Scalable and reproducible genome analysis in the age of next-generation genome sequencing

by

Daniel Scott Standage

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Volker Brendel, Co-major Professor
Amy Toth, Co-major Professor
Karin Dorman
Xiaoqui Huang
Jonathan Wendel

Iowa State University
Ames, Iowa
2016

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

To go here.

CHAPTER 1. INTRODUCTION

Overview

In the 2000s, the advent of new nucleotide sequencing strategies based on ion semiconductors (Ion Torrent), pyrosequencing (454), and sequencing-by-synthesis (Illumina) provided new tools for studying genomes of both model and non-model organisms at unprecedented scale, resolution, and cost effectiveness. These technologies continue to evolve, the latest innovations involving single-molecule long read sequencing (Pacific Biosciences SMRT and Oxford Nanopore). By the 2010s, these so-called *next-generation sequencing (NGS)* technologies had made genome sequencing accessible to essentially any scientist with even a modest research budget. This "democratization of sequencing" has precipitated a tremendous increase in the number of published genome projects and draft genome sequences (http://www.ncbi.nlm.nih.gov/genome/browse/), as well as genome-scale data sets profiling gene expression, chromatin accessibility, transcription initiation, and a multitude of other genomic characteristics.

During this same time frame, however, the democratization of genome analysis has occurred to a much lesser extent. The availability of high-quality model reference genomes has changed very little. And as newly acquired data continues to flood into public databases, many scientists struggle to effectively manage the data and critically evaluate downstream research products. The proliferation of new algorithms and software tools for analyzing NGS data is a mixed blessing for scientists who now have both the flexibility and the burden of selecting suitable tool(s) for a particular analysis.

The complexity and difficulty of genome assembly [1, 2, 3, 4], annotation [5, 6, 7], and analysis [8, 9, 10] has been reported in various recent studies and community projects. The consistent, resounding message from this growing body of work is that genomics data quality varies considerably across data sets (and even within a single data set), and that the performance of state-of-the-art algorithms is difficult to predict on new data. Without a well-funded and well-staffed research consortium to manage the painstaking work of gap-filling each new genome assembly and carefully curating its contents, the new reality is that most reference genome assemblies will remain fragmented and unfinished, and that provisionally annotated genome features will fall on a wide spectrum of reliability. Doing principled and reproducible science in this setting requires disciplined quality control and data evaluation. The focus of this dissertation has been the development of a framework—and associated software tools—to enable robust annotation and analysis of NGS-based genomics data, as motivated by research problems I encountered in genomics research projects.

Polistes dominula genome project

My first encounter with many of these issues came from my genomics studies of the paper wasp *Polistes dominula*. This wasp is an important model system for studying the evolution of social behavior, as it exhibits an intermediate level of social complexity, with no morphological differentiation between castes and frequent competition for reproductive opportunities within colonies [11]. The molecular mechanisms underlying caste differentiation across all social insects are still poorly understood, although various mechanisms have been proposed to play primary roles, such as differential gene expression [12, 13, 14, 15, 16, 17, 18], differential splicing [19], DNA methylation patterns [19, 20], and recently emerged species- or clade-specific genes [21, 22]. These hypotheses have been tested in bees and ants, but prior to this study no genome resources were

available for any species from the major vespid wasp lineage. This project was funded to sequence the genome, transcriptome, and methylome of the wasp, to facilitate investigation of the molecular basis of caste differentiation in *Polistes*, and to provide an additional important data source for comparative analysis of all social insects.

Our initial work was driven largely by questions of genome composition. How large is the *Polistes* genome? What is its nucleotide composition? How many genes does it encode? Are the small handful of well-known "social behavior" genes present in the genome as expected? In short, is there anything that immediately distinguishes the paper wasp genome from genomes of related species?

The next set of questions were driven by a comparative genomics perspective. How does the size and composition of the *Polistes* genome compare to the bees and the ants? What proportion of annotated gene models are well-conserved within the Hymenoptera? Can conserved single-copy orthologs provide any insight into the unresolved evolutionary lineage of the bees, ants, and wasps?

At the same time, we investigated questions of functional genomics and epigenomics. How many *Polistes* genes show differential expression between the queen and worker castes? What is the extent of alternative splicing in *Polistes*, and do any alternative splicing events exhibit caste-related bias? What is the extent of DNA methylation in *Polistes*, and does the genome encode a full complement of methylation-related genes?

This last question led us to one of the highlight discoveries of the study: that *Polistes* lacks a critical DNA methyltransferase (*Dnmt3*) and has essentially no DNA methylation genome-wide. But more in line with the focus of this dissertation, the *Polistes dominula* genome project exposed me to the challenges of creating genomic data resources *de novo* for a non-model research system, and the corresponding challenges of data quality assessment and management. These challenges motivated the development of methods and tools to facilitate comparison, evaluation, and analysis of genome annotations, with additional applications to studying genome organization.

ParsEval: comparison of distinct annotation sources

Automated genome annotation typically relies on integrating tools for *ab initio* gene prediction, transcript and protein spliced alignment, and evaluation of support for gene structural components. Each class of tools comes with a variety of parameter settings, and it can be difficult to predict in advance the influence these parameters will have on the final annotation product. When annotating a non-model genome *de novo*, it is often necessary to refine parameter selection on a small subset of the available data before proceeding to annotate the entire genome. Subsequent re-annotation is often necessary when additional data (such as ESTs or RNA-Seq reads) become available, or when improved gene prediction methods are published. And in some cases, the scientists annotating a particular genome have a vested interest in improving the accuracy of the annotation software itself. In each of these scenarios, a primary objective is to identify similarities and differences between annotations derived from different parameter settings or alternative workflows, to facilitate evaluation of the annotations.

Development of the CpGAT annotation pipeline within our research group provided the initial motivation for ParsEval. Evaluating CpGAT's performance in comparison to a gold-standard annotation was impractical manually, and existing software tools [23, 24] lacked important features such as locus-scale resolution. I created the ParsEval program to address the need for genome-scale evaluations with locus-scale detail. Later during the preliminary stages of the *Polistes dominula* genome project, ParsEval proved a valuable data assessment tool as we tuned our genome annotation workflow.

The ParsEval paper introduces a precise operational definition for a *gene locus* based on the locations of annotated gene models. The objective was to define a parsing of the genome into distinct units that are complete and can be analyzed independently. Generalizing this concept and applying it as an organizational principle for genome analyses has been a major focus of this dissertation.

iLoci: an organizational framework

The ParsEval tool proved valuable throughout the *Polistes dominula* genome project, but we subsequently encountered issues that required additional attention and development. Although our *Polistes* assembly and annotation compared favorably to other published hymenopteran genomes, our data was subject to the same quality inconsistencies that characterize any NGS-based genome project. Answering questions related to gene expression and genome composition therefore required careful consideration of, for example, how precisely to handle overlapping gene models and how to distinguish differences rooted in biology from technical artifacts. It was in addressing these issues that we extended the *gene locus* definition introduced by ParsEval and developed the *interval locus (iLocus)* as a more generalized organizational framework for genome analyses.

iLoci define an unambiguous parsing of an annotated genome sequence into distinct regions, each encapsulating the genomic context of a gene or intergenic space. This parsing provides a complete and granular decomposition of the genome. In parallel with the *P. dominula* genome project in which iLoci played a prominent role, we investigated general applications of iLoci. We discuss their utility for describing the contents of a genome and for applying quality control when calculating diagnostic characteristics of a genome. We investigate the utility of iLoci for characterizing genome organization, within a single genome and between multiple genomes in a clade of species. Finally, we report the stability of iLoci across distinct assembly and annotation versions, highlighting their utility as reproducible units of analysis.

