On the other hand, I detect significant transcriptional sex-bias for Nasonia's sex-lethal only in the adult stage, and its only detected isoform shows male-bias in the pupal stage. These findings are in accordance with current literature, which does not report a role for *Nasonia*'s sex-lethal. The clustering of female-biased genes around sex-lethal is thus unlikely to be due to the formation of a pseudo-sexual chromosome. Conversely, modeling studies predict that genomic clustering of genes involved in local adaptations may provide a substantial fitness advantage in populations that experience heterogeneous spatial environments (Yeaman, 2013). This scenario would be congruent with *Nasonia*'s ecology, which is characterized by patchy environments with widely varying local optima for sex-ratios (Werren, 1980). Robust modelling of the interplay between sex-biased linkage, haplodiploid genetics, sex-biased dispersal and skewed sex-ratios is necessary in order to assess the biological significance of this linkage group.

Non-adaptive explanations for clustering of co-expressed genes are also possible. In particular, co-expression of closely related genes may be arising only as a side effect of tandem duplication, which can generate a large number of closely located genes which may share expression pattern because of identity by descent. This scenario should be relatively easy to identify by checking whether the sex-biased genes in the region are paralogs. This seems to be the case for the male-biased linkage group 4.1, in which a series of tandem duplications for male-biased "venom allergen" proteins (orthologous group EOG8W9MM2) is present.

3.5.3. Heterochrony in Gametogenesis Drives Developmental Sex-Bias Shifts

Shifts between male and female bias in different developmental stages are observed only in 3 clusters, suggesting that developmental sexual conflict might be a prominent constraint

of sex-biased gene expression. The only 3 clusters that show sex-bias changes total 324 genes (1.1% of the whole network) and all three shift from male bias in pupae to female bias in adults. By comparison, *Drosophila* studies report sex-bias changes in 4.9% of autosomal genes and 2.9% of X-linked genes, with a stronger propensity to shift between female to male bias (Perry et al., 2014).

I find three possible reasons for the sex-bias shifts in these three clusters. First, the increase of female expression in adults could be due to the greater proportional mass of gonads present in adult females compared to males. Since RNA extractions were performed on whole animals, this could lead to increased tissue representation in females rather than males and consequent overestimation of gonadal gene expression in females. Results from differential correlation analyses show male correlation bias in four clusters that appear to be over-expressed in females. Characterization of these clusters reveals enrichment in testes-related processes, leading me to believe that, at least in these cases, tissue bias in adults is sufficiently strong to reverse measured gene expression bias. If correct, these findings would require a conservative reinterpretation of adult sex-bias estimates and prompt their validation via tissue-specific transcriptomic analyses. To date, only one study separated gonads from carcassess in males and females before sequencing (Tennessen et al., 2014). While their findings are similar to those obtained by whole organism sequencing in Wang et al. (2015) the study focused on genes with at least 100 fold expression differences: an amount of sex-bias most likely sufficient to overcome tissue bias.

A second possibility is that these shifts in sex-bias direction may be attributed to an adaptive heterochronic shift in *Nasonia*'s gametogenesis. *Nasonia* spermatogenesis peaks during pupation while its oogenesis occurs primarily during the adult stage (Whiting, 1968). It follows that genes involved in gametogenesis will be under selection for earlier peak expression in males than females. This scenario would lead to developmental genomic conflict on the timing of gametogenesis-related genes but not on their function, since they are likely to be involved in the same process (gametogenesis) in both sexes, albeit in different developmental stages. As such, I would expect rapid evolution of their regulation,

but little if any impact on their protein evolutionary rates.

A third possibility is that these clusters may indeed be subject to developmental sexual conflict by being involved in different sex-specific processes at different developmental times. Such patterns have been reported in the silkworm *Bombyx mori* (Zhao et al., 2011) and *Drosophila melanogaster* (Perry et al., 2014), although in both cases the shift observed is from female to male bias. Female to male bias shifts are present in the *Nasonia* developmental transcriptome and comprise a significant proportion of pre-pupal bias when testing transcripts individually but do not form coexpressed clusters. Finally, I point out that transcripts with pre-pupal sex-bias are significantly more prone to shifts in sex-bias direction, suggesting that early developmental stages may possess greater flexibility in their gene regulation.

3.5.4. Sex-Bias in Early Development

I identified several cryptic early regulatory events using complementary analyses based on both differential expression and correlation. Embryonic stages in particular show little differentiation between sexes when relying exclusively on differential expression, but reveal several hidden co-regulatory events when analyzed using differential correlation methods. For instance, only one cluster comprising 79 genes is differentially expressed in early female embryos, compared to three differentially correlated clusters containing a total of 373 genes. Differential cluster expression identifies 156 male-biased genes in late embryos, whereas differential cluster integration reveals 661 male-biased genes. This suggests that small proportional changes in the expression of multiple transcripts may play a previously unrecognized role in early sexual differentiation. Intriguingly, two clusters with early sex-biased correlation show male-biased expression in adults and varying degrees of testicular enrichment. More detailed analysis of the genes included in these clusters reveals a clear enrichment for putative male fertility factors as well as developmental regulators.

Histone and histone-modification enzymes are enriched and occupy hub positions in the

only early embryonic cluster that shows sex-bias according to both our measures. While overexpression of histones in diploid females is expected due to the higher amount of DNA in their cells, the female-specific increase in correlation suggests that histones and their modification enzymes may be involved in sex-specific interactions in early embryogenesis. This result is especially interesting in light of the ongoing debate on *Nasonia*'s sex-determination mechanism. While there is now consensus on the need for a silencing mechanism of maternal *Feminizer* expression (Verhulst et al., 2010a, 2013), investigations to identify which mechanism is involved have so far been inconclusive. Several recent papers aimed at investigating the role of DNA methylation have shown that genes subject to DNA methylation show less variation across evolutionary and developmental space (Park et al., 2011; Wang et al., 2013) and there is very limited evidence for sex-biased differential methylation in adults (Wang et al., 2015). Our study reinforces a lack of support on DNA methylation as a mechanism for sex-biased genome imprinting, supporting the modification of specific histones as a possible alternative. Since the genome copy carried by sperm is bound by sperm-specific protamines (Tennessen et al., 2014) rather than histones, such a mechanism would provide a robust means of erasing only paternal imprinting without the need for divergent histone markings in adults. Histone-mediated wasp-specific control of sex-determination would also be consistent with the findings in section §1, which identify histone genes as a primary target of lineage-specific gene family expansions in the wasp clade and potentially with those of Xiao et al. (2013), which find a consistent enrichment of histone-related GO terms in genes specific to the fig wasp *Ceratosolen solmsi* compared to an older and significantly less complete version of the *Nasonia* gene set.

3.5.5. Network Structure of Sex-Biased Clusters

Sex-biased clusters show high proportions of wasp-specific genes (figure 14); which occupy different positions within their networks (table 11). In differentially correlated clusters, *Nasonia*-specific genes are highly connected but have low hub scores. This result is consistent with my hypothesis that dense clusters would be more constrained in the

evolution of new regulators due to pre-existing regulatory overlap between their members. New genes would thus be likely to correlate strongly with several genes rather than forming specific interactions. The high density of differentially correlated clusters would also impede the rise of internal regulators since each pair of nodes is more likely to be already connected than in a sparse network, reducing the need and impact of novel co-ordinators.

Differentially expressed clusters are more hierarchically organized, as measured by their lower density and higher centralization. One of my initial hypotheses is that both sparsity and hierarchy may facilitate the emergence of new regulators. *Nasonia*-stratum genes in differentially expressed clusters are sparsely connected and show the highest hub-scores. Their preferential position between groups of not otherwise connected nodes is characteristic of gene regulators and reveals a propensity of differentially expressed clusters to incorporate new genes in control positions. While a sparser network would increase the odds of a new node to become a regulator, the prevalence of *Nasonia*-stratum nodes in hub positions remains surprising when compared to that of equally sparse non sex-biased clusters, which are instead occupied by nodes from the Arthropoda and Insecta strata.

As the closest available genomes for phylostratigraphic comparisons belong to either bees or ants, *Nasonia*-stratum genes could have arisen at any point after the split between wasps and the other hymenopteran lineages (~180 Mya, Werren et al., 2010; Misof et al., 2014). The method of phylostratigraphic dating has also been shown to be prone to bias (Moyers and Zhang, 2014), particularly when attempting to detect deep matches for short or rapidly evolving genes whose sequence similarity rapidly degenerates below homology criteria. Considering that sex-biased genes have indeed often been observed to have higher evolutionary rates (Wang et al., 2015) it is likely that a portion of *Nasonia*-stratum genes will be consisting of rapidly diverging genes from older strata.

Depending on the extent of phylostratigraphic bias, I can interpret these findings in two ways. Either new genes are indeed more readily integrated in key regulatory positions within differentially expressed networks (low phylostratigraphic bias scenario) or genes in key positions in differentially expressed networks in the *Nasonia* clade have rapidly

mutated beyond homology criteria (high phylostratigraphic bias scenario). Both scenarios imply that genes in sex-biased clusters show significant evolutionary differences compared to non-biased ones, and that those differences are closely related to the genes' positions within the regulatory network. Rapid integration of novel genes into regulatory positions of sex-specific networks has already been documented multiple times in *Drosophila* for mechanisms as diverse as male fertility (Ding et al., 2010; Chen et al., 2012) and courtship specificity (Dai et al., 2008), whereas over 75% of the caste-biased genes in the wasp *Polistes canadensis* lack homology outside of the species (Ferreira et al., 2013).

The pattern of rapid acquisition I observe in differentially expressed clusters in particular is consistent with Developmental Systems Drift (DSD, True and Haag, 2001; Haag, 2014), an evolutionary model which allows for the change of the underlying regulatory pathways *via* stochastic drift while conserving the final result through the repeated emergence and loss of redundant regulators. A similar pattern is already observed for the primary regulators of sex-determinations across Insecta and Hymenoptera (Verhulst et al., 2010b; Koch et al., 2014) and could be indicative of a general feature of sexual development. With rapid rates of molecular evolution and a strong constraint for retaining two functional phenotypes, sexually dimorphic development might indeed be the optimal scenario for the prevalence of DSD.

3.6. Conclusions

The characterization of *Nasonia*'s sexual development offers a powerful tool for future inquiries in insect biology and reveals numerous interesting properties about the evolution of sexual dimorphism in this haplodiploid species. I provide for the first time a detailed comparison of the interplay between transcription and splicing over *Nasonia's* sexual development, assessing the prevalence of transcription and noting instances of splicing which are most likely to mediate sexual conflict. My analyses of early developmental expression reveal that differentially correlated sets of transcripts could play a previously unrecognized role in the onset of sexual differentiation and possibly sex-determination

itself. Despite the lack of genetic sex-determination, I find at least three genomic regions enriched in sex-biased clusters, one of which includes homologs of key sex-determinants. Several scenarios could explain their presence, spanning from selective advantage of their co-inheritance to non-adaptive linkage hitchhiking. Discriminating between these options will require modelling that integrates knowledge about *Nasonia*'s genome with its ecology and taxonomy.

Compared with other species, *Nasonia*'s sex-bias is strongly developmentally restricted, with few transcripts showing sex-bias in multiple stages. While I observe several cases of male to female bias transitions between stages, they remain mostly confined to meiosis-related genes or contrasts between pre and post-pupation stages. The recurrence of sex-bias in the same direction in the majority of transcripts supports strong constraint as the same gene will tend towards the same sex-bias direction across different stages. The prevalence of stage-specific sex-bias and the fact that transcript which shift in sexd-bias direction do so during pupation underscores the importance of treating different life-stages as factors of interest in order to correctly understand gene expression evolution.

Finally, my characterization of two main classes of sex-biased clusters via network analyses better understanding of the role of fast and novel genes within co-regulated clusters. While all sex-biased clusters showed enrichment for novel genes, I find that they occupy fundamentally different positions in their networks, acting as potential regulators only in differentially expressed clusters. This finding provides a first empirical confirmation for hypotheses on how sparsity and hierarchy can facilitate the rapid evolution of regulatory structures, but should be critically re-examined to determine whether this effect is general or rather restricted to specific conditions. Comparative studies on the evolution of pseudoparasitism in wasps would be especially useful, as this ecological shift is also known to involve rapid genomic restructuring and may explain a sizeable portion of its lineage-specific genes and possibly interact with sex-biased development.

Nevertheless, the observation that novel genes can be incorporated into pivotal regulatory positions in sex-biased clusters poses a critique to the evo-devo assumption that regulators

are conserved over time, supporting instead the model of phenotypic stasis and regulatory reshaping that characterizes developmental system drift.

4. General Conclusions

In this thesis, I develop and apply tools for the study of sexual dimorphism in the development *Nasonia vitripennis*. My investigation provides different contributions to different communities. In the case of *Nasonia* and Hymenoptera biology, I provide a previously missing characterization of developmental expression patterns in general and sex-bias pattern in the specific. In the case of sex-bias and sexual conflict studies, I provide a detailed analysis of the different means by which an organism without sex chromosomes induces sexual dimorphism and compare them to the literature on models with genetic sex-determination. In the case of Systems Biology, I provide a proof of concept set of analyses that demonstrate the power of explicitly integrating evolutionary thinking in investigations.

In chapter one, I assess the quality of an improved gene set and employ it to detect genome evolution's peculiarities in the wasp branch of the tree of life. The results from phylogenetic expansion analyses in particular are consistent with my findings in chapter three, where I identify a function of the *Nasonia*-specific histone genes in early sexual differentiation and possibly sex-determination. Intriguingly, the histone genes involved in early sex-biased expression patterns are not the same identified by *Nasonia*-specific family expansions or faster evolution along the wasp branch. While their sequence homology is sufficient to place them firmly among histones, they all appear to have either arisen after the split from the nearest species or mutated rapidly enough to fall outside of orthology assignment criteria, further adding interest to the functions of this protein family in wasps.

In chapter two, I design a simple algorithm for the detection of novel splice events based on experimental data. Developing this algorithm was necessary for several reasons. The dataset used in chapter three had been generated using microarrays: a platform capable of producing large amounts of data for competitive costs but whose data analyses methods are mainly gene-based rather than splicing-oriented. As my interest lies in comparing the role of both splicing and transcription in polyphenisms, development of a specific pipeline able to detect experiment-specific events and disentangle the two types of signal

was necessary. Although the use of the FESTA algorithm as a stand-alone method for transcript annotation would be inadvisable, data-based estimation of novel alternative splicing events is crucial in any experiment which previously undescribed transcripts may be generated; as relying on previously published data may cause false negatives. When paired with stringent downstream quality control and further testing for involvement in the biological process of interest, the FESTA algorithm enabled me to detect that splicing plays a rather minor component in *Nasonia*'s sexual development. More importantly, the use of splicing ratios allowed me to represent potential splice events as statistically independent from parent gene expression, enabling the construction of a hybrid transcription and splicing network in chapter three and consequently the analysis of both processes and their interactions' role in sexual development.

In chapter three, I employ the methods developed in the rest of this thesis to tackle the unanswered questions posed by sexual dimorphism *via* network analyses of *Nasonia*'s developmental transcriptome. Alongside the methods already mentioned, I also developed and implemented a permutation-based algorithm to detect sex-biased changes in correlations among genes, based on the assumption that novel functions can be exerted not only through the expression of different genes, but also by establishing specific interactions and combinations among transcripts. The results of differential correlation and differential expression analyses are mostly convergent, yet I find that differential correlation can complement differential expression as it appears to be less sensitive to tissue bias and more powerful in detecting small coordinated changes in groups of transcripts. Through the joint application of differential expression and correlation at the cluster level, I distinguish between two categories of sex-biased clusters, each with specific topological properties. Differentially coexpressed clusters appear to be small, dense and democratic. Differentially expressed clusters are instead sparser and more hierarchical. Both classes of sex-biased clusters show preferential integration of novel genes compared to non sex-biased clusters, but each incorporates them in different positions in their networks. New genes in differentially correlated clusters occupy lower-level positions, with several connections but low hub

potential, consistent with a large overlap in the regulation of cluster members. New genes in differentially expressed clusters on the other hand show the highest regulatory potential of all gene ages, suggesting highly selective regulatory interactions in which the regulators can either be efficiently replaced by younger genes with similar regulatory potential or rapidly evolve as long as their regulatory effect remains unchanged.

The final picture of sex-bias and sexual conflict in *Nasonia* is, perhaps unsurprisingly, one of novelty and rapid evolution. The intersection of sexual dimorphism with haplodiploid genetics and the absence of sex chromosomes places an intense selective pressure on loci involved in the differentiation between sexes without the "safe havens" provided by non-recombining sex-specific genomic regions or recessive loci. Fast and novel gene families are all overrepresented among sex-biased clusters. When looking at the whole of *Nasonia*'s development I was also able to discover that most transcriptional sex-bias is highly restricted to specific developmental stages. Developmental sex-bias restriction could conceivably play a role in rapid evolution, as developmentally restricted genes tend to evolve faster due to lower pleiotropic constraints. Given the incremental nature of development, transient sex-bias may be pivotal also in generating large-scale dimorphism by causing alterations in the starting conditions which propagate throughout morphogenesis long after the initial triggers have disappeared. Both of these factors would be missed if we focused exclusively on adult expression. Finally, the prevalence of novel genes in regulatory positions of sex-biased networks demonstrates how network and evolutionary biology can work in tandem to reveal how the molecular mechanisms that give rise to alternative phenotypes can arise across phylogenies.

A. Appendix

A.1. Additional Figures and Tables

Table 8: Histone genes present in OGS2.0 annotated with presence or absence of lineage-specific expansions. NA entries were not assigned to orthologous groups at the level of Hymenoptera.

Name	ODB6 OG ID	Expanded?
histone deacetylase 3 (92%a)	EOG6N8PM3	No
histone-lysine N-methyltransferase SETDB1 (65%a)	EOG6BRV1R	No
histone acetyltransferase Tip60 (83%a)	EOG6H44K2	No
jmjC domain-containing Histone demethylation protein 3B (66%A)	EOG669P90	No
Histone-lysine N-methyltransferase NSD2 (45%A)	EOG60GB5Q	No
Histone acetyltransferase MYST4 (66%U)	EOG63N5W3	No
Histone-lysine N-methyltransferase. H3 lysine-79 specific (60%U)	EOG612JM9	No
nucleosomal Histone kinase 1 (49%A)	EOG6H18B6	No
jmjC domain-containing Histone demethylation protein (63%A)	EOG69CNPT	No
Histone deacetylase complex subunit SAP130 (49%U)	EOG644J2T	No
Histone-lysine N-methyltransferase. H3 lysine-9 specific 5 (Fragment) (55%U)	EOG6WSTRM	No
Histone-lysine N-methyltransferase PR-set7 (63%A)	EOG6H9W32	Yes
Histone-lysine N-methyltransferase PR-set7 (56%A)	EOG6H9W32	Yes
Non-histone protein 10 (79%U)	EOG65X6CB	No
Histone-lysine N-methyltransferase NSD2 (45%A)	NA	NA
Histone deacetylase 4 (79%U)	EOG64QRGN	No
jmjC domain-containing histone demethylation protein 1. putative (69%a)	EOG6KSN0T	No
histone chaperone asf1 (88%a)	EOG64TMRB	No
Histone RNA hairpin-binding protein (57%A)	EOG625491	No
histone acetyltransferase type B catalytic subunit. putative (70%a)	EOG69ZW4X	No
Histone-lysine N-methyltransferase. H4 lysine-20 specific. putative (19%U)	NA	NA
set1/Ash2 histone methyltransferase complex subunit ASH2 (77%a)	EOG66DJHV	No
JmjC domain-containing histone demethylation protein 1D (60%U)	EOG6NK99J	No
Histone demethylase UTX (92%U)	EOG6D51FJ	Yes
Histone demethylase UTX (94%U)	EOG6D51FJ	Yes
Histone-lysine N-methyltransferase Suv4-20 (56%A)	EOG64MW6W	No
histone acetyltransferase MYST1. putative (80%a)	EOG65QFV4	No
Histone-lysine N-methyltransferase SETMAR (46%A)	EOG61894T	No
Histone demethylase JARID1A (73%U)	EOG6X69Q3	No
jmjC domain-containing Histone demethylation protein 2B (50%A)	EOG62JM67	No
jmjC domain-containing Histone demethylation protein 2C (81%A)	EOG64J10T	No
Histone-lysine N-methyltransferase SUV39H2 (46%A)	EOG663XT6	No
histone-binding protein Caf1. putative (99%a)	EOG6RBP0X	No
histone deacetylase Rpd3 (87%a)	EOG6N2Z47	No
histone deacetylase complex subunit SAP18. putative (81%a)	EOG6BZKK4	No
lysine-specific Histone demethylase 1A (77%A)	EOG61C5C8	No

Table 8: Histone genes present in OGS2.0 annotated with presence or absence of lineage-specific expansions. NA entries were not assigned to orthologous groups at the level of Hymenoptera.

Name	ODB6 OG ID	Expanded?
lysine-specific histone demethylase 1A (83%a)	EOG6905R9	No
Histone-lysine N-methyltransferase pr-set7 (Fragment) (23%U)	NA	NA
Histone deacetylase (51%A)	EOG6DJHBF	No
Histone-lysine N-methyltransferase Suv4-20 (52%A)	NA	NA
histone-arginine methyltransferase CARMER. putative (89%a)	EOG68SF85	Yes
histone-lysine N-methyltransferase E(z) (84%a)	EOG6BZKHM	No
histone-arginine methyltransferase CARMER. putative (87%a)	EOG68SF85	Yes
Histone-lysine N-methyltransferase Suv4-20 (52%A)	NA	NA
sin3 histone deacetylase corepressor complex component SDS3. putative (84%a)	EOG6VX0NM	No
Histone-lysine N-methyltransferase SETD1B (Fragment) (57%U)	EOG6F7M0V	Yes

Paralog locus consensus	Inparalogs	Uniquepar
Count of paralog families (first locus)	874	441
Paralogs on different scaffold	1,795	686
Paralogs >10kb distant on same scaffold	64	15
Paralogs <10kb, same orientation, non overlap	119	70
Paralogs <10kb, reversed orientation	27	23
Gene spans overlap (CDS overlap uncertain)	19	8

Table 9: **Consensus in the location of the OGS2 gene set on the genome assemblies of sibling species Nasonia *longicornis* and N. *giraulti*, including recent, high identity paralogs.** Almost all OGS2 genes are located on 2 sibling species draft assemblies Werren et al. (2010), using GMAP Wu and Watanabe (2005) transcript mapping. Paralog locus consensus patterns are tabulated for inparalogs (sharing orthology to other species) and uniquepar (lacking strong homology to other species). Of the total paralog families, each with several genes, most paralogs are on different scaffolds for all species. The counts of tandem paralogs with different separations are indicated.

Gene set	Average homology bitscore	Protein size deviation from median	Percent shorter than 2 SD from median
Nasonia OGS2	727.6	-7.70	3.2
Nasonia NCBI	722.3	-7.80	2.7
Nasonia OGS1.2	683.5	-12.7	4.0
Apis	733.9	-0.30	2.4
Harpegnathos	694.3	-30.0	7.3
Tribolium	552.0	-26.1	4.5
Drosophila	508.7	54.5	1.3

Table 10: **Gene set quality measurements**. Including deviation of protein size from the group median, and maximal bit score per species in pairwise comparisons within the arthropod orthology groups. The bit score measures both gene model artefacts of alternative gene sets within species and evolutionary divergence. Protein sizes may be more evolutionarily conserved, and may detect artefacts across and within species. See materials and methods for details on how each score is generated.

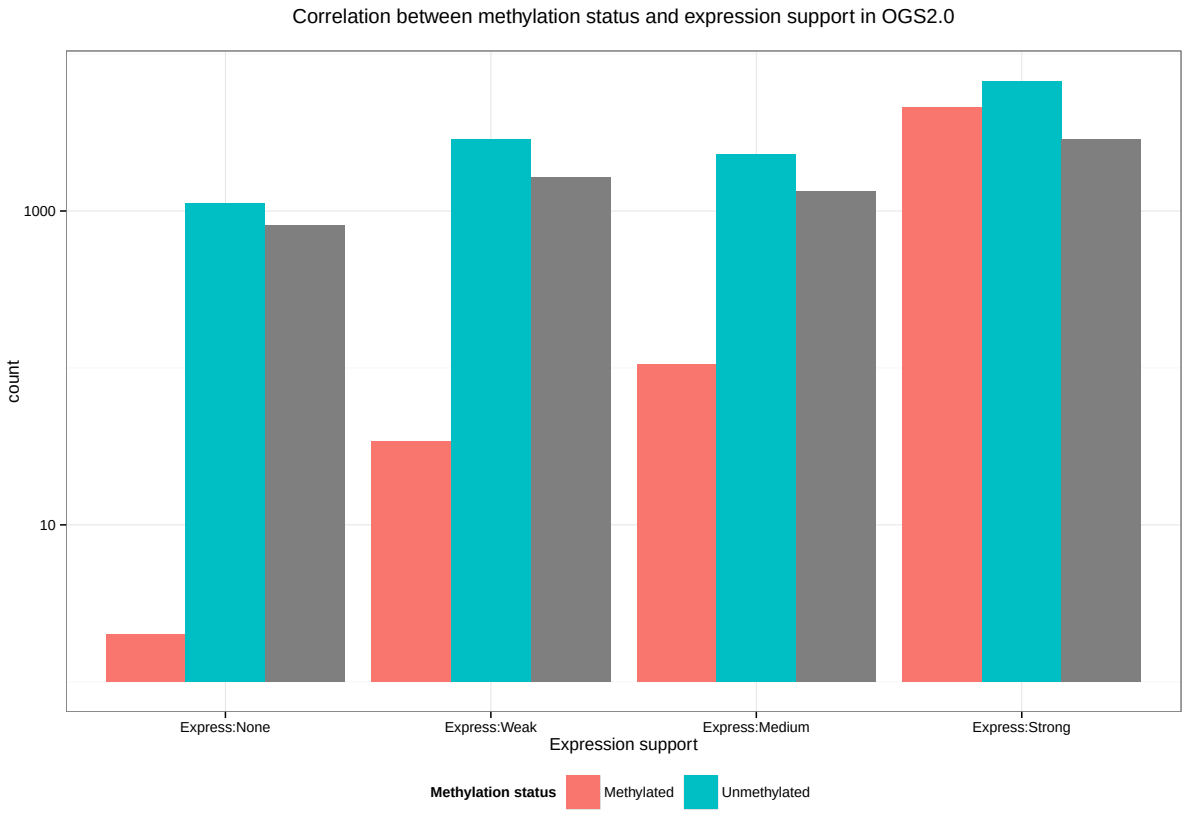


Figure 16: **Log counts of methylated and unmethylated genes in different classes of expression support**.

