2021

Identification and Sequence Analysis of SET Superfamily Genes in Hymenopteran Insects

Tanima Joshi
Scripps College

Identification and Sequence Analysis of SET Superfamily Genes in Hymenopteran Insects

A Thesis Presented by

Tanima Joshi

To the Keck Science Department

Of Claremont McKenna, Pitzer, and Scripps Colleges

In partial fulfillment of

The degree of Bachelor of Arts

Senior Thesis in Molecular Biology

January 27th, 2021

Table of Contents

ABSTRACT

Post-translational chemical modifications of histones, the proteins that package DNA, are a key component of the "language" of epigenetics. SET (Suppressor of Variegation 39, Enhancer of Zeste, Trithorax) domain containing proteins are crucial enzymes that establish histone methylations as a major part of the histone code. Studies using the fruit fly, *Drosophila melanogaster*, as a model organism have provided most of the known information on SET genes/protein diversity and function in insects. To expand our knowledge of this important protein family, I identified all SET domain containing genes in a select group of insects belonging to the Hymenoptera order: wasps, bees, ants and sawflies. I investigated whether the number of SET genes correlated with factors that contributed to genome complexity. Species with sex chromosomes and greater number of total genes within their genome tended to have a greater number of SET genes, while overall genome size (in Mb of DNA) and number of chromosomes did not. An outlier to these trends was the jewel wasp, *Nasonia vitripennis*, with far more SET genes than any other examined species. Phylogenetic comparisons between the jewel wasp and species of varying relation to it revealed that *SET and MYND* (**My**eloid, **N**ervy, and **D**EAF-1) and *Lysine Methyltransferase 5* (KMT5) genes expanded extensively in the jewel wasp compared to other species, suggesting that diversification of these gene families has played an important role in genome function in this organism. Studying chromatin remodelers such as SET domain containing proteins can further our understanding of the dynamics of genome evolution and factors that impact epigenetics and life-threatening diseases such as cancer.

INTRODUCTION

The genome is packaged with histone proteins and other non-histone proteins into a complex known as chromatin. There are two main types of chromatin: heterochromatin and euchromatin; however there also exists many different subtypes (Lee & Orr-Weaver, 2001). Heterochromatin is defined by very tight packaging of DNA around histones, such that genes located adjacent to or within these regions are transcriptionally silenced (Allshire & Madhani, 2018). Euchromatin is believed to be a "less dense" form of chromatin that allows for the genes in these regions to be expressed (Lee & Orr-Weaver, 2001). At its most basic form, chromatin consists of DNA wrapped around proteins known as histones to form a structure known as the nucleosome. As mentioned earlier, it is thought that the density or level of compactness of chromatin determines whether a gene is expressed in an organism, which heavily relies on chemical modifications to histones. These histone modifications can result in types of signals that govern certain processes of chromatin, including gene silencing, transcriptional activation or deactivation, chromatin restructuring, DNA recombination, and compaction into resolved chromosomes for segregation during division (Mariño-Ramírez et al., 2005). These chromatin states as dictated by chemical modifications to histones are a defining feature of epigenetics. Epigenetics is described, literally, as "in addition to changes in genetic sequence," but it also includes processes that result in changes to gene expression without alterations to DNA that can be passed on to daughter cells. Some examples of epigenetic processes included acetylation, phosphorylation, ubiquitination, and methylation (which is the focus of my study) (Epigenetics: The Science of Change, 2006).

Methylation of histones is an important aspect of chromatin remodeling and has been shown to affect gene transcription (Trievel et al., 2002). SET domain containing genes are the

"writers" of these methylation marks. The name "SET" is an abbreviation of the first three genes found to contain this domain in the fruit fly, *Drosophila melanogaster*, named: Suppressor of Variegation 39 (Su(Var)3-9), Enhancer of Zeste, and Trithorax. The function of this conserved domain is to transfer a methyl group from *S*-adenosyl-L-methionine (AdoMet) to the amino group of a lysine or arginine residue on a histone forming the cofactor, *S*-adenosyl-L-homocysteine (AdoHcy) in the process (Dillon, S.C., et al., 2005). Genes containing a SET domain can influence the histone code and, therefore, chromatin states. These changes can affect gene expression, illustrating that SET genes play a central role in epigenetics.

While the function of the SET domain is conserved among a number of genes across multiple species, conservation of amino acids sequences in genes containing SET domains is moderate. Studies have shown that there is a greater existence of conservation in the structural conformation of genes that contain the domain (Yeates, 2002). The similarities in structure between genes containing SET domains are what leads to similarities in their function. Solved protein structures of genes that contain SET domains show the existence of a knot-like structure, commonly referred to as a "pseudoknot," formed by the C-terminus of the protein threaded through a loop in the protein, surrounded by many beta-sheets. This "pseudoknot" forms an active site near the binding site of the methyl donor and peptide binding cleft. A channel through the SET domain allows for transfer of the methyl group from AdoMet to the amino group on the lysine by connecting their binding sites (Couture et al., 2009). Another key feature of the pseudoknot is that it is formed by the most conserved amino acid sequences motifs found in studied SET genes: RFINHXCXPN and ELXFDY (Qian & Zhou, 2006; Zhang, & Bruice, 2008). Alterations to this catalytic site such as deletions or mutations could lead to severe consequences in activity.

Mutated SET domains can promote the formation of disease and other serious problems. Many studies have shown the implications of mutated SET genes on cancer. For example, the SET domain containing Mixed Lineage Leukaemia (MLL) family of genes, specifically the gene *MLL3* was found to be deleted in myeloid leukemia patients, and therefore is called a tumor suppressor gene. Additionally, glioblastoma, melanoma, pancreatic and breast cancer patients also had deletions in *MLL3* (Weirich et al., 2015). Another study discovered that, in the gene *SET1A*, a deleted SET domain triggered apoptosis and hindered proliferation of embryonic stem cells (Sze et al., 2017). The eggless SET gene product (known as SETDB1 in humans and *Nasonia vitripennis*) is required for trimethylation of the histone 3 at lysine 9 in the fruit fly, *Drosophila melanogaster*. One study found that upon deletion of the SET domain in this gene, oogenesis was arrested in early stages (Clough et al., 2007). Furthermore, discoveries have shown correlations between histone methyltransferases and cancer parthenogenesis. In 2004, small molecule inhibitors of histone lysine methyltransferases, called methylases, were discovered and found to exhibit selected cancer cell killing (Liu & Wang, 2016). Together, these studies illustrate that SET genes are vital to the survival of an organism. Studying patterns in these genes across multiple species can further help to shed light on the overall dynamics of genome evolution.