GeneAnnoLogy: annotation version control

Continual refinement of our genome assembly and annotation hindered progress on our *P. dominula* research at times. This issue is certainly not unique to our work. For example, the honeybee *Apis mellifera* (the model social insect) had three official

annotation versions in concurrent use during our work on the *Polistes* genome project, with some studies even making their own unpublished refinements to an annotation [19, 25]. What seemed to be lacking was a precise way to refer to a particular gene or genomic region, as annotated at a particular time, and to make statements about its expression, or conservation, or methylation status, or any number of additional characteristics.

The idea of tracking annotations over time is not new [6], and some well-supported communities provide tools for mapping annotations from an older assembly version to updated assembly [26, 27]. However, iLoci provide an alternative solution to these issues and, along with the additional benefits previously described, furnish a complete framework for organization, quality control, and reproducibility for provisional genome projects.

At an early stage in my training, I prototyped a tool called GeneAnnoLogy for maintaining an annotation version history, leveraging existing version control tools to track changes to individual iLoci over time. I later built in features that facilitate filtering iLoci based on their length, nucleotide composition, annotation quality, or any number of additional characteristics that can be computed from or attached to the annotation. The GeneAnnoLogy tool encapsulates the culmination of concepts and principles investigated in this dissertation.

Dissertation Organization

This dissertation is organized into six chapters. Chapter 1 provides an overview of the dissertation, a motivation for the work, and a brief discussion of relevant literature. Chapters 2 through 5 are presented as complete manuscripts: chapter 2 is a research paper published in *Molecular Ecology* describing the genome, transcriptome, and methylome of the paper wasp *Polistes dominula*, highlighting its reduced DNA methylation system, several hundred loci with caste-related differential expression, and the lack of

any detectable caste-related differential splicing in the adult organism; chapter 3 is a paper published in *BMC Bioinformatics* describing *ParsEval*, a tool for comparing two alternate sources of annotation for a genome sequence; chapter 4 is a methodology paper to be submitted to *BMC Bioinformatics/Genomics*[1], describing the use of *interval loci (iLoci)* as an organizational framework for reproducible genome analysis; chapter 5 is a manuscript slated for submission to *A Journal* describing *GeneAnnoLogy*, an *iLocus*-based tool for quality control and version control of genome annotations. Chapter 6 provides brief concluding remarks and suggestions for further research.

[1] Will be submitted within a couple of weeks.

CHAPTER 2. GENOME, TRANSCRIPTOME, AND METHYLOME SEQUENCING OF A PRIMITIVELY EUSOCIAL WASP REVEAL A GREATLY REDUCED DNA METHYLATION SYSTEM IN A SOCIAL INSECT

A manuscript published in *Molecular Ecology*. **Supporting Information (SI)** available online: doi:10.1111/mec.13578.

Standage DS, Berens AJ, Glastad KM, Severin AJ, Brendel VP, Toth AL

Abstract

Comparative genomics of social insects has been intensely pursued in recent years with the goal of providing insights into the evolution of social behavior and its underlying genomic and epigenomic basis. However, the comparative approach has been hampered by a paucity of data on some of the most informative social forms (e.g. incipiently and primitively social) and taxa (especially members of the paper wasp family Vespidae) for studying social evolution. Here we provide a draft genome of the primitively eusocial model insect *Polistes dominula*, accompanied by analysis of caste-related transcriptome and methylome sequence data for adult queens and workers. *P. dominula* possesses a fairly typical hymenopteran genome, but shows very low genome-wide GC content and some evidence of reduced genome size. We found numerous caste-related differences in gene expression, with evidence that both conserved and novel genes are related to caste

differences. Most strikingly, these –omics data reveal a major reduction in one of the major epigenetic mechanisms that has been previously suggested to be important for caste differences in social insects: DNA methylation. Along with a conspicuous loss of a key gene associated with environmentally responsive DNA methylation (the de novo DNA methyltransferase *Dnmt3*), these wasps have greatly reduced genome-wide methylation to almost zero. In addition to providing a valuable resource for comparative analysis of social insect evolution, our integrative –omics data for this important behavioral and evolutionary model system call into question the general importance of DNA methylation in caste differences and evolution in social insects.

Introduction

The rapidly increasing availability of genomic resources for non-traditional model organisms with well-developed social behavior has incited great interest in the genomic basis of complex social life, or sociogenomics (Robinson *et al.* 2005). Sociogenomic studies on a variety of species, from rodents to fish to insects, have provided a wealth of information about the transcriptomic and genomic characters associated with different forms of derived social behavior, from affiliative behavior, to aggression, to division of labor within animal societies (Robinson *et al.* 2008). To date, however, these studies have focused on relatively few species separately, and current data have not allowed a comprehensive comparative and phylogenetic approach to understanding the genomic changes that accompany the evolution of sociality. As advances in sequencing technology have greatly facilitated the generation of genome-scale data for emerging model species (Rokas & Abbot 2009), it is an exciting time to seek integration of genomic, transcriptomic, and epigenomic data from species at key transitional points in the evolution of sociality from solitary behavior.

The eusocial insects are one of the most important sociogenomic model groups, com-

prising a diverse and ecologically successful group of animals with a highly derived form of social behavior characterized by the presence of reproductive and non-reproductive castes (Hlldobler & Wilson 2009). Eusocial insects are excellent models for understanding the evolution of complexity as the switch from solitary to eusocial life marks one of the major transitions in evolution due to the shift from individual selection to colony level selection (Maynard Smith & Szathmry 1997). Currently, published genome sequences are available for three parasitic *Nasonia* wasps (Werren *et al.* 2010), which provide a solitary outgroup for all of the social Hymenoptera species, and 20 eusocial insect genomes—ten bees (Kapheim *et al.* 2015; Kocher *et al.* 2013; Sadd *et al.* 2015; Weinstock *et al.* 2006) and nine ants (Bonasio *et al.* 2010; Nygaard *et al.* 2011; Oxley *et al.* 2014; Smith *et al.* 2011a; Smith *et al.* 2011b; Suen *et al.* 2011; Wurm *et al.* 2011), and very recently, one paper wasp (Patalano *et al.* 2015). These studies provide important baseline data on genomic characters associated with eusociality.

Importantly however, prior comparative sociogenomic analyses within the social insects have suffered from two major deficits. First, there have been relatively scant genomic resources available for one of the three major hymenopteran eusocial lineages, the paper wasp family Vespidae (**Figure 1A**). Although social wasps, bees, and ants evolved from a common ancestor over 100 million years ago, these societies have independently evolved many convergent features, including the presence of female castes in the form of queens and workers. Second, there have been relatively few genome sequences available for species in key transitional stages between solitary and eusocial forms (notably, there has been a large recent advance in this area (Kapheim *et al.* 2015; Kocher *et al.* 2013; Patalano *et al.* 2015; Sadd *et al.* 2015)). Here, we expand the potential for comparative genomics of eusocial Hymenoptera by describing the first complete genome sequence of *Polistes dominula*, a behavioral model species within the family Vespidae that exhibits an intermediate form of social behavior, making it highly informative for studying the evolution of sociality (Jandt & Toth 2015).