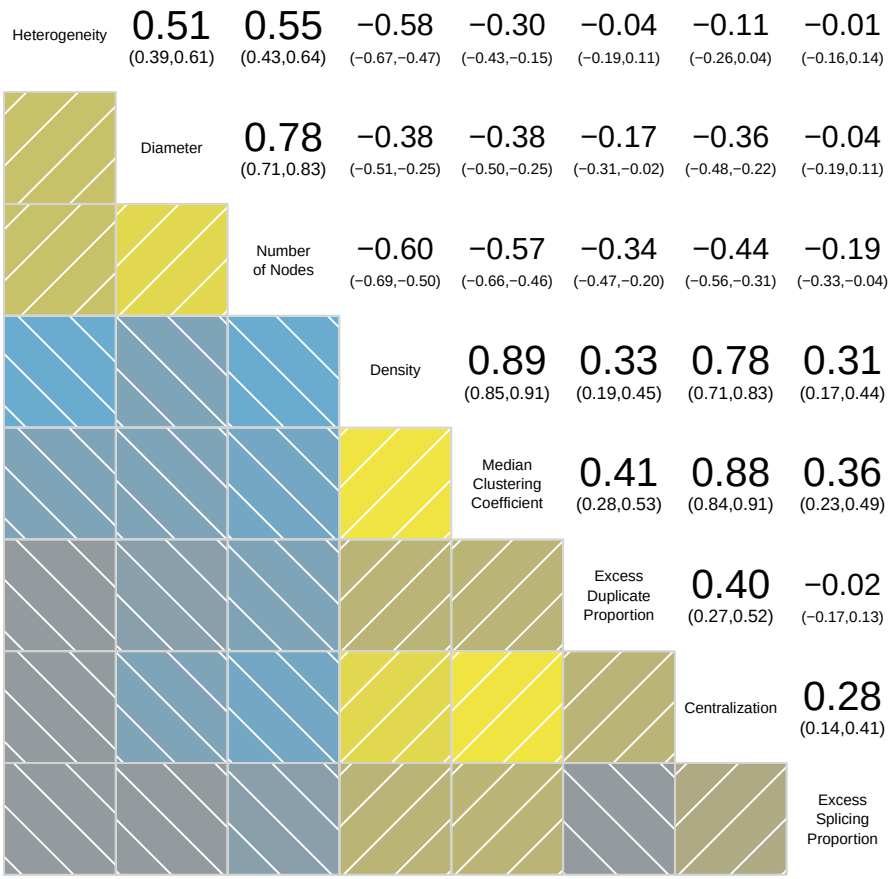


Figure 17: **Correlations between different cluster parameters in the *Nasonia* developmental network**
Yellow squares in the bottom left corner indicate positive correlations, blue ones negative. Lighter shades are more significant than darker ones. Numbers at the top right corner indicate the Pearson correlation score with confidence intervals in parentheses.

Factor	Estimate	RI
DE	-7.30	0.97
DI	0.26	1.00
nNodes	0.14	1.00
Hymenoptera	-4.25	1.00
Insecta	-4.96	1.00
Arthropoda	-7.89	1.00
Metazoa	0.68	1.00
DE:Hymenoptera	3.38	0.95
DE:Insecta	6.85	0.95
DE:Arthropoda	10.31	0.95
DE:Metazoa	7.72	0.95
DI:Hymenoptera	-2.29	0.91
DI:Insecta	-4.31	0.91
DI:Arthropoda	-2.83	0.91
DI:Metazoa	-7.31	0.91

(a)

Factor	Estimate	RI
DE	0.197	1.00
DI	0.167	1.00
nNodes	-0.003	1.00
Hymenoptera	0.172	1.00
Insecta	0.226	1.00
Arthropoda	0.339	1.00
Metazoa	0.022	1.00
DE:Hymenoptera	-0.122	1.00
DE:Insecta	-0.256	1.00
DE:Arthropoda	-0.416	1.00
DE:Metazoa	-0.259	1.00
DI:Hymenoptera	0.021	0.23
DI:Insecta	0.046	0.23
DI:Arthropoda	0.024	0.23
DI:Metazoa	0.051	0.23

(b)

Table 11: **Predictors of Connection Density (11a) and Hub Scores (11b)** with Model-Averaged Effect Sizes and Relative Importances, or probability that the factor in question is included in the best model. Stratum coefficients are relative to the *Nasonia* stratum. See section 3.3.9 for details.

- Additional File 1: Model selection table for models comprising different combinations of factors with a putative role in characterizing genes with and without annotated isoforms.

- Additional File 2: OrthoDB6 BUSCO (Benchmarking Universal Single Copy Orthologs) genes present in multiple copies in OGS2

- Additional File 3: OGS2 genes whose ortholog groups are characterized by lineage-specific expansions or contractions.

- Additional File 4: Protein evolutionary distances of OGS2.0 genes compared to ant and bee lineages, residuals distances after model fitting and fast/slow evolving categorization at the 5th and 20th quantile threshold.

- Additional File 5: Protein alignment of the OG EOG6R4ZDK (hymenopteran histone H3). Clipped to include only residues shared between all genes.

- Additional File 6: Genes with more than 10 isoforms present in OGS2
- Additional File 7: Script containing the functions used in the FESTA algorithm
- Additional File 8: Complete annotation of each transcript in the *Nasonia* developmental transcriptome network
- Additional File 9: Complete annotation of each cluster in the *Nasonia* developmental transcriptome network

A.1.1. Code for the FESTA algorithm

Listing 1: Code for the two main function of the FESTA algorithm

```
## FESTA algorithm
# Load required packages and utility functions
require(amap)
require(plyr)

##### Splicing detection
##### Cluster genes according to reciprocal correlations, then iteratively cut tree (bottom up) until
     one cluster is the most expressed or tied for expression across all samples. Uses hcluster from
     amap package for clustering

## Required input data:
## data.frame with one row per exon and one column per sample, plus
## "geneID" column with unique gene identifier
## "exonID" column with unique exon identifier

# Example data
# geneID <- paste("gene",100:500, sep = "")
# exonID <- paste(merge(geneID,c(1:10))[,1], merge(geneID,c(1:10))[,2],sep = "exon")
# exprData <- matrix(log2(rbinom(n = 4010*10, size = 1000, prob = .3)), ncol = 10)
# exampleData <- data.frame(geneID = geneID,
#                           exonID = exonID,
#                           Evalue = exprData)

## Parameter description
## exceptions:
## signDigits: number of digits rounded from expression scores for ranking calculations
## distMethod: distance metric used by the clustering algorithm (default: correlation), see function
     hcluster from package amap for more information
## link: agglomeration method used by the clustering algorithm (default: complete), see function
     hcluster from package amap for more information
## nbproc: number of subprocess for parallelization (default: 1), see function hcluster from package
     amap for more information

ClusterExons <- function(data = NULL, exceptions = ceiling(x = (ncol(data)-2)*.1), signDigits = 3,
     distMethod = "correlation", link = "complete", nbproc = 1) {
  require(amap)
  require(plyr)
  ExonAssTable <- list()
  exceptions <- exceptions/(ncol(data)-2)
  row.names(data) <- data$exonID
  for (gID in unique(data$geneID)){
    Evalues<-data[which(data$geneID%in%gID),-grep(pattern = "ID",x = names(data))]
    if (nrow(Evalues)<2) {
      ExonAssTable[[gID]]<-as.data.frame(matrix(row.names(Evalues), ncol = 1))
      names(ExonAssTable[[gID]])<-"exonID"
      ExonAssTable[[gID]]$splicing_category<-"single_expressed_exon"
      ExonAssTable[[gID]]$clusters<-0
      ExonAssTable[[gID]]$clusterranks<-1
      ExonAssTable[[gID]]<-ExonAssTable[[gID]][c("clusters", "exonID", "clusterranks", "
     splicing_category")]
    } else {
      tree<-hcluster(Evalues, method = distMethod, link = link, nbproc = nbproc)
          # evaluate relative times it is ranked as first
      for (trans in nrow(Evalues):1){
        Evalues$clusters<-cutree(tree, k = trans)
        clusterranks<-ddply(Evalues, .(clusters), colwise(median))
        if (nrow(clusterranks)==1){ # there are no subranking isoforms
          ExonAssTable[[gID]]<-as.data.frame(matrix(row.names(Evalues), ncol = 1))
          names(ExonAssTable[[gID]])<-"exonID"
          ExonAssTable[[gID]]$clusters<-0
          ExonAssTable[[gID]]$clusterranks<-1
          ExonAssTable[[gID]]$splicing_category<-"unspliced"
          ExonAssTable[[gID]]<-ExonAssTable[[gID]][c("clusters", "exonID", "clusterranks", "
     splicing_category")]
          break} else {
            clusterranks<-apply(clusterranks[,-1], 2, function(x){
                        rank(-round(x, digits = signDigits),
                        ties.method = "min", na.last = T)
                        })
            row.names(clusterranks)<-unique(Evalues$cluster)
            clusterranks<-apply(clusterranks, 1, function(x){sum(x==1)/ncol(clusterranks)})
                        # control that there is only one exon group which consistently ranks 1st or
     tied allowing for exceptions
            if (sum((clusterranks)>=(1-exceptions))==1) {
              # create table with exon subcluster assignments
                          matchmaker<-Evalues[,"clusters", drop=F]
                          # store subcluster ranks
              clusters<-as.data.frame(clusterranks)
                          # annotate subcluster ranks with subcluster IDs
              clusters$clusters<-c(1:nrow(clusters))
                          # merge exon with subcluster ID and ranks
              matchmaker<-join(matchmaker, clusters, by="clusters", type="left")
                          # add exon names
              row.names(matchmaker)<-row.names(Evalues)
              matchmaker$exonID<-row.names(matchmaker)
              ExonAssTable[[gID]]<-matchmaker
              ExonAssTable[[gID]]$splicing_category<-"spliced"
              ExonAssTable[[gID]]<-ExonAssTable[[gID]][c("clusters", "exonID", "clusterranks", "
     splicing_category")]
```

```
            break} else {next}}}
    }
  }

  # this step causes problem in R below version 2
  ExonAssTable<-ldply(ExonAssTable,rbind)
  names(ExonAssTable)[1]<-"geneID"
  # assign constitutive/specific status
  ExonAssTable$constitutive<-ifelse((ExonAssTable$clusterranks>=(1-exceptions)),"constitutive","
     facultative")
  # merge cluster assignments into unique IDs with designation of constitutiveness
  ExonAssTable$transcriptID<-as.factor(paste0(ExonAssTable$geneID, "_t", ExonAssTable$clusters, ifelse(
     ExonAssTable$constitutive=="constitutive", "_con","_fac"), sep=""))
  # code ID variables as factors
  ExonAssTable$geneID <- as.factor(ExonAssTable$geneID)
  ExonAssTable$splicing_category <- as.factor(ExonAssTable$splicing_category)
  ExonAssTable$constitutive <- as.factor(ExonAssTable$constitutive)
  ExonAssTable
}


##### Average expression values based on unique eigenexon IDs

## Required input data:
## data.frame with one row per exon and one column per sample, plus
## "geneID" column with unique gene identifier
## "transcriptID" column with unique exon identifier as per assigned via ClusterExons
## "constitutive" column identifying which transcripts are from constitutive nodes as per assigned via
     ClusterExons

## Parameter description:
## splicingRatios: Logical. If FALSE, expression from all entries is reported on the same scale. If
     TRUE, expression from splicing entries is normalized by their gene's constitutive expression score
     , generating splicing ratios
## NAcorrection: Logical. Applicable only if splicingRatios is TRUE. If TRUE, splicing ratios higher
     than 1 are set to 1 and NA/NaN/infinity values to 0. This accounts for experimental error in
     measurements.

AverageExons <- function(data = NULL, splicingRatios = F, NAcorrection = F){
  if (splicingRatios == F) {
    out <-ddply(.data = data, .variables = .(transcriptID), numcolwise(median), na.rm = T)
    out[order(out$transcriptID),]
  } else {
    # split into constitutives and facultatives
    ConTranscripts <- data[which(data$constitutive=="constitutive"),]
    FacTranscripts <- data[which(data$constitutive!="constitutive"),]
    # average exon values within transcripts
    ConTranscripts <- ddply(ConTranscripts, .variables = .(geneID, transcriptID), numcolwise(median),
     na.rm = T)
    FacTranscripts <- ddply(FacTranscripts, .variables = .(geneID, transcriptID), numcolwise(median),
     na.rm = T)
    FacSplicing <- apply(FacTranscripts, 1, function(Fac){
      Spl <- as.numeric(Fac[-grep(pattern = "ID", x = names(FacTranscripts))])
      Con <- ConTranscripts[which(Fac["geneID"]==ConTranscripts$geneID),-grep(pattern = "ID", x = names
     (ConTranscripts))]
      Spl/Con
    })
    FacSplicing <- cbind(FacTranscripts[,c("geneID","transcriptID")],ldply(FacSplicing))
    if (NAcorr == T) {
      # set NA/NaN and infinity scores to 0, set scores greater than 1 to 1
      FacSplicing[,sapply(FacSplicing, is.numeric)]<-apply(FacSplicing[,sapply(FacSplicing, is.numeric)
     ],c(1,2),function(x){
            as.numeric(ifelse(is.na(x), 0, x))
            })
      FacSplicing[,sapply(FacSplicing, is.numeric)]<-apply(FacSplicing[,sapply(FacSplicing, is.numeric)
     ],c(1,2),function(x){
            as.numeric(ifelse(x>1, 1, x))
            })
    }
    out <- rbind(ConTranscripts,FacSplicing)
    out[order(out$transcriptID),]
  }
}
```

B. Publication 1: Function and evolution of DNA methylation in Nasonia vitripennis

OPEN ACCESS Freely available online

PLOS | GENETICS

Function and Evolution of DNA Methylation in *Nasonia vitripennis*

Xu Wang[1,2], David Wheeler[3¤], Amanda Avery[3], Alfredo Rago[4], Jeong-Hyeon Choi[5], John K. Colbourne[4], Andrew G. Clark[1,2]*, John H. Werren[3]*

1 Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, United States of America, **2** Cornell Center for Comparative and Population Genomics, Cornell University, Ithaca, New York, United States of America, **3** Department of Biology, University of Rochester, Rochester, New York, United States of America, **4** School of Biosciences, The University of Birmingham, Birmingham, United Kingdom, **5** Cancer Center, Department of Biostatistics and Epidemiology, Georgia Regents University, Augusta, Georgia, United States of America

Abstract

The parasitoid wasp *Nasonia vitripennis* is an emerging genetic model for functional analysis of DNA methylation. Here, we characterize genome-wide methylation at a base-pair resolution, and compare these results to gene expression across five developmental stages and to methylation patterns reported in other insects. An accurate assessment of DNA methylation across the genome is accomplished using bisulfite sequencing of adult females from a highly inbred line. One-third of genes show extensive methylation over the gene body, yet methylated DNA is not found in non-coding regions and rarely in transposons. Methylated genes occur in small clusters across the genome. Methylation demarcates exon-intron boundaries, with elevated levels over exons, primarily in the 5′ regions of genes. It is also elevated near the sites of translational initiation and termination, with reduced levels in 5′ and 3′ UTRs. Methylated genes have higher median expression levels and lower expression variation across development stages than non-methylated genes. There is no difference in frequency of differential splicing between methylated and non-methylated genes, and as yet no established role for methylation in regulating alternative splicing in *Nasonia*. Phylogenetic comparisons indicate that many genes maintain methylation status across long evolutionary time scales. *Nasonia* methylated genes are more likely to be conserved in insects, but even those that are not conserved show broader expression across development than comparable non-methylated genes. Finally, examination of duplicated genes shows that those paralogs that have lost methylation in the *Nasonia* lineage following gene duplication evolve more rapidly, show decreased median expression levels, and increased specialization in expression across development. Methylation of *Nasonia* genes signals constitutive transcription across developmental stages, whereas non-methylated genes show more dynamic developmental expression patterns. We speculate that loss of methylation may result in increased developmental specialization in evolution and acquisition of methylation may lead to broader constitutive expression.

Citation: Wang X, Wheeler D, Avery A, Rago A, Choi J-H, et al. (2013) Function and Evolution of DNA Methylation in *Nasonia vitripennis*. PLoS Genet 9(10): e1003872. doi:10.1371/journal.pgen.1003872

Editor: Claude Desplan, New York University, United States of America

Received May 17, 2013; **Accepted** August 27, 2013; **Published** October 10, 2013

Funding: This work was supported in part by a Meinig Family Investigator Award to AGC and NIH R24 GM084917 and ARRA supplement to JHW. JHW thanks the Wissenschaftskolleg zu Berlin for support during his sabbatical stay. Additional support for the transcriptome experiments was provided to the Center for Genomics and Bioinformatics, supported in part by the Indiana METACyt Initiative of Indiana University, Lilly Endowment, Inc., and Indiana 21st Century Research and Technology Fund. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ac347@cornell.edu (AGC); werr@mail.rochester.edu (JHW)

¤ Current address: Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand.

Introduction

DNA methylation is an important epigenetic modification found in many plants and animals [1–5]. In mammals, DNA methylation is associated with important epigenetic processes such as genomic imprinting [6], histone modifications and X chromosome inactivation [7,8], and plays an important role in brain development [9]. Clusters of CpG sites (CpG islands or CGIs) are often found in the 5′ regulatory regions including the promoter regions in mammals [10,11]. Methylation at the promoter will typically result in silencing of the gene [12]. The promoters of transposable elements (TEs) are also often repressed by DNA methylation [13]. Non-CpG methylation has been observed in mammals, with high percentages in embryonic stem cells [14].

DNA methylation is also widespread in invertebrates [4,15–26]. In contrast with mammals, methylation typically occurs over gene bodies, and is correlated with elevated gene expression [4,15,16,18,19,22,27], rather than gene inactivation. Consistent with gene activation, several studies of invertebrate methylation have reported that methylated genes tend to have "house-keeping functions", whereas non-methylated genes are more tissue-specific [18,28,29].

DNA methylation is not universal among invertebrates [30,31]. For example, the fruit fly *Drosophila melanogaster* lacks DNA methylation in adults due to the loss of two of three DNA methyltranferases (*Dnmt1 and Dnmt3*), and the reported DNA methylation found in early embryonic stages [32,33] may be due to bisulfite conversion artifacts [31]. Nevertheless, in insects a combination of insect genome sequencing, identification of a full

Author Summary

Insects use methylation to modulate genome function in a different manner from vertebrates. Here, we quantified the global methylation profile in a parasitic wasp species, *Nasonia vitripennis*, a model with some advantages over ant and honeybee for functional and genetic analyses of methylation, such as short generation time, inbred lines, and inter-fertile species. Using a highly inbred line permitted us to precisely characterize DNA methylation, which is compared to gene expression variation across developmental stages, and contrasted to other insect species. DNA methylation is almost exclusively on the 5′-most 1 kbp coding exons, and ~1/3 of protein coding genes are methylated. Methylated genes tend to occur in small clusters in the genome. Unlike many organisms, *Nasonia* leaves nearly all transposable element genes non-methylated. Methylated genes exhibit more uniform expression across developmental stages for both moderately and highly expressed genes, suggesting that DNA methylation is marking the genes for constitutive expression. Among pairs of differentially methylated duplicated genes, the paralogs that lose DNA methylation after duplication in the *Nasonia* lineage show lower expression and greater specialization of expression. Finally, by comparative analysis, we show that methylated genes are more conserved at three different time scales during evolution.

complement of DNMTs, and indirect or direct quantification of methylation, has uncovered genome-wide methylation in many species. A common indirect computational method for identifying genome-wide methylation is gene specific depletion of expected frequencies of CpG relative to observed (CpG O/E), which occurs in methylated genes due to mutational biases of methylated C to T [16]. This approach yielded evidence of genome-wide methylation in a number of insects, including the honeybee *Apis mellifera*, parasitoid wasp *Nasonia vitripennis*, pea aphid *Acrythiosiphon pisum*, and others [16,26,34–36]. Direct methods that have been used to quantify genome-wide methylation in insects include methylation sensitive restriction enzymes [37] and methylated DNA immunoprecipitation (MeDIP) [21]. However, to achieve single-base resolution of methylation in the genome requires bisulfite conversion coupled with high throughput sequencing, which has so far only been reported for honeybee (*Apis mellifera*) [15,19], silkworm (*Bombyx mori*) [15,22] and ants *Camponotus floridanus*, *Harpegnathos saltator* [18] and *Solenopsis invicta* [38].

Most of the work on arthropod DNA methylation has focused on the social insects (honeybees and ants) where alternative castes drive an interest in developmental processes that modulate caste determination [18,28,29]. Investigations in honeybee and ants have suggested an association between alternate splicing and methylation [18,39]. *N. vitripennis* is a non-social haplodiploid parasitoid wasp with a well annotated reference genome [34,40,41]. Prior studies have revealed DNA methylation in *Nasonia* [20] and the presence of requisite DNA methyltransferases, including three members of *Dnmt1* [34]. Here, we report findings of a whole-genome bisulfite sequencing (WGBS-seq) study that provides base-pair resolution of the genome of *Nasonia vitripennis*, a non-social Hymenopteran species [34,40,41]. The highly inbred strain of *Nasonia* used here allows for precisely mapping of WGBS-seq reads and CpG methylation calls to the genome without the complications caused by SNP variation found within heterologous DNA samples from variable strains or populations. We analyze whole genome patterns of DNA methylation in *N. vitripennis*, including the relationship between methylation, gene expression, expression breadth, and gene length, clustering of methylated CpG sites and methylated genes in the genome, patterns of methylation among transposons, non-CpG methylation, methylome comparisons with *Apis*, and changes in gene expression correlated with changes in methylation among paralogs in the *Nasonia* lineage. The *Nasonia* methylome helps to shed light on the function(s) and evolution of DNA methylation in insects.

Results

A. Base-pair resolution profile of CpG DNA methylation in *Nasonia vitripennis*

To profile the *Nasonia* methylome, we performed Illumina whole-genome bisulfite sequencing (WGBS-seq) in adult female samples with 25× haploid genome coverage (Figure S1) and 16.2× average CpG coverage (Figure S2). From the control lambda DNA alignments, the bisulfite conversion efficiency was 99.7% (Table S1), indicating highly efficient conversion. Additional quality control metrics and procedures to assure the high quality of this methylome are described in Materials and Methods.

Across the 8 million CpG sites in the *Nasonia* genome covered by our data, the average percentage methylation is 1.45%, and 1.6% of sites are defined as methylated CpG sites (mCpG) based on our criteria of the site having at least 10× coverage and >10% methylation (see Materials and Methods, Table 1 and Table S2). The percentage of methylation is not uniform across mCpG sites – those with 100% methylated sites are highly enriched, and the distribution is biased toward highly methylated sites with >75% methylation (Figure S3). In other words, CpG sites tend to either be largely non-methylated or highly methylated. We established that genome-wide bisulfite sequencing correctly identifies methylated and non-methylated CpGs by sequencing multiple clones from bisulfite converted DNA from three randomly chosen methylated genes and three non-methylated genes (Figures S4 S5, S6, S7, S8, S9 and Text S1).

Below we describe some of the striking patterns observed in the methylome of *Nasonia*.

A.1. CpG methylation occurs on gene bodies and is enriched in the 5′ coding region. DNA methylation in *Nasonia* predominantly occurs over gene bodies, and in particular over exons (Figure 1). While only containing 10% of 14 million CpGs, the annotated coding regions in *Nasonia* OGS2 (Official Gene Set v2; see Evidential Genes for *Nasonia vitripennis* at http://arthropods.eugenes.org/genes2/nasonia/) [42] are significantly enriched for mCpGs (61.4%, P-value$<2.2\times10^{-16}$, Chi-squared test; Figure 1A). Overall, 11.9% of CpGs located in exons are methylated. By contrast, the intergenic (0.2%), intronic (0.7%) and 1 kbp flanking regions of genes (1%) are depleted of methylated CpGs (Figure 1A). mCpGs are also clustered in the *Nasonia* genome, 78.5% of which are found in 5,440 clusters (Table 1 and Text S2). 98.8% of mCpG clusters are in gene regions (Table 1), which is consistent with gene body methylation. Furthermore, among the 65 mCpG clusters in "intergenic" regions, we found detectable expression in adult female RNA-seq data for 42 (Table S3). We therefore conclude that methylated CpG islands in *Nasonia* occur almost exclusively within transcribed genes.

To compare *Nasonia* mCpG clusters to mammalian-type CpG islands, we ran predictions of CpG islands in the *Nasonia* genome using the same criteria as in mammals [20] (see Materials and Methods). Of 9,265 CpG islands, 36.8% occurred outside of gene bodies and these were nearly universally not methylated (0.15% mCpGs). Methylation also shows a clear pattern of being enriched

Table 1. Summary of DNA methylation status for CpG islands and methylated CpG clusters.

	CpG islands	methylated CpG clusters	Genome
Criteria	200 bp–10 kbp, GC% >50%, CpG O/E >0.6	mCpG/covered CpG >80%, average methylation% >40%	-
Counts/Average length	9265/723 bp	5440/1.2 kbp	-
Total length (% of genome)	6,701,356 (2.3%)	6,596,158 (2.2%)	295.1 Mbp
Total number of CpGs (% of genome)	609,994 (4.35%)	109,676 (0.78%)	14,024,488
CpG density (fold of genome average)	9.1% (1.90)	1.7% (0.35)	4.8%
Number of covered CpGs (% of genome)	139,484 (1.8%)	97,310 (1.2%)	7,818,889
Number of mCpGs (% of genome)	177 (0.15%)	91,803 (78.5%)	116,929
Methylation percentage (mCs/all CpG reads)	0.16% (4,405/2,814,740)	64.91% (2,205,276/3,397,307)	1.45%
In intergenic regions	3,412 (36.8%)	65 (1.20%)	-

mCpG: methylated CpG sites; mC: methylated cytosines.
doi:10.1371/journal.pgen.1003872.t001

at the beginning of genes (Figure 1C,F,G). Based on this pattern, we define genes with >10% methylated CpGs in the first 1 kbp coding region as methylated genes, and genes with ≤10% methylated CpGs in the first 1 kpb as non-methylated genes (see Materials and Methods). Methylation is largely absent in genes defined as non-methylated (0.31% mCpGs) (Figure 1B–I). In methylated genes, the highest levels of methylation occur in the 5′ exons of genes classified as methylated, and decline toward the 3′ region of the gene (Figure 1G–I). Exon methylation in methylated genes peaks at exon 2 or 400–500 bp into the coding region (Figure 1D–H); intron methylation was observed around the exon-intron junctions and also peaks at intron 2 (Figure 1B,C,H). In *N. vitripennis*, 26.7% of protein-coding genes are methylated among the 17,726 genes for which we have sufficient coverage to score methylation status. Excluding 1,540 expressed transposon genes, 4,739 (29.3%) of protein-coding gene are methylated. Our genome-wide investigation also confirms an association between the ratio of observed to expected CpG (CpG O/E) and DNA methylation status, a pattern that was predicted earlier based on bisulfite sequencing of 18 individual genes [20] (Figure S10 and Text S3).

A.2. Transposons are rarely methylated. Among 17,726 annotated genes in OGS2 with adequate uniquely mapped read coverage, 1,540 are expressed transposable element genes (expressed TE genes). The TE genes were characterized in OGS2 with detectable expression level in at least one developmental stage [42]. In adult females, 99.8% of these TE genes are non-methylated (Figure 1F). However, because many TEs occur in multiple copies in the genome with insufficient divergence to be uniquely mapped, we also quantified the DNA methylation percentages in 839 repetitive TEs annotated in the *Nasonia* genome paper [34] that were not covered by uniquely mapped reads (see Materials and Methods). Among the 803 elements with adequate WGBS-seq coverage, only five (GYPSY, SPRINGER, SNAKEHEAD, IFAC and BLASTOPIA) have >5% methylation averaging across CpG positions, and the top three are highly expressed in adult female RNA-seq data (Table S4). Therefore, we can conclude that TEs are rarely methylated and when they are, it can be associated with activation rather than inactivation. This finding contrasts sharply with methylation patterns in plants and mammals, in which methylation of TEs is involved in transcription suppression [13,43].

A.3. CpG methylation shows a strong exon/intron pattern, and "marks" the beginning and end of protein-coding regions. There is a strong exon/intron patterning to methylation, with significantly heavier methylation levels occurring over exons, and declining in adjacent introns (Figure 1H, I). For example, there is significantly higher methylation in both the leading (P-value$<2.2\times10^{-16}$, Wilcoxon Matched-Pairs Signed Ranks Test - WMSRT) and trailing coding exons (P-value$<2.2\times10^{-16}$, WMSRT) relative to the intervening intron between the first two 5′ coding exons. The pattern persists even as overall methylation level decreases toward the 3′ regions of genes (Figure 1H, I).