Studies in insects have produced much of the current knowledge of SET gene function. Most of what is known about SET genes in insects comes from studies in *Drosophila melanogaster*. Thus, little is known about SET gene function and diversity in other insect groups. In this study, I have investigated SET gene diversity in a number of insects belonging to the order Hymenoptera. Hymenoptera is extremely diverse and dates back to the Triassic period. Species within this order fall into four main groups of insects: sawflies, bees, wasps and ants,

totaling around 150,000 characterized species and one million estimated species (Peters et al., 2017). Additionally, Hymenoptera do not have sex chromosomes and instead reproduce by haplo-diploidy meaning that males are haploid and develop from unfertilized eggs, while females are diploid and develop from fertilized eggs (Branstetter et al., 2018).

Genome complexity can change and vary across the four hymenopteran groups over evolutionary time. Differences in genome composition such as variations in amounts of heterochromatin or euchromatin contribute to the level of complexity in a given species genome. However, little is known about how the machinery that dictates chromatin states responds to changes that occur in the genome throughout evolution. Of the many factors that contribute to genome complexity, I have chosen four to focus on in my study. These include whether a species has sex chromosomes, its number of chromosomes, its total number of genes, and its genome size. If a given species has more genes, a larger genome, or more chromosomes it is likely that that its genome is relatively more complex. Sex chromosomes are thought to be more complex that autosomes given the unique nature of their inheritance. Studies argue that sex chromosomes are responsible for the rise of intralocus conflict, in which alleles determine fitness based on the sex they are found in, and intragenomic conflict, in which the phenotype of selfish genetic elements promote their own transmission (Mank et al., 2014). Therefore, the presence of sex chromosomes is a key aspect to observe when studying genome evolution.

In this study, I have identified all SET genes in the jewel wasp, *Nasonia vitripennis*, and other selected Hymenopteran insects. I first observed trends relating number of SET genes of each species to factors contributing to genome complexity. Results from this process revealed an outlier: the jewel wasp, which had far more SET genes than any other species. Therefore, I created phylogenies between hymenopteran insects with varying levels of relatedness to *Nasonia*

vitripennis and itself to analyze patterns of divergence in SET genes, as measured by number of gene copies or duplications.

METHODS

NCBI Protein BLAST

NCBI protein BLAST was used to identify SET domain containing genes in select Hymenopteran and insect outgroup species. I used a list of amino acid sequences from the fruit fly, *Drosophila melanogaster*, as a query sequence for all other protein BLAST searches, since the SET domain was first discovered in this species. This list of query sequences was obtained from the resource PFAM and cross-checked with a list of genes gathered using names and references from scientific articles to be as exhaustive as possible. The query list consisted of 36 SET domain containing genes.

SET genes were identified in two species from each of the Hymenoptera group: sawflies, wasps, bees, and ants, and in three insect outgroup species, which were chosen based on diversity. Protein BLAST was used to identity SET genes for the following ingroup species: *Bombus terrestris, Linepithema humile, Orussus abietinus, Certosolen solmsi marchali,* and *Athalia rosae*. It was also used for the following outgroup species: *Lepisma saccharina*. All other species SET domain lists were collected from PFAM. Ingroup species include: *Nasonia vitripennis, Apis mellifera* and *Atta cephalotes*. Outgroup species included: *Drosophila melanogaster, Tribolium castaneum*.

NCBI protein BLAST was carried out as follows: each query sequence was blasted against the listed species above. Once a list of gene hits appeared, genes were selected only if they contained a SET domain and if they were unique to genes recorded in prior searches. To determine whether a given gene contained a SET domain, I used the NCBI conserved protein domain search. Sequences were also double-checked for amino acids patterns known to be conserved in SET domains. Secondly, if different isoforms of the same gene appeared, only the

longest amino acid sequence was selected. Location and direction of similar genes were studied to determine whether genes were unique. If two or more genes were similar in direction, location or length, NCBI protein BLAST alignment was used to directly compare the genes to one another. Genes that were then substantially different from each other (i.e. they had *e* value near 0 and a query cover and score were below 50 percent), were then included in the list of new genes. Both named, predicted and uncharacterized genes were identified for species, and their identification number (XP number), and amino acids sequence were entered in an Excel spreadsheet.

Phylogenetic Tree Analysis

Phylogenies were created using the SET gene lists recorded from NCBI protein BLAST searches. The Excel file containing the lists to be included were converted to a text file, which was then uploaded to the online software Clustal Omega. In this software, the sequence alignment and tree file were created. The tree file was in neighbor-joining tree format. This file was then uploaded to Interactive Tree of Life (iTOL) to create a cladogram such as those presented in Figures 1-4. Branch lengths were omitted to showcase the phylogeny more clearly.

RESULTS

NCBI Protein BLAST Identification of SET Genes in Hymenoptera

My initial goal was to identify all of the SET domain containing genes in the genome of the jewel wasp, *Nasonia vitripennis*, and a handful of other hymenopteran species. To do this, I used BLAST searches of the NCBI protein databases for these organisms using *D. melanogaster* SET genes as query sequences. In order to be as exhaustive as possible with these searches, I combined my retrieved genes for these species with those listed for the same species in the PFAM protein database.