Polistes wasps form small "primitively eusocial" societies containing queens and altruistic workers, but unlike honey bees, their colonies are characterized by prominent conflict over reproduction (Pardi 1948). Queens and workers engage in dominance interactions and there is constant competition between females for reproductive opportunities. In addition, *Polistes* have small colonies with a relatively small number of individuals, and colonies are started anew annually by founding queens (Reeve 1991). These characteristics have made *Polistes* one of the main systems for testing hypotheses about the evolution of altruistic behavior (West-Eberhard 1996). For example, studies of cooperation and conflict in small groups of *Polistes* wasps have provided some key tests of how genetic relatedness can facilitate cooperation (Hamilton 1964). In addition, observations of *Polistes* behavior led to new hypotheses about the evolution of altruistic behavior from maternal behavior (West-Eberhard 1996), which have derived some support from transcriptomic studies (Toth *et al.* 2010; Toth *et al.* 2007). A genome sequence for *Polistes dominula*, the best-studied member of the model genus *Polistes,* greatly enhances our power to study the genetics of social behavior *via* comparative genomic and transcriptomic analyses, allowing for the identification of protein coding changes, regulatory regions, and epigenetic modifications associated with sociality. To facilitate future comparative analyses, we provide a high quality draft *P. dominula* genome and describe informative features of this genome in reference to other previously published bee, ant, and *Nasonia* wasp genomes. Our genome sequence was derived from an invasive population (from Pennsylvania, USA) of *Polistes dominula*, a temperate species that is native to Europe. Our genome represents the second published paper wasp genome, the first being the very recently published genome of the Neotropical paper wasp *Polistes canadensis* (Patalano *et al.* 2015). Although both species are primitively eusocial, they are not closely related (split between Old and New World *Polistes* at 10-80 million years (Ezenwa *et al.* 1998)) and have several differences in their ecology and social biology. Thus, we provide comparisons between these congeners, confirming many con-

served *Polistes* genome characteristics but also highlighting some conspicuous differences between the two paper wasp genomes.

One of the most active recent areas of research in insect sociogenomics centers on the role of epigenetics in the regulation and evolution of eusociality. Recent studies suggest epigenetic modifications to DNA are ubiquitous within the social Hymenoptera (Kronforst *et al.* 2008; Weiner *et al.* 2013), and furthermore, various authors have suggested that differential DNA methylation during larval development contributes to caste differential gene expression and alternative splicing (Li-Byarlay *et al.* 2013; Lyko *et al.* 2010), and differences in caste-related phenotypes (Kucharski *et al.* 2008) in both honey bees and ants. We previously hypothesized that DNA methylation might also be important for caste differential expression and behavioral and physiological caste differences in primitively eusocial species such as *Polistes dominula* (Weiner *et al.* 2013; Weiner & Toth 2012). However, recent studies suggest DNA methylation may be less important for primitively eusocial species, including *Polistes canadensis* (Kapheim *et al.* 2015; Patalano *et al.* 2015). Therefore, we looked for evidence of a functional DNA methylation system by investigating the presence of a full complement of DNA methylation enzymes in the *Polistes dominula* genome. In addition, we performed RNA-sequencing and whole genome bisulfite sequencing (methylome sequencing) on a set of adult queen and worker samples in order to examine caste-associated differential expression and DNA methylation in the *Polistes dominula* genome. These experiments provide a valuable point of comparison to other social insects for assessing whether DNA methylation is a shared, general mechanism related to sociality in insects.

Materials and Methods

Detailed protocols are provided in **SI Materials and Methods**.

Sample collection and sequencing.

Five Illumina paired-end whole genome shotgun libraries were prepared from a single pupal male collected from an invasive population in State College, Pennsylvania. The libraries, ranging in insert size from <200bp to 8Kbp (see **Table S1**), were sequenced on the Illumina HiSeq 2000 platform.

Same-nest pairs of six adult workers and six egg-laying adult queens and from six different colonies were collected for transcriptome and methylome sequencing; these twelve individuals were from the same State College, Pennsylvania population as the pupal male used for genome sequencing. The head of each adult was cut in half, with RNA extracted from one half for transcriptome sequencing and DNA extracted from the other half for methylome sequencing.

Genome assembly and annotation

. Five whole genome shotgun libraries were assembled using the AllPaths-LG genome assembler (Gnerre *et al.* 2011), with the smallest library designated as the *fragment library* and the other four libraries designated as *jumping libraries*. The assembled scaffolds were then screened for repetitive DNA, masked, and annotated by the MAKER pipeline (Cantarel *et al.* 2008). The annotation workflow incorporated evidence from spliced alignments of transcripts from three *Polistes* species, spliced alignments of reference proteins from *Apis mellifera* and *Drosophila melanogaster*, gene models produced by three *ab initio* gene predictors, and manual gene annotations contributed *via* the PdomGDB genome browser's community annotation portal.

Transcriptome assembly and annotation

. Twelve Illumina paired-end RNA-Seq libraries were assembled using the Trinity assembler (Grabherr *et al.* 2011) with the --CuffFly algorithm and the --jaccard_clip setting enabled. Assembled transcripts were then post-processed to discard contaminants, split

transcript chimeras, and annotate transcript functions by similarity to known proteins and miRNAs. Two previously published *Polistes* transcriptomes (Berens *et al.* 2015b; Ferreira *et al.* 2013) were processed with the same procedure.

Differential expression analysis

. Twelve Illumina paired-end RNA-Seq libraries, six from queens and six from workers, were sequenced and reads were mapped individually to the genome using Bowtie (Langmead *et al.* 2009). Preliminary examination of the alignments revealed an extremely wide dynamic range of expression values (<10 reads to millions of reads mapped per replicate) and in some cases considerable variation between replicates. To account for these observations we discarded loci with too many or too few reads mapped (normalized by sequence length) or with a high coefficient of variation.

Expression was quantified using RSEM (Li & Dewey 2011), and the EBSeq package (Leng *et al.* 2013) was used for identifying loci with caste differential expression at a false discovery rate of < 0.05. A complete description of the number of raw reads, data filtering procedure, and software parameters used is available in **SI Methods and Results**.

Methylome analysis

. Two DNA samples (one pooled sample from workers and one pooled sample from queens, derived from the same six individuals used for transcriptome sequencing) were subjected to bisulfite treatment, and each sample was used to generate separate Illumina libraries for sequencing. The Bismark software (Krueger & Andrews 2011) was used for read mapping and methylation calls. Highly supported methylation sites were determined as sites with significant number of methylation calls based on a binomial probability model with Bonferroni correction at the 1% significance level (assuming a 99.5% conversion rate in the treatment). We also reanalyzed existing datasets from from *Polistes*

canadensis (Patalano *et al.* 2015), honey bees (Lyko *et al.* 2010), the ants *Harpegnathos saltator* and *Camponotus floridanus* (Bonasio *et al.* 2012), and the parasitoid wasp *Nasonia vitripennis* (Wang *et al.* 2013) in order to compare our *P. dominula* results to previously analyzed Hymenoptera (see **SI Methods**). Numbers of methylation sites were determined with the BWASP workflow (`http://brendelgroup.github.io/BWASP/`) using pooled reads from published data sets for each species and caste. In-depth descriptions of the comparative data, analysis pipelines, and comparative results are provided in a companion paper (Toth AL, Sankaranarayanan S, and Brendel VP, *in preparation*).