In addition, the protein-coding regions of methylated genes are enriched for methylation relative to flanking untranslated regions (UTRs) of the same genes. For methylated genes, only 3.0% of the covered CpGs are methylated in the 5′ UTRs (Figure 1B) compared to 35.5% in the first coding exons (Figure 1D). Levels of methylation increase following the start codon for protein-coding genes, with significantly lower levels of mCpGs within 500 bp 5′ of the start codon (1336/26350 or 5.1% mCpGs) relative to 500 bp 3′ of the start codon (30544/46513 or 65.7% mCpGs; (Figure 1G; P-value$<2.2\times10^{-16}$, Chi-squared test). We are confident in the UTR identifications for *Nasonia* OGS2 because they are based on extensive RNA sequencing and tiling array data (see http://arthropods.eugenes.org/genes2/nasonia/), and consistent with our own adult RNA-seq data.

For smaller genes (*e.g.* with coding region <1 kbp), the end of the protein-coding region after the stop codon is also "marked" by reduced methylation level (Figure 1H). Comparison of methylation levels 100 bp before and 100 bp after the stop codon (with ≥4 covered CpGs) shows a significant decline in methylation level of the 3′UTR in genes with protein coding regions <1 kbp (11.4% mean before, 5.7% mean after, P-value = 0.003, WMSRT). The same does not hold, however, for genes of greater length (1.6% mean before, 1.6% mean after, P-value = 0.97, WMSRT). For genes shorter than 1 kbp, the relative number of genes with higher mCpG percentage before the stop codon is also significantly greater than those with lower mCpG percentage (P-value = 4.2×10^{-7}, Chi-squared test), but this is not the case for larger genes (P-value = 0.21, Chi-squared test). Implications of the apparent tagging of the protein-coding exons and start codon from methylated genes are explored in the Discussion.

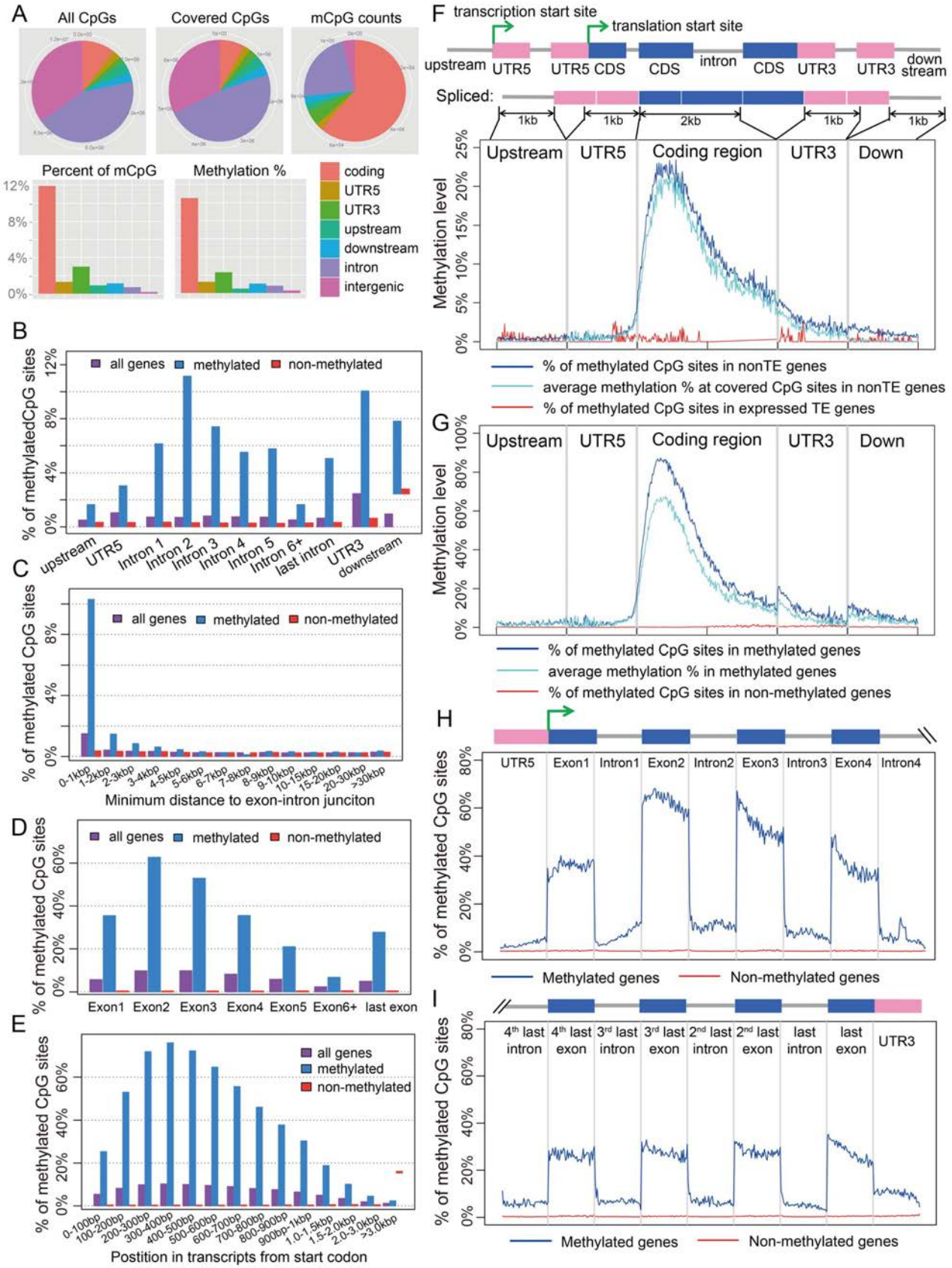

A
All CpGs
Covered CpGs
mCpG counts
Percent of mCpG
Methylation %
coding
UTR5
UTR3
upstream
downstream
intron
intergenic
B
all genes
methylated
non-methylated
% of methylatedCpG sites
upstream
UTR5
Intron 1
Intron 2
Intron 3
Intron 4
Intron 5
Intron 6+
last intron
UTR3
downstream
C
% of methylated CpG sites
Minimum distance to exon-intron junciton
D
Exon1
Exon2
Exon3
Exon4
Exon5
Exon6+
last exon
E
Postition in transcripts from start codon
F
transcription start site
translation start site
upstream
UTR5
UTR5 CDS
CDS
intron
CDS
UTR3
UTR3
down stream
Spliced:
1kb
2kb
Upstream
Coding region
Down
Methylation level
% of methylated CpG sites in nonTE genes
average methylation % at covered CpG sites in nonTE genes
% of methylated CpG sites in expressed TE genes
G
% of methylated CpG sites in methylated genes
average methylation % in methylated genes
% of methylated CpG sites in non-methylated genes
H
Exon1
Intron1
Exon2
Intron2
Exon3
Intron3
Exon4
Intron4
Methylated genes
Non-methylated genes
I
4th last intron
4th last exon
3rd last intron
3rd last exon
2nd last intron
2nd last exon
last intron
last exon

Figure 1. Distribution of CpG DNA methylation in the *Nasonia* genome across protein-coding genes. (A) Distributions across genomic features for all 14 million CpG sites (Top left), 8 million covered CpG sites (Top middle) and methylated CpG sites (mCpGs, Top right). Plotted in the bottom panel are the distributions for percentage of mCpGs and methylation percentage at covered CpG sites. (B) Percentage of mCpGs in the 1 kbp upstream, 1 kbp downstream, UTR and intronic regions for methylated (blue), non-methylated (red) and all genes (purple). (C) Percentage of mCpGs in introns for methylated (blue), non-methylated (red) and all genes (purple), binned by the nearest distance to the exon-intron junctions. (D) Percentage of mCpGs across exons for methylated (blue), non-methylated (red) and all genes (purple). (E) Percentage of mCpGs in the coding region starting from first codon for methylated (blue), non-methylated (red) and all genes (purple). (F) Methylation level in 1 kbp upstream, 1 kbp 5′-UTR, first 2 kbp coding, 1 kbp 3′-UTR and 1 kbp downstream regions for 1,540 expressed transposable element genes (TE genes) and 16,186 non-TE genes. Dark blue line: percentage of mCpGs for non-TE genes; light blue line: average methylation percentage across covered CpGs for non-TE genes; red line: percentage of mCpGs for TE genes. (G) Methylation level in 1 kbp upstream, 1 kbp 5′-UTR, first 2 kbp coding, 1 kbp 3′-UTR and 1 kbp downstream regions for 4,751 methylated non-TE genes and 12,975 non-methylated non-TE genes. Dark blue line: percentage of mCpGs for methylated genes; light blue line: average methylation percentage across covered CpGs for methylated genes; red line: percentage of mCpGs for non-methylated genes. (H–I) Plot of Percentage of methylated CpG sites in the 5′UTR, the first four exons and introns (H) and 3′UTR, the last four exons and introns (I) for methylated (blue) and non-methylated genes (red). All exons, introns and UTRs were rescaled to the same length.
doi:10.1371/journal.pgen.1003872.g001

A.4. Correlation of transcript length and methylation is driven by 5′ bias in methylation. In *Nasonia*, we initially observed a significant negative correlation (Spearman's rank correlation coefficient $\rho = -0.52$, $P\text{-value} < 2\times10^{-16}$) between transcript length and the percentage of mCpGs in methylated genes, when the entire transcribed region is used (Figure 2A). However, the majority of DNA methylation is located in the first 1 kbp of the coding region (Figure 1E,G). When we examined the relationship using one kbp 5′ of coding regions (Figure 2A), the correlation disappeared (Spearman's $\rho = -0.03$ $P\text{-value} > 0.05$). Therefore, the correlation with gene length is a byproduct of the 5′ bias to the distribution of methylation within genes.

A.5. Methylated genes are clustered in the *Nasonia* genome. Tandem methylated genes (MM) and non-methylated gene pairs (NN) are significantly over-represented compared to MN and NM pairs (Figure S11A; $P\text{-value} < 2.2\times10^{-16}$, Chi-squared test), suggesting that methylated genes are clustered in the genome. The average distance between MM gene pairs (4.8 kbp) is much shorter than the expected distance under random distribution of methylated genes, and the distance for NN gene pairs is significantly longer (18.5 kbp) (Figure 2B; $P\text{-value} < 2.2\times10^{-16}$, Mann-Whitney U Test). Moreover, consecutive runs of methylated genes (1M, 2M, 3M, etc.), are longer than expected by chance (Figure 2C; $P\text{-value} < 2.2\times10^{-16}$, Chi-squared test), with a mean cluster size of 2.48. Neighboring genes within distances <1 kbp and coding on opposite strands (*i.e.*, in head-head and tail-tail formations) are enriched among methylated genes, and head-head formations comprise the highest fraction of methylation gene pairs (Figure S11B). This observed pattern of neighboring genes significantly sharing their methylation status (MM or NN) suggests potential co-regulation of methylation (Figure S11B). In conclusion, methylated genes tend to occur in small clusters within the genome.

A.6. The *Nasonia* genome lacks non-CpG DNA methylation. Non-CpG DNA methylation is rarely observed in the *Nasonia* genome; only 0.18% of Cs among the 60 million non-CpG positions with adequate read-depth are methylated (Table S5, Figure S12 and Text S4), which is less than the unconverted Cs in the lambda DNA used as a bisulfite control (Table S1). Therefore, many of these counts are likely experimental artifacts of bisulfite conversion or nucleotide mismatches in the reference genome (Table S6 and Figure S13). For example, of 28 top candidate non-CpG methylation sites with >30% unconverted Cs, eight (4 in top 10) are actually methylated at CpG sites, but were misidentified as non-CpG methylation due to sequence errors in the reference genome sequence (Table S6, Figure S13 and Text S4). Only one candidate non-CpG methylation site out of four examined was verified within the coding region of a gene (Figure S14 and Text S4).

B. CpG methylation and gene expression

We next investigated associations between DNA methylation and gene expression, using a combination of RNA-seq data from adult females and genome-wide tiling microarray data from five different developmental stages: early embryo, late embryo, larva, pupa, and adult (Figure S15 and Dataset S1, See Materials and Methods). Here, we compare expression patterns across developmental stages, and also examine copies of duplicated genes that differ in their methylation status.

B.1. Methylated genes show higher median expression levels. The relationship between methylation status and gene expression level was investigated using two different data sets – RNA-seq data for adult females and tiling microarray data for 5 different developmental stages (early embryo, late embryo, larva, pupa, adult). The RNA-seq results displayed a bimodal distribution of gene expression level in adult females (Figure 3A, $P\text{-value} < 2.2\times10^{-16}$, Hartigans' dip test for unimodality) [44,45]. Methylated genes have significantly higher expression level than non-methylated genes ($P\text{-value} < 2.2\times10^{-16}$, Mann-Whitney U Test) and they showed markedly different patterns. The distribution of gene expression levels for methylated genes was unimodal ($P\text{-value} = 1$) and is generally composed of the higher expressed genes (Figure 3A), whereas the expression of the non-methylated genes is bimodal in distribution, with the moderately expressed set of genes overlapping with the expression levels observed from the methylated genes (Figure 3A, $P\text{-value} = 0.03$). Examination of the expression level for all genes reveals that non-methylated genes constitute the vast majority of low expressed genes. Furthermore, the non-methylated genes account for 99% of the genes that were not found to be expressed in the adult female RNA-seq data (FPKM <1).

In conclusion, DNA methylation in adult females is positively correlated with gene expression level in adult females, and most methylated genes are more highly expressed than typical for non-methylated genes (Figure S16). Nevertheless, methylation status is clearly not the only determinant for high gene expression, as many non-methylated genes also show high expression levels. The same general pattern was observed in tiling array data using median expression level across development (Figure S17). To examine whether there is a simple linear relationship between gene expression level and the percentage of mCpGs in methylated genes, we tested the difference of expression level for genes in different classes of mCpG percentage (Figure 3B). Among the methylated genes, we observed no positive correlation between methylation percentage and expression level (Spearman's $\rho = -0.08$). Therefore, gene expression is correlated with methylation status (methylated vs. non-methylated), but does not increase with increasing methylation level among methylated genes.

B.2. Methylated genes are constitutively expressed during development. Two metrics of gene expression change across development were calculated from the genome-wide tiling path

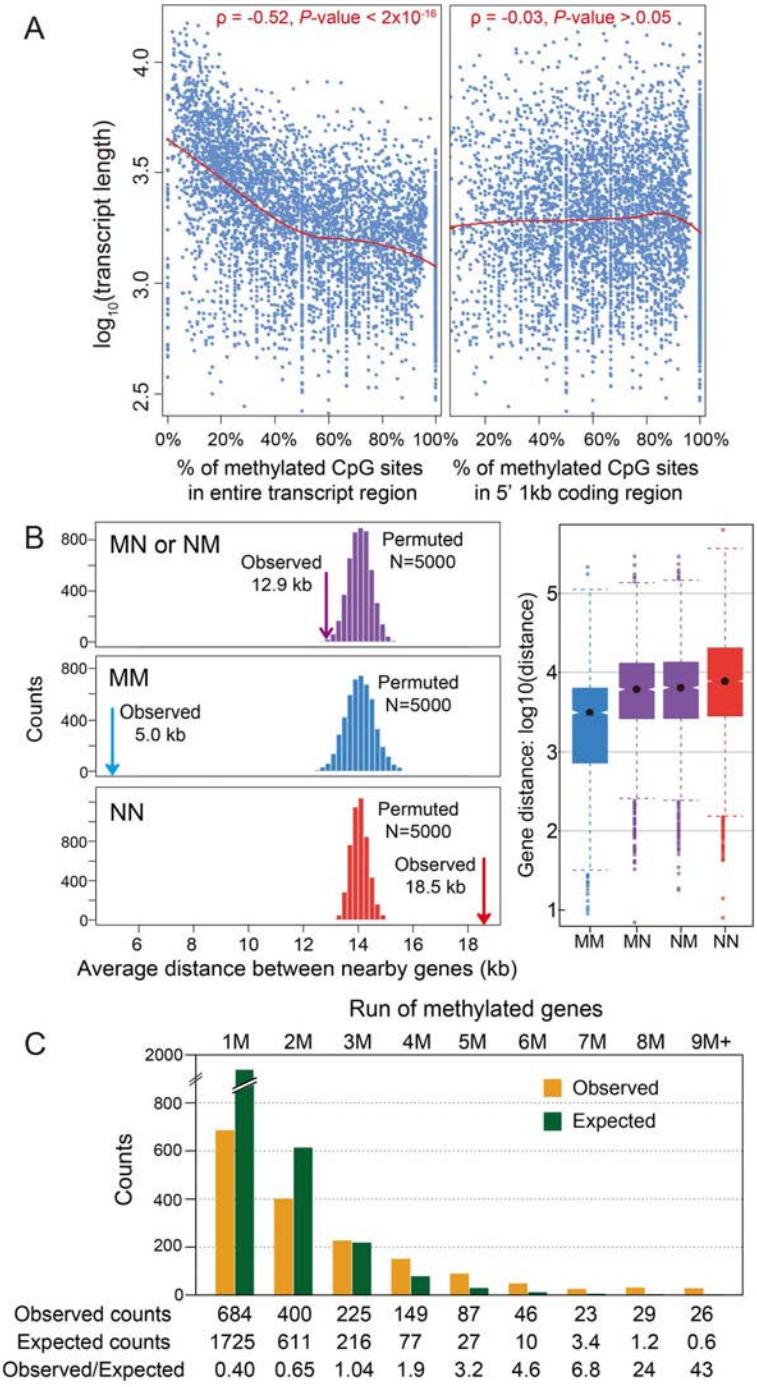


Figure 2. DNA methylation and gene length, exon number and gene locations. (A) Scatterplot for gene length (log_{10}) and percentage of methylated CpG sites for methylation genes in the entire transcript region (left) and in 5′ 1 kbp coding region (right). The fitted lines using non-parametric local regression are shown in red. (B) Left: Distance between neighboring methylated genes (MM), non-methylated genes (NN) and methylated-non-methylated genes (MN or NM). The expected distributions for the three classes calculated by permuting the methylation status (N = 5,000) were plotted (MM: blue; NN: red; MN or NM: purple). The observed mean distance for each group was shown using arrows. Right: Distribution of the distance for the four classes (MM, NN, MN and NM). (C) Distribution of observed (orange) and expected (blue) counts for consecutive run of methylated genes. The expected counts were computed assuming the methylation status is randomly distributed.
doi:10.1371/journal.pgen.1003872.g002

microarray data: a coefficient of expression-level variation across the five developmental stages (expression CV), and the number of stages when gene expression is detected above baseline (see Methods & Materials).

While mean expression CV is lower in methylated (5.07) than non-methylated (6.07) genes, it is clear that both CV and median expression level across development covary (Figure 3C), which is confirmed in a logistic regression analysis (Text S5, Table S7 and

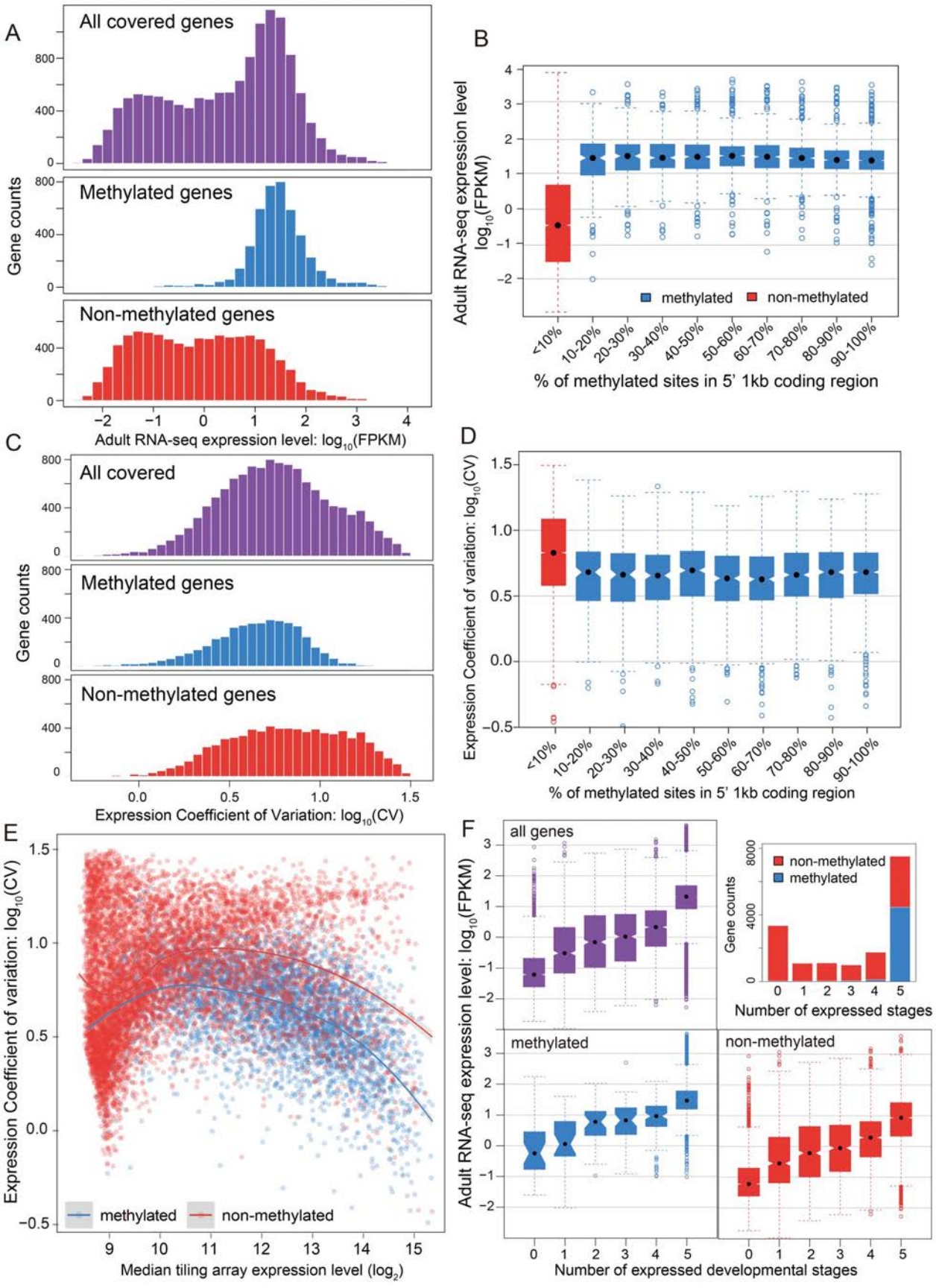

A
All covered genes
Methylated genes
Non-methylated genes
Gene counts
Adult RNA-seq expression level: log10(FPKM)
B
Adult RNA-seq expression level log10(FPKM)
methylated
non-methylated
<10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
90-100%
% of methylated sites in 5' 1kb coding region
C
All covered
Methylated genes
Non-methylated genes
Gene counts
Expression Coefficient of Variation: log10(CV)
D
Expression Coefficient of variation: log10(CV)
% of methylated sites in 5' 1kb coding region
E
Expression Coefficient of variation: log10(CV)
methylated
non-methylated
Median tiling array expression level (log2)
F
all genes
Adult RNA-seq expression level: log10(FPKM)
non-methylated
methylated
Gene counts
Number of expressed stages
Number of expressed developmental stages

Figure 3. DNA methylation, gene expression and expression breadth. (A) Distribution of RNA-seq expression level (log_{10} FPKM) in adult female for methylated (blue), non-methylated (red) and all genes (purple). (B) Distribution of RNA-seq expression level (log_{10} FPKM) in adult female for groups of genes binned by percentage of methylated CpG sites in 5′ 1 kbp coding region. Red: non-methylation genes; blue: methylated genes. (C) Histograms for distribution of expression coefficient of variation (log_{10} expression CV) in five developmental stages (early embryo, late embryo, larvae, pupae and adult) for methylated (blue), non-methylated (red) and all genes (purple). (D) Distribution of expression breadth measurement (log_{10} expression CV) in six developmental stages for groups of genes binned by percentage of methylated CpG sites in 5′ 1 kbp coding region. Red: non-methylation genes; blue: methylated genes. (E) Scatterplot of expression breadth (log_2 expression CV) on y-axis against median expression level (log_2 signal intensity) in tiling array on x-axis, color-coded by adult female methylation status (blue: methylated genes; red: non-methylated genes). Fitted lines using non-parametric local regression are shown for methylated and non-methylated genes respectively. (F) Top right panel: Stacked barplot for expressed methylated and non-methylated genes with 0 to 6 expressed stages. Red: unmethylation genes; blue: methylated genes. Top left and bottom panel: boxplot for distribution of adult female RNA-seq expression level (log_{10} FPKM) for methylated (in blue), non-methylated (in red) and all genes (in purple) expressed in 0–5 developmental stages.
doi:10.1371/journal.pgen.1003872.g003

S8). Because CV varies as a function of expression level, we examined the expression CV against the median expression level across development (Figure 3E). Excluding genes with very low expression (level<9) because there are too few methylated genes to make a proper comparison, we find that methylated genes have lower expression variation than non-methylated genes across a wide range of median expression levels. Dividing median expression level into three categories (9–11, 11–13, >13), methylated genes show significantly lower CV than do non-methylated genes for all categories (P-value$<2.2\times10^{-16}$ for all three categories, Mann-Whitney U Test; Table S9). The same trend is obtained when adult RNA-seq expression level is used in place of median expression across development (Figure S18, S19). Methylated genes have lower CVs across a broad range of median expression levels, indicating that they are expressed more evenly across development.

We next investigated the relationship between the methylation status of genes and the number of developmental stages with nonzero gene expression (see Materials and Methods). The majority of methylated genes (95%) are expressed broadly in all five developmental stages and less than 0.8% of the methylated genes have expression values below 9 in all five stages (Figure 3F). In contrast, only 28% of the non-methylated genes are expressed in all five stages, and 30% are absent (expression value <9) in all stages (Figure 3F). However, there is still a good proportion of non-methylated genes (3062 genes or 28%) that are expressed in all five stages, allowing us to compare expression breadth to median expression level across stages. Whereas it is not the case that all non-methylated genes are stage-specific, most methylated genes show broad expression across developmental stages, even when their median level of expression is relatively low.

The number of expressed stages is also correlated with the gene expression level. Genes present in more life stages tend to have a higher expression level (Figure S20). Taken together, these results strongly suggest that methylation is a general signal for constitutive expression of genes across development, and that this applies both to moderately expressed and highly expressed genes. Studies in *Apis* [28,29], found that methylated genes are more broadly expressed across tissue/cell types. Here we show that methylated genes in *Nasonia* are more broadly expressed across developmental stages.

B.3. Methylated genes are enriched for basal cellular functions. We used blast2go (v2.6.0) to explore the enrichment of Gene Ontology (GO) term categories among methylated genes in *Nasonia*. This analysis reveals that methylated genes were generally enriched for basal cellular functions, such as translation, mRNA processing, and post-translational modifications (Table S10 and Figure S21). As the expression of methylated genes is distributed to the right of the median genome expression, we were concerned that the GO-term enrichment may be confounded by expression level differences between methylated and non-methylated genes. To adjust for this, we carried out a second analysis using gene lists restricted to low-, medium-, and high-expressed genes (See Materials and Methods). The GO-term enrichment among low-expressed methylated genes (Table S11) closely reflected those observed for all methylated genes (Table S10), however, for the medium- and high-expressed methylated genes (Table S12 and S13), cellular component terms became significantly enriched, specifically terms related to intracellular organelles. Both results are consistent with the conclusion that methylated genes in *Nasonia* are typically involved in cellular "house-keeping" functions, especially those involving translation, transcription and organelles.