In these searches, I identified SET genes within two species from each major group of the Hymenoptera insect order, which includes wasps, bees, ants and sawflies. My analyses were restricted to species with sequenced and publicly archived genomes. In wasps, I chose *Nasonia vitripennis* and *Ceratosolen solmsi marchali*, the fig wasp, for which I found 46 and 34 SET genes, respectively. In bees, I chose *Bombus terrestris*, the buff-tailed bubble bee, and *Apis mellifera*, the honeybee, and found that they had 26 and 29 SET genes, respectively. I found 30 and 32 SET genes in the two sawflies, *Orussus abietinus*, the parasitic wood wasp (sawfly), and *Athalia rosae*, the turnip sawfly, respectively. Finally, the two ant species *Atta cephalotes*, the leaf cutter ant, and *Linepithema humile*, the Argentine ant, had 23 and 29 SET genes, respectively. Of the Hymenopteran species that I examined, *Nasonia vitripennis* had the most SET genes (46) while *Atta cephalotes* had the least (23) (Table 1). Moreover, wasps had a greater number of SET genes in comparison to other hymenopteran groups, while ants had fewer. Sawflies and bees had intermediate numbers of the SET genes relative to the other groups (Table 1). The difference in SET gene count between the jewel wasp and fig wasp varied the most,

relative to SET gene number differences between species within the other three groups. Ants, bees and sawflies had similar SET gene numbers within their respective groups (Table 1).

Additionally, I identified SET genes in three non-hymenopteran insect species: *Lepisma saccharina*, the silverfish, *Tribolium castaneum*, the red flour beetle, and *Drosophila melanogaster*, the fruit fly. I identified 36 SET genes in the fruit fly, 35 in the red flour beetle and 1 in the silverfish. Of all the species that I studied (ingroup and outgroup), the silverfish had the least number of SET genes (only 1), while the jewel wasp still had the greatest (46) (Table 1).

Comparison of SET gene number to genome complexity

I next wanted to test if the number of SET genes in the hymenopteran species correspond with any of several different genome characteristics, including total gene number per genome, genome size (in Megabases of DNA), presence or absence of sex chromosomes, and the total number of chromosomes. I found that the red flour beetle and the fruit fly both have sex chromosomes and a greater number of total genes within their genome than the other species in this study. With the exception of the jewel wasp, these two species also have more SET genes than the others. Both bee species, the buff-tailed bumble bee and the honeybee, have the greatest number of chromosomes (18 and 16, respectively), but they have fewer SET genes than all other species except the ants. Additionally, both bees as well as the leafcutter ant and the fig wasp have greater genome sizes (over 200 Mb) than the other species (Table 1). Yet, three of these species (both bees and the leafcutter ant) have less than 30 SET genes, which is a smaller number of genes than what was identified in species with smaller genome size. The fruit fly and jewel wasp have the least number of chromosomes but have the two highest SET gene counts amongst

all species. Taken together, these findings show that larger genome size and greater number of chromosomes seems to correspond loosely (at least with the species examined in my study) with smaller numbers of SET genes. Conversely, a greater number of total genes and presence of sex chromosomes corresponds with greater numbers of SET genes. *N. vitripennis* is an outlier to these trends because it doesn't have any sex chromosomes and intermediate number of total genes but has the highest number of SET genes.

Species Name	**Common name**	**Number of SET Genes**	**Total number of genes**	**Genome size (Mb)**	**Sex chromosomes (yes/no)**	**Number of chromosomes**
Drosophila melanogaster	fruit fly	36	15,682	143.73	Yes	4
Tribolium castaneum	Red flour beetle	35	16,593	163.93	Yes	10
Lepisma saccharina	silverfish	1				
Bombus terrestris	Buff-tailed bumble bee	26	11,875	433	No	18
Apis mellifera	honey bee	29	10,600	236	No	16
Orussus abietinus	parasitic wood wasp (sawfly)	30	11,063	186.4	No	x
Athalia rosae	turnip sawfly	32	12,365	155.84	No	8
Linepithema humile	argentine ant	29	13,084	219.5	No	8
Atta cephalotes	leafcutter ant	23	11,117	317.69	No	
Ceratosolen solmsi marchali	fig wasp	34	10,189	277.06	No	X
Nasonia vitripennis	Jewel wasp	46	13,141	137.84	No	5

Table 1. Genome information for Hymenopteran species used phylogenies displaying scientific name, common name, number of SET domain containing genes, number of total genes, genome size (Mb), presence of sex chromosomes for each species displayed in phylogenies. This data was collected through NCBI protein BLAST and from the NCBI species genome database.

Phylogenetic Comparison of SET Genes in Species Among Hymenopteran Insects

I reasoned that the unusually high number of SET genes in *N. vitripennis* (Table 1) may result from gene duplication events that occurred during the evolution of this insect group. It is possible that such expansions occurred by multiple duplications of one or a few particular genes, perhaps due to certain chromatin related needs that arose in the *N. vitripennis* genome. Another possibility is that these gene number differences came about due to small numbers of duplications of individual genes across the SET gene family in *N. vitripennis*. To distinguish between these possibilities, I performed pairwise comparisons between the SET gene families of *N. vitripennis* with those of several select insects, ranging from more to less evolutionary distance from *N. vitripennis*. In principle, if there is no variation in the number and identities of SET genes between two given species, the resulting cladogram should exhibit a 1:1 correspondence of gene from each species on each branch (*i.e.*, a "doublet" of the two genes on each branch). In contrast, a gene duplication would result in a branch containing multiple gene copies (true copies or paralogs) from one species alongside a matched gene from another species.