Examination of interfamilial relationships *via* gene tree analysis

. Genes with conserved single-copy orthologs in *Apis mellifera*, *Harpegnathos saltator*, *Polistes dominula*, and *Nasonia vitripennis* were identified with protein clustering (see **SI Methods** for clustering criteria). For each gene, a multiple sequence alignment of the four corresponding protein sequences was computed, and from that alignment a phylogenetic tree was inferred *via* maximum likelihood. After all gene trees had been constructed, each tree was analyzed to note its topology and collect a tally of the three possible topologies observed.

Results and Discussion

Genome description and assessment of *de novo* genome quality.

We used Illumina technology to sequence genomic DNA from a single, haploid pupal male *P. dominula* from an invasive population in State College, Pennsylvania, USA. The DNA was used to generate five Illumina genomic DNA libraries of varying insert size (**Table S1**), each sequenced on a single channel on an Illumina HiSeq instrument. This generated a total of 78.6 Gb of raw sequence, which was filtered using Trimmomatic (Bailey *et al.* 2009) to remove sequencing adapters and low quality base calls. The groomed

data were assembled using AllPaths-LG (Gnerre *et al.* 2011), producing 1,483 scaffolds with an N50 of 1.63 Mb and a combined length of approximately 208 Mb. The genomic reads provide approximately 319x coverage of the genome. This genome assembly compares favorably to Illumina- and 454-based assemblies of other social Hymenoptera, in particular to that of *Polistes canadensis*, which despite a higher level of fragmentation provides a consistent estimate of the *Polistes* genome size. **Table 1** shows *P. dominula* in comparison to a few representative previously published paper wasp, bee, ant, and non-social wasp genomes (taxon selection described in **SI Methods**).

The genome assembly appears to be quite complete, on par with other Illumina-based draft insect genomes (Bonasio *et al.* 2010; Nygaard *et al.* 2011). CEGMA analysis (Parra *et al.* 2007) showed 246 (99.2%) of 248 ultra-conserved core eukaryote genes to be present in the genome assembly. The total assembled length of the genome approaches both *in silico* estimates of the *P. dominula* genome size based on *k*-mer distributions in the sequence data (246 Mb, see **SI Results**) and earlier estimates based on flow cytometry (300 Mb, (Johnston *et al.* 2004)). The gene space of the *P. dominula* genome therefore appears to be almost completely represented in the assembly, suggesting that the unrepresented portions of the genome are likely highly repetitive regions that are difficult to assemble with Illumina technology.

Automated annotation of *P. dominula* genes was based on a specifically trained MAKER workflow (Campbell *et al.* 2014) and incorporated protein evidence from *Apis mellifera* (NCBI release 102 and OGS 3.2) and *Drosophila melanogaster* (FlyBase r5.55) and transcript data from *P. dominula* (described below), *P. metricus* (Berens *et al.* 2015a), and *P. canadensis* (Ferreira *et al.* 2013). Also integrated into the annotation were 180 gene models that, during preliminary stages of annotation, were manually curated and refined using the yrGATE portal (Wilkerson *et al.* 2006). This resulted in 11,819 predicted gene models, designated as release 1.2 (see **DATA ACCESIBILITY**). Similarity searches using BLAST revealed most of these predicted genes—10,755 out of

11,819 (91%)—have hits to the NCBI non-redundant database, whereas 1,064 show no significant similarity to known proteins. Of the genes with predicted homologs, most (10,504, or 98%) have best hits to other Hymenoptera annotated proteins (**Figure S3**). These gene models represent the first whole-genome annotation of a vespid wasp, and include thousands of high-quality conserved genes enabling more detailed comparative analysis of hymenopteran genomes, as well as many species-specific gene models with which to investigate for evidence of novel clade-specific genes.

Composition of *P. dominula* genome shows a combination of typical hymenopteran as well as unique features.

Comparisons of the *P. dominula* genome assembly to those of other Hymenoptera revealed the assembled genome size and proportion of the genome occupied by transposable elements to be within the range of the other species. Published hymenopteran genomes show variety in the types and amounts of transposable element (TE) and other repetitive content, with *Apis mellifera* harboring almost exclusively a small number of *mariner* class transposons, and *Nasonia vitripennis* on the other hand harboring diverse repetitive elements constituting approximately a quarter of its genome (Honeybee Genome Sequencing Consortium 2006; Werren *et al.* 2010). The *Polistes dominula* and *P. canadensis* genomes contain a fairly low level of repetitive DNA; 11-14% of the genome assemblies (24-30 Mb) are estimated to be repetitive, the majority of which represents simple repeats and low complexity sequence. The two *Polistes* genomes harbor a very similar cohort of TEs, dominated in both genomes primarily by L2/CR1/Rex and R1/LOA/Jockey LINEs, Gypsy/DIRS1 LTRs, and Tc1-IS630-Pogo DNA elements (**Table S4**).

Characteristics of genome structure were further investigated by parsing the *P. dominula* genome into 17,888 *interval loci* (iLoci), each iLocus capturing the local genomic context of a single gene (11,713 iLoci), a cluster of overlapping genes (205 iLoci), or an

intergenic region (5,970 iLoci; see **SI Methods**). In order to compare a set of comparable genes across species and rule out differences due to annotation artifacts, homologous iLoci were determined by computing iLoci for several additional insect species and clustering their protein products. A comprehensive comparison included 17 hymenopteran species in total, but for illustrative purposes, seven representative species (*P. dominula*, *P. canadensis*, two bees, two ants, and a non-aculelate, non-social hymenopteran outgroup to the social insects, *Nasonia vitripennis*) are shown for comparison in this and subsequent analyses; see **SI Methods**.

At 11,918, the number of protein-coding iLoci (genes or gene clusters) in the *P. dominula* genome is well within the range observed in other Hymenoptera (see **Figure S5**). However, gene iLoci occupy only 73.0 Mb (35.1%) of the *P. dominula* genome (and a similar proportion was found in *P. canadensis*); this is much less compared to other species of Hymenoptera in which genes occupy between 140-160 Mb (and 50-65%) of the assembled genome (**Figure 1B**). This difference is due primarily to the annotation of fewer long genes, and in particular, long introns: while other Hymenoptera have 600-700 gene iLoci 50 kb in length or greater, *P. dominula* has only 84 (see **Figure 1B**). Further comparative genomic analyses can help resolve whether this observed reduction in long introns in the *P. dominula* genome is a truly unique characteristic of the genome sequence, or whether it stems from differences in annotation workflows.

The *Polistes dominula* genome is also characterized by an extremely biased nucleotide composition (**Figure 1C**). With a genome-wide GC content of 30.8%, the *P. dominula* genome is the most biased genome yet reported in Hymenoptera (Bonasio *et al.* 2010; Nygaard *et al.* 2011; Smith *et al.* 2011a; Smith *et al.* 2011b; Suen *et al.* 2011; Weinstock *et al.* 2006; Werren *et al.* 2010) and one of the most biased known in any animal (`http://www.ncbi.nlm.nih.gov/genome/browse/`). At the resolution of individual gene loci, however, *P. dominula* is not as GC-poor as *Apis mellifera* (29.0% and 24.7% median GC content, respectively), the primary factor being the extreme bias of introns in *A.*

mellifera (21.3% median GC content for *P. dominula* versus 17.3% median GC content for *A. mellifera*; the distribution of intron length is nearly identical for *P. dominula* and *A. mellifera*). The composition of the *P. canadensis* genome is slightly less biased than *P. dominula* at all levels of resolution, but all trends in comparison to *A. mellifera* are consistent (see **SI Results**). The biased composition of *Polistes* genomes raises some compelling questions about the evolution of genome composition and potential contributing factors such as bias in DNA mismatch repair and other genome maintenance mechanisms, as well as the possibility of historically high levels of CpG methylation and cytosine deamination.