B.4. Methylation is not required for differential splicing. We investigated patterns of DNA methylation in genes showing alternative splicing, to determine whether a signal of the alternative splice forms is apparent. We found no genome-wide correlation between methylation status and alternative splicing in adult females (Figure 4A–C and Figures S22 and S23, See Materials and Methods). Genes showing differential splicing are not more likely to be methylated than expected by chance (Figure 4A; P-value = 0.49, Chi-squared test), and there was no significant difference in the degree of alternative splicing between methylated and non-methylated genes, quantified by the fraction of major spliced forms (Figure 4B, P-value = 0.65, Kolmogorov-Smirnov test). In methylated genes with multiple methylated CpG clusters, we found in most cases that alternative exons within the first 1 kbp of the coding region do retain methylation (Text S6, Table S14 and Figure S24). However, as non-methylated genes also show extensive alternative splicing (Figure 4C), DNA methylation is clearly not required for differential splicing in *Nasonia*.

C. Comparative Genomics of Methylated Genes

C.1. Methylated genes are more conserved in evolution. To check the conservation status for methylated genes, we investigated 5,039 *Nasonia* single-copy genes covered in our WGBS-seq data that have either one or zero orthologs in each of seven other insect species (*Apis mellifera*, *Tribolium castaneum*, *Bombyx mori*, *Anopheles gambiae*, *Drosophila melanogaster*, *Pediculus humanus* and *Acyrthosiphon pisum*; see Materials and Methods). For these genes, we compared the methylation status in *Nasonia* to other factors among three gene conservation categories: genes present in single copy in all eight insect species (conserved genes), genes present in honeybee and *Nasonia* but not in other species (Hymenoptera-specific genes) and genes present only in *Nasonia* (*Nasonia*-specific genes) (Figure 5A). *Nasonia* methylated genes account for 71% of the genes present in all species, compared to 27% of Hymenoptera-specific genes and 14% of *Nasonia*-specific genes (Figure 5B). Therefore, *Nasonia* methylated genes were highly enriched in the conserved gene class (P-value$<2\times10^{-16}$, Chi-square test). The degree of methylation in methylated genes,

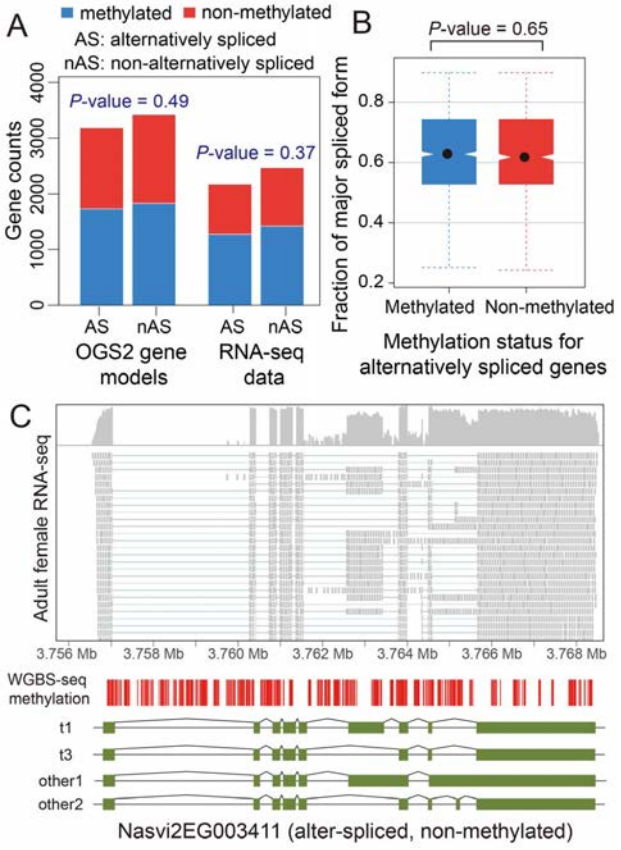


Figure 4. DNA methylation and alternative splicing. (A) Counts of alternatively spliced and non-alternatively spliced genes with different methylation status from OGS2 gene models (left) and RNA-seq data (right). AS: alternatively spliced; nAS: non-alternatively spliced. Methylated is shown in blue and non-methylated shown in red. (B) Distribution of fraction of major spliced forms for alternatively spliced methylated (blue) and non-methylated genes (red). (C) Gene expression, DNA methylation and alternative splicing profile for a non-methylated gene Nasvi2EG003411. Plotted at the top is the IGV browser screenshot showing adult female RNA-seq coverage (on log scale) and read alignments in the gene region. Plotted at the bottom are the CpG methylation profile at covered CpG sites from WGBS-seq data and the exon model of the alternatively spliced transcripts from OGS2 gene models. A vertical bar was drawn for each CpG at its position in the gene, color-coded by the methylation percentage in proportion to the bar length (blue: methylated Cs; red: non-methylated Cs). All 587 covered CpGs in the gene region were non-methylated. Two of the three OGS2 transcript variants, Nasvi2EG003411t1 (labelled as t1) and Nasvi2EG003411t3 (labelled as t3), were covered in the RNA-seq data with 47% and 41% of the transcript abundance, respectively. Two of the remaining minor transcript variants (other1 and other2) were also plotted.
doi:10.1371/journal.pgen.1003872.g004

measured by percentages of methylated CpG sites, was not significantly different among the three classes (*P*-value = 0.64, Kruskal-Wallis rank sum test) (Figure 5C).

Methylated genes have higher expression level (in *Nasonia*) in all three conservation classes (*P*-value$<10^{-10}$, Mann-Whitney U Test), which is consistent with the methylation-expression correlation we observed (Figure 5D). In addition, Hymenoptera-specific and *Nasonia*-specific genes have lower expression levels compared to the conserved genes (*P*-value$<2.2\times10^{-16}$, Mann-Whitney U Test); however, the over-representation of non-methylated genes among them was not purely due to the expression difference. For all *Nasonia* single copy genes, non-methylated genes have higher expression variability across the five life stages (*P*-value$<2.2\times10^{-16}$, Mann-Whitney U Test, one-side). This is also true for the conserved genes (the "all species" category; *P*-value = 9.1×10^{-16}, Mann-Whitney U Test, one-side). However, there is no significant difference in expression variability for hymenopteran-specific genes (*P*-value = 0.17, Mann-Whitney U Test, one-side), while *Nasonia*-specific genes showed the opposite pattern with (*P*-value = 0.018, Mann-Whitney U Test, one side) (Figure 5E). The reverse pattern found in *Nasonia*-specific methylated genes is relatively weak, although statistically significant.

Because expression level of non-methylated genes declines with decreasing conservation (Figure 5D) and CV co-varies with expression level (Figure 3E), CV is not the best index of expression breadth when comparing methylated and non-methylated genes of different conservation levels. We therefore examined how broadly genes are expressed across development for different conservation levels and methylation status (Figure 5F). Methylated genes are expressed more broadly than non-methylated genes for all three conservation categories. Conserved non-methylated genes (i.e. present in all species) are expressed in 4 stages on average, but the number dropped to 3.1 for hymenopteran-specific genes and further dropped to 2.5 for *Nasonia*-specific genes; methylated genes showed a much less dramatic decline, from 4.97 for all species to 4.81 for hymenoptera-specific genes and 4.60 for *Nasonia*-specific genes (Figure 5F). The median values were significantly different for all three categories (Table S15). These results show that methylated genes are more broadly expressed than non-methylated genes across conservation categories, and therefore indicate that even more recently evolved methylated genes acquire broader constitutive expression across development than comparable non-methylated genes.

C.2. There is significant conservation of gene methylation status between *Nasonia* and *Apis*. We next compared patterns between *Nasonia* and *Apis*, each being a representative of two major groups of Hymenoptera that have diverged approximately 180 MYA [34]. The honeybee (*Apis*) methylome data were available in the literature [15]. There were 3,206 *Nasonia-Apis* 1:1 orthologous gene-pairs with methylation status called in both species. Of these, 71.9% are methylated in *Nasonia* compared to 47.7% in *Apis*. Note that the calling of methylation status is different between the *Nasonia* and *Apis*, as data on the distribution of methylated sites within genes (*i.e.* 5′ to 3′) was not available to us for *Apis* (see Materials and Methods). Despite these methodological limitations, there is a strong positive correlation in gene methylation status between *Apis* and *Nasonia* (P-value$<2.2\times10^{-16}$, Chi-squared test), with 42.2% of genes methylated in both species, compared to an expected 34.3%. Furthermore, when we calculated the % of methylated CpGs across the entire gene (the same as done for *Apis*), only 5% of the non-methylated genes changed status to methylated, and the finding of general conservation of methylation status was still found. These findings, based on genome-wide methylation criteria, are consistent with an earlier study showing conservation in gene methylation between *Nasonia* and *Apis*, based on inferred methylation from CpG O/E [20].

C.3. Methylated genes evolve more slowly within the *Nasonia* clade. We also examined methylation status and gene conservation at a shorter evolutionary time scale among *Nasonia* species. The nucleotide substitution rates in ~7,000 genes were compared across three *Nasonia* species: *N. longicornis*, *N. giraulti* and *N. vitripennis* (see Materials and Methods). In all comparisons, methylated genes have lower nucleotide substitution rates (*P*-value$<2.2\times10^{-16}$, Mann-Whitney U Test) (Figure 5H).

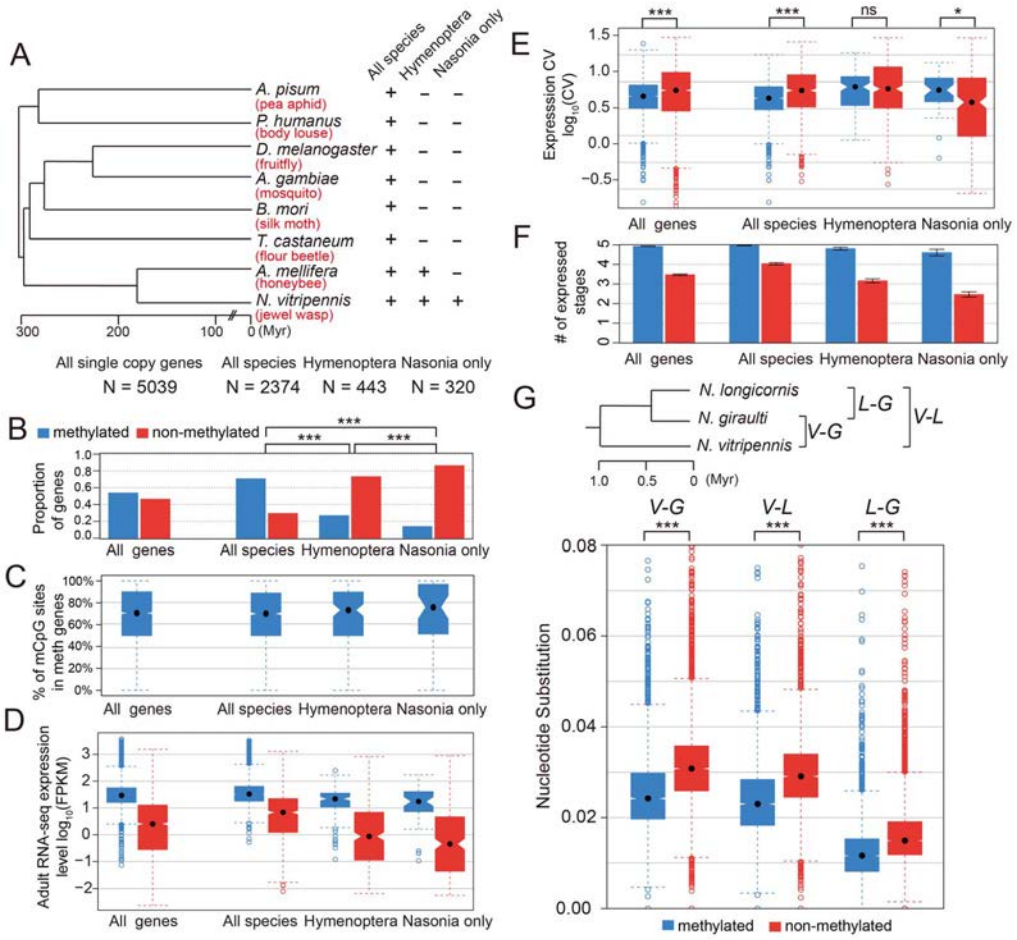


Figure 5. DNA methylation and gene conservation. (A) Phylogenetic tree of eight insect species: *Nasonia vitripennis*, *Apis mellifera*, *Tribolium castaneum*, *Bombyx mori*, *Anopheles gambiae*, *Drosophila melanogaster*, *Pediculus humanus* and *Acyrthosiphon pisum*. The methylation status and correlating factors were plotted in (B–F) for four groups of genes: all 5,039 *Nasonia* single-copy genes with one or zero ortholog in seven other insect species, 2,374 genes with one orthologs in all eight insect species, 443 genes with one orthologs in *Apis* and *Nasonia* but missing in other six species, and 320 genes present only in *Nasonia*. The *y*-axes plotted in (B–F) are (B): proportion of methylated (blue) and non-methylated genes (red); (C): percentage of methylated CpG sites in methylated genes; (D): adult RNA-seq expression levels (log_{10}FPKM); (E): coefficient of variation of expression level in tiling array across six developmental stages; (F): number of expressed tissues. (G) Top: Phylogenetic tree of three *Nasonia* species: *N. longicornis* (L), *N. girault* (G) and *N. vitripennis* (V). Bottom: boxplots of nucleotide substitution rates between V–L, V–G and L–G.
doi:10.1371/journal.pgen.1003872.g005

C.4. *Apis* to *Nasonia* differences in methylation associate with gene ontology. To investigate whether GO-categories of methylated genes are conserved between *Apis* and *Nasonia*, we identified all 1-to-1 orthologs between *Nasonia* and *Apis* for which we had confident methylation status calls (3206 loci), and tested for enrichment of GO-terms where methylation status was either conserved or diverged between these two hymenopteran species. Once again, the most significantly enriched GO-terms for genes methylated in both *Nasonia* and *Apis* (1354 loci) are in categories associated with basal cellular processes such as metabolism and organelle function (Table S16). Next we restricted our lists to genes showing lineage-specific methylation within either *Nasonia* or *Apis*. For genes with methylation only in *Nasonia* (682 loci), ribonucleoprotein complex was enriched at the 5% FDR cutoff level (Table S17). No GO term enrichment was observed at this cutoff for genes methylated in *Apis* only (176 loci); however, processes related to sensory system were enriched at a more permissive FDR cutoff (results not shown).

C.5. When duplicated genes lose methylation, they evolve more quickly and become more developmentally specialized. Finally, we investigated the patterns of evolution of genes that have undergone gene duplications in the clade leading to *Nasonia*. A total of 145 orthologous gene sets were identified that are present in a single copy in all other Hymenoptera (OrthoDB, 13 taxa examined) [46], but which have undergone a gene duplication in the *Nasonia* clade. Methylation status in both *Nasonia* duplicates and the *Apis* paralog was available for 33 of these. In 9 (27%), the *Apis* ortholog and both *Nasonia* paralogs are methylated. In 8 cases, one of the *Nasonia* paralogs was non-methylated (N) whereas the other paralog and *Apis* gene was methylated (M) (Table S18). Those 8 gene pairs are present in a single copy across all Hymenoptera, with the exception of *Nasonia*. We therefore infer that they underwent a lineage-specific duplication, followed by loss of methylation in one of the paralogs. We examined gene expression and rates of divergence in each M to N conversion, using *Apis* as the outgroup. Despite the small sample size, several striking patterns are observed. First, in 7 of 8 cases, the N paralog has lower median expression across developmental stages than does the M (P-value = 0.016,

WMSRT). In one case, the N paralog expression was close to the minimum detection level at all five developmental stages in the tiling array data, and we therefore excluded it as possible pseudogene. In the remaining 7 cases, the N paralogs showed significantly lower median expression levels (Figure 6A; *P*-value = 0.031, WMSRT). As is apparent in Figure 6, whereas the median expression is lower for the N genes, they show a greater variation of expression (*P*-value = 0.016, WMSRT in coefficient of variation) and greater maximum expression difference than do their M paralogs (*P*-value = 0.031, WMSRT), indicating that the N genes have maintained or evolved high expression within certain life stages. Finally, the N genes have significantly longer branch lengths (Figure 6B, *P*-value = 0.016, WMSRT) than their M paralogs, indicating more rapid evolution. These results suggest that loss of methylation status following gene duplication correlates with loss of constitutive expression across developmental stages, and possibly increased evolution and specialization of the duplicated gene.

Discussion

In this study, we profiled the genome-wide methylation at base-pair resolution in *Nasonia* and found several striking features. First, 1.6% of covered CpG sites are methylated in the *Nasonia* genome, and the methylated CpGs are clustered along the genome. As found in several other invertebrates [15,18,19], DNA methylation is located mainly in the gene bodies in *Nasonia*, with coding genes falling into two distinct groups: around 30% of genes are methylated and show strong CpG methylation in 5′ exons, while DNA methylation is largely absent in the remaining genes. To compare the global methylation level across hymenopteran species, we calculated the percentage of methylated CpGs (mC/C) in *Nasonia*, *Apis* and ants (Text S7). Although it is difficult to compare genome-wide methylation levels due to differences in methodology, it appears that *Nasonia* (1.6%) has a higher overall methylation level than is found in honeybees (0.8%) or ants (1.05% in *Camponotus* and 0.68% in *Harpegnathos*).

Unlike mammals, where methylation is associated with suppression of transposon gene expression, with rare exceptions TEs are not methylated in *Nasonia*. The finding is in concordance with honeybee TE methylation profile [19], and suggests that DNA methylation is not required for TE repression in insects. In ants, TE methylation is at the genomic background level, but certain types of TE are hypermethylated and the pattern is species-specific [18]. In our data, we found five retrotransposon families with >5% methylation across CpG sites. The top three methylated TE types (SNAKEHEAD, GYPSY and SPRINGER) are highly expressed in the adult female RNA-seq data (Table S4), suggesting that DNA methylation may actually enhance expression of these elements. We do not know how this is accomplished, but it is possible that certain TEs may contain (or land near) sequence signals that promote DNA methylation. But globally the vast majority of TEs show no methylation in *Nasonia*.

Close examination of methylation in coding genes revealed a striking matching of methylation with the transcription unit. Methylation is low in 5′ UTR and increases rapidly near the transcription start site. Methylation is then consistently higher on exons and decreases significantly on introns, resulting in a clear delineation of exon-intron boundaries by methylation "tagging". Finally, at least for methylated genes <1 kbp in length, methylation also declines significantly in the 3′UTR (after the stop codon). These patterns across the gene region suggest that DNA methylation provides "tags" that mark exons and targets introns for excision during transcription, but also that mark location of translational start and stop, even though translation occurs in the cytoplasm and is not directly associated with the DNA. If methylation affects the rate of transcription, then it is possible that methylation-induced transcriptional pausing at the exon-intron boundary could play a role in splicing [47]. However, how would the DNA methylation signal result in tagging of mature mRNA to demarcate translational initiation and termination? One possibility is through directing mRNA base modifications. For example, in mammals methylation of the N6 position of adenosine (m6A) has been shown to accumulate at stop codons and 3′UTR [48], suggesting a possible signal for translation termination.

It has been hypothesized that in insects DNA methylation regulates alternative splicing [19]; however, a direct causal relationship between methylation and differential splicing remains unsubstantiated. In *Nasonia*, we found no global correlation between methylation status and alternative splicing, although methylation changes across exon/intron boundaries suggested a potential link between DNA methylation and splicing. We should emphasize that DNA methylation is not required for either intron splicing or coding region demarcation, as non-methylated genes show both. Nevertheless, it is possible that methylation expedites these signals for a subset of methylated house-keeping genes, which we have shown to be expressed constitutively and at higher levels. Investigating these mechanisms is an interesting avenue for future research.

In *Nasonia*, the exon-intron pattern is augmented by a strong 5′ bias in level of methylation. The majority of DNA methylation was within the first 1 kbp coding exons and clearly drops beyond that in *Nasonia*, although an exon-intron distinction is still discernible in larger genes. A similar 5′-biased DNA methylation pattern has been observed in ants [18]. Studies in honeybee have reported a negative correlation between gene length and methylation status [49] and we observed the same pattern in *Nasonia* when the methylation percentage across the entire gene was used; however, this pattern disappears in *Nasonia* when the score of methylation level is restricted to the first 1 kbp of the coding region. We found little evidence for non-CpG methylation in *Nasonia*, but were able to confirm a single case. Therefore, non-CpG methylation is present, but it is extremely rare in *Nasonia*. Most candidate non-CpG methylation sites were located in genes nested in CpG methylation clusters. These findings suggest that non-CpG methylation may result from the inaccurate methylation at non-CpG sites by the CpG methylation machinery. It may strengthen the CpG methylation cluster, but the biological significance remains an open question.

In mammals, DNA methylation at promoter regions is often associated with suppression of gene expression [50,51]. However, in insects, DNA methylation has been shown to be positively correlated with expression level in silkworm and ants [18,22]. Here, we also observed a strong positive correlation between methylation and gene expression level; however, methylation is more strongly associated with constitutive expression across development independent of expression level. The distribution of expression levels for methylated genes is unimodal, matching the high expression class. Non-methylated genes show a bimodal distribution, with a mixture of both low and moderate expression, indicating DNA methylation is not the only factor affecting expression level. Other epigenetic marks such as histone modifications are likely to play a role in expression differences among non-methylated genes.

By comparing gene expression levels across five developmental stages, we found that methylated genes show more even expression across stages, and this pattern applies to both highly- and moderately-expressed methylated genes. The finding complements

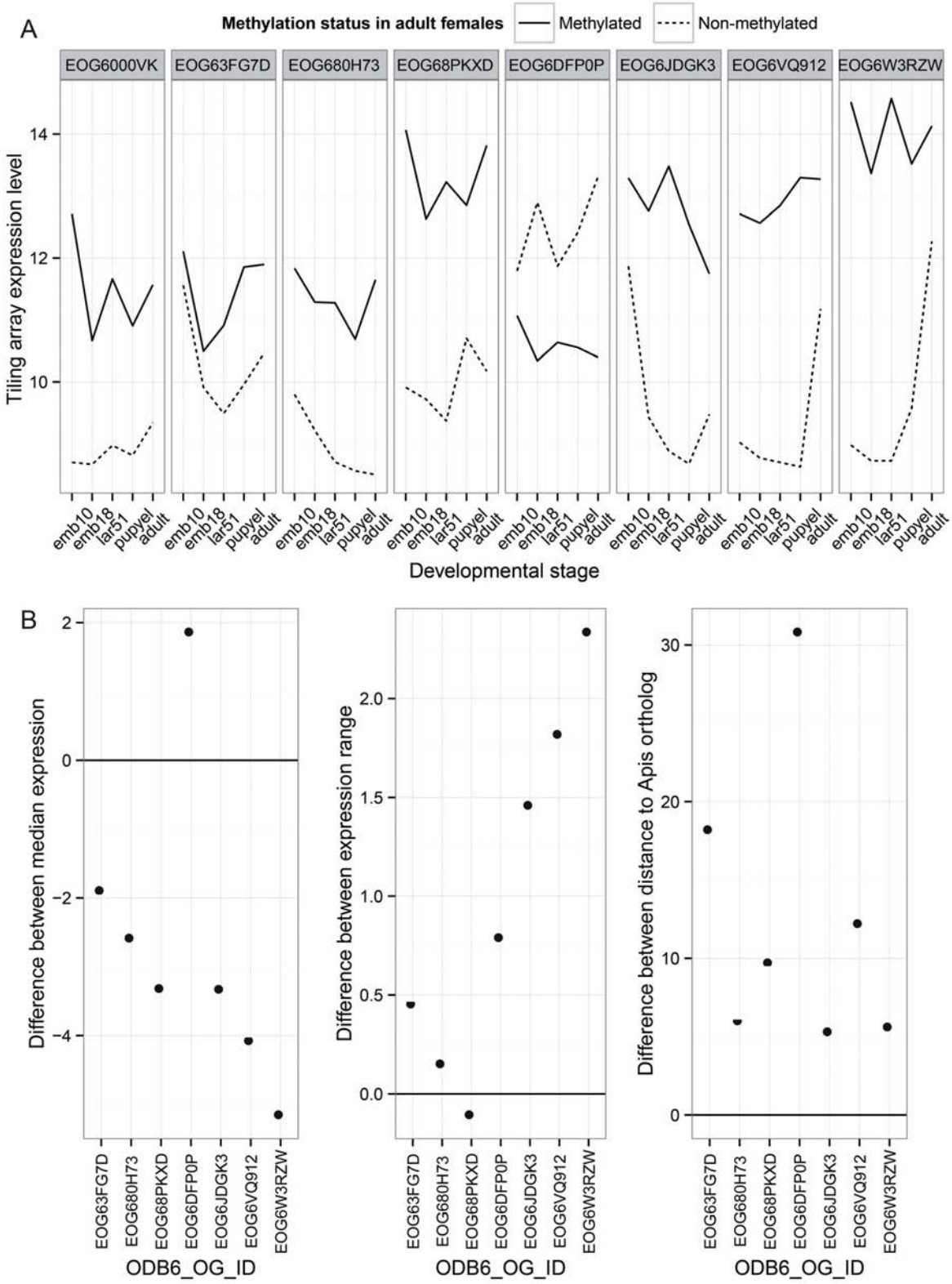


Figure 6. Paralog analysis. Differences between two paralogs that have changed in methylation status in the Nasonia lineage are shown. (A) Comparisons of expression pattern across developmental stages for duplicated genes in the *Nasonia* lineage where one gene is methylated (M) and the other lost methylation (N). These genes have 1:1 orthologs in other hymenopteran species, and the ortholog is methylated in *Apis*. (B) Those paralogs that lost methylation show significant reductions in median expression level across development relative to the M paralog (N–M), significant increases in the range of expression level (N–M), and significantly greater divergence from the *Apis* ortholog (N–M).
doi:10.1371/journal.pgen.1003872.g006

studies in honeybee, which found methylated genes to be expressed across multiple tissues, whereas non-methylated genes showed a more spatially restricted expression pattern [28,29]. In both cases, methylation appears to be more prevalent in genes that are constitutively expressed across development and tissue types. GO-term analysis showed that methylated genes in *Nasonia* are enriched for genes with housekeeping functions, as observed in honeybee and ants [18,19,21]. Furthermore, genes methylated in *Nasonia* tend to be more evolutionarily conserved, as also found in recent studies in ants and other invertebrates [27,52]. Housekeeping genes tend to be expressed in most tissue and cell types, which may explain the low expression variability for methylated genes across stages.

Further support for the role of methylation in constitutive expression of genes comes from the study of duplicated genes that have lost methylation relative to their paralog in the *Nasonia* lineage. Comparing non-methylated and methylated paralogs reveals both a marked median reduction in expression level, and evolution toward more developmental stage-specific expression patterns in the non-methylated genes. Functional category enrichment analysis showed that methylated genes are enriched for basic cellular functions, such as transcription and translation, as also found in honeybee and ants [18,19,21]. Our comparative genomic analysis also shows that many genes have maintained their methylation status across the long evolutionary time scale from *Apis* to *Nasonia*. This probably reflects the role of methylation in constitutive expression of basal housekeeping genes. We also find that methylated genes are enriched among the class of genes that are conserved among insects, while non-methylated genes are enriched among Hymenoptera-specific and *Nasonia*-specific genes. Nevertheless, methylated genes are expressed more broadly across development than are non-methylated genes for each of these conservation categories. Even the more recently evolved "*Nasonia*-specific" methylated genes show broad expression across developmental stages (median 4.60), considerably greater than for non-methylated genes (median 2.5). This suggests that broader constitutive expression is a hallmark of methylated genes whether they are conserved or recently evolved.