I first examined a phylogeny that included the complete array of SET genes from *N. vitripennis* and the fig wasp, *Ceratosolen solmsi marchali*. It was apparent from this phylogeny that a group of genes known as SMYD, standing for SET and MYND (**my**eloid, **N**ervy, and **D**EAF-1) domain-containing genes, underwent multiple gene duplication events in the *N. vitripennis* genome (Figure 1). The KMT5 gene, also known as Suv4-20 sub-group, also underwent substantial expansion in this lineage. *N. vitripennis* contained a total of 18 SMYD-like genes and 5 KMT5 genes, compared with 11 and 0 in *C. solmsi marchali,* respectively. This

seems like a striking difference in the number of these specific SET genes between these relatively closely related species. I also note that a similar pattern was observed in SET genes between *N. vitripennis* and the bumblebee, *Bombus terrestris*, which contained 10 SMYD genes and 1 KMT5 (Figure 2). I also phylogenetically compared SET genes between *N. vitripennis* and a more distantly related hymenopteran, *Athalia rosae*, and the dipteran fruit fly, *Drosophila melanogaster*. *D. melanogaster* contained 14 SMYD genes and only 1 KMT5 gene, while *A. rosae* contained 16 SMYD genes and no KMT5 genes (Figures 3 & 4). Lineage specific expansion occurred in the KMT5 gene in *N. vitripennis* relative to all species with which it was compared. However, the number of SMYD genes seemed to increase in number in more distantly related species. On closer inspection, I observed that the number of SMYD gene sister pairings that showed a 1:1 correspondence between *N. vitripennis* and the outgroup species increased with more closely related outgroup species. There were no SMYD gene copies in either *C. solmsi marchali* or *B. terrestris* SMYD genes (meaning that each gene had a corresponding *N. vitripennis* sister pair) (Figure 1 & 2) , while both *D. melanogaster* and *A. rosae* had many (Figure 3 & 4).

In summary, my phylogenetic analysis has revealed that two SET gene subgroups – KMT5 and SMYD – seem to have expanded substantially within the *N. vitripennis* lineage but not in the other analyzed hymenopteran insects, thus explaining the larger total number of SET genes in this species compared to the others.

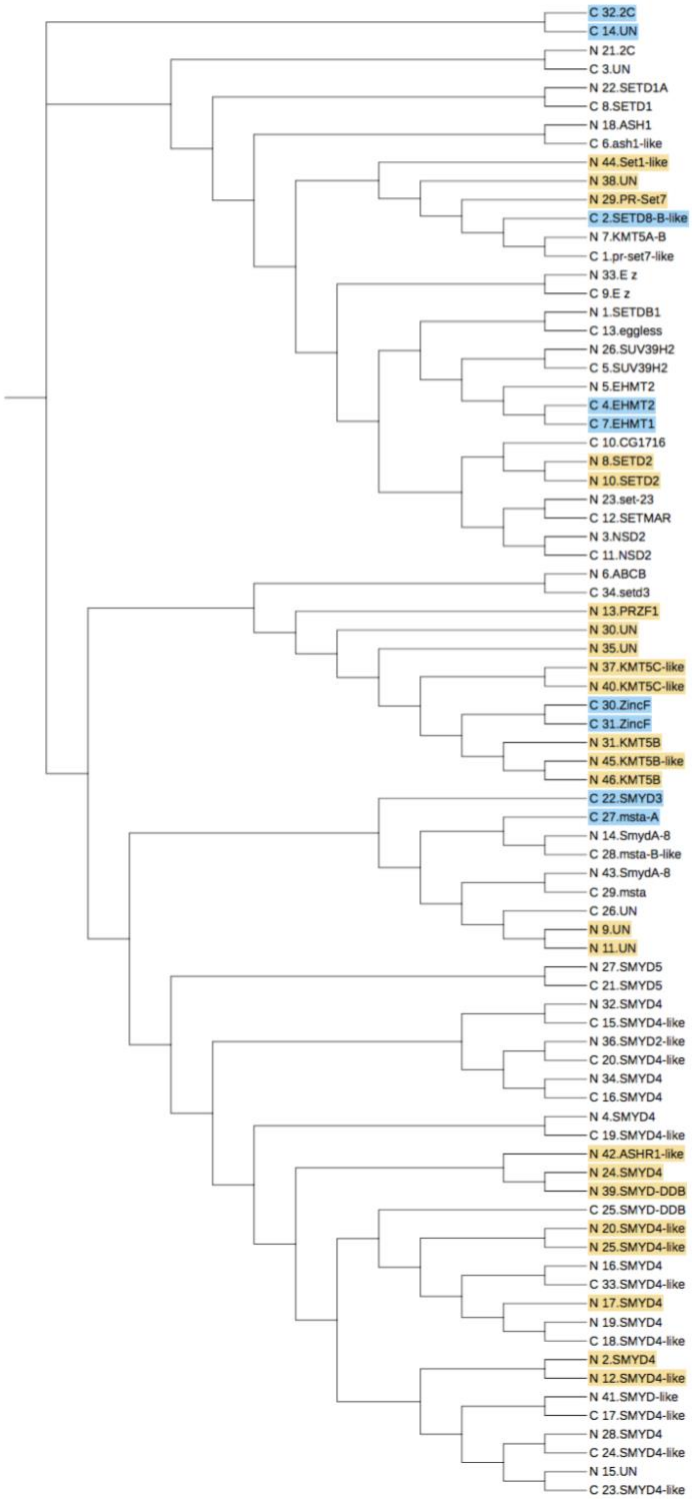

Figure 1. Cladogram of SET genes from *Nasonia vitripennis* and *Ceratosolen solmsi marchali*. Phylogeny was created using the neighbor-joining method without distance corrections. Sequence alignments and tree data were created using Clustal Omega software. Tree data was imported to Interactive Tree of Life (iTOL) to create a cladogram.