Caste-related transcriptome reveals differentially expressed conserved and novel genes.

Distinct queen and worker castes arise from the same genome, a phenomenon known as caste polyphenism that is characteristic of many social insects (Smith *et al.* 2008). Differences in gene expression and alternative splicing between castes has been a topic of intense research interest because it provides a striking example of environmentally-induced phenotypic plasticity. Caste-differential expression has been widely investigated in advanced eusocial honey bees (Chen *et al.* 2012; Grozinger *et al.* 2007; Whitfield *et al.* 2003) and ants (Ometto *et al.* 2011; Simola *et al.* 2013a). More recently, high-throughput RNA sequencing technology (RNA-Seq) has been applied to profile expression in species representing a wider array of insect sociality, including an incipiently social small carpenter bee, an intermediately social bumble bee (Harrison *et al.* 2015) and two primitively eusocial species of *Polistes* wasps (Berens *et al.* 2015b; Ferreira *et al.* 2013). New RNA-Seq data described in this study represent transcriptome data for a third *Polistes* species, facilitating the discovery not only of caste-differentially expressed genes in *P. dominula*, but also of conserved *Polistes*-specific genes.

We performed two lanes of Illumina paired-end RNA-Seq on mRNA isolated from

heads of individual adult workers and active egg-laying queens (six replicates per group, from the same population as the male used for genomic DNA sequencing). The RNA-Seq reads were then mapped to 17,888 iLoci using Bowtie (Langmead *et al.* 2009), with most libraries mapping at an efficiency of 80%, after which iLocus abundances were estimated using RSEM (Li & Dewey 2011) and tested for differential expression using EBSeq (Leng *et al.* 2013) (methods and quality control described in detail in **SI Methods**). We identified 381 iLoci differentially expressed between queens and workers (**Figure 2A**), 100 lacking annotated gene models, 276 containing a single annotated gene, and 5 containing multiple genes. The majority of the 381 differentially expressed iLoci (231; 60%) are up-regulated in workers. Other reports that also focused on head or brain gene expression from *Polistes* (Berens *et al.* 2015b; Ferreira *et al.* 2013) and honey bees (Grozinger *et al.* 2007) also found the majority of genes are worker-biased in expression. The skew towards worker-biased expression could reflect differences in behavioral flexibility and/or cognitive demands of workers compared to egg-laying queens (O'Donnell *et al.* 2011).

Differentially expressed iLoci are significantly enriched for Gene Ontology functions in fatty acid metabolism, neurotransmitter activity, and amino acid metabolism when compared to the background set of all *P. dominula* gene models (**Figure 2B**). Previous studies on caste-related gene expression in other *Polistes* species have also identified consistent differences in the expression of genes related to lipid metabolism (Berens *et al.* 2015b; Sumner *et al.* 2006; Toth *et al.* 2010). These data contribute to a growing base of information suggesting the expression of deeply conserved genes (i.e. a "genetic toolkit") related to metabolism is related to caste differences and may play a key role in the evolution of caste-containing insect societies (Toth & Robinson 2007).

All previously examined insects show evidence of large amounts of alternative splicing, including other social insect genomes (Flores *et al.* 2012; Li-Byarlay *et al.* 2013). As expected, we found evidence for alternative splicing in 1,743 of the *P. dominula* gene

models. In particular, *via* transcript mapping and scanning for the two major types of alternative splicing (intron retention and exon skipping, see **SI Methods**) we uncovered 1,616 intron retention events in 1,135 genes and 1,720 exon skipping events in 884 genes (see **SI Results**). 859 genes show only intron retention, 608 genes show only exon skipping, and 276 genes show both. However, analysis with Cufflinks and Cuffdiff {Trapnell, 2013 #520}(Trapnell *et al.* 2013) reported no cases of caste differential splicing, suggesting alternative splicing is not related to adult caste differences, at least in heads, of *P. dominula.*

Recently, there has been growing interest in the potential for "novel", or taxonomically restricted genes in the evolution of novel phenotypes and in particular, eusociality (Sumner 2014). We also used our transcriptome data, in conjunction with previously published data for other *Polistes* species, to search for well-supported *Polistes*-specific genes. Our data represent the third published transcriptome dataset for a *Polistes* species, together with the transcriptomes of two New World species, *Polistes metricus* (Berens *et al.* 2015b) and the Neotropical *Polistes canadensis* (Ferreira *et al.* 2013). In the *Polistes canadensis* study, the authors identified a large number of novel transcripts (approximately 50% of sequenced transcripts) with no similarity to any known sequence and suggest that novel genes may be related to the evolution of caste differences in social insects (Sumner 2014). We performed a more in-depth exploration of the three transcriptomes in order to identify *Polistes*-specific transcripts that were shared by all three species. Such transcripts are much more likely to represent true protein-coding genes because they are conserved across species and there is evidence of their expression in multiple species. Considering *P. dominula* transcripts with an open reading frame of at least 80 aa, we found 19,173 transcripts with no significant similarity to Hexapoda sequences. Only 144 of these transcripts have translation products that are conserved between all three *Polistes* transcriptomes. Of the 144 conserved shared transcripts, 118 are found in the annotated genome assembly, aligning to 93 different iLoci (**Figure 2C**).

Only 10 of these 93 iLoci also have evidence of *Polistes*-specific genes from the genome annotation (in the form of gene models without matches to protein databases), and even in these 10 cases there is little agreement between transcript alignment structure and predicted gene structure (**Table S11**). These results suggest that while single lines of evidence may offer hints of clade-specific genes, very few cases are well supported when subjected to multiple lines of inquiry. This confirms a recent study in ants that uncovered evidence for very few shared, genus-specific genes and more unique species-specific gene, some of which are likely bioinformatics artifacts (Simola *et al.* 2013a).

There are conflicting reports on the association of novel transcripts in caste differences in *Polistes*. Transcriptomic comparisons from *P. canadensis* adults suggested novel transcripts are more likely to be caste-biased in expression (Ferreira *et al.* 2013), whereas novel transcripts from *P. metricus* larvae did not show this caste-bias (Berens *et al.* 2015b). In the current study, we found that 77 out of 93 iLoci (83%) associated with the 144 well-supported *Polistes*-specific transcripts are caste differentially expressed (significantly overrepresented, Fisher's Exact Test $p < 2.2e\text{-}16$), 34 of which (44%) are up-regulated in workers. In addition, eight of the 10 iLoci containing both unmatched transcripts and unmatched gene models are caste differentially expressed, with 4/8 up-regulated in workers. These results are consistent with data from *P. canadensis* (Ferreira *et al.* 2013) suggesting novel genes are more likely to be caste-biased in expression in adults. The fact that a similar relationship between caste-biased expression and novel genes was not found in *P. metricus* larvae suggests there could be different functions for novel genes across species and/or life stages.

DNA methylation system is greatly reduced in *P. dominula*.

There has been great interest in the role of epigenetics in eusociality, and data from honey bees has generated considerable interest in the potential role of DNA methylation in the regulation of gene expression during the development of queen and worker castes

(Lyko & Maleszka 2011). We used the aforementioned genome and transcriptome data from *P. dominula*, along with newly generated whole genome bisulfite sequencing (methylome) data, to probe the presence and extent of caste-association of DNA methylation in the independently evolved social paper wasps.