Bisulfite sequencing and expression profiling in our study were done on whole insects. Therefore, it could be argued that the correlation between methylation status and expression level occurs because genes that are methylated in more tissues show both higher levels of methylation and higher expression. In other words, tissue specific changes in methylation regulate tissue-specific gene expression, and this creates a correlation between methylation status and gene expression in whole animals. Although a possibility, we found that among methylated genes there is no correlation between level of methylation and level of expression (Figure 3B), which would be expected if the proportion of tissues in which the gene is methylated was driving the pattern. Future work will help resolve whether some genes are being differentially regulated by changes in methylation status. However, it appears that in general DNA methylation is a hallmark of genes that are constitutively "turned on", at least across developmental stages.

In some eusocial organisms such as honeybee and ants, DNA methylation was shown to be related to caste determination [18,19,53]. In *Nasonia*, we have no evidence as yet that changes in methylation regulate specific developmental programs. In contrast, the general data reported above suggest that its primary role is in maintaining constitutive (and perhaps higher) expression of a subset of important cellular "house-keeping" genes, whereas non-methylated genes are more involved in stage-specific differences in expression.

Investigating the role of methylation in epigenetic processes (*e.g.* sexual differentiation, tissue-specific gene expression) will motivate the future study of establishment, maintenance, epigenetic reprogramming and interactors of DNA methylation in *Nasonia* and other insects. Comparison among closely related *Nasonia* also provides the opportunity to study the microevolution of DNA methylation. In addition, the ability to genetically dissect species differences in *Nasonia* through inter-fertile crosses [41,54,55] could provide tools for the genetic investigation of *cis*-regulatory mechanisms of DNA methylation.

Materials and Methods

Sample collection, genomic DNA and total RNA extraction

Genomic DNA samples were extracted from a pool of 50 24 h adult females from the standard *N. vitripennis* strain AsymCX using DNeasy Blood & Tissue Kit (Qiagen, CA). This is the same strain used for the *Nasonia* genome project [34] and is cured of the intracellular bacterium *Wolbachia*.

For RNA-seq, total RNA samples were extracted from adult females ~24 h following eclosion from pupation, using RNeasy Plus mini kit (Qiagen, CA) following the manufacturer's protocol. The DNA, RNA concentration and the A260 nm/A280 nm absorption ratios were measured by NanoDrop ND-1000 Spectrophotometer (Thermo Scientific, DE) to assess quality. RNA integrity was checked using the Agilent 2100 Bioanalyzer (Agilent Technologies, CA). All of the samples had a RIN (RNA integrity number) in the range 9.8–10.0 ($RIN_{max} = 10.0$).

For tiling microarrays, total RNA was extracted from samples of 5 different life stages, 0–10 h embryos, 18–30 h embryos, 51–57 h larvae, day yellow pupae (little to no red eye pigment), and 1 day post eclosion adults. To generate the samples, mated females were first singly given two *Sarcophaga bullata* hosts for 48 h and then given one host for 6 hours, with access to the host restricted to one end for ease of embryo collection. Embryos or larvae were then collected from the hosts. Under this experimental design, females typically produce 85–95% female offspring, and these percentages were confirmed using control hosts where the offspring were permitted to complete development. Therefore, the wasps from these samples are predominantly female, although individual embryos or larvae were not sexed. For pupal collections, hosts were opened and female pupae from the "yellow pupal" stage were collected. Adult females were collected for RNA extraction ~24 h after eclosion from the pupal stage. Six replicates per sample were used, averaging 400 individuals per replicate for embryos, 300 for larvae, 20 for pupae and 20 for adults. Samples were extracted in Trizol (Invitrogen, cat#15596-026) and then sent to the Indiana University Center for Genomics and Bioinformatics for sample preparation and tiling microarray analysis using previously published methods.

WGBS-seq and mRNA-seq library preparation and Illumina sequencing

20 μg of female *Nasonia* genomic DNA and 5 μg non-methylated control lambda DNA (catalog #: D1521, Promega, WI) were sheared by Covaris S2 system (Covaris, MA) for 480 second with 10% duty cycle, level 5 intensity and 200 cycles per burst. The DNA fragments were purified with Zymo DNA Clean & Concentrator-5 columns (Zymo Research, CA), size-selected for 130–180 bp with E-Gel system (Life technologies, CA) and QIAquick Gel Extraction Kit (Qiagen, CA), end-repaired with NEBnext end repair module and NEBnext dA tailing module (New England Biolabs Inc., MA), ligated with Illumina methylated

PE adapter oligo (part #1005560, Illumina, CA) and then purified with Agencourt AMPure XP beads (Beckman Coulter, CA). We performed bisulfite conversion on purified *Nasonia* adult DNA and lambda control DNA using Qiagen EpiTect Bisulfite kit with 2× bisulfite conversion cycles to improve the conversion efficiency and then purified the elute by AMPure XP beads. The purified converted DNA was amplified with PfuTurbo Cx Hotstart DNA Polymerase (Agilent Technologies, CA) using 15 cycles. The final libraries were purified again using AMPure XP beads and the library concentration was measured by Qubit (Life technologies, CA). The library size distribution was checked by Agilent 2100 Bioanalyzer (Agilent Technologies, CA).

We mixed 0.5% of the lambda control DNA library in the *Nasonia* DNA WGBS-seq library, and performed Illumina short-read sequencing in one 84 bp lane on Genome Analyzer IIx (GAIIx) and one 101 bp paired-end lane on HiSeq2000 instrument. Image analysis and base calling were performed by the Illumina instrument software. In total, we obtained 27,766,713 reads from the GAIIx lane and 89,739,445 reads from the HiSeq2000 lane. Illumina WGBS-seq data have been deposited in GEO under accession no. GSE43423.

The mRNA-Seq library was made from 3.5 μg total RNA samples from 24 h adult females, using TruSeq RNA Sample Preparation Kits v2 (Illumina Inc., CA). The library was sequenced on an Illumina HiSeq2000 instrument and we obtained 65,334,896 reads. Illumina RNA-seq data in this study have been deposited in GEO under accession no. GSE43422.

WGBS-seq and mRNA-seq read alignments and data analysis

The Illumina quality score and nucleotide distribution were checked by the FASTX toolkits (http://hannonlab.cshl.edu/fastx_toolkit/index.html). The adapter sequences were removed from the raw reads by custom scripts (0.7% in GAIIx lane and 0.9% in HiSeq lane). To include only high quality bases in our analysis, the sequence reads were trimmed to 75 bp. After trimming, the GAIIx and HiSeq (read 1 only) data gave us 8.75 Gbp of sequences or 25× coverage of the haploid genome, assuming 350 Mbp genome size.

We first aligned the reads to the plus and minus strands of non-methylated lambda genome (NCBI reference sequence NC_001416) with all Cs converted to Ts, using BWA with 4 mismatches [56]. A total of 746,736 (0.64%) reads were uniquely mapped to the lambda genome without indels, resulting 1155× coverage of lambda genome. We estimated the unconverted Cs to be 0.31% by subtracting the background T→C sequence error from the remaining unconverted Cs, therefore the final bisulfite conversion efficiency, at 99.69%, was ideal for downstream analysis. The Illumina sequencing error rates for each type of nucleotide in the GAIIx and HiSeq lane were also estimated from the lambda control alignments (Table S1).

From the *N. vitripennis* reference scaffolds [34], we built C→T converted reference genomes for both the Watson (+) and Crick (−) strand separately, with all Cs in CpGs context remains Cs (meth_genome) and all CpG Cs converted to Ts (unmeth_genome). The rest Illumina sequencing reads were aligned to the converted genomes with BWA [56] with a maximum of 4 mismatches, and summarized in a single BAM file (Figure S1). We tested 4, 6, 8 and 10 mismatches and found 4 mismatches will give the best mapping percentage without ambiguity due to reduced genome complexity after bisulfite conversion. ~80% of the reads could be mapped to the converted *Nasonia* reference genome. To get accurate methylation estimation, we only used uniquely mapped reads without any indel (60% of total reads) for the methylation quantification. CpG methylation percentages were estimated from the proportion of remaining Cs in CpG context (Table S2). Non-CpG methylation was also quantified (Table S5).

We aligned adult female RNA-seq reads to the *Nasonia* reference scaffolds using TopHat v1.4.1 [57] with a maximum of three mismatches. 94% of the reads were uniquely mapped to the genome. Total expression level (FPKM: Fragments Per Kilobase-pair of exon Model) was calculated using Cufflinks v1.3.0 [58] based on all mapped reads from the TopHat alignments. The multiple mapped reads were weighted using the "-u" parameter in Cufflinks. The RNA-seq alignments were viewed in the IGV browser [59,60].

CpG methylation quantification and gene methylation analysis

Among the 14,024,488 CpG sites in *Nasonia* haploid genome, we covered >90% with 2 or more uniquely aligned reads and >55% with 10 or more reads. The average coverage at CpG sites is 16.2× (Figure S2). To obtain accurate quantification of the methylation percentages, we only included ~8 M CpGs sites with 10 or more coverage (covered CpGs). To quantify the CpG methylation levels, we used two metrics: percentage of methylated CpGs (percentage of mCpGs) and average methylation percentage in covered CpGs (methylation percentage). Methylated CpGs (mCpGs) are defined as CpG sites with >10% methylated Cs and ≥10 coverage. This definition requires at least two unconverted C containing reads to call a site methylated, therefore a single T→C Illumina sequence error will not results a spurious methylated site.

Methylation percentage is the average methylated percentage over all CpGs in a particular region, which is the total number of unconverted Cs divided by the total number of reads at CpG sites. The methylated CpG sites were annotated using both the *Nasonia* OGS1.2 (official gene set) and OGS2 gene models [42]. OGS2 gene models incorporated both whole genome tiling expression array and RNA-seq data from multiple tissues at multiple developmental time points, proving high quality support for 5′- and 3′-UTR annotation. Among the 14,024,488 CpGs, 1,159,303 were located in overlapped gene models and were excluded from the analysis. To determine the gene methylation status, we calculated the percentage of mCpG among the covered CpG sites (depth ≥10) in both the first 1 kbp coding region and in the entire transcript region. Since the majority of the mCpGs are located in the first 1 kbp coding region and the methylation level is under the UTR level beyond 2 kbp (Figure 1G), long genes with heavy methylation at the beginning will be averaged out if the entire transcript length was used. Therefore, we inferred the gene methylation status using the percentage of mCpG in the first 1 kbp coding region. Because single or sparse mCpG could be spuriously generated by T→C sequencing error, local incomplete bisulfite conversion or alignment problems, we applied arbitrary cut-off and genes with at least four covered CpGs and >10% mCpG in the first 1 kbp coding region are classified as methylated genes; genes with ≤10% mCpG are defined as non-methylated genes.

To quantify the DNA methylation in repetitive elements and retrotransposons, we built a non-redundant repeat sequence database for the repeat library and retroid elements annotation from the *Nasonia* genome project [34]. From the 1195 sequences in the repeat library, 763 that are >100 bp in length and contain 4 or more CpGs were kept. Simple repeats and STRs were excluded from the analysis. The longest element in each of the 76 retroid families was included in repetitive elements database. We aligned the unmapped and non-uniquely mapped WGBS-seq reads to the database, and quantified the methylation percentage at CpG

positions. Elements with average read depth four or more were included in the analysis.

Characterizing the CpG islands and methylated CpG clusters in the *Nasonia* genome

To search for mammalian type CpG islands (CGIs) in the *Nasonia* genome, we ran predictions of CGIs in the *Nasonia* genome using the same criteria as in mammals [10]: GC percent >50%, CpG O/E (observed/expected CpGs) ratio >0.6, and greater than 200 bp in length. 9,265 CGIs were found in the *Nasonia* genome. We define methylated CpG clusters (mCpGCLs) as regions with >80% methylated CpGs and >40% average methylation percentage, and we found 5,440 mCpGCLs in the *Nasonia* genome.

Analysis of clustering of methylated genes in the *Nasonia* genome

To determine whether the methylated genes are clustered or randomly distributed in the *Nasonia* genome, we analyzed the frequency and distance between neighboring gene pairs (MM: methylated-methylated; MN: methylated-nonmethylated; NM: nonmethylated-methylated; NN: nonmethylated-nonmethylated), as well as the consecutive runs of methylated genes. Scaffold rather than the chromosomal locations were used for the analysis because neighboring genes on two different scaffolds are not in proximity. To eliminate the effect of short scaffolds with few genes in them, only the top 100 largest scaffolds were included for the analysis, containing 11,683 genes with methylation status.

Validation of methylated and non-methylated genes using cloning and sequencing method

To confirm methylation status of individual genes, DNA from 20 pooled 24–27 h virgin *Nasonia vitripennis* (strain Asymcx) females was extracted using the Qiagen DNeasy Blood and Tissue Kit (Cat No. 69504). The bisulfite conversion was performed by the Qiagen EpiTect Bisulfite Kit (Cat No. 59104) with 1.5 μg of starting DNA. Bisulfite PCR primers for six selected genes were designed using Methyl Primer Express software v1.0 (Applied Biosystems by Life Technologies, CA). The amplified PCR product was gel purified and cloned using Promega pGEM-T Easy Vector System II (Cat No. A1380). Direct PCR from the *E. coli* "white" colonies with T7 and SP6 primers was used to select colonies with the right insert size, which were then inoculated in LB broth with ampicillin and the plasmid was extracted using the QIAprep Miniprep kit (Cat No. 27104). Prism BigDye Terminator v3.1 Cycle Sequencing Ready Run Kit (Applied Biosystems) was used to prepare the products for sequencing. BigDye clean-up was completed using ABgene Dye Terminator Removal Kit (Cat No. AB-0943). Sequencing was completed at the Function Genomic Center at the University of Rochester.

Tiling microarray sample preparation

We used NimbleGen high-density 2 (HD2) arrays for transcriptome investigations. The custom 4-array (chip) set consisted of 8.4 million isothermal long-oligonucleotide probes that are 50–60 nt in length and that span the *Nasonia* genome sequence at overlapping intervals of 33 bp, on average. Each slide contained 27,000 Markov model random probes that are not represented in the genome for setting background level thresholds. All probes were designed using NimbleGen's ArrayScribe software and the quality assurance tests of the probes were conducted using Indiana University's Centre for Genomics and Bioinformatics in-house algorithms. Signal to background ratios were determined by first calling probes that fluoresced at intensities greater than 99% of the random probes' signal intensities; therefore only 1% of fluorescing probes are likely to be false positives. The arrays reliably produced high signal to background ratios; log2 ratios of eight were observed for signal over background.

We conducted three replicates each using RNA from independent biological extractions of female early embryo (0–10 h), late embryo (18–30 h), 1st instar larvae, and pupae. Additional experiments were performed comparing transcription in testis and the female reproductive tract. Samples were prepared at 25°C as follows: Approximately 100 *N. vitripennis* (AsymCX) virgins were collected as black pupae. After eclosion, females were provided with males and allowed to mate overnight. Females were initially provisions 15–20 *Sarchophaga bullata* hosts in groups of 20 females for 24 h to induce production of eggs. The hosts were then removed and females were left overnight (~18 h). Mated females produced 85% female progeny under the design used here, and therefore the embryo and larval collections are predominantly female offspring. To collect embryos, individual females were given access to a host at one end (to restrict the oviposition site) and allowed to lay eggs for 6–10 h before being removed. Embryos were then harvested immediately (early embryos), 18 h later (late embryos), or 51 h later (1st instar larvae). All embryos and larvae were collected in an RNase free environment. The host was cracked open and the "cap" removed to expose the embryo. Dissecting needles were used to gently scrape embryos from the surface of the host and transfer them into a 1.5 ml tube pre-chilled on dry ice. Samples were stored at −80°C. If at any time the host was punctured or embryos were exposed to host hemolymph, they were discarded. Estimates of the number of embryos per replicate (three per life stage/sex) were recorded; early embryos ranged from 300–900, late embryo 140–500, 1st instar larvae 245–520. Since sex cannot be determined at larval stage, some of the mated female hostings were allowed to mature to adulthood then males and females were counted to determine the sex ratio. Early larvae showed an average of 82.9% females and late larvae had an average of 84.2%. Pupae collections were made among the progeny of mated females provided with hosts for 48 hrs. They were sorted by sex and stage (early yellow, red-eye, half black, and black pupae). Equal numbers (*S20*) of pupae from each stage were then pooled prior to RNA extraction. Female reproductive tracts (30 per replicate) were removed from 1–3 days post eclosion virgin females and transferred to a tube on dry ice prior to RNA extraction.

Tissue was disrupted and homogenized using Trizol reagent (Invitrogen), and extracted RNA was purified using the Qiagen RNeasy protocol with optimal, on column DNase treatment from specific tissues. Beginning with at least 0.5 μg of total RNA (for early to late embryo) or at least 1.0 μg (for other tissue types), a single round of amplification using MessageAmp II aRNA kit (Ambion) produced between 30 and 45 μg of cRNA for embryo RNA and greater than 100 μg for all other tissue types. Starting with 10 μg of cRNA, double strand cDNA synthesis was carried out using the Invitrogen SuperScript Double-Stranded cDNA Synthesis kit using random hexamer primer followed by DNA labeling using 1 O.D. CY-labeled random nonomer primer and 100 U Klenow fragment (3>5 exo) per 1 μg double-stranded cDNA. The use of random primers ensured that all transcripts hybridize to the array, which contains probes designed solely from a single strand of the DNA sequence. Both sexes for each tissue type were alternatively labelled and a dye-swap was included among the replicate experiments. Dual-color hybridization, post-hybridization washing and scanning were done according to the manufacturer's instructions. Images were acquired using a GenePix 4200A scanner with GenePix 6.0 software. The data from these arrays were extracted using the

software NimbleScan 2.4 (Roche NimbleGen). The normalized tiling array data can be found in Dataset S1.

Tiling array data analysis

The data analysis was performed using the statistical software package R (http://www.r-project.org/) and Bioconductor (http://www.bioconductor.org/) [61]. The signal distributions across chips, samples and replicates were adjusted to be equal according to the mean fluorescence of the random probes on each array. All probes including random probes were quantile normalized across replicates. Scores were assigned for each predicted OGS v2 gene, for each sample, based on the median log2 fluorescence over background intensity of probes falling within the boundaries of the largest gene transcript. The genes were deemed to be transcribed only when greater than ½ or their tiled length was expressed. On average, the 23,161 interrogated genes were tiled by 95.4±1.1 probes. Genes validated by tiling array or EST data are available online at http://www.hymenopteragenome.org/nasonia/?q=sequencing_and_analysis_consortium_datasets.

Analysis of alternative splicing

We used two methods to obtain the alternative splicing status for *Nasonia* transcripts. First, we used the alternative splicing status from the OGS2 gene models with good intron information support. Genes with more than one OGS2 transcripts per gene were considered as alternatively spliced genes, and genes with a single form in OGS2 were considered as non-spliced genes. We also inferred the alternative splicing status from the adult female RNA-seq data using Cufflinks software. Moderately and highly expressed genes with expression level FPKM>2 were included in the study because sufficient RNA-seq coverage is needed to detect the alter-spliced forms in the RNA-seq data. Genes with the percentage of second most abundant forms greater than 10% were considered as alternatively spliced genes.

Methylation conservation and GO-term enrichment analysis

For inference about conservation of methylation status of genes, loci were called *Nasonia*-specific if they did not have a homolog in OrthoDB BLASTp homolog (1e-5) to a database containing Human, Mouse, Xenopus, *Apis mellifera*, *Drosophila melanogaster*, and *Anopheles gambiae*. Arthropod-specific loci were those *Nasonia* sequences that had strong BLASTp hits (1e-5) to *Apis mellifera, Drosophila melanogaster, Anopheles gambiae*, but had no homology to proteins from Human, Mouse or Xenopus. GO term enrichment analysis was performed using blast2go [62] with the *Nasonia* OGS2 protein sequences and a BLASTp cut-off score of 1E-3 for assigning terms. Enrichment was determined using Fisher exact test as implemented by blast2go, with the cut-off for enrichment set to a 5% false discovery rate. The background gene set was restricted to the 17726 *Nasonia* genes with a known adult female methylation status as determined by bisulfite sequencing. For enrichment across different expression levels, genes were divided into low (9–11), medium (11–13) and high expression (13–15) based on median array expression (Table S11, S12, S13), with the background restricted to all genes with known methylation status that fell within that expression range. For GO-term analysis of genes with conserved methylation status between *Apis* and *Nasonia*, 1:1 orthologs were selected based on their known methylation status for *Apis* (taken from [15]).

Comparative genomic analysis of methylated genes

The orthology status for thirteen Hymenoptera insect species (*Acromyrmex echinatior, Apis florea, Apis mellifera, Atta cephalotes, Bombus impatiens, Bombus terrestris, Camponotus floridanus, Harpegnathos saltator, Linepithema humile, Megachile rotundata, Nasonia vitripennis, Pogonomyrmex barbatus, and Solenopsis invicta*) was obtained from OrthoDB [46]. The updated Official Gene Set 2.0 (OGS2) for *Nasonia vitripennis* was used in this analysis (http://arthropods.eugenes.org/genes2/nasonia/). The honeybee methylation status was from Zemach *et al.* 2010 [15]. The nucleotide substitution rates between three *Nasonia* species (*N. longicornis*, *N. giraulti* and *N. vitripennis*) were from the *Nasonia* genome project [34]. Analysis of paralogs that had undergone changes in methylation status was accomplished by first identifying all genes that had 1:1 orthologs in thirteen sequenced hymenopteran genomes, but are duplicated in *N. vitripennis*, using the OrthoDB database [46]. These were then divided into categories based on methylation status. Rates of evolution of the *Nasonia* genes relative to the *Apis* orthologs were measured by comparing pairwise distances of protein alignments scores obtained from the AllAll tool (available at http://www.cbrg.ethz.ch/services/AllAll). Median expression level, range in expression and largest difference in expression were calculated using tiling microarray data.

Statistical analyses

The logistic regression analysis of the effect of expression level and expression breadth on gene methylation status was performed using the `LOGISTIC` procedure in SAS 9.1 (Text S5). The statistical software R (version 2.13.0, www.r-project.org) was used for the rest of the statistical tests. Comparisons between matched gene samples were conducted using the Wilcoxon Matched-Pairs Signed Ranks Test (WMSRT) implemented in `wilcox.test()` function in the `stats` package. The test *P*-value of unimodality of gene expression distribution for methylated and non-methylated genes was calculated using the Hartigans' dip test for unimodality (`dip` package).

Supporting Information

Dataset S1 Tiling array expression level for female developmental stages.
(XLSX)

Figure S1 Illumina WGBS-seq alignment strategies.
(TIF)

Figure S2 Illumina WGBS-seq coverage distribution and summary at CpG sites.
(TIF)

Figure S3 Distribution of methylation percentages for methylated CpG sites with methylation percentage >10%.
(TIF)

Figure S4 Validation of CpG methylation status for non-methylated gene Nasvi2EG001314 in adult females. (A) IGV browser screenshot of the WGBS-seq alignments in a 277 bp region on SCAFFOLD2, showing the CpG sites in non-methylated gene Nasvi2EG001314. All 65 covered CpGs in 5′ 1 kbp transcript region were non-methylated in the WGBS-seq data for this gene. (B) Zoom-in view for the boxed region in (A), demonstrating that all CpG were converted to TpGs in the WGBS-seq read alignments. (C) Plots of the gene model, translation start site and CpG methylation profile for Nasvi2EG001314. A vertical bar was drawn for each CpG at its position in the gene, color-coded by the methylation percentage in proportion to the bar length (blue: methylated Cs; red: non-methylated Cs). There are 143 covered CpGs in the gene region. (D) Bisulfite sequencing verification results for the 16 CpGs sites in

the 201 bp amplicon at the 5′-coding region (shown in A) using the cloning method with 25 clones sequenced. The estimated methylation percentages at each CpG site from the WGBS-seq and single-gene bisulfite sequencing were shown on the top.
(TIF)

Figure S5 Validation of CpG methylation status for non-methylated gene Nasvi2EG000207 in adult females. (A) IGV browser screenshot of the WGBS-seq alignments in a 237 bp region on SCAFFOLD1 (1519410–1519646), showing the CpG sites in non-methylated gene Nasvi2EG000207. All 72 covered CpGs in 5′ 1 kbp transcript region were non-methylated in the WGBS-seq data for this gene. (B) Zoom-in view for the boxed region in (A), demonstrating that all CpG were converted to TpGs in the WGBS-seq read alignments. (C) Plots of the gene model, translation start site and CpG methylation profile for Nasvi2EG000207. A vertical bar was drawn for each CpG at its position in the gene, color-coded by the methylation percentage in proportion to the bar length (blue: methylated Cs; red: non-methylated Cs). There are 232 covered CpGs in the gene region. (D) Bisulfite sequencing verification results for the 17 CpGs sites in the 237 bp 5′-coding region (shown in A) using the cloning method with 22 clones sequenced. The estimated methylation percentages at each CpG site from the WGBS-seq and single-gene bisulfite sequencing were shown on the top. "?" stands for missing data at the end of the sequences.
(TIF)

Figure S6 Validation of CpG methylation status for non-methylated gene Nasvi2EG006064 in adult females. (A) IGV browser screenshot of the WGBS-seq alignments in a 403 bp region on SCAFFOLD9 (3192664–3193066), showing the CpG sites in non-methylated gene Nasvi2EG006064. All 75 covered CpGs in 5′ 1 kbp transcript region were non-methylated in the WGBS-seq data for this gene. (B) Zoom-in view for the boxed region in (A), demonstrating that all CpG were converted to TpGs in the WGBS-seq read alignments. (C) Plots of the gene model, translation start site and CpG methylation profile for Nasvi2EG006064. A vertical bar was drawn for each CpG at its position in the gene, color-coded by the methylation percentage in proportion to the bar length (blue: methylated Cs; red: non-methylated Cs). There are 137 covered CpGs in the gene region. (D) Bisulfite sequencing verification results for the 43 CpGs sites in the 403 bp 5′-coding region (shown in A) using the cloning method with 30 clones sequenced. The estimated methylation percentages at each CpG site from the WGBS-seq and single-gene bisulfite sequencing were shown on the top.
(TIF)

Figure S7 Validation of CpG methylation status for methylated gene Nasvi2EG002725 in adult females. (A) IGV browser screenshot of the WGBS-seq alignments in a 296 bp region on SCAFFOLD3 (3229802–3230097), showing the CpG sites in methylated gene Nasvi2EG002725. All 20 covered CpGs in 5′ 1 kbp transcript region were methylated in the WGBS-seq data for this gene. (B) Zoom-in view for the boxed region in (A), demonstrating that the C in CpG context remains a C after bisulfite conversion. (C) Plots of the gene model, translation start site and CpG methylation profile for Nasvi2EG002725. A vertical bar was drawn for each CpG at its position in the gene, color-coded by the methylation percentage in proportion to the bar length (blue: methylated Cs; red: non-methylated Cs). There are 125 covered CpGs in the gene region. (D) Bisulfite sequencing verification results for the 9 CpGs sites in the 296 bp 5′-coding region (shown in A) using the cloning method with 25 clones sequenced. The estimated methylation percentages at each CpG site from the WGBS-seq and single-gene bisulfite sequencing were shown on the top. Percentages of mCpG labeled in gray in the WGBS-seq data are the ones with less than 10 read coverage.
(TIF)

Figure S8 Validation of CpG methylation status for methylated gene Nasvi2EG000295 in adult females. (A) Plots of the gene model, translation start site and CpG methylation profile for Nasvi2EG000295. A vertical bar was drawn for each CpG at its position in the gene, color-coded by the methylation percentage in proportion to the bar length (blue: methylated Cs; red: non-methylated Cs). There are 102 covered CpGs in the gene region. (B) Bisulfite sequencing verification results for the 7 CpGs sites in the 357 bp 5′-coding region using the cloning method with 20 clones sequenced. The estimated methylation percentages at each CpG site from the WGBS-seq and single-gene bisulfite sequencing were shown on the top. Percentages of mCpG labeled in gray in the WGBS-seq data are the ones with less than 10 read coverage.
(TIF)