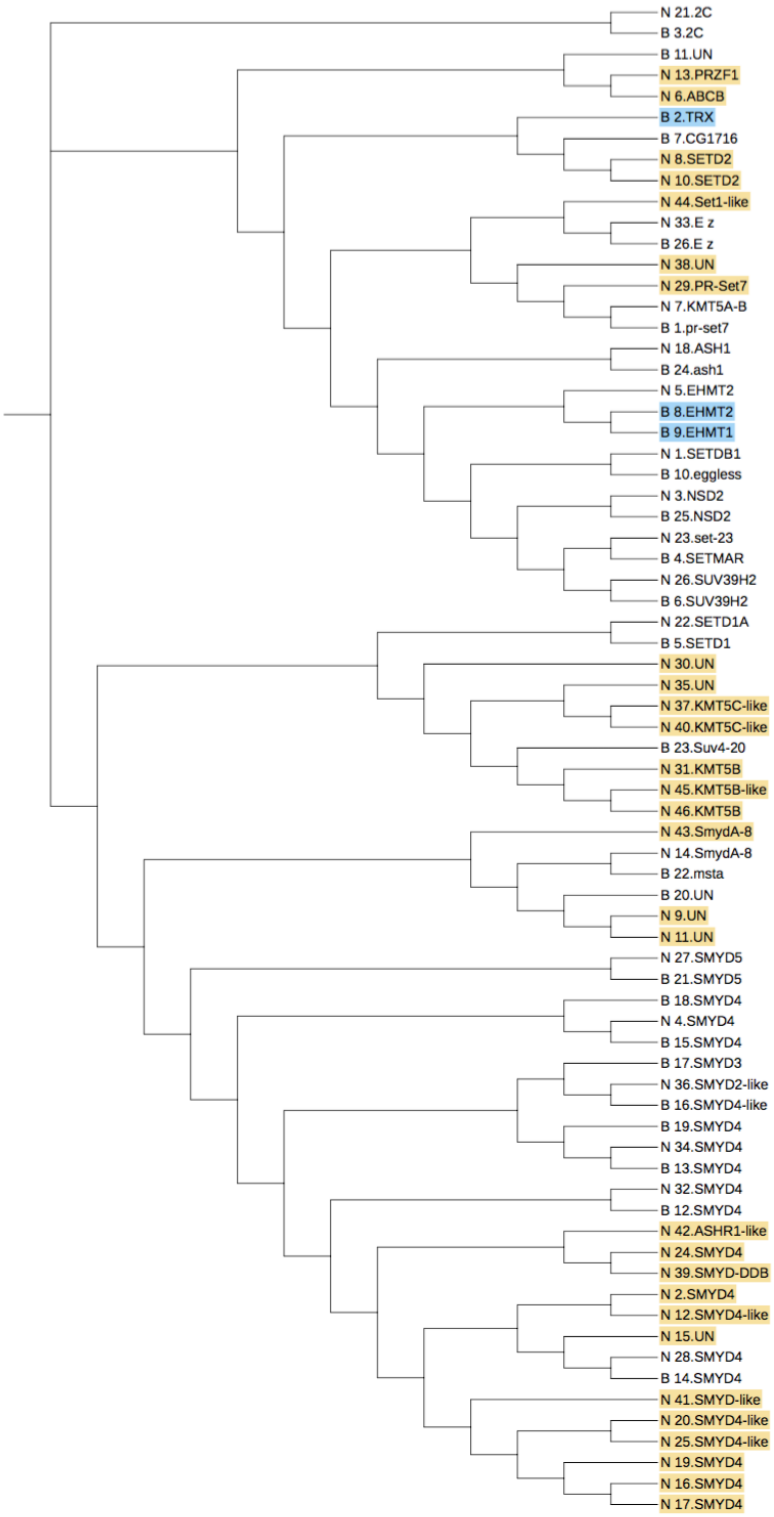

Figure 2. Cladogram of SET genes from *Nasonia vitripennis* and *Bombus terrestris*. Phylogeny was created using the neighbor-joining method without distance corrections. Sequence alignments and tree data were created using

Clustal Omega software. Tree data was imported to Interactive Tree of Life (iTOL) to create a cladogram.