A full complement of DNA methyltransferases, *Dnmt1, 2,* and *3,* is considered to be necessary for a fully functional DNA methylation system (Lyko & Maleszka 2011). *Dnmt1* is typically considered as the "maintenance" methyltransferase involved in maintaining consistent methylation across cell divisions and generations (Lyko & Maleszka 2011). *Dnmt2* is thought to be involved mainly in the methylation of transfer RNAs. *Dnmt3* is the "de novo" methyltransferase, and has been suggested to be related more to environmentally-responsive DNA methylation that occurs within the lifetime of an individual (Lyko & Maleszka 2011). Other canonical methylation-related proteins include MBD (Methyl-CpG-binding domain protein) and the demethylation enzyme TET (Ten-eleven translocation methylcytosine dioxygenase) (Lyko & Maleszka 2011). We used BLAST to identify sets of homologs for each of these genes and subjected these sets of sequences to molecular phylogeny analysis to determine copy numbers of each of these five major DNA methylation related genes.

All previously sequenced Hymenoptera possess a full complement of DNA methyltransferases (Yan *et al.* 2015), except the recently sequenced *Polistes canadensis* which lacks *Dnmt3* (Patalano *et al.* 2015). Our MAKER annotation workflow (augmented by manual annotations as well as low stringency similarity searches for potentially incomplete or highly diverged homologs) also uncovered no *Dnmt3* gene, but did identify one *Dnmt1* gene (as in ants (Bonasio *et al.* 2010), as opposed to two in honey bees (Wang *et al.* 2006)) and one *Dnmt2* gene, as well as genes encoding MBD and TET homologs (summarized in **Figure 3A**). To further investigate whether the absence of a *Dnmt3* gene model might represent a gene loss, we examined available Hymenoptera genomes for shared synteny in the region harboring *Dnmt3* in *Apis mellifera*. Results show that

the *Dnmt3* locus is within a syntenic block encompassing at least an additional two genes upstream and two genes downstream, conserved in bee and ant genomes. Synteny analyses were conducted with the SynFind and associated tools within the CoGe platform (https://genomevolution.org/coge/; Tang *et al.* (2015)). Sample genome alignments are shown in **Figure 3B**. The first upstream and the two downstream genes co-localize in a 90kb region on scaffold0086 of our *P. dominula* assembly, preserving the syntenic block (the leftmost gene of the bee-ant syntenic block is preserved on scaffold0049 but would be at least 235 kb away if this scaffold were to align upstream of scaffold0086). However, there is a conspicuous absence of any similarity to *Dnmt3* in the syntenic region of the *P. dominula* genome (**Figure 3B**). Intriguingly, across the remaining Hymenoptera species, the region upstream of the annotated *Dnmt3* genes encoding the conserved C-terminus of the methyltransferase is highly variable in size and gene structure annotation is unclear (including annotation of the possibly overlapping upstream gene). The *Nasonia vitripennis Dnmt3* protein is highly diverged at the N-terminus, and although the other genes of the bee-ant syntenic block are highly conserved in *Nasonia*, they are widely spread in the genome. These results suggest that the *Dnmt3* locus, and by extension, perhaps some functional aspects of DNA methylation systems in general, are not highly conserved in different lineages of Hymenoptera.

In addition to the lack of genome sequence evidence for a functional *Dnmt3* gene in *Polistes dominula*, we found no significant similarity between Hymenoptera *Dnmt3* sequences and transcripts from any of the three *Polistes* species' transcriptomes (Berens *et al.* 2015b; Ferreira *et al.* 2013) (tblastn search with -evalue 1e-8). The lack of any *Dnmt3* transcripts in the three congeners strongly suggests this gene has indeed been lost across the genus *Polistes.* Further work on additional species will be necessary to determine whether this loss is common to the entire paper wasp family Vespidae.

Along with the loss of *Dnmt3,* whole genome bisulfite sequencing of one pool of queen and one pool of worker heads revealed a dramatic reduction in DNA methyla-

tion in *P. dominula* compared to other Hymenoptera. This is in contrast to previous reports suggesting the presence of typical amounts of DNA methylation in *P. dominula*, but these reports used a less reliable method for estimating DNA methylation based on a methylation-sensitive restriction enzyme assay (Kronforst *et al.* 2008; Weiner *et al.* 2013). Through a complete and uniform reanalysis of previously published Hymenoptera bisulfite sequencing data (from honey bees, two ant species, *Polistes canadensis,* and *Nasonia vitripennis*) we were able to make a reliable comparison of DNA methylation levels in *P. dominula* to those of other Hymenoptera (described in **SI Methods**). Overall, levels of DNA methylation in *P. dominula* are more than two orders of magnitude lower than in other Hymenoptera. This includes the congeneric *Polistes canadensis*, which, although showing lower levels of methylation than ants and bees, still has two orders of magnitude more methylated CpG sites than *P. dominula* (**Figure 3C**, **SI Results**). We uncovered only 124 and 158 CpG methylated sites, respectively in the queen and worker samples; this is in stark contrast to tens of thousands of sites uncovered in all of the other Hymenoptera species (**Figure 3C**). Similar to other insects (Rasmussen & Amdam 2015), most methylated sites were found within genes (74 and 89 sites in queen and worker samples, respectively); see **SI Results**. Strikingly, methylation was targeted to the same seven genes in both queen and worker samples (**Figure 3D**), and several of these have putative functions related to DNA binding. Thus, there were zero caste differentially methylated genes, and great similarity between castes, even at the level of which cytosines within the seven genes were methylated. Of the 101 total methylated cytosines within the seven genes, 62 (61%) of the same methylated cytosines were shared between both castes (**Figure 3D**). This result is again in contrast to studies from both bees (Lyko *et al.* 2010) and ants (Bonasio *et al.* 2012), which reported hundreds of caste differentially methylated genes between queen and worker castes. The fact that nearly identical methylation patterns were found in just a few genes, but consistently across castes, suggests the extremely low level of DNA methylation we describe in *P. dominula*

is real and may be of some functional significance. We suggest that, despite a massive reduction in *de novo* methylation in paper wasps, there may have been selection to retain "maintenance methylation", likely *via* the action of of *Dnmt1,* for a few key genes. This idea is supported by the observation that five out of the seven *P. dominula* methylated genes also showed strong methylation in *P. canadensis* (**SI Results**), and homologs of three of these seven genes in *Apis mellifera* are methylated consistently across multiple independently published experiments.

Our methylome data from *P. dominula* also suggest no clear connection to dynamic gene expression patterns: none of the seven methylated genes is differentially expressed between castes; two out of the seven show some evidence of alternative splicing, but not caste differential splicing (PdomGENEr1.2-09385 and PdomGENEr1.2-09184). Although *P. canadensis* also shows some evidence of a reduced methylation system (loss of *Dnmt3* and fewer methylated CpG sites and methylated genes than bees and ants (Patalano *et al.* 2015)), the reduction in *P. dominula* is much more striking. This suggests reduced DNA methylation systems may be a general characteristic of paper wasps, but that there has been even further reduction of these systems in the *P. dominula* lineage relative to some of its congeners.