Figure S9 Validation of CpG methylation status for methylated gene Nasvi2EG003593 in adult females. (A) IGV browser screenshot of the WGBS-seq alignments in a 283 bp region on SCAFFOLD4 (5219843–5220125), showing the CpG sites in methylated gene Nasvi2EG003593. All 19 covered CpGs in 5′ 1 kbp transcript region were methylated in the WGBS-seq data for this gene. (B) Zoom-in view for the boxed region in (A), demonstrating that the C in CpG context remains a C after bisulfite conversion. (C) Plots of the gene model, translation start site and CpG methylation profile for Nasvi2EG003593. A vertical bar was drawn for each CpG at its position in the gene, color-coded by the methylation percentage in proportion to the bar length (blue: methylated Cs; red: non-methylated Cs). There are 48 covered CpGs in the gene region. (D) Bisulfite sequencing verification results for the 8 CpGs sites in the 283 bp 5′-coding region (shown in A) using the cloning method with 14 clones sequenced. The estimated methylation percentages at each CpG site from the WGBS-seq and single-gene bisulfite sequencing were shown on the top. Percentages of mCpG labeled in gray in the WGBS-seq data are the ones with less than 10 read coverage.
(TIF)

Figure S10 DNA methylation and observed/expected CpG ratios (CpG O/E). (A) Histograms for distribution of CpG O/E ratios in the 5′ 1 kbp coding region for methylated (blue), non-methylated (red) and all genes (purple). (B) Distribution of CpG O/E ratios in classes of genes with different percentage of methylated CpG sites in 5′ 1 kbp coding region. Red: non-methylated genes; blue: methylated genes. (C) Top: Stacked barplot GC content in methylated (blue) and non-methylated genes (red). Middle: scatterplot of GC percent and CpG O/E ratios in methylated genes. Bottom: scatterplot of GC percent and CpG O/E ratios in non-methylated genes.
(TIF)

Figure S11 Clustering of methylated genes in *Nasonia* genome. (A) Fourfold plot of the neighboring methylated-methylated genes (MM), non-methylated-non-methylated genes (NN) and methylated-non-methylated genes (MN) and non-methylated-methylated (NM). (B) Middle panel: Counts of non-overlapping close neighboring genes (<1 kb distance) in four orientation categories (Head-Head, Tail-Tail, Head-Tail and Tail-Head). Top panel: Percentage of methylated genes for the first gene (orange) and second gene (green) gene in the four categories (HH, TT, HT and TH). The red horizontal line is the genome average. Bottom panel:

barplot of methylation status for HH, TT and HT/TH groups. The expected percentages for each category were plotted as a block dot.
(TIF)

Figure S12 Distribution of percentages of unconverted Cs at non-CpG sites.
(TIF)

Figure S13 Eight candidate non-CpG methylation sites in which the methylation is actually in CpG context due to reference sequence error or paralogous sequences in the genome. The IGV browser screenshot was shown for each candidate non-CpG methylation sites. The unconverted Cs were in CpG context instead of non-CpG context. (A) A spurious non-CpG methylation site due to reference genome sequencing error. (B–H) seven examples of spurious non-CpG methylation sites due to paralogous sequences in the genome.
(TIF)

Figure S14 Validation of non-CpG methylation site in gene Nasvi2EG004247 in adult females. (A) IGV browser screenshot of the WGBS-seq alignments (top) and RNA-seq coverage (bottom) for Nasvi2EG004247 gene region on SCAFFOLD6 (1765528–1768213), showing the CpG sites methylation in this gene. The candidate non-CpG methylation site at position 1767201 is labeled in the red box. (B) Zoom-in view for the boxed region in (A), demonstrating that the non-CpG methylation in CAT context on the minus strand, with 42% methylated Cs estimated from the WGBS-seq reads. (C) Plots of the gene model, translation start site and CpG methylation profile for Nasvi2EG004247. A vertical bar was drawn for each CpG at its position in the gene, color-coded by the methylation percentage in proportion to the bar length (blue: methylated Cs in CpGs; red: unmethylated Cs in CpGs). There are 44 covered CpGs in the gene region. The 252 bp target region for bisulfite sequencing validation of the non-CpG methylation is labeled at the bottom. (D) Bisulfite sequencing verification results at the candidate non-CpG methylation site (site #14). The estimated methylation percentages at all C positions from the WGBS-seq and single-gene bisulfite sequencing were shown on the top. There are one CpG C (site #20) and 27 non-CpG Cs in this region. 10/19 (53%) clones have a C at the CpG C position, which is consistent with the methylation status in WGBS-seq data. Among the rest of the 27 non-CpG Cs, only the candidate non-CpG site has unconverted C in more than one clone. The non-CpG methylation at site #14 was confirmed and 3/19 (16%) clones have unconverted Cs.
(TIF)

Figure S15 Distribution of the normalized tiling array expression values in five developmental stages. Plotted on the *x*-axis is the normalized tiling array expression value ($\log_2$). The *y*-axis is the gene count for each stage. The median expression value for each stage is labeled with the red vertical line.
(TIF)

Figure S16 Stacked barplot for expressed methylated and non-methylated genes. Stacked barplot of methylated and non-methylated genes with adult RNA-seq expression level FPKM $\geq$1, binned by different expression level categories. Red: non-methylated genes; blue: methylated genes.
(TIF)

Figure S17 DNA methylation status and tiling array median expression level. Distribution of median tiling array expression level ($\log_2$) for methylated (blue), non-methylated (red) and all genes (purple).
(TIF)

Figure S18 Expression breadth and the adult female RNA-seq expression level for methylation and non-methylated genes. Plotted on the *y*-axis is the $\log_{10}$ coefficient of variation (CV) for tiling array expression values in five developmental stages. On the *x*-axis is the RNA-seq expression level in adult female samples ($\log_{10}$ FPKM). The methylated genes were represented with blue dot and non-methylated genes with red dot. The fitted curve and confidence interval using non-parametric local regression for methylated and non-methylated genes were plotted in blue and red curve, respectively.
(TIF)

Figure S19 Expression breadth and the adult female tiling array expression level for methylated and non-methylated genes. Scatterplot of expression breadth (log2 expression CV) on *y*-axis against adult female gene expression level (log2 signal intensity) in tiling array on *x*-axis, color-coded by adult female methylation status (blue: methylated genes; red: non-methylated genes). Fitted lines using non-parametric local regression are shown for methylated and non-methylated genes respectively.
(TIF)

Figure S20 DNA methylation status and gene expression level, expression breadth and number of expressed tissues. Relationship between DNA methylation status, gene expression level, expression CV and number of expressed stages. Plotted on the *y*-axis is the average expression CV, and on the *x*-axis is the average gene expression level. Methylated (in blue) and non-methylated genes (in red) present in 0–5 developmental stages are plotted as separate round dot. The size of the area is in proportion to the number of genes in each category.
(TIF)

Figure S21 Enriched Gene Ontology categories for methylated gene in *Nasonia* genome.
(TIF)

Figure S22 Distribution of RNA-seq expression level for the four methylation-alternative splicing classes. Plotted here is the distribution of adult female RNA-seq expression level ($\log_{10}$ FPKM) for alternatively spliced methylated, alternatively spliced non-methylated, non-alternatively spliced methylated, non-alternatively spliced non-methylated genes (from left to right). For methylated genes, the expression levels of alternatively spliced genes were not significantly higher than the non-alternatively spliced ones (P-value = 0.67, Kolmogorov-Smirnov test, one side). For non-methylated genes, the expression levels of alternatively spliced genes were significantly higher than the non-alternatively spliced ones (P-value$<2.2\times10^{-16}$, Kolmogorov-Smirnov test, one side).
(TIF)

Figure S23 Correlation between percentage of mCpGs and fraction of major spliced form in alternatively spliced methylated genes. Scatterplot for percentage of methylated CpGs and fraction of major spliced form in alternatively spliced methylated genes. The fitted lines using non-parametric local regression are shown in red.
(TIF)

Figure S24 Gene expression, DNA methylation and alternative splicing profile for three methylated genes. (A) Nasvi2EG000107 showing differential 5′-exon usage. (B) Nasvi2EG013697 showing differential middle exon usage. (C) Nasvi2EG022273 showing intron retention. For each panel, plotted at the top is the IGV browser screenshot showing adult female RNA-seq coverage (on log scale) and read alignments in the gene region. Plotted at the bottom are the CpG methylation profile at covered CpG sites from WGBS-seq data and the exon model of the alternatively spliced transcripts from OGS2 gene models. The locations of methylated

CpG clusters were shown as blue horizontal boxes in (A). A vertical bar was drawn for each CpG at its position in the gene, color-coded by the methylation percentage in proportion to the bar length (blue: methylated Cs; red: non-methylated Cs). OGS2 transcript variants detected in the RNA-seq data with high abundance were plotted at the bottom. The remaining minor forms were not shown in this figure.
(TIF)

Table S1 Summary of Illumina sequencing error rates estimated from lambda control DNA alignments.
(DOC)

Table S2 Summary for methylated and non-methylated CpGs in *Nasonia* genome.
(DOC)

Table S3 Methylated CpG clusters in genic region in *Nasonia* OGS1.2 but absent in OGS2.
(DOC)

Table S4 Top five methylated repetitive TE families and the adult female RNA-seq coverage.
(DOC)

Table S5 Summary for unconverted Cs in non-CpG context in WGBS-seq data.
(DOC)

Table S6 Candidate non-CpG methylation sites with >30% unconverted Cs in non-CpG context.
(DOC)

Table S7 Logistic regression analysis results for methylation status with gene expression and expression breadth as predictors.
(DOC)

Table S8 Logistic regression analysis for methylation status with gene expression or expression breadth as single predictor.
(DOC)

Table S9 Mean/median expression CV for methylated and non-methylated genes in median array expression level categories.
(DOC)

Table S10 Twenty most significant GO terms enriched amongst all methylated *Nasonia* genes.
(DOC)

Table S11 Enriched GO terms amongst methylated genes with median array expression levels 9–11 (low expression).
(DOC)

Table S12 Enriched GO terms amongst methylated genes with median array expression levels 11–13 (medium expression).
(DOC)

Table S13 Enriched GO terms amongst methylated genes with median array expression levels 13–15 (high expression).
(DOC)

Table S14 Methylated genes containing two or more methylated CpG clusters (mCGCLs) with differential 5′exon usage.
(DOC)

Table S15 Statistical significance of expression breadth difference between methylated and non-methylated genes in three conservation categories.
(DOC)

Table S16 Ten most significant enriched GO terms for conserved genes methylated in both *Nasonia* and *Apis*.
(DOC)

Table S17 Enriched GO terms amongst *Nasonia* methylated genes that are conserved but not methylated in *Apis*.
(DOC)

Table S18 Duplicated genes in *Nasonia* and their methylation status.
(DOC)

Text S1 Validation of methylation status estimated from WGBS-seq with single-gene bisulfite sequencing using cloning method.
(DOC)

Text S2 mCpGs are organized in methylated CpG clusters while non-coding and CpG islands are non-methylated.
(DOC)

Text S3 DNA methylation and observed/expected CpG ratios.
(DOC)

Text S4 Quantification and validation of non-CpG methylation in the *Nasonia* genome.
(DOC)

Text S5 Logistic regression analysis of the effect of expression and expression breadth on gene methylation status.
(DOC)

Text S6 Potential links between DNA methylation and alternative splicing.
(DOC)

Text S7 Calculation of global methylation (mC/C) in *Nasonia*, *Apis* and ants.
(DOC)

Acknowledgments

We thank Jenny Xiang, Ying (Diana) Shao, Amanda Manfredo and Li (Grace) Chi for assistance with library preps and Illumina sequencing. Rachel Edwards is thanked for sample preparation and J. Lopez for processing of the tiling microarrays. We thank members of the *Nasonia* community for their assistance in generating the blast2go data file. D. Gilbert is thanked for assistance in aspects of microarray data analysis. Dr. Soojin Yi is thanked for assistance with the *Apis* methylome data. Drs. R. Libbrecht and L. Keller and two anonymous reviewers are thanked for valuable suggestions on the manuscript.

Author Contributions

Conceived and designed the experiments: XW JKC AGC JHW. Performed the experiments: XW AA. Analyzed the data: XW DW AR JHC JKC AGC JHW. Contributed reagents/materials/analysis tools: JKC AGC JHW. Wrote the paper: XW JKC AGC JHW.

References

1. He XJ, Chen T, Zhu JK (2011) Regulation and function of DNA methylation in plants and animals. Cell Res 21: 442–465.
2. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13: 484–492.
3. Klose RJ, Bird AP (2006) Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci 31: 89–97.
4. Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9: 465–476.
5. Bird A (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16: 6–21.
6. Paulsen M, Ferguson-Smith AC (2001) DNA methylation in genomic imprinting, development, and disease. J Pathol 195: 97–110.

7. Hellman A, Chess A (2007) Gene body-specific methylation on the active X chromosome. Science 315: 1141–1143.
8. Norris DP, Brockdorff N, Rastan S (1991) Methylation status of CpG-rich islands on active and inactive mouse X chromosomes. Mamm Genome 1: 78–83.
9. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, et al. (2013) Global Epigenomic Reconfiguration During Mammalian Brain Development. Science 341(6146):1237905.
10. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261–282.
11. Antequera F, Bird A (1993) Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A 90: 11995–11999.
12. Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. Genes Dev 25: 1010–1022.
13. Jones PA, Takai D (2001) The role of DNA methylation in mammalian epigenetics. Science 293: 1068–1070.
14. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315–322.
15. Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328: 916–919.
16. Glastad KM, Hunt BG, Yi SV, Goodisman MA (2011) DNA methylation in insects: on the brink of the epigenomic era. Insect Mol Biol 20: 553–565.
17. Gavery MR, Roberts SB (2010) DNA methylation patterns provide insight into epigenetic regulation in the Pacific oyster (Crassostrea gigas). BMC Genomics 11: 483.
18. Bonasio R, Li Q, Lian J, Mutti NS, Jin L, et al. (2012) Genome-wide and Caste-Specific DNA Methylomes of the Ants Camponotus floridanus and Harpegnathos saltator. Curr Biol 22: 1755–1764.
19. Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, et al. (2010) The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS Biol 8: e1000506.
20. Park J, Peng Z, Zeng J, Elango N, Park T, et al. (2011) Comparative analyses of DNA methylation and sequence evolution using Nasonia genomes. Mol Biol Evol 28: 3345–3354.
21. Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, et al. (2011) The genome of the fire ant Solenopsis invicta. Proc Natl Acad Sci U S A 108: 5679–5684.
22. Xiang H, Zhu J, Chen Q, Dai F, Li X, et al. (2010) Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat Biotechnol 28: 516–520.
23. Gao F, Liu X, Wu XP, Wang XL, Gong D, et al. (2012) Differential DNA methylation in discrete developmental stages of the parasitic nematode Trichinella spiralis. Genome Biol 13: R100.
24. Tweedie S, Charlton J, Clark V, Bird A (1997) Methylation of genomes and genes at the invertebrate-vertebrate boundary. Mol Cell Biol 17: 1469–1475.
25. Bracht JR, Perlman DH, Landweber LF (2012) Cytosine methylation and hydroxymethylation mark DNA for elimination in Oxytricha trifallax. Genome Biol 13: R99.
26. Walsh TK, Brisson JA, Robertson HM, Gordon K, Jaubert-Possamai S, et al. (2010) A functional DNA methylation system in the pea aphid, Acyrthosiphon pisum. Insect Mol Biol 19 Suppl 2: 215–228.
27. Sarda S, Zeng J, Hunt BG, Yi SV (2012) The Evolution of Invertebrate Gene Body Methylation. Mol Biol Evol 29: 1907–16.
28. Foret S, Kucharski R, Pittelkow Y, Lockett GA, Maleszka R (2009) Epigenetic regulation of the honey bee transcriptome: unravelling the nature of methylated genes. BMC Genomics 10: 472.
29. Elango N, Hunt BG, Goodisman MA, Yi SV (2009) DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. Proc Natl Acad Sci U S A 106: 11206–11211.
30. Simpson VJ, Johnson TE, Hammen RF (1986) Caenorhabditis elegans DNA does not contain 5-methylcytosine at any time during development or aging. Nucleic Acids Res 14: 6711–6719.
31. Raddatz G, Guzzardo PM, Olova N, Fantappie MR, Rampp M, et al. (2013) Dnmt2-dependent methylomes lack defined DNA methylation patterns. Proc Natl Acad Sci U S A 110: 8627–8631.
32. Lyko F, Ramsahoye BH, Jaenisch R (2000) DNA methylation in Drosophila melanogaster. Nature 408: 538–540.
33. Gowher H, Leismann O, Jeltsch A (2000) DNA of Drosophila melanogaster contains 5-methylcytosine. EMBO J 19: 6918–6923.
34. Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, et al. (2010) Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science 327: 343–348.
35. Srinivasan DG, Brisson JA (2012) Aphids: a model for polyphenism and epigenetics. Genet Res Int 2012: 431531.
36. (2006) Insights into social insects from the genome of the honeybee Apis mellifera. Nature 443: 931–949.
37. Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, et al. (2011) Draft genome of the red harvester ant Pogonomyrmex barbatus. Proc Natl Acad Sci U S A 108: 5667–5672.
38. Hunt BG, Glastad KM, Yi SV, Goodisman MA (2013) Patterning and regulatory associations of DNA methylation are mirrored by histone modifications in insects. Genome Biol Evol 5: 591–598.
39. Flores KB, Wolschin F, Allen AN, Corneveaux JJ, Huentelman M, et al. (2012) Genome-wide association between DNA methylation and alternative splicing in an invertebrate. BMC Genomics 13: 480.
40. Beukeboom L, Desplan C (2003) Nasonia. Curr Biol 13: R860.
41. Werren JH, Loehlin DW (2009) The parasitoid wasp Nasonia: an emerging model system with haploid male genetics. Cold Spring Harb Protoc 2009: pdb emo134.
42. Munoz-Torres MC, Reese JT, Childers CP, Bennett AK, Sundaram JP, et al. (2011) Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. Nucleic Acids Res 39: D658–662.
43. Saze H, Tsugane K, Kanno T, Nishimura T (2012) DNA methylation in plants: relationship to small RNAs and histone modifications, and functions in transposon inactivation. Plant Cell Physiol 53: 766–784.
44. Hartigan JA, Hartigan PM (1985) The Dip Test of Unimodality. Annals of Statistics 13: 70–84.
45. Hartigan PM (1985) Computation of the Dip Statistic to Test for Unimodality. Applied Statistics-Journal of the Royal Statistical Society Series C 34: 320–325.
46. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res 41: D358–365.
47. Andersen PK, Jensen TH (2010) A pause to splice. Mol Cell 40: 503–505.
48. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, et al. (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell 149: 1635–1646.
49. Zeng J, Yi SV (2010) DNA methylation and genome evolution in honeybee: gene length, expression, functional enrichment covary with the evolutionary signature of DNA methylation. Genome Biol Evol 2: 770–780.
50. Wolffe AP, Matzke MA (1999) Epigenetics: regulation through repression. Science 286: 481–486.
51. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39: 457–466.
52. Simola DF, Wissler L, Donahue G, Waterhouse RM, Helmkampf M, et al. (2013) Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. Genome Res 23: 1235–1247.
53. Herb BR, Wolschin F, Hansen KD, Aryee MJ, Langmead B, et al. (2012) Reversible switching between epigenetic states in honeybee behavioral subcastes. Nat Neurosci 15: 1371–1373.
54. Loehlin DW, Werren JH (2012) Evolution of shape by multiple regulatory changes to a growth gene. Science 335: 943–947.
55. Niehuis O, Buellesbach J, Gibson JD, Pothmann D, Hanner C, et al. (2013) Behavioural and genetic analyses of Nasonia shed light on the evolution of sex pheromones. Nature 494: 345–348.
56. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.
57. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111.
58. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511–515.
59. Thorvaldsdottir H, Robinson JT, Mesirov JP (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14: 178–92.
60. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nat Biotechnol 29: 24–26.
61. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
62. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

References

Abrams, Peter A. The Evolution of Predator-Prey Interactions: Theory and Evidence. Annual Review of Ecology and Systematics . 2000 nov;31(1):79–105. doi:10.1146/annurev.ecolsys.31.1.79. iii

Akbari, Omar S, Igor Antoshechkin, Bruce A Hay, Patrick M Ferree. Transcriptome profiling of Nasonia vitripennis testis reveals novel transcripts expressed from the selfish B chromosome, paternal sex ratio. G3: Genes Genomes Genetics . 2013;3(9):1597–605. doi:10.1534/g3.113.007583. 66, 67, 68

Altschul, S F, W Gish, W Miller, E W Myers, D J Lipman. Basic local alignment search tool. Journal of molecular biology . 1990;215(3):403–410. doi:10.1016/S0022-2836(05)80360-2. 12

Altschul, Stephen F, Thomas L Madden, Alejandto A Schäffer, J Zhang, Zheng Zhang, Webb Miller, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research . 1997 sep;25(17):3389–402. doi:10.1093/nar/25.17.3389. 12

Ament, Seth A, Charles a Blatti, Cedric Alaux, Marsha M Wheeler, Amy L Toth, Yves Le Conte, et al. New meta-analysis tools reveal common transcriptional regulatory basis for multiple determinants of behavior. Proceedings of the National Academy of Sciences . 2012 jun;109(26):E1801–E1810. doi:10.1073/pnas.1205283109. 46

Arnold, Arthur P., Atila van Nas, Aldons J. Lusis. Systems biology asks new questions about sex differences. Trends in endocrinology and metabolism . 2009 dec;20(10):471–6. doi:10.1016/j.tem.2009.06.007. 46

Ayroles, Julien F, Mary Anna Carbone, Eric a Stone, Katherine W Jordan, Richard F Lyman, Michael M Magwire, et al. Systems genetics of complex traits in Drosophila melanogaster. Nature genetics . 2009 mar;41(3):299–307. doi:10.1038/ng.332. 35

Badyaev, Alexander V. Growing apart: an ontogenetic perspective on the evolution of sexual size dimorphism. Trends in Ecology and Evolution . 2002 aug;17(8):369–378. doi:10.1016/S0169-5347(02)02569-7. 44

Badyaev, Alexander V. Stress-induced variation in evolution: from behavioural plasticity to genetic assimilation. Proceedings of the Royal Society B: Biological Sciences . 2005 may;272(1566):877–886. doi:10.1098/rspb.2004.3045. vi

Badyaev, Alexander V. Evolutionary significance of phenotypic accommodation in novel environments: an empirical test of the Baldwin effect. Philosophical transactions of the Royal Society of London Series B, Biological sciences . 2009 apr;364(1520):1125–41. doi:10.1098/rstb.2008.0285. vi

Baker, Richard H, Apurva Narechania, Philip M Johns, Gerald S Wilkinson. Gene duplication, tissue-specific gene expression and sexual conflict in stalk-eyed flies (Diopsidae). Philosophical transactions of the Royal Society of London Series B, Biological sciences . 2012 aug;367(1600):2357–75. doi:10.1098/rstb.2011.0287. xvi

Baldwin, J Mark. A New Factor in Evolution. The American Naturalist . 1896a;30(354):441–451. vi

Baldwin, J Mark. A New Factor in Evolution (Continued). The American Naturalist . 1896b;30(355):536–553. vi

Barabási, Albert-László, Zoltán N Oltvai. Network biology: understanding the cell's functional organization. Nature reviews Genetics . 2004 feb;5(2):101–13. doi:10.1038/nrg1272. iii, 53

Barribeau, S M, N M Gerardo. An evolutionarily and ecologically focused strategy for genome sequencing efforts. Heredity . 2012 may;108(5):577–80. doi:10.1038/hdy.2011.109.3

Barton, Kamil. MuMIn: Multi-model inference. 2011. 16, 60, 61

Bates, Douglas, Martin Maechler, Ben Bolker. lme4: Linear mixed-effects models using S4 classes. 2013. 16

Berger, Bonnie, Jian Peng, Mona Singh. Computational solutions for omics data. Nature Reviews Genetics . 2013 apr;14(5):333–346. doi:10.1038/nrg3433. viii

Bertossa, Rinaldo C, Louis van de Zande, Leo W Beukeboom. The Fruitless gene in Nasonia displays complex sex-specific splicing and contains new zinc finger domains. Molecular biology and evolution . 2009 jul;26(7):1557–69. doi:10.1093/molbev/msp067. 27, 45

Bettencourt-Dias, M., A. Rodrigues-Martins, L. Carpenter, M. Riparbelli, L. Lehmann, M. K. Gatt, et al. SAK/PLK4 is required for centriole duplication and flagella development. Current Biology . 2005;15(24):2199–2207. doi:10.1016/j.cub.2005.11.042. 66

Beukeboom, L. W., John H. Werren. The paternal-sex-ratio (PSR) chromosome in natural populations of Nasonia (Hymenoptera: Chalcidoidea). Journal of Evolutionary Biology . 2000 nov;13(6):967–975. doi:10.1046/j.1420-9101.2000.00231.x. xiii

Beukeboom, Leo W, Louis Van De Zande. Genetics of sex determination in the haplodiploid wasp Nasonia vitripennis (Hymenoptera: Chalcidoidea). Journal of Genetics . 2010 sep;89(3):333–339. doi:10.1007/s12041-010-0045-7. xiv

Bittleston, Leonora S., Naomi E. Pierce, Aaron M. Ellison, Anne Pringle. Convergence in Multispecies Interactions. Trends in Ecology and Evolution . 2016 apr;31(4):269–280. doi:10.1016/j.tree.2016.01.006. ii, viii

Blank, Diana, Luise Wolf, Martin Ackermann, Olin K Silander. The predictability of molecular evolution during functional innovation. Proceedings of the National Academy of Sciences . 2014 feb;doi:10.1073/pnas.1318797111. iii

Bloom, Joshua S, Ian M Ehrenreich, Wesley T Loo, Thúy-Lan Võ Lite, Leonid Kruglyak. Finding the sources of missing heritability in a yeast cross. Nature . 2013 feb;494(7436):234–237. doi:10.1038/nature11867. i

Bonasio, Roberto, Qiye Li, Jinmin Lian, Navdeep S. Mutti, Lijun Jin, Hongmei Zhao, et al. Genome-wide and Caste-Specific DNA Methylomes of the Ants Camponotus floridanus and Harpegnathos saltator. Current Biology . 2012 oct;22(19):1755–1764. doi:10.1016/j.cub.2012.07.042. 28

Bonduriansky, Russell, Stephen F Chenoweth. Intralocus sexual conflict. Trends in Ecology and Evolution . 2009 may;24(5):280–8. doi:10.1016/j.tree.2008.12.005. xv

Bordenstein, SR, JJ Uy, John H. Werren. Host genotype determines cytoplasmic incompatibility type in the haplodiploid genus Nasonia. Genetics . 2003;. xii, 3

Bossdorf, Oliver, Christina L Richards, Massimo Pigliucci. Epigenetics for ecologists. Ecology letters . 2008 feb;11(2):106–15. doi:10.1111/j.1461-0248.2007.01130.x. ii

Botero, Carlos a., Franz J. Weissing, Jonathan Wright, Dustin R. Rubenstein. Evolutionary tipping points in the capacity to adapt to environmental change. Proceedings of the National Academy of Sciences . 2015 jan;112(1):184–189. doi:10.1073/pnas.1408589111. iii