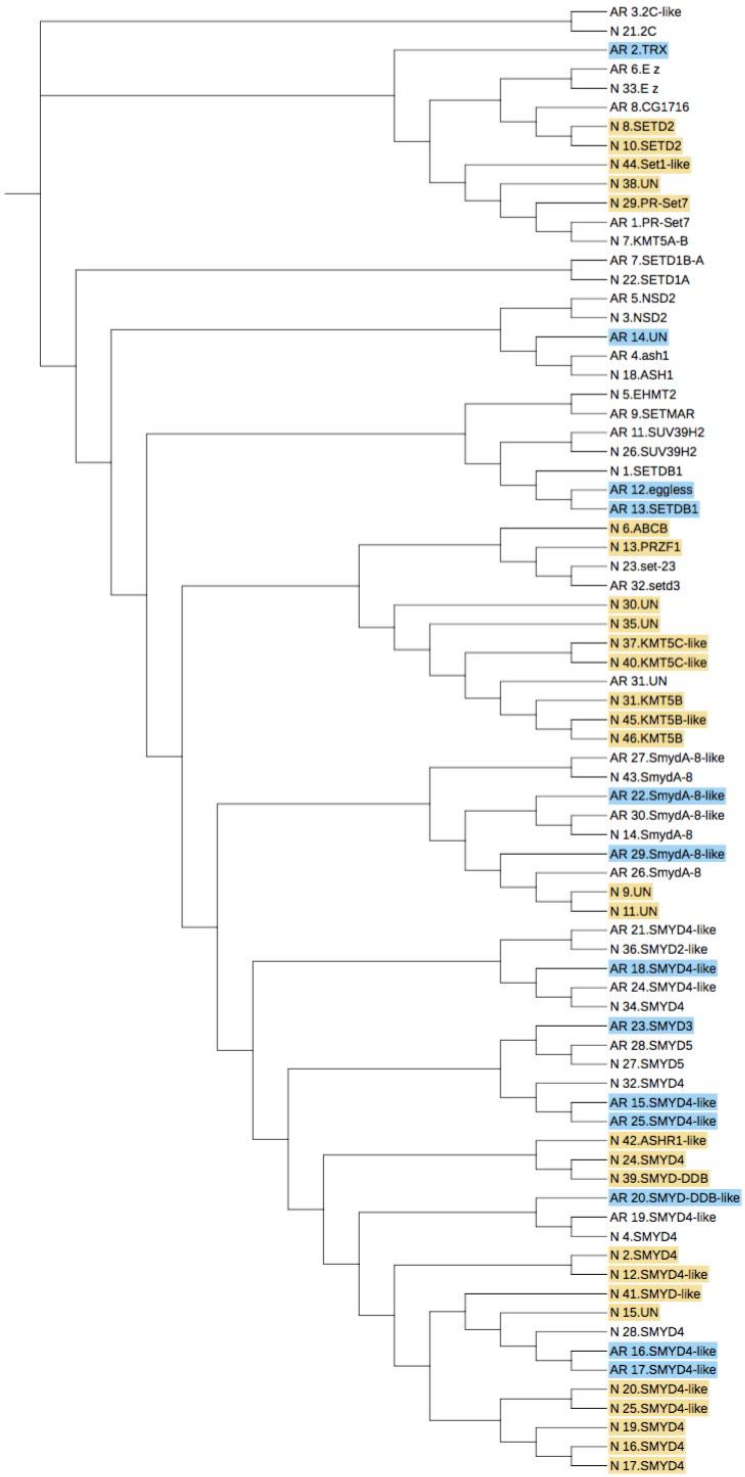

Figure 3. Cladogram of SET genes from *Nasonia vitripennis* and *Athalia Rosae*. Phylogeny was created using the neighbor-joining method without distance corrections. Sequence alignments and tree data were created using Clustal Omega software. Tree data was imported to Interactive Tree of Life (iTOL) to create a cladogram.

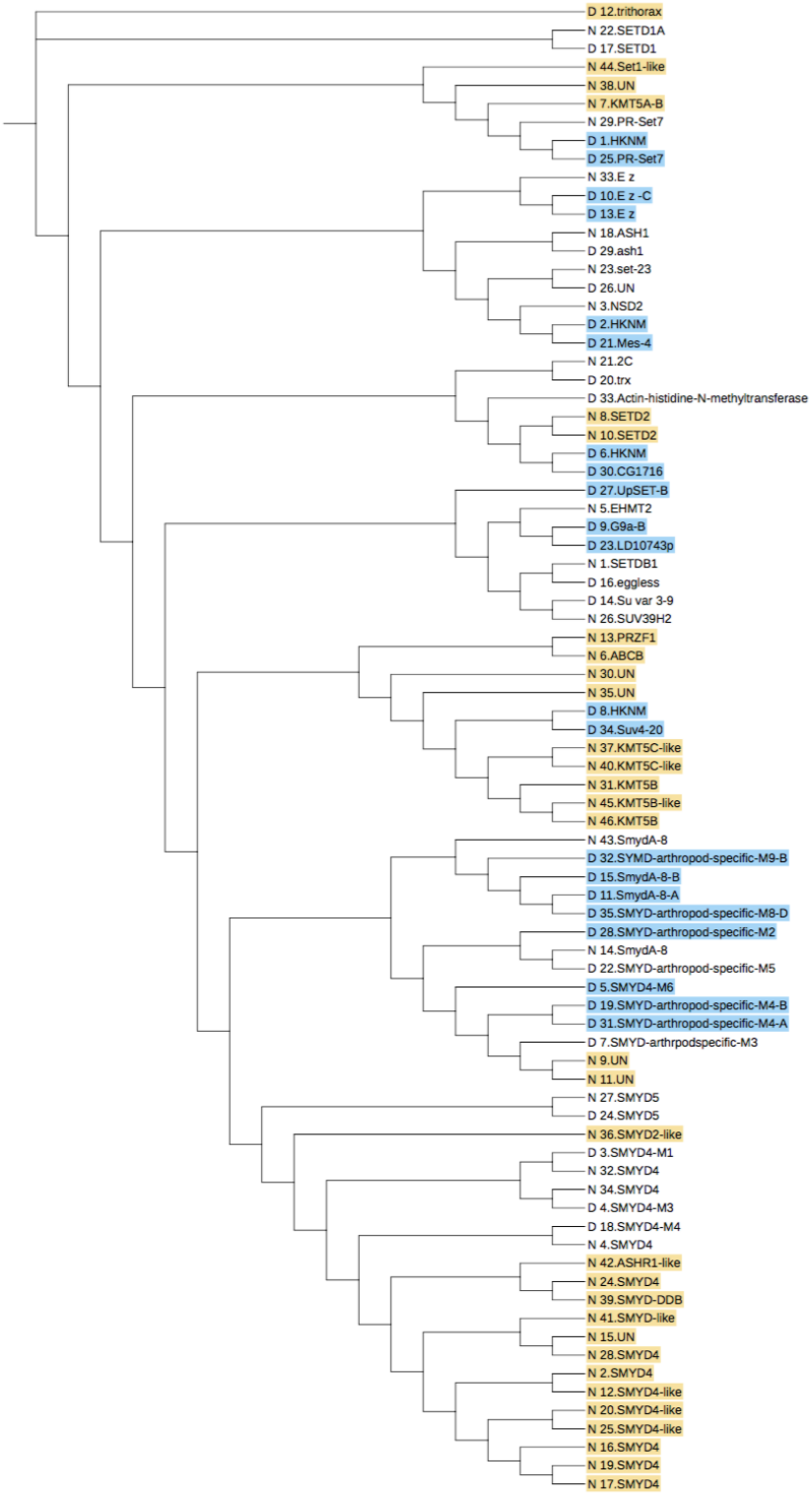

Figure 4. Cladogram of SET genes from *Nasonia vitripennis* and *Drosophila melanogaster*. Phylogeny was created using the neighbor-joining method without distance corrections Sequence alignments and tree data were created using Clustal Omega software. Tree data was imported to Interactive Tree of Life (iTOL) to create a cladogram.

DISCUSSION

In this study, the main goal was to analyze how the SET (Suv(ar)39, Enhancer of Zeste, Trithorax) gene superfamily changes (or doesn't) over evolutionary time. These genes encode histone methyltransferases that are involved in chromatin dynamics and gene expression, such as the protein *egg*, the product of the SET gene *eggless*, which is required for oogenesis in *D. melanogaster* (Brody, 2006) or the protein *lysine-specific methyltransferase 2H*, the product of the SET gene *ASH1* in humans (Inglis & Johnson, 2002) . Specifically, I used the jewel wasp, *Nasonia vitripennis* as a starting point because our lab uses this insect as a model organism for genetic, genomic, and developmental experiments. Additionally, our lab is interested in certain aspects of chromatin dynamics, such as how selfish genetic elements disrupt chromatin processes for their own benefit. I first identified all SET genes within this species. Afterward, I identified SET genes in a number of other select hymenopteran insects – those with published genomes. I compared different factors of genome complexity to the number of SET genes found within each species. The overarching expectation was that a given species with a more complex genome would need a more expansive assembly of chromatin remodelers and other chromatin-associated proteins to "package" the genome.