These data raise intriguing questions about the importance and function of DNA methylation in insects. DNA methylation systems have also been dramatically reduced in other insect lineages (e.g. *Drosophila* flies and *Tribolium* beetles), the shared feature being a loss of *Dnmt3* and large reduction in overall levels of DNA methylation (Glastad *et al.* 2011). Furthermore, there are other insects where *Dnmt3* is not present, but moderate levels of DNA methylation remain (Mita *et al.* 2004; Patalano *et al.* 2015). Thus, DNA methylation is not clearly related to gene regulation in some insects (Glastad *et al.* 2014) and even some social insects, suggesting other types of epigenetic mechanisms such as histone modifications (Simola *et al.* 2013b) or microRNAs may be more important. Our data also highlight the surprising lability of epigenetic mechanisms even

within an insect lineage (Hymenoptera) and do not support the idea that phenotypic plasticity afforded by DNA methylation is required for the evolution of castes in social insects (Weiner & Toth 2012).

We also examined patterns of occurrence of CpG dinucleotides in the *P. dominula* genome, because segmental ratios of observed to expected (o/e) CpG frequency have been used as an indicator of regional DNA methylation status, based on the assumption that highly methylated regions are characterized by mutational loss of methylated cytosines (Yi & Goodisman 2009). The distribution of CpG o/e in *P. dominula* is similarly broad as that of other Hymenoptera, but lacking the bimodal distribution characteristic of the measure in bee coding regions (**Figure S8**). Use of this measure as an indicator of methylation status in *P. dominula* would have incorrectly inferred the presence of numerous methylated genes. Thus, the CpG o/e measure does not accurately reflect the true methylation status of the *P. dominula* genome based on bisulfite sequencing, a much more direct and sensitive method for detecting actual site-specific methylation. It is conceivable that CpG depletion is still correlated with historical (not modern) patterns of DNA methylation, and this is reflected in the fairly typical CpG o/e distribution in *P. dominula* (**Figure S8**). Because appreciable levels of DNA methylation are found in a wide variety of other Hymenoptera (**Figure 3A**), it is likely that reduced DNA methylation is a derived condition in *Polistes*, but more data on additional species are needed to understand when and why reduced DNA methylation evolved in vespid wasps.

Aculeate Phylogeny.

The genome of *Polistes dominula* provides a beneficial complement to the genomes of several species of aculeate Hymenoptera already published. Together, these genomes are a powerful comparative genomics resource for identifying what is conserved and what is unique among the primary aculeate lineages. A delineation of the phylogenetic relationships between these lineages is a fundamental component for analysis and interpretation

of evolved traits, and yet consensus regarding the phylogeny of Aculeata remains elusive (Johnson *et al.* 2013; Pilgrim *et al.* 2008). A study using molecular data from 4 loci in 64 taxa placed bees (superfamily Apoidea) as sister to scoliid and bradynobaenid wasps (Pilgrim *et al.* 2008), while a more recent study involving analysis of 308 genes from 19 taxa found ants (family Formicidae) to be sister to bees (Johnson *et al.* 2013).

Because our current work describes the one of the first published complete genomes of a vespid wasp, we sought to use these data to investigate the phylogenetic grouping of *Polistes* proteins relative to orthologs in bees, ants, and the non-aculeate wasp *Nasonia* as an outgroup. Using conserved single-copy orthologs present in *Apis mellifera* (bee), *Harpegnathos saltator* (ant), *Polistes dominula* (paper wasp), and *Nasonia vitripennis* (non-aculeate outgroup), we inferred a phylogenetic tree for each gene using these four representative protein sequences (see **SI Methods**). We observed all possible topologies in the 2,077 gene trees: bees and ants as closest neighbors in 889 trees (43%), *Polistes* and bees as closest in 696 trees (34%), and *Polistes* and ants as closest in 492 trees (24%). Although the most common topology (bees and ants as closest) agrees with the results of the most recent, transcriptome-based phylogenetic analysis of aculeates (45), there was definitely not a clear consensus from our data. Based on our analysis, the protein-coding genomes of the published Hymenoptera do not yet provide a definitive answer to the question of the phylogenetic relationship of bees, ants, and vespid wasps. Additional aculeate genomes, including more representatives of the Vespidae (Jandt & Toth 2015) and other wasp families, may help to better resolve aculeate relationships in the future.

Conclusions

This paper provides valuable and comprehensive genomic resources for one of the major lineages of eusocial insects, the vespid wasps, represented by the behavioral model

species *Polistes dominula*. The *P. dominula* genome is a relatively compact (250Mb) genome with little repetitive DNA, as well as low GC content, in comparison to other Hymenoptera. Transcriptomic analyses revealed several hundred genes with caste-related expression, with functions related to fatty acid and amino acid metabolism and neurotransmitter activity. In addition, we identified several *Polistes*-specific genes, several of which also show differential expression between queen and worker castes. Together, these data provide some support for the roles of both conserved genes and novel genes in the evolution and maintenance of caste differences in social wasps (Sumner 2014; Toth & Robinson 2007). The most surprising finding from our *P. dominula* –omics data was clear evidence of a striking reduction in the DNA methylation system. *P. dominula* have a reduced complement of DNA methylation enzymes, including a loss of the *de novo* methyltransferase *Dnmt3*, as well as extremely reduced levels of DNA methylation in the genome—with evidence for just over 100 methylated sites in only seven genes. In addition, there was no relationship between DNA methylation and caste- related gene expression, methylation, nor alternative splicing. There has been great interest and research activity related to the potential role of DNA methylation in the regulation of caste differences and caste evolution in eusocial insects (Kucharski *et al.* 2008; Weiner & Toth 2012). Our data are novel in that they suggest *P. dominula* possesses the most reduced DNA methylation system known for any eusocial insect, but there are other examples of non-social insects with similarly reduced methylation systems, including *Drosophila*. These data add to growing evidence for a surprising amount of lability of epigenetic mechanisms in insects, and suggest DNA methylation *per se* is not generally related to the evolution of castes in social insects. These genomic, transcriptomic, and epigenomic data on a primitively eusocial vespid wasp open up exciting new possibilities for comparative genomics of social evolution. Comparisons both across eusocial lineages and within lineages have the potential to provide new insights into the roles of conserved genes and pathways, novel genes, and epigenetic mechanisms in social evolution (Rehan & Toth

2015).

Acknowledgments

This work was supported in part by the U.S.A. National Science Foundation grant NSF-IOS-1311512 and a grant from the Iowa State University Center for Integrated Animal Genomics, both awarded to AT. DS was supported in part by NSF award #1221984 to VB.

The authors would like to thank Amy Geffre for assistance with DNA and RNA extractions, Susan Weiner for preliminary work on DNA methyltransferase genes, and members of the Toth lab for reviewing the manuscript. We also thank Michael Goodisman and Brendan Hunt for discussions about methylome sequencing and DNA methylation patterns. We also thank Christina Grozinger, Stefano Turillazzi, Gene Robinson, and Joan Strassmann for helpful discussions and support in planning stages of this project, and GR and JS for comments on the manuscript. We also thank our manual annotation team: Arian Avalos, Seth Barribeau, Katherine Noble, Sandra Rehan, Fabio Manfredini, Griffin Smith, Amy Geffre, Adam Dolezal, Jimena Carrillo-Tripp, Alexander Walton, and Jennifer Jandt for submitting manual gene annotations; and Jon Duvick for reviewing and curating annotation submissions.