Boyle, Alan P., Carlos L. Araya, Cathleen Brdlik, Philip Cayting, Chao Cheng, Yong Cheng, et al. Comparative analysis of regulatory information and circuits across distant species. Nature . 2014 aug;512(7515):453–456. doi:10.1038/nature13668. 46

Breeuwer, Johannes A. J., John H. Werren. Microorganisms associated with chromosome destruction and reproductive isolation between two insect species. Nature . 1990 aug; 346(6284):558–560. doi:10.1038/346558a0. 3

Brites, Daniela, Carlo Brena, Dieter Ebert, Louis Du Pasquier. More than one way to produce protein diversity: Duplication and limited alternative splicing of an adhesion molecule gene in basal arthropods. Evolution . 2013;67(10):2999–3011. doi:10.1111/evo.12179. 32

Brown, Carolyn J, John M Greally. A stain upon the silence: genes escaping X inactivation. Trends in Genetics . 2003 aug;19(8):432–438. doi:10.1016/S0168-9525(03)00177-X. v

Brown, James B., Nathan Boley, Robert Eisman, Gemma E. May, Marcus H. Stoiber, Michael O. Duff, et al. Diversity and dynamics of the Drosophila transcriptome. Nature . 2014 mar;512(7515):393–399. doi:10.1038/nature12962. 5, 35, 45, 74, 75

Brucker, Robert M RM, SR Seth R Bordenstein. The hologenomic basis of speciation: gut bacteria cause hybrid lethality in the genus Nasonia. Science . 2013 aug;466(6146):667–670. doi:10.1126/science.1240659. 3

Buchta, Thomas, Orhan Ozüak, Dominik Stappert, Siegfried Roth, Jeremy A Lynch. Patterning the dorsal-ventral axis of the wasp Nasonia vitripennis. Developmental biology . 2013 sep;381(1):189–202. doi:10.1016/j.ydbio.2013.05.026. 20

Bull, Alice Louise. Stages of living embryos in the jewel wasp Mormoniella (nasonia) vitripennis (walker) (hymenoptera: pteromalidae). International Journal of Insect Morphology and Embryology . 1982 jan;11(1):1–23. doi:10.1016/0020-7322(82)90034-4. 45, 50

Bull, James J, Others. Evolution of sex determining mechanisms. The Benjamin/Cummings Publishing Company, Inc., 1983. 44

Burggren, Warren W, David Crews. Integrative and Comparative Biology Epigenetics in Comparative Biology : Why We Should Pay Attention. Integrative and comparative biology . 2014 apr;54(1):7–20. doi:10.1093/icb/icu013. i, iv

Cao, Mengfei, Christopher M. Pietras, Xian Feng, Kathryn J. Doroschak, Thomas Schaffner, Jisoo Park, et al. New directions for diffusion-based network prediction of protein function: Incorporating pathways with confidence. Bioinformatics . 2014;30(12):219–227. doi:10.1093/bioinformatics/btu263. 57

Carthew, Richard W, Erik J Sontheimer. Origins and Mechanisms of miRNAs and siRNAs. Cell . 2009 feb;136(4):642–55. doi:10.1016/j.cell.2009.01.035. 4

Chang, Peter L, Joseph P Dunham, Sergey V Nuzhdin, Michelle N Arbeitman. Somatic sex-specific transcriptome differences in Drosophila revealed by whole transcriptome sequencing. BMC genomics . 2011 jan;12(1):364. doi:10.1186/1471-2164-12-364. 44, 45

Charlesworth, Deborah, Judith E Mank. The birds and the bees and the flowers and the trees: lessons from genetic mapping of sex determination in plants and animals. Genetics . 2010 sep;186(1):9–31. doi:10.1534/genetics.110.117697. 76

Chen, Feng-Chi. Are all of the human exons alternatively spliced? Briefings in bioinformatics . 2013 may;15(4):542–551. doi:10.1093/bib/bbt025. 38

Chen, Kevin, Nikolaus Rajewsky. The evolution of gene regulation by transcription factors and microRNAs. Nature reviews Genetics . 2007 feb;8(2):93–103. doi:10.1038/nrg1990. 23

Chen, Shuang, Pengcheng Yang, Feng Jiang, Yuanyuan Wei, Zongyuan Ma, Le Kang. De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. PloS one . 2010 jan;5(12):e15633. doi:10.1371/journal.pone.0015633. vii

Chen, Sidi, Xiaochun Ni, Benjamin H Krinsky, Yong E Zhang, Maria D Vibranovski, Kevin P White, et al. Reshaping of global gene expression networks and sex-biased gene expression by integration of a young gene. The EMBO journal . 2012 jun;31(12):2798–809. doi:10.1038/emboj.2012.108. 81

Chippindale, a K, J R Gibson, W R Rice. Negative genetic correlation for adult fitness between sexes reveals ontogenetic conflict in Drosophila. Proceedings of the National Academy of Sciences . 2001 feb;98(4):1671–5. doi:10.1073/pnas.041378098. xv

Cho, Soochin, Zachary Y. Huang, Jianzhi Zhang. Sex-specific splicing of the honeybee doublesex gene reveals 300 million years of evolution at the bottom of the insect sex-determination pathway. Genetics . 2007;177(3):1733–1741. doi:10.1534/genetics.107.078980. 44

Civelek, Mete, Aldons J Lusis. Systems genetics approaches to understand complex traits. Nature reviews Genetics . 2014;15(1):34–48. doi:10.1038/nrg3575. ii

Conesa, Ana, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, Montserrat Robles. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics . 2005;21(18):3674–3676. 14, 47

Cotton, Travis B., Hien H. Nguyen, Joseph I. Said, Zhengyu Ouyang, Jinfa Zhang, Mingzhou Song. Discerning mechanistically rewired biological pathways by cumulative interaction heterogeneity statistics. Scientific Reports . 2015;5:9634. doi:10.1038/srep09634. viii

Dai, Chao, Wenyuan Li, Juan Liu, Xianghong Jasmine Zhou. Integrating many co-splicing networks to reconstruct splicing regulatory modules. BMC systems biology . 2012 jul;6 Suppl 1(Suppl 1):S17. doi:10.1186/1752-0509-6-S1-S17. 38

Dai, Hongzheng, Ying Chen, Sidi Chen, Qiyan Mao, David Kennedy, Patrick Landback, et al. The evolution of courtship behaviours through the origination of a new gene in Drosophila. Proceedings of the National Academy of Sciences . 2008;105(21):7478–7483. doi:10.1073/pnas.0800693105. 81

Darby, A C, Jeong-Hyeon Choi, T Wilkes, M a Hughes, John H. Werren, G D D Hurst, et al. Characteristics of the genome of Arsenophonus nasoniae , son-killer bacterium of the wasp Nasonia. Insect Molecular Biology . 2010 feb;19:75–89. doi:10.1111/j.1365-2583.2009.00950.x. xiii

Davidson, E. H. A Genomic Regulatory Network for Development. Science . 2002 mar;295(5560):1669–1678. doi:10.1126/science.1069883. ii

Davies, Nathaniel J., Eran Tauber. WaspAtlas: a Nasonia vitripennis gene database and analysis platform. Database . 2015 oct;2015(8):bav103. doi:10.1093/database/bav103. 5, 61

de la Fuente, Alberto. From differential expression to differential networking: identification of dysfunctional regulatory networks in diseases. Trends in Genetics . 2010 jul;26(7):326–333. doi:10.1016/j.tig.2010.05.001. 46

De Robertis, E M, J B Gurdon. Gene activation in somatic nuclei after injection into amphibian oocytes. Proceedings of the National Academy of Sciences . 1977;74(6):2470–4. doi:10.1073/pnas.74.6.2470. iv

Dean, Derek M, Luana S Maroja, Sarah Cottrill, Brent E Bomkamp, Kathleen A Westervelt, David L Deitcher. The wavy Mutation Maps to the Inositol 1,4,5-Trisphosphate 3-Kinase 2 (IP3K2) Gene of Drosophila and Interacts with IP3R to Affect Wing Development. G3: Genes Genomes Genetics . 2016 feb;6(2):299–310. doi:10.1534/g3.115.024307. 67

Dean, R., J. E. Mank. The role of sex chromosomes in sexual dimorphism: Discordance between molecular and phenotypic data. Journal of Evolutionary Biology . 2014;27(7):1443–1453. doi:10.1111/jeb.12345. 44

Dean, Rebecca, Jennifer C Perry, Tommaso Pizzari, Judith E Mank, Stuart Wigby. Experimental evolution of a novel sexually antagonistic allele. PLoS genetics . 2012 jan;8(8):e1002917. doi:10.1371/journal.pgen.1002917. 44

Desjardins, C. A., J. Gadau, J. A. Lopez, O. Niehuis, A. R. Avery, D. W. Loehlin, et al. Fine-Scale Mapping of the Nasonia Genome to Chromosomes Using a High-Density Genotyping Microarray. G3: Genes Genomes Genetics . 2013;3(2):205–215. doi:10.1534/g3.112.004739. xii, 56

Ding, Yun, Li Zhao, Shuang Yang, Yu Jiang, Yuan Chen, Ruoping Zhao, et al. A young drosophila duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. PLoS Genetics . 2010;6(12):1–12. doi:10.1371/journal.pgen.1001255. 81

Domazet-Lošo, Tomislav, Diethard Tautz. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature . 2010 dec;468(7325):815–8. doi:10.1038/nature09632. 31, 60

Dong, Jun, Steve Horvath. Understanding network concepts in modules. BMC systems biology . 2007 jan;1:24. doi:10.1186/1752-0509-1-24. 59

Du, W, O Elemento. Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. Oncogene . 2015 jun;34(25):3215–3225. doi:10.1038/onc.2014.291. iii

Eckalbar, Walter L, Elizabeth D Hutchins, Glenn J Markov, April N Allen, Jason J Corneveaux, Kerstin Lindblad-Toh, et al. Genome reannotation of the lizard Anolis carolinensis based on 14 adult and embryonic deep transcriptomes. BMC genomics . 2013 jan;14:49. doi:10.1186/1471-2164-14-49. 5

Eddy, Sean R. The ENCODE project: Missteps overshadowing a success. Current Biology . 2013 apr;23(7):R259–R261. doi:10.1016/j.cub.2013.03.023. i

Ellegren, Hans, John Parsch. The evolution of sex-biased genes and sex-biased gene expression. Nature reviews Genetics . 2007 sep;8(9):689–98. doi:10.1038/nrg2167. xvi, 44

Elsik, Christine G, Kim C Worley, Anna K Bennett, Martin Beye, Francisco Camara, Christopher P Childers, et al. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC genomics . 2014 jan;15(1):86. doi:10.1186/1471-2164-15-86. 4

Enright, a J, S V Dongen, C a Ouzounis. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research . 2002;30(7):1575–1584. 12

Ferreira, Pedro G, Solenn Patalano, Ritika Chauhan, Richard Ffrench-Constant, Toni Gabaldón, Roderic Guigó, et al. Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. Genome Biology . 2013;14(2):R20. doi:10.1186/gb-2013-14-2-r20. 81

Flores, Kevin, Florian Wolschin, Jason J Corneveaux, April N Allen, Matthew J Huentelman, Gro V Amdam. Genome-wide association between DNA methylation and alternative splicing in an invertebrate. BMC genomics . 2012 jan;13(1):480. doi:10.1186/1471-2164-13-480. 28, 30, 31

Gallach, Miguel, Esther Betrán. Intralocus sexual conflict resolved through gene duplication. Trends in Ecology and Evolution . 2011 may;26(5):222–8. doi:10.1016/j.tree.2011.02.004. xvi, 44

Gempe, Tanja, Martin Beye. Function and evolution of sex determination mechanisms, genes and pathways in insects. BioEssays : news and reviews in molecular, cellular and developmental biology . 2011 jan;33(1):52–60. doi:10.1002/bies.201000043. xiv, 75

Gerstein, Mark B., Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, et al. Architecture of the human regulatory network derived from ENCODE data. Nature . 2012;489(7414):91–100. doi:10.1038/nature11245. ii

Gerstein, Mark B., Joel Rozowsky, Koon-Kiu Yan, Daifeng Wang, Chao Cheng, James B. Brown, et al. Comparative analysis of the transcriptome across distant species. Nature . 2014 aug;512(7515):445–448. doi:10.1038/nature13424. 5, 24

Gibson, J D, O Niehuis, B R E Peirson, E I Cash, J Gadau. Genetic And Developmental Basis Of F2 Hybrid Breakdown In Nasonia Parasitoid Wasps. Evolution . 2013 jul; 67(7):2124–2132. doi:10.1111/evo.12080. 3

Gilbert, Donald G. euGenes: a eukaryote genome information system. Nucleic acids research . 2002;30(1):145–148. 6

Gilbert, Donald G. Gene-omes built from mRNA-seq not genome DNA. In: 7th annual arthropod genomics symposium. Notre Dame, 2013. 35

Gilbert, Donald G, John K Colbourne, John H. Werren. Evidential Genes for Nasonia vitripennis. 2012. 5

Gilbert, Scott F, David Epel. Ecological developmental biology: integrating epigenetics, medicine, and evolution. Sinauer Associates Sunderland, 2009. v, vi

Giniger, Edward, Kathleen Tietje, LY Jan, YN Jan. lola encodes a putative transcription factor required for axon growth and guidance in Drosophila. Development . 1994; 1398:1385–1398. 26

Godfray, H C J. Nasonia: a jewel among wasps. Heredity . 2010 mar;104(3):235–6. doi:10.1038/hdy.2010.3. 45

Golan, David, Eric S Lander, Saharon Rosset. Measuring missing heritability: Inferring the contribution of common variants. Proceedings of the National Academy of Sciences . 2014 dec;111(49):E5272–E5281. doi:10.1073/pnas.1419064111. i

Gurdon, JB. Nuclear transplantation in eggs and oocytes. Journal of Cell science . 1986; 318:287–318. iv

Gurdon, JB, TR Elsdale, M Fischberg. Sexually mature individuals of Xenopus laevis from the transplantation of single somatic nuclei. Nature . 1958;. iv

Haag, Eric S. The Same but Different: Worms Reveal the Pervasiveness of Developmental System Drift. PLoS Genetics . 2014;10(2):1–3. doi:10.1371/journal.pgen.1004150. 81

Hamilton, WD. Extraordinary sex ratios. Science . 1967;(April):477–488. xiii

Hardy, Sara, Pierre-Etienne Jacques, Nicolas Gévry, Audrey Forest, Marie-Eve Fortin, Liette Laflamme, et al. The euchromatic and heterochromatic landscapes are shaped by antagonizing effects of transcription on H2A.Z deposition. PLoS genetics . 2009 oct; 5(10):e1000687. doi:10.1371/journal.pgen.1000687. 24

Hartmann, Britta, Robert Castelo, B. Minana, E. Peden, M. Blanchette, D. C. Rio, et al. Distinct regulatory programs establish widespread sex-specific alternative splicing in Drosophila melanogaster. RNA . 2011 mar;17(3):453–468. doi:10.1261/rna.2460411. 44, 45, 75

Heimpel, George E, Jetske G de Boer. Sex determination in the hymenoptera. Annual review of entomology . 2008 jan;53:209–30. doi:10.1146/annurev.ento.53.103106.093441. xiv, 45

Hemani, Gibran, Konstantin Shakhbazov, Harm-Jan Westra, Tonu Esko, Anjali K Henders, Allan F McRae, et al. Detection and replication of epistasis influencing transcription in humans. Nature . 2014 apr;508(7495):249–53. doi:10.1038/nature13005. i

Hendry, Ap. Key questions in the genetics and genomics of eco-evolutionary dynamics. Heredity . 2013;111(6):456–466. doi:10.1038/hdy.2013.75. iii

Herzig, B., Toma a. Yakulov, Kathrin Klinge, U. Gunesdogan, H. Jackle, Alf Herzig. Ballchen is required for self-renewal of germline stem cells in Drosophila melanogaster. Biology Open . 2014 jun;3(6):510–521. doi:10.1242/bio.20147690. 67

Hoedjes, Katja M, Hans M Smid, Elio Gwm Schijlen, Louise Em Vet, Joke Jfa van Vugt. Learning-induced gene expression in the heads of two Nasonia species that differ in long-term memory formation. BMC Genomics . 2015;16(1):1–13. doi:10.1186/s12864-015-1355-1. 69

Hollis, Brian, David Houle, Zheng Yan, Tadeusz J Kawecki, Laurent Keller. Evolution under monogamy feminizes gene expression in Drosophila melanogaster. Nature communications . 2014 jan;5:3482. doi:10.1038/ncomms4482. xvi

Horvath, Steve, Jun Dong. Geometric interpretation of gene coexpression network analysis. PLoS computational biology . 2008 jan;4(8):e1000117. doi:10.1371/journal.pcbi.1000117. 47, 59

Hsu, Chia-lang, Hsueh-Fen Juan, Hsuan-Cheng Huang. Functional Analysis and Characterization of Differential Coexpression Networks. Scientific Reports . 2015;5:13295. doi:10.1038/srep13295. 46, 53, 57

Huang, Wen, Stephen Richards, Mary Anna Carbone, Dianhui Zhu, Robert R H Anholt, Julien F Ayroles, et al. Epistasis dominates the genetic architecture of Drosophila quantitative traits. Proceedings of the National Academy of Sciences . 2012 sep; 109(39):15553–15559. doi:10.1073/pnas.1213423109. i

Huang, Yueping, Zhiping Liu, Yikang S. Rong. Genome Editing: From Drosophila to Non-Model Insects and Beyond. Journal of Genetics and Genomics . 2016 may;43(5):263–272. doi:10.1016/j.jgg.2016.04.007. xi

Hudson, Nicholas J., Antonio Reverter, Brian P. Dalrymple. A Differential Wiring Analysis of Expression Data Correctly Identifies the Gene Containing the Causal Mutation. PLoS Computational Biology . 2009 may;5(5):e1000382. doi:10.1371/journal.pcbi.1000382. 46

Hunt, Brendan G., Lino Ometto, Laurent Keller, Michael A D Goodisman. Evolution at two levels in fire ants: The relationship between patterns of gene expression and protein sequence evolution. Molecular Biology and Evolution . 2013;30(2):263–271. doi:10.1093/molbev/mss234. iv, vii

Innocenti, Paolo, Edward H Morrow. The sexually antagonistic genes of Drosophila melanogaster. PLoS biology . 2010 mar;8(3):e1000335. doi:10.1371/journal.pbio.1000335. xvi, 44

Jaquiéry, Julie, Claude Rispe, Denis Roze, Fabrice Legeai, Gaël Le Trionnaire, Solenn Stoeckel, et al. Masculinization of the x chromosome in the pea aphid. PLoS genetics . 2013 jan;9(8):e1003690. doi:10.1371/journal.pgen.1003690. 44

Jeong, Hawoong, B Tombor, Reka Albert, Zoltan N Oltvai, A. L. Barabasi. The large-scale organization of metabolic networks. Nature . 2000 oct;407(6804):651–654. doi:10.1038/35036627. 47, 53

Joron, Mathieu, Lise Frezal, Robert T Jones, Nicola L Chamberlain, Siu F Lee, Christoph R Haag, et al. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature . 2011 sep;477(7363):203–6. doi:10.1038/nature10341. 63, 75

Jousselin, Emmanuelle, Simon Van Noort, Jaco M. Greeff. Labile male morphology and intraspecific male polymorphism in the Philotrypesis fig wasps. Molecular Phylogenetics and Evolution . 2004;33(3):706–718. doi:10.1016/j.ympev.2004.08.008. vii

Kamping, Albert, Vaishali Katju, Leo W Beukeboom, John H. Werren. Inheritance of gynandromorphism in the parasitic wasp Nasonia vitripennis. Genetics . 2007 mar;175(3):1321–33. doi:10.1534/genetics.106.067082. 45

Kauffman, Stuart A. Developmental logic and its evolution. BioEssays . 1987;6(2):82–87. doi:10.1002/bies.950060211. 47

Kaufman, Phillip E, Stefan J Long, Donald a Rutz, J. Keith Waldron. Parasitism Rates of Muscidifurax raptorellus and Nasonia vitripennis (Hymenoptera: Pteromalidae) After Individual and Paired Releases in New York Poultry Facilities. Journal of Economic Entomology . 2001 apr;94(2):593–598. doi:10.1603/0022-0493-94.2.593. xii

Khatri, Purvesh, Marina Sirota, Atul J. Butte. Ten years of pathway analysis: Current approaches and outstanding challenges. PLoS Computational Biology . 2012;8(2). doi:10.1371/journal.pcbi.1002375. 47

Knight, Christopher G., John W. Pinney. Making the right connections: Biological networks in the light of evolution. BioEssays . 2009;31(10):1080–1090. doi:10.1002/bies.200900043. iii

Knoll, a. H. Early Animal Evolution: Emerging Views from Comparative Biology and Geology. Science . 1999 jun;284(5423):2129–2137. doi:10.1126/science.284.5423.2129. 23

Koch, Vasco, Inga Nissen, Björn D. Schmitt, Martin Beye. Independent Evolutionary Origin of fem Paralogous Genes and Complementary Sex Determination in Hymenopteran Insects. PLoS ONE . 2014 apr;9(4):e91883. doi:10.1371/journal.pone.0091883. 81

Kopelman, Naama M, Doron Lancet, Itai Yanai. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nature genetics . 2005;37(6):588–589. doi:10.1038/ng1575. 32

Kovalick, Gae E., Donna L. Griffin. Characterization of the SCP/TAPS gene family in Drosophila melanogaster. Insect Biochemistry and Molecular Biology . 2005 aug; 35(8):825–835. doi:10.1016/j.ibmb.2005.03.003. 64

Kriventseva, Evgenia V, Fredrik Tegenfeldt, Tom J Petty, Robert M Waterhouse, F. A. Simao, Igor A Pozdnyakov, et al. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Research . 2015 jan; 43(D1):D250–D256. doi:10.1093/nar/gku1220. 5

Kulmuni, Jonna, Pekka Pamilo. Introgression in hybrid ants is favored in females but selected against in males. Proceedings of the National Academy of Sciences of the United States of America . 2014 aug;111(35). doi:10.1073/pnas.1323045111. xvi

Kunte, K, W Zhang, a Tenger-Trolander, D H Palmer, a Martin, R D Reed, et al. Doublesex Is a Mimicry Supergene. Nature . 2014 mar;507(7491):229–32. doi:10.1038/nature13112. 75

Langfelder, Peter, Steve Horvath. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics . 2008 jan;9:559. doi:10.1186/1471-2105-9-559. 46, 53

Li, Li. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research . 2003 sep;13(9):2178–2189. doi:10.1101/gr.1224503. 10

Lindquist, S, E A Craig. The Heat-Shock Proteins. Annual Review of Genetics . 1988 dec; 22(1):631–677. doi:10.1146/annurev.ge.22.120188.003215. v

Liu, Xiaoping, Yuetong Wang, Hongbin Ji, Kazuyuki Aihara, Luonan Chen. Personalized characterization of diseases using sample-specific networks. Nucleic Acids Research . 2016;44(22):gkw772. doi:10.1093/nar/gkw772. 57

Loehlin, D. W., John H. Werren. Evolution of Shape by Multiple Regulatory Changes to a Growth Gene. Science . 2012 feb;335(6071):943–947. doi:10.1126/science.1215193. 3

Lopez, Jacqueline a, John K Colbourne. Dual-Labeled Expression-Tiling Microarray Protocol for Empirical Annotation of Genome Sequences. CGB Technical Report . 2011; doi:http://dx.doi.org/10.2506/cgbtr-201101. 50

Lowdon, Rebecca F., Hyo Sik Jang, Ting Wang. Evolution of Epigenetic Regulation in Vertebrate Genomes. Trends in Genetics . 2016;32(5):269–283. doi:10.1016/j.tig.2016.03.001. iii, ix

Lyko, Frank, Sylvain Foret, Robert Kucharski, Stephan Wolf, Cassandra Falckenhayn, Ryszard Maleszka. The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS biology . 2010 jan;8(11):e1000506. doi:10.1371/journal.pbio.1000506. 24, 28

Lynch, Jeremy A. The Expanding Genetic Toolbox of the Wasp Nasonia vitripennis and Its Relatives. Genetics . 2015;199(4):897–904. doi:10.1534/genetics.112.147512. 3, 5

Lynch, Jeremy A, Claude Desplan. A method for parental RNA interference in the wasp Nasonia vitripennis. Nature protocols . 2006 jan;1(1):486–94. doi:10.1038/nprot.2006.70. xii

Ma, Haisu, Eric E. Schadt, Lee M. Kaplan, Hongyu Zhao. COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. Bioinformatics . 2011;27(9):1290–1298. doi:10.1093/bioinformatics/btr136. 57

Mackay, Trudy F C, Robert R H Anholt. Of flies and man: Drosophila as a model for human complex traits. Annual review of genomics and human genetics . 2006 jan;7:339–67. doi:10.1146/annurev.genom.7.080505.115758. i

Mackay, Trudy F C, Eric a Stone, Julien F Ayroles. The genetics of quantitative traits: challenges and prospects. Nature reviews Genetics . 2009 aug;10(8):565–77. doi:10.1038/nrg2612. 35

Mank, Judith E., Kiwoong Nam, B. Brunstrom, Hans Ellegren. Ontogenetic Complexity of Sexual Dimorphism and Sex-Specific Selection. Molecular Biology and Evolution . 2010 jul;27(7):1570–1578. doi:10.1093/molbev/msq042. 44, 75

Mank, Judith E, Nina Wedell, David J Hosken. Polyandry and sex-specific gene expression. Philosophical transactions of the Royal Society of London Series B, Biological sciences . 2013;368(1613):20120047. doi:10.1098/rstb.2012.0047. 44

Marbach, Daniel, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, et al. Wisdom of crowds for robust gene network inference. Nature Methods . 2012 jula;9(8):796–804. doi:10.1038/nmeth.2016. iii

Marbach, Daniel, Sushmita Roy, Ferhat Ay, Patrick E. Meyer, Rogerio Candeias, Tamer Kahveci, et al. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. Genome Research . 2012b;22(7):1334–1349. doi:10.1101/gr.127191.111. iii

McManus, C Joel, Joseph D Coolon, Jodi Eipper-Mains, Patricia J Wittkopp, Brenton R Graveley. Evolution of splicing regulatory networks in Drosophila. Genome research . 2014 may;24(5):786–96. doi:10.1101/gr.161521.113. 38

Mengistu, Henok, Joost Huizinga, Jean-Baptiste Mouret, Jeff Clune. The Evolutionary Origins of Hierarchy. PLOS Computational Biology . 2016 jun;12(6):e1004829. doi: 10.1371/journal.pcbi.1004829. 48

Meyer, David, Achim Zeileis, Kurt Hornik. vcd: Visualizing Categorical Data. 2014. 16, 61

Minelli, Alessandro, Giuseppe Fusco. On the Evolutionary Developmental Biology of Speciation. Evolutionary Biology . 2012 may;39(2):242–254. doi:10.1007/s11692-012-9175-6. iv

Misof, Bernhard, S. Liu, K. Meusemann, R. S. Peters, A. Donath, C. Mayer, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science . 2014 nov; 346(6210):763–767. doi:10.1126/science.1257570. 3, 80

Moyers, Bryan a., Jianzhi Zhang. Phylostratigraphic Bias Creates Spurious Patterns of Genome Evolution. Molecular Biology and Evolution . 2014;32(1):258–267. doi: 10.1093/molbev/msu286. 33, 80

Mozhui, Khyobeni, Lu Lu, William E. Armstrong, Robert W. Williams. Sex-specific modulation of gene expression networks in murine hypothalamus. Frontiers in Neuroscience . 2012;6(MAY):1–18. doi:10.3389/fnins.2012.00063. 46