Although my sample size of species analyzed in this study are relatively small, I found a possible pattern where more species with larger genomes and more genes tended to have more SET genes. Several species in my study had sex chromosomes, and they tended to have more SET genes. Although this pattern was not upheld perfectly, my results imply that this pattern could be real, which makes sense given what we expected in this regard.

Another interesting finding was that the silverfish, *Lepisma saccharina,* had only one SET gene. It is possible that I may not have found all SET genes. However, even if my analyses failed

to find a few additional SET genes that were otherwise present in this genome, the total number of SET genes would be very low. Very little is known about genome organization in this organism. Nevertheless, my result suggests that the genome of a very ancient insect may be quite simple in chromatin organization. This hypothesis can be further tested if it turns out to be possible for silverfish to be cultured in lab, allowing researchers the chance to examine different aspects of chromatin through simple approaches like microscopy. It would also be interesting to look at the genomes of other silverfish species to see if *Lepisma saccharina* is an outlier or it represents a common pattern in this ancient insect group.

Through phylogenetic comparisons of SET genes from *N. vitripennis* to other insects in my study, I found that *N. vitripennis* had strikingly more SET genes than the other insects, even another wasp species. By examining the phylogenies in my study, it became apparent that the large number of SET genes in *N. vitripennis* could be explained by the expansion of two types of SET genes: SMYD (SET and Myeloid-Nervy-DEAF1) and KMT5 (Lysine Methyltransferase 5). I do not think that these gene numbers are losses since they are recurrent when observed through comparative phylogenies of *N. vitripennis* to a number of different insects. Instead, it seems more likely that these expansions occurred at different times since the divergence of the *N. vitripennis* lineage from the other hymenopterans because phylogenetic comparisons highlight extensive duplications and copies of these gene families in *N. vitripennis*, which are not observed to the same degree in compared species.

An important question from this finding is why these two gene families and not others have expanded in *N. vitripennis*. Very few studies have currently been conducted on genes in either family; what has been done comes from studies in human cell lines and in *D. melanogaster*. The SMYD family of genes are thought to function in transcriptional control and in cell proliferation

(Spellmon et al., 2015). What separates SMYD genes from other SET genes is that the catalytic SET region of the gene is interrupted by an MYND region, which is a zinc finger domain mostly made up of cysteine and histidine and found to be involved in DNA binding and protein-protein interactions (Calpena et al., 2015). A total of 8 types of SMYD genes have been discovered to date; these genes are numbered 1 through 8 (Spellmon et al., 2015). The SMYD genes 1-3, have been studied in human cell lines and were found to be expressed in the mesoderm, involved in heart and muscle development, and correlate with cancer cell proliferation (Calpena et al., 2015). Previous research has also shown that SMYD4 and 5 have expanded vastly in the insects compared to 1-3 (Calpena et al., 2015). This pattern is consistent with my findings in this study. SMYD4 genes are distinct from others as they have two tetratricopeptide (TRP) repeats (present on C and N terminus) compared with others which contain one (C terminus) or even zero. In *D. melanogaster*, SMYD4 was found to be expressed in muscles tissues, and SMYD5 is involved in genetic programming of immune response (Calpena et al., 2015).

Additionally, the KMT5 (Lysine Methyltransferase 5) family of genes are known to methylate the lysine 20 residue on histone 4 (H4k20), and in doing so, are involved in a wide range of chromatin related processes including transcriptional activation and repression, DNA damage repair, cell cycle progression and DNA replication. These genes also play a central role in pericentric and telomeric heterochromatin maintenance (Mohan et al., 2012). Trimethylation of H4k20 is a tag for epigenetic transcriptional silencing (Mohan et al., 2012).

While we do not know how many SMYD or KMT5 genes are actually expressed in *N. vitripennis*, it is possible that there is some functional basis to their expansion. SMYD genes were found to play a role in regulation of chromatin that deals with muscle function in human cell lines and in *D. melanogaster* (Calpena et al., 2015). Although it is not known, it is possible

that the link between SMYD genes and muscle function can be explained by a possible need of these histone methyltransferases as regulators of muscle-specific genes. Based on my analyses, the jewel wasp's genome contains more genes than the genomes of the other insects in my study. This factor may necessitate a larger number of SMYD genes, especially if these genes play specific roles in gene regulation. Based on the literature, KMT5 genes most likely play a role in maintaining heterochromatin (Mohan et al., 2012). Given that the jewel wasp's number of chromosomes are small relative to species with which it was compared, it is unusual that it would need more so many more genes that regulate chromatin. However, its relatively high number of genes can explain the extensive regulation of chromatin needed to control gene transcription. It is possible that *N. vitripennis* has a higher content of heterochromatin, or perhaps more specific subtypes of heterochromatin, than the other analyzed insects. Currently such differences have not been explored.

The results from my study have opened up further questions about the assembly of genes needed to package and remodel the genome during development and in the different cell types. Why have SMYD4 genes expanded in insects, and even further in *N. vitripennis*? Could TPR repeats play a role in this expansion and how to they contribute to the function of the protein? Why have KMT5 genes expanded in only the jewel wasp? Fortunately, a growing number of genetic tools, such as systemic RNA interference, exist in N. vitripennis, thus making it possible for further experimentations to address these questions.

ACKNOWLEDGEMENTS

I would like to sincerely thank Professor Patrick Ferree for helping me navigate the many challenges of completing an online senior thesis during the COVID19 pandemic. His support, advice, words of wisdom and encouragement were incredibly valuable to me. I am very lucky to have learned biology from such a brilliant professor.