Data Accessibility

Raw Illumina sequences are available from the NCBI Short Read Archive under the following accessions: accession SAMN02584905 for whole genome shotgun reads; accessions SAMN03940809-SAMN03940820 for RNA sequence reads; and accessions SAMN03946123 and SAMN03946134 for bisulfite sequence reads. The genome assembly is available from GenBank under the accession GCA_001465965.1, and the transcriptome shotgun assembly is available from GenBank under the accession GEDB00000000.1.

Additional supporting data, analysis documentation, and supporting code have been deposited in the fig**share** archive and in several GitHub repositories, as described at https://pdomgenomeproject.github.io/.

Sequences, annotations, and alignments are also available through PdomGDB, an integrated data resource including a genome browser, a BLAST server, and a community annotation portal (see http://goblinx.soic.indiana.edu/PdomGDB).

Author Contributions

All computational analyses of the data were done by DS and VB at Indiana University. AB assisted with the assembly of the *P. dominula* transcriptome and with differential expression analysis. AS contributed to preliminary identification of transposable and repetitive elements, and to analysis of telomere-related genes and sequence motifs. KG contributed template Perl code for preliminary analyses of CpG o/e. AT conceived of and oversaw the project, interpreted results, and collected and processed wasp samples. DS, VB, and AT wrote the manuscript.

Tables and Figures

Figure 1. **A**. Best supported molecular phylogeny of the eusocial aculeate Hymenoptera based on recent transcriptome studies and analysis of 2,077 conserved genes reported in this study. **B**. Stacked bar plot showing genome content of *Polistes dominula* broken down by the proportion occupied by various categories of gene content and conservation, compared to another paper wasp (green labels), two bees (black labels), two ants (red labels), and the outgroup *Nasonia vitripennis*. See also **Figure S5**. **C**. Stacked rug plot showing nucleotide composition of long genomic sequences from 3 bees (in black), 2 ants (in red), the outgroup *Nasonia vitripennis* (in blue), and 2 paper wasps (in green). Each vertical bar represents a chromosome, linkage group, or scaffold at least

1 Mb in length. Photo credits: *N. vitripennis* by E. Cash and J. Gibson; *P. dominula* by S. McCann; *S. invicta* and *A. mellifera* by A. Wild.

Figure 2. **A**. Heatmap of expression values of 367 differentially expressed interval loci (iLoci). The blue color indicates overexpression, while the yellow indicates underexpression. 212 iLoci (58%) of the differentially expressed iLoci are overexpressed in workers. **B**. Bar chart showing the representation of eight GO functional categories as a proportion of differentially expressed iLoci (blue bars) versus all iLoci (red bars), determined by an enrichment analysis to be overrepresented in differentially expressed iLoci. **C**. Putative *Polistes*-specific genes, defined as unmatched transcripts with significant pairwise protein-level similarity among three Polistes species (Pd for *P. dominula*, Pc for *P. canadensis*, and Pm for *P. metricus*). Because of variation in gene copy number and number of alternative transcript isoforms, the number of transcripts in each intersection of the diagram is different for each species, and only the smallest number is shown. For example, the 3-species intersection consists of 144 transcripts from *P. dominula*, 136 transcripts from *P. canadensis*, and 95 transcripts from *P. metricus*.

Figure 3. **A**. Copy number of five methylation-related genes in the primary Hymenoptera lineages. **B**. Evidence for shared synteny around the *Dnmt3* locus in bee, paper wasp, and ant genomes. Colored bars represent conserved coding regions between *Polistes dominula* (top track) and *Apis mellifera* (black blocks), *Bombus terrestris* (grey blocks), and *Camponotus floridanus* (red blocks). Regions of similarity are largely collinear (shown for the *Apis mellifera* to *Camponotus floridanus* comparison by the blue lines connecting the similar blocks). Gene models are shown by arrow structures with coding exons in green, UTRs in blue, and introns as thin lines. Each gene is denoted by a numbered box as follows: 1) GenBank protein entries XP_006568814 (26S proteasome non-ATP regulatory subunit 6-like), 2) XP_006568813 (uncharacterized), XP_00658716 (Dnmt3), 4) XP_006568806 (polynucleotide 5'-hydroxyl-kinase NOL9-like), and 5) XP_0065688112 (histone lysine demethylase PHF8-like) in *A. mellifera* (linkage

group LG2, GenBank NC_007071), from left to right. The lack of similarity blocks around 50K on the *P. dominula* scale demonstrates the postulated loss of *Dnmt3* in this species. **C**. Bar chart showing the number of CpGs with a high level of support for methylation from bisulfite sequence data. The number of highly supported methylation sites for each species is based on pooled reads for all available samples, and is shown for one bee (black), two ants (red, note that "worker" refers to "minor workers" in *C. floridanus*), two paper wasps (green), and *N. vitripennis* (blue). **D.** Gene models for the 7 *P. dominula* methylated genes (with corresponding Gene ID numbers and putative annotations based on best BLAST hits). Approximate locations of each of the 124 highly supported methylation sites within each gene are indicated with a line-dot symbol, with sites methylated in both queen and worker samples (n=76) indicated in black, sites methylated in just the queen sample in green (n=18), and sites methylated in just the worker sample (n=30) in orange.

Table 1. Genome assembly summary for *Polistes dominula* and six related Hymenoptera.

CHAPTER 3. PARSEVAL: PARALLEL COMPARISON AND ANALYSIS OF GENE STRUCTURE ANNOTATIONS

A paper published in *BMC Bioinformatics*: doi:10.1186/1471-2105-13-187.

Standage DS, Brendel VP

Abstract

Background: Accurate gene structure annotation is a fundamental but somewhat elusive goal of genome projects, as witnessed by the fact that (model) genomes typically undergo several cycles of re-annotation. In many cases, it is not only different versions of annotations that need to be compared but also different sources of annotation of the same genome, derived from distinct gene prediction workflows. Such comparisons are of interest to annotation providers, prediction software developers, and end-users, who all need to assess what is common and what is different among distinct annotation sources. We developed ParsEval, a software application for pairwise comparison of sets of gene structure annotations. ParsEval calculates several statistics that highlight the similarities and differences between the two sets of annotations provided. These statistics are presented in an aggregate summary report, with additional details provided as individual reports specific to non-overlapping, gene-model-centric genomic loci. Genome browser styled graphics embedded in these reports help visualize the genomic context of the annotations. Output from ParsEval is both easily read and parsed, enabling systematic

identification of problematic gene models for subsequent focused analysis.

Results: ParsEval is capable of analyzing annotations for large eukaryotic genomes on typical desktop or laptop hardware. In comparison to existing methods, ParsEval exhibits a considerable performance improvement, both in terms of runtime and memory consumption. Reports from ParsEval can provide relevant biological insights into the gene structure annotations being compared.

Conclusions: Implemented in C, ParseEval provides the quickest and most feature-rich solution for genome annotation comparison to date. The source code is freely available (under an ISC license) at http://parseval.sourceforge.net/.

Background

It was only a decade ago when annotating a eukaryotic genome required years of extensive collaboration and millions of dollars of investment. Since then, the tremendous rate at which the cost of DNA sequencing has been dropping as well as increased accessibility to gene prediction software are placing genome sequencing and annotation well within the reach of most single investigator biology laboratories. As a result, proliferation of distinct annotation sets corresponding to the same genomic sequences is becoming increasingly common. Annotation sets for a particular genome can accumulate in a variety of scenarios. When developing gene prediction software, it is common to test the software on a genomic