Müller, Jürg, Craig M. Hart, Nicole J. Francis, Marcus L. Vargas, Aditya Sengupta, Brigitte Wild, et al. Histone Methyltransferase Activity of a Drosophila Polycomb Group Repressor Complex. Cell . 2002 oct;111(2):197–208. doi:10.1016/S0092-8674(02)00976-5. 24

Munoz-Torres, Monica C., Justin T. Reese, Christopher P. Childers, Anna K. Bennett, Jaideep P. Sundaram, Kevin L. Childs, et al. Hymenoptera Genome Database: Integrated community resources for insect species of the order Hymenoptera. Nucleic Acids Research . 2011;39(SUPPL. 1):658–662. doi:10.1093/nar/gkq1145. 5, 15

Nawaz, Hossain M, Per Kylsten, Noriko Hamada, Daisuke Yamamoto, C I Edvard Smith, Jessica M Lindvall. Differential evolutionary wiring of the tyrosine kinase Btk. PloS one . 2012 jan;7(5):e35640. doi:10.1371/journal.pone.0035640. 47

Niehuis, Oliver, Jan Buellesbach, Joshua D Gibson, Daniela Pothmann, Christian Hanner, Navdeep S Mutti, et al. Behavioural and genetic analyses of Nasonia shed light on the evolution of sex pheromones. Nature . 2013 feb;494(7437):345–8. doi:10.1038/nature11838. xii, 3, 5, 22

Nijhout, H Frederik. Development and evolution of adaptive polyphenisms. Evolution & development . 2003;5(1):9–18. doi:ede03003[pii]. vii

Olesnicky, Eugenia C, Claude Desplan. Distinct mechanisms for mRNA localization during embryonic axis specification in the wasp Nasonia. Developmental biology . 2007 jun; 306(1):134–42. doi:10.1016/j.ydbio.2007.03.012. 4

Olson-Manning, Carrie F., Maggie R. Wagner, Thomas Mitchell-Olds. Adaptive evolution: evaluating empirical support for theoretical predictions. Nature Reviews Genetics . 2012 nov;13(12):867–877. doi:10.1038/nrg3322. i

Ometto, Lino, DeWayne Shoemaker, Kenneth G. Ross, Laurent Keller. Evolution of gene expression in fire ants: the effects of developmental stage, caste, and species. Molecular Biology and Evolution . 2011 apr;28(4):1381–1392. doi:10.1093/molbev/msq322. 75

Opsahl, Tore. Structure and Evolution of Weighted Networks. University of London (Queen Mary College), London, UK, 2009, 104–122 . 55

Pannebakker, B. A., R. Watt, S. A. Knott, S. A. West, D. M. Shuker. The quantitative genetic basis of sex ratio variation in Nasonia vitripennis: A QTL study. Journal of Evolutionary Biology . 2011 jan;24(1):12–22. doi:10.1111/j.1420-9101.2010.02129.x. xii

Pannebakker, Bart a., Urmi Trivedi, Mark L Blaxter, Mark a Blaxter, Rebekah Watt, David M Shuker. The transcriptomic basis of oviposition behaviour in the parasitoid wasp Nasonia vitripennis. PloS one . 2013 jan;8(7):e68608. doi:10.1371/journal.pone.0068608. 3, 5

Papakostas, Spiros, L Asbjørn Vøllestad, Matthieu Bruneaux, Tutku Aykanat, Joost Vanoverbeke, Mei Ning, et al. Gene pleiotropy constrains gene expression changes in fish adapted to different thermal conditions. Nature communications . 2014 jan;5(0316):4071. doi:10.1038/ncomms5071. 47

Park, Jungsun, Zuogang Peng, Jia Zeng, Navin Elango, Taesung Park, Dave Wheeler, et al. Comparative analyses of DNA methylation and sequence evolution using Nasonia genomes. Molecular biology and evolution . 2011 dec;28(12):3345–54. doi:10.1093/molbev/msr168. 79

Parker, D J, a Gardiner, M C Neville, M G Ritchie, S F Goodwin. The evolution of novelty in conserved genes; evidence of positive selection in the Drosophila fruitless gene is localised to alternatively spliced exons. Heredity . 2013 oct;112(3):300–306. doi:10.1038/hdy.2013.106. 45

Parsch, John, Hans Ellegren. The evolutionary causes and consequences of sex-biased gene expression. Nature reviews Genetics . 2013 feb;14(2):83–7. doi:10.1038/nrg3376. xvi, 44

Patrick, Ellis, Michael Buckley, Yee Hwa Yang. Estimation of data-specific constitutive exons with RNA-Seq data. BMC bioinformatics . 2013 jan;14:31. doi:10.1186/1471-2105-14-31. 38

Pauli, Andrea, John L. Rinn, Alexander F. Schier. Non-coding RNAs as regulators of embryogenesis. Nature Reviews Genetics . 2011 feb;12(2):136–149. doi:10.1038/nrg2904. 4

Pauls, S, B Geldmacher-Voss, J a Campos-Ortega. A zebrafish histone variant H2A.F/Z and a transgenic H2A.F/Z:GFP fusion protein for in vivo studies of embryonic development. Development genes and evolution . 2001 dec;211(12):603–10. doi:10.1007/s00427-001-0196-x. 24

Payne, Joshua L., Jason H. Moore, Andreas Wagner. Robustness, Evolvability, and the Logic of Genetic Regulation. Artificial Life . 2014 jan;20(1):111–126. doi:10.1162/ARTL_a_00099. v

Payne, Joshua L, Andreas Wagner. Latent phenotypes pervade gene regulatory circuits. BMC systems biology . 2014 jana;8:64. doi:10.1186/1752-0509-8-64. vi

Payne, Joshua L, Andreas Wagner. The robustness and evolvability of transcription factor binding sites. Science . 2014 febb;343(6173):875–7. doi:10.1126/science.1249046. iv, v

Perry, Jennifer C., Peter W. Harrison, Judith E. Mank. The ontogeny and evolution of sex-biased gene expression in drosophila melanogaster. Molecular Biology and Evolution . 2014;31(5):1206–1219. doi:10.1093/molbev/msu072. 44, 75, 77, 78

Pers, Daniel, Thomas Buchta, Orhan Özüak, Selma Wolff, Jessica M. Pietsch, Mohammad Bilal Memon, et al. Global analysis of dorsoventral patterning in the wasp Nasonia reveals extensive incorporation of novelty in a regulatory network. BMC Biology . 2016; 14(1):63. doi:10.1186/s12915-016-0285-y. 45, 47

Pierson, Emma, Daphne Koller, Alexis Battle, Sara Mostafavi. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. PLoS Computational Biology . 2015; 11(5):1–19. doi:10.1371/journal.pcbi.1004220. 47

Pirkkala, L, P Nykanen, L Sistonen. Roles of the heat shock transcription factors in regulation of the heat shock response and beyond. FASEB journal : official publication of the Federation of American Societies for Experimental Biology . 2001;15(7):1118–1131. v

Poissant, Jocelyn, Alastair J Wilson, David W Coltman. Sex-specific genetic variance and the evolution of sexual dimorphism: A systematic review of cross-sex genetic correlations. Evolution . 2010 jan;64(1):97–107. doi:10.1111/j.1558-5646.2009.00793.x. xv

Pultz, M a, K K Zimmerman, N M Alto, M Kaeberlein, S K Lange, J N Pitt, et al. A genetic screen for zygotic embryonic lethal mutations affecting cuticular morphology in the wasp Nasonia vitripennis. Genetics . 2000 mar;154(3):1213–29. xii

Pultz, Mary Anne, David S Leaf. The jewel wasp Nasonia: querying the genome with haplo-diploid genetics. Genesis (New York, NY : 2000) . 2003 mar;35(3):185–91. doi: 10.1002/gene.10189. xii, 45

Pultz, Mary Anne, Lori Westendorf, Samuel D Gale, Kyle Hawkins, Jeremy A Lynch, Jason N Pitt, et al. A major role for zygotic hunchback in patterning the Nasonia embryo. Development . 2005 aug;132(16):3705–15. doi:10.1242/dev.01939. xii

Queitsch, Christine, Todd a Sangster, Susan Lindquist. Hsp90 as a capacitor of phenotypic variation. Nature . 2002 jun;417(6889):618–24. doi:10.1038/nature749. vi

Quicke, Donald L J, Others. Parasitic wasps. Chapman & Hall Ltd, 1997. 3

R Core Team. R: A Language and Environment for Statistical Computing. 2013. 16, 61

Rago, Alfredo, Donald G. Gilbert, Jeong-Hyeon Choi, Timothy B. Sackton, Xu Wang, Yogeshwar D. Kelkar, et al. OGS2: genome re-annotation of the jewel wasp Nasonia vitripennis. BMC Genomics . 2016;17(1):5303. doi:10.1186/s12864-016-2886-9. 1, 8, 51, 67

Raychoudhury, R, Christopher a Desjardins, J Buellesbach, D W Loehlin, B K Grillenberger, L Beukeboom, et al. Behavioral and genetic characteristics of a new species of Nasonia. Heredity . 2010;104(3):278–288. doi:10.1038/hdy.2009.147. 3

Rhee, David Y., Dong Yeon Cho, Bo Zhai, Matthew Slattery, Lijia Ma, Julian Mintseris, et al. Transcription factor networks in Drosophila melanogaster. Cell Reports . 2014; 8(6):2031–2043. doi:10.1016/j.celrep.2014.08.038. iii

Rivers, DB, DL Denlinger. Venom-induced alterations in fly lipid metabolism and its impact on larval development of the ectoparasitoid Nasonia vitripennis (Walker)(Hymenoptera:. Journal of Invertebrate Pathology . 1995;. xii

Roux, Julien, Marc Robinson-Rechavi. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. Genome Research . 2011;21(3):357–363. doi:10.1101/gr.113803.110. 32

Roy, Meenakshi, Namshin Kim, Y Xing, Christopher Lee. The effect of intron length on exon creation ratios during the evolution of mammalian genomes. RNA . 2008 sep; 14(11):2261–2273. doi:10.1261/rna.1024908. 30

Roy, Sushmita, Jason Ernst, Peter V Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L Eaton, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science . 2010 dec;330(6012):1787–97. doi:10.1126/science. 1198374. iii, x, 35

Rutherford, S L, S Lindquist. Hsp90 as a capacitor for morphological evolution. Nature . 1998 nov;396(6709):336–42. doi:10.1038/24550. vi

Sackton, Timothy B, John H. Werren, Andrew G Clark. Characterizing the infection-induced transcriptome of Nasonia vitripennis reveals a preponderance of taxonomically-restricted immune genes. PloS one . 2013 jan;8(12):e83984. doi:10.1371/journal.pone. 0083984. 3, 5, 15, 60

Schliep, K P. phangorn: phylogenetic analysis in R. Bioinformatics . 2011;27(4):592–593. 13, 16

Shin, T, D Kraemer, Jane Pryor, Ling Liu, James Rugila. Cell biology: a cat cloned by nuclear transplantation. Nature . 2002;415(February):2002. doi:10.1038/nature723. Supplementary. iv

Shine, Richard. Ecological causes for the evolution of sexual dimorphism: a review of the evidence. The Quarterly review of biology . 1989 dec;64(4):419–61. vii

Shukla, Sanjeev, Ersen Kavak, Melissa Gregory, Masahiko Imashimizu, Bojan Shutinoski, Mikhail Kashlev, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature . 2011 nov;479(7371):74–9. doi:10.1038/nature10442. 28

Simão, Felipe A., Robert M Waterhouse, Panagiotis Ioannidis, Evgenia V Kriventseva, Evgeny M Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics . 2015 oct;31(19):3210–3212. doi:10.1093/ bioinformatics/btv351. 10

Simon, Jean-Christophe, Michael E Pfrender, Ralph Tollrian, Denis Tagu, John K Colbourne. Genomics of environmentally induced phenotypes in 2 extremely plastic arthropods. The Journal of heredity . 2011;102(5):512–25. doi:10.1093/jhered/esr020. vii

Skinner, SW. Maternally inherited sex ratio in the parasitoid wasp Nasonia vitripennis. Science . 1982;215(4536):1133–1134. xiii

Smyth, Gordon K. Limma: linear models for microarray data. In: Gentleman, R, V Carey, S Dudoit, R Irizarry, W Huber, editors, Bioinformatics and Computational Biology Solutions Using {R} and Bioconductor, New York: Springer, 2005. 397–420. 56

Snell-Rood, Emilie C, James David Van Dyken, Tami Cruickshank, Michael J Wade, Armin P Moczek. Toward a population genetic framework of developmental evolution: the costs, limits, and consequences of phenotypic plasticity. BioEssays : news and reviews in molecular, cellular and developmental biology . 2010 jan;32(1):71–81. doi:10.1002/bies.200900132. vii

Soshnev, Alexey A., Steven Z. Josefowicz, C. David Allis. Greater Than the Sum of Parts: Complexity of the Dynamic Epigenome. Molecular Cell . 2016 jun;62(5):681–694. doi:10.1016/j.molcel.2016.05.004. ii

Spitz, Francois, Eileen E M Furlong. Transcription factors: from enhancer binding to developmental control. Nature Reviews Genetics . 2012;13(9):613–626. doi:10.1038/nrg3207. 46

Standen, Emily M., Trina Y. Du, Hans C. E. Larsson. Developmental plasticity and the origin of tetrapods. Nature . 2014 aug;doi:10.1038/nature13708. vi

Stazic, D., B. Voß. The complexity of bacterial transcriptomes. Journal of Biotechnology . 2016 aug;232:69–78. doi:10.1016/j.jbiotec.2015.09.041. ii

Sterck, Lieven, Stephane Rombauts, Klaas Vandepoele, Pierre Rouzé, Yves Van de Peer. How many genes are there in plants (... and why are they there)? Current Opinion in Plant Biology . 2007;10(2):199–203. doi:10.1016/j.pbi.2007.01.004. 4

Strimmer, Korbinian. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics (Oxford, England) . 2008 jun;24(12):1461–2. doi:10.1093/bioinformatics/btn209. 56

Stuart, Joshua M. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science . 2003 oct;302(5643):249–255. doi:10.1126/science.1087447. iii

Su, Zhixi. Evolution of alternative splicing after gene duplication. Genome Research . 2005 dec;16(2):182–189. doi:10.1101/gr.4197006. 32

Su, Zhixi, Xun Gu. Revisit on the evolutionary relationship between alternative splicing and gene duplication. Gene . 2012;504(1):102–106. doi:10.1016/j.gene.2012.05.012. 32

Subramanian, Aravind, Pablo Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences . 2005 oct;102(43):15545–15550. doi:10.1073/pnas.0506580102. 47

Sun, Bao-Fa, Yong-Xing Li, Ling-Yi Jia, Li-Hua Niu, Robert W. Murphy, Peng Zhang, et al. Regulation of transcription factors on sexual dimorphism of fig wasps. Scientific Reports . 2015;5:10696. doi:10.1038/srep10696. 44

Suzuki, Y. Evolution of a Polyphenism by Genetic Accommodation. Science . 2006; 311(5761):650–652. doi:10.1126/science.1118888. vi

Takahashi, Kazuo H. Multiple capacitors for natural genetic variation in Drosophila melanogaster. Molecular ecology . 2013 mar;22(5):1356–65. doi:10.1111/mec.12091. vi

Takahashi, Kazuo H, Phillip J Daborn, Ary a Hoffmann, Toshiyuki Takano-Shimizu. Environmental stress-dependent effects of deletions encompassing Hsp70Ba on canalization and quantitative trait asymmetry in Drosophila melanogaster. PloS one . 2011 jan; 6(4):e17295. doi:10.1371/journal.pone.0017295. vi

Takahashi, Kazuo H, Yasukazu Okada, Kouhei Teramura. Deficiency screening for genomic regions with effects on environmental sensitivity of the sensory bristles of drosophila melanogaster. Evolution . 2012 sep;66(9):2878–2890. doi:10.1111/j.1558-5646.2012.01636.x. vi

Talavera, David, Christine Vogel, Modesto Orozco, Sarah a. Teichmann, Xavier De La Cruz. The (In)dependence of alternative splicing and gene duplication. PLoS Computational Biology . 2007;3(3):0375–0388. doi:10.1371/journal.pcbi.0030033. 32

Talbert, Paul B, Steven Henikoff. Histone variants–ancient wrap artists of the epigenome. Nature reviews Molecular cell biology . 2010 apr;11(4):264–75. doi:10.1038/nrm2861. 24

Telonis-Scott, M., A. Kopp, M. L. Wayne, S. V. Nuzhdin, L. M. McIntyre. Sex-Specific Splicing in Drosophila: Widespread Occurrence, Tissue Specificity and Evolutionary Conservation. Genetics . 2008 nov;181(2):421–434. doi:10.1534/genetics.108.096743. 44, 45, 75

Tennessen, Jason M, Nicolas M Bertagnolli, Janelle Evans, Matt H Sieber, James Cox, Carl S Thummel. Coordinated metabolic transitions during Drosophila embryogenesis and the onset of aerobic glycolysis. G3 (Bethesda, Md) . 2014 mar;4(5):839–50. doi:10.1534/g3.114.010652. 77, 79

Tesson, Bruno M, Rainer Breitling, Ritsert C Jansen. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. BMC bioinformatics . 2010; 11:497. doi:10.1186/1471-2105-11-497. 46, 57

Thibaud-Nissen, Françoise, Alexander Souvorov, Terence Murphy, Michael DiCuccio, Paul Kitts. Eukaryotic Genome Annotation Pipeline. 2013. 6

Thompson, M J, C D Jiggins. Supergenes and their role in evolution. Heredity . 2014 mar; 113(1):1–8. doi:10.1038/hdy.2014.20. 63, 75

Trapnell, Cole, Brian a Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology . 2010 may;28(5):511–5. doi:10.1038/nbt.1621. 38

Trent, C, C Crosby, J Eavey. Additional evidence for the genomic imprinting model of sex determination in the haplodiploid wasp Nasonia vitripennis: isolation of biparental diploid males after X-ray mutagenesis. Heredity . 2006 may;96(5):368–76. doi:10.1038/sj.hdy.6800810. xiv

True, John R., Eric S. Haag. Developmental system drift and flexibility in evolutionary trajectories. Evolution and Development . 2001;3(2):109–119. doi:10.1046/j.1525-142X.2001.003002109.x. 81

Van de Peer, Yves, Steven Maere, Axel Meyer. The evolutionary significance of ancient genome duplications. Nature reviews Genetics . 2009 oct;10(10):725–32. doi:10.1038/nrg2600. 17

van den Berg, Bart H J, Fiona M McCarthy, Susan J Lamont, Shane C Burgess. Re-annotation is an essential step in systems biology modeling of functional genomics data. PloS one . 2010 jan;5(5):e10642. doi:10.1371/journal.pone.0010642. xi

van Dongen, Stijn. Graph Clustering by Flow Simulation. Ph.D. thesis, University of Utrecht. 2000. 12

Van Nas, Atila, Debraj Guhathakurta, Susanna S. Wang, Nadir Yehya, Steve Horvath, Bin Zhang, et al. Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks. Endocrinology . 2009;150(3):1235–1249. doi:10.1210/en.2008-0563. 46

Verhulst, Eveline C, Leo W Beukeboom, Louis van de Zande. Maternal Control of Haplodiploid Sex Determination in the Wasp Nasonia. Science . 2010 apra;328(5978):620–623. doi:10.1126/science.1185805. xiv, xv, 79

Verhulst, Eveline C, Jeremy A Lynch, Daniel Bopp, Leo W Beukeboom, Louis van de Zande. A new component of the nasonia sex determining cascade is maternally silenced and regulates transformer expression. PloS one . 2013 jan;8(5):e63618. doi:10.1371/journal.pone.0063618. xiv, 45, 46, 79

Verhulst, Eveline C, Louis van de Zande. Insect sex determination: a cascade of mechanisms. Sexual development : genetics, molecular biology, evolution, endocrinology, embryology, and pathology of sex determination and differentiation . 2014 jan;8(1-3):5–6. doi:10.1159/000358405. 44

Verhulst, Eveline C, Louis van de Zande, Leo W Beukeboom. Insect sex determination: it all evolves around transformer. Current opinion in genetics & development . 2010 augb;20(4):376–83. doi:10.1016/j.gde.2010.05.001. xiv, 45, 81

Vibranovski, Maria D., Hedibert F. Lopes, Timothy L. Karr, Manyuan Long. Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. PLoS Genetics . 2009;5(11). doi:10.1371/journal.pgen.1000731. 44

Villar, Diego, Paul Flicek, Duncan T Odom. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nature reviews Genetics . 2014 apr;15(4):221–33. doi:10.1038/nrg3481. viii

Waddington, CH. Canalization of development and the inheritance of acquired characters. Nature . 1942;(3811):563–565. v

Waddington, CH. Genetic assimilation of an acquired character. Evolution . 1953; 7(2):118–126. vi

Wagner, A, D A Fell. The small world inside large metabolic networks. Proceedings of the Royal Society B: Biological Sciences . 2001 sep;268(1478):1803–1810. doi:10.1098/rspb.2001.1711. 53

Walley, Aj, P Jacobson, M Falchi, L Bottolo, Jc Andersson, E Petretto, et al. Differential coexpression analysis of obesity-associated networks in human subcutaneous adipose tissue. International Journal of Obesity . 2012;36(1):1–11. doi:10.1038/ijo.2011.22. 57

Wang, Xu, John H. Werren, Andrew G. Clark. Genetic and epigenetic architecture of sex-biased expression in the jewel wasps Nasonia vitripennis and giraulti. Proceedings of the National Academy of Sciences . 2015 jul;112(27):E3545–E3554. doi:10.1073/pnas.1510338112. 5, 26, 45, 63, 77, 79, 80

Wang, Xu, David Wheeler, Amanda Avery, Alfredo Rago, Jeong-Hyeon Choi, John K Colbourne, et al. Function and evolution of DNA methylation in Nasonia vitripennis. PLoS genetics . 2013 oct;9(10):e1003872. doi:10.1371/journal.pgen.1003872. 3, 5, 15, 28, 31, 79

Waterhouse, Robert M. A maturing understanding of the composition of the insect gene repertoire. Current Opinion in Insect Science . 2015;7(January):15–23. doi:10.1016/j.cois.2015.01.004. 4

Waterhouse, Robert M, Fredrik Tegenfeldt, Jia Li, Evgeny M Zdobnov, Evgenia V Kriventseva. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic acids research . 2013 jan;41(Database issue):D358–65. doi:10.1093/nar/gks1116. 5, 10, 12

Weinstock, George M., Gene E. Robinson, Richard a. Gibbs, Kim C. Worley, Jay D. Evans, Ryszard Maleszka, et al. Insights into social insects from the genome of the honeybee Apis mellifera. Nature . 2006 oct;443(7114):931–949. doi:10.1038/nature05260. 4

Werren, John H. Sex ratio adaptations to local mate competition in a parasitic wasp. Science . 1980;208(JUNE):1157–1159. xii, 76

Werren, John H. The paternal-sex-ratio chromosome of Nasonia. American Naturalist . 1991;:392–402. xiii

Werren, John H., Lorna B. Cohen, Juergen Gadau, Rita Ponce, Emmanuelle Baudry, Jeremy A. Lynch. Dissection of the complex genetic basis of craniofacial anomalies using haploid genetics and interspecies hybrids in Nasonia wasps. Developmental Biology . 2015;415(2):391–405. doi:10.1016/j.ydbio.2015.12.022. 3

Werren, John H., David W Loehlin. The parasitoid wasp Nasonia: An emerging model system with haploid male genetics. Cold Spring Harbor Protocols . 2009;4(10):1–31. doi:10.1101/pdb.emo134. 3

Werren, John H., Stephen Richards, Christopher a Desjardins, Oliver Niehuis, Jürgen Gadau, John K Colbourne, et al. Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science . 2010 jan;327(5963):343–8. doi: 10.1126/science.1178028. 8, xi, 3, 17, 21, 50, 80, 89

Whiting, Phineas W. The chromosomes of Mormoniella. The Journal of heredity . 1968; 59(1):19–22. 77

Wickham, Hadley. Reshaping Data with the {reshape} Package. Journal of Statistical Software . 2007;21(12):1–20. 16, 61

Wickham, Hadley. ggplot2: elegant graphics for data analysis. Springer New York, 2009. ISBN 978-0-387-98140-6. 16, 61

Wickham, Hadley. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software . 2011;40(1):1–29. 16, 61

Williams, Tim D, Nil Turan, Amer M Diab, Huifeng Wu, Carolynn Mackenzie, Katie L Bartie, et al. Towards a system level understanding of non-model organisms sampled from the environment: a network biology approach. PLoS computational biology . 2011 aug;7(8):e1002126. doi:10.1371/journal.pcbi.1002126. ii

Wu, Thomas D., Colin K. Watanabe. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics . 2005;21(9):1859–1875. doi:10.1093/ bioinformatics/bti310. 8, 17, 89

Wyman, Minyoung J, Asher D Cutter, Locke Rowe. Gene duplication in the evolution of sexual dimorphism. Evolution . 2012 may;66(5):1556–1566. doi:10.1111/j.1558-5646. 2011.01525.x. xvi

Xiao, Jin-Hua, Zhen Yue, Ling-Yi Jia, Xin-Hua Yang, Li-Hua Niu, Zhuo Wang, et al. Obligate mutualism within a host drives the extreme specialization of a fig wasp genome. Genome biology . 2013 jan;14(12):R141. doi:10.1186/gb-2013-14-12-r141. 79

Xue, Shifeng, Siqi Tian, Kotaro Fujii, Wipapat Kladwang, Rhiju Das, Maria Barna. RNA regulons in Hox 5' UTRs confer ribosome specificity to gene regulation. Nature . 2014 nov;517(7532):33–38. doi:10.1038/nature14010. 4

Yakulov, Toma, Ufuk Günesdogan, Herbert Jäckle, Alf Herzig, H. Aihara, T. Nakagawa, et al. Bällchen participates in proliferation control and prevents the differentiation of Drosophila melanogaster neuronal stem cells. Biology open . 2014 oct;3(10):881–6. doi:10.1242/bio.20148631. 67

Yang, Jing, Hui Yu, Bao Hong Liu, Zhongming Zhao, Lei Liu, Liang Xiao Ma, et al. DCGL v2.0: An R package for unveiling differential regulation from differential co-expression. PLoS ONE . 2013;8(11):4–11. doi:10.1371/journal.pone.0079729. 46, 57

Yeaman, S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. Proceedings of the National Academy of Sciences . 2013 apr;doi:10.1073/pnas.1219381110. 76

Yu, Hui, Bao-Hong Liu, Zhi-Qiang Ye, Chun Li, Yi-Xue Li, Yuan-Yuan Li. Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. BMC bioinformatics . 2011;12(1):315. doi:10.1186/1471-2105-12-315. 57

Zampieri, Mattia, Nicola Soranzo, Claudio Altafini. Discerning static and causal interactions in genome-wide reverse engineering problems. Bioinformatics (Oxford, England) . 2008 jul;24(13):1510–5. doi:10.1093/bioinformatics/btn220. 46

Zhang, Bin, Steve Horvath. A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology . 2005 jan;4(1). doi:10.2202/1544-6115.1128. 53, 54

Zhao, Min, Xing-Fu Zha, Jin Liu, Wen-Ji Zhang, Ning-Jia He, Dao-Jun Cheng, et al. Global expression profile of silkworm genes from larval to pupal stages: Toward a comprehensive understanding of sexual differences. Insect Science . 2011 dec;18(6):607–618. doi:10.1111/j.1744-7917.2010.01392.x. 44, 78

Zwier, M V, E C Verhulst, R D Zwahlen, L W Beukeboom, Louis van de Zande. DNA methylation plays a crucial role during early Nasonia development. Insect molecular biology . 2012 feb;21(1):129–38. doi:10.1111/j.1365-2583.2011.01121.x. xv, 46