Additionally, I would like to thank Professor Findley Finseth for her thoughtful feedback and positivity towards my project. She was my very first professor of biology in college, so I would also like to thank her for inspiring me early in my college career to pursue the subject.

Finally, I would like to thank my advisor Professor Jenna Monroy for her consistent support and guidance throughout my undergraduate career.

WORKS CITED

Brody, T. (2006). *Eggless*. The Interactive Fly. https://www.sdbonline.org/sites/fly/genebrief/eggless.htm.

Allshire, R. C., & Madhani, H. D. (2018). Ten principles of heterochromatin formation and function. *Nature Reviews Molecular Cell Biology*, *18*, 229–244. https://doi.org/10.1038/nrm.2017.119

Branstetter, M., Childers, A. K., Cox-Foster, D., Hopper, K. R., Kapheim, K. M., Toth, A. L., & Worley, K. C. (2018). Genomes of the Hymenoptera. *Current Opinion in Insect Science*, *25*, 65–75. https://doi.org/10.1016/j.cois.2017.11.008

Calpena, E., Palau, F., Espinós, C., & Galindo, M. I. (2015). Evolutionary History of the Smyd Gene Family in Metazoans: A Framework to Identify the Orthologs of Human Smyd Genes in Drosophila and Other Animal Species. *Plos One*, *10*(7). https://doi.org/10.1371/journal.pone.0134106

Clough, E., Moon, W., Wang, S., Smith, K., & Hazelrigg, T. (2007). Histone methylation is required for oogenesis in Drosophila. *Development*, *134*(1), 157–165. https://doi.org/10.1242/dev.02698

Couture, J.-F., Dirk, L. M. A., Brunzelle, J. S., Houtz, R. L., & Trievel, R. C. (2008). Structural origins for the product specificity of SET domain protein methyltransferases. *Proceedings of the National Academy of Sciences*, *105*(52), 20659–20664. https://doi.org/10.1073/pnas.0806712105

Dillion, S. C., Zhang, X., Trieval, R. C., & Cheng, X. (20AD). The SET-domain protein superfamily: protein lysine methyltransferases. *Genome Biology*, *6*(227). https://doi.org/https://doi.org/10.1186/gb-2005-6-8-227

Epigenetics: The Science of Change. (2006). *Environmental Health Perspectives*, *114*(3), 160–167. https://doi.org/10.1289/ehp.114-a160

Inglis, D. O., & Johnson, A. D. (2002). Ash1 Protein, an Asymmetrically Localized Transcriptional Regulator, Controls Filamentous Growth and Virulence of Candida albicans. *Molecular and Cellular Biology*, *22*(24), 8669–8680. https://doi.org/10.1128/mcb.22.24.8669-8680.2002

Lee, J. Y., & Orr-Weaver, T. L. (2001). Chromatin. *Encyclopedia of Genetics*, 340–343. https://doi.org/10.1006/rwgn.2001.0199

Liu, Q., & Wang, M.-wei. (2016). Histone lysine methyltransferases as anti-cancer targets for drug discovery. *Acta Pharmacologica Sinica*, *37*(10), 1273–1280. https://doi.org/https://doi.org/10.1038/aps.2016.64

Mank, J. E., Hosken, D. J., & Wedell, N. (2014). Conflict on the Sex Chromosomes: Cause, Effect, and Complexity. *Cold Spring Harbor Perspectives in Biology*, *6*(12). https://doi.org/10.1101/cshperspect.a017715

Mariño-Ramírez, L., Kann, M. G., Shoemaker, B. A., & Landsman, D. (2005). Histone structure and nucleosome stability. *Expert Review of Proteomics*, *2*(5), 719–729. https://doi.org/10.1586/14789450.2.5.719

Mohan, M., Herz, H.-M., & Shilatifard, A. (2012). SnapShot: Histone Lysine Methylase Complexes. *Cell, 149*, 498. https://doi.org/10.1016/j.cell.2012.03.025

Peters, R. S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., … Niehuis, O. (2017). Evolutionary History of the Hymenoptera. *Current Biology*, *27*(7), 1013–1018. https://doi.org/10.1016/j.cub.2017.01.027

Qian, C., & Zhou, M.-M. (2006). SET domain protein lysine methyltransferases: Structure, specificity and catalysis. *Cellular and Molecular Life Sciences*, *63*(23), 2755–2763. https://doi.org/10.1007/s00018-006-6274-5

Spellmon, N., Holcomb, J., Trescott, L., Sirinupong, N., & Yang, Z. (2015). Structure and Function of SET and MYND Domain-Containing Proteins. *International Journal of Molecular Sciences*, *16*(1), 1406–1428. https://doi.org/10.3390/ijms16011406

Sze, C. C., Cao, K., Collings, C. K., Marshall, S. A., Rendleman, E. J., Ozark, P. A., … Shilatifard, A. (2017). Histone H3K4 methylation-dependent and -independent functions of Set1A/COMPASS in embryonic stem cell self-renewal and differentiation. *Genes & Development*, *31*(17), 1732–1737. https://doi.org/10.1101/gad.303768.117

Trievel, R., Beach, B., Dirk, L., Houtz, R., & Hurley, J. (2002). Structure and Catalytic Mechanism of a SET Domain Protein Methyltransferase. *Cell*, *111*(1), 91–103. https://doi.org/10.2210/pdb1mlv/pdb

Weirich, S., Kudithipudi, S., Kycia, I., & Jeltsch, A. (2015). Somatic cancer mutations in the MLL3-SET domain alter the catalytic properties of the enzyme. *Clinical Epigenetics*, *7*(1). https://doi.org/10.1186/s13148-015-0075-3

Yeates, T. O. (2002). Structures of SET Domain Proteins. *Cell*, *111*(1), 5–7. https://doi.org/10.1016/s0092-8674(02)01010-3

Zhang, X., & Bruice, T. C. (2008). Enzymatic mechanism and product specificity of SET-domain protein lysine methyltransferases. *Proceedings of the National Academy of Sciences*, *105*(15), 5728–5732. https://doi.org/10.1073/pnas.0801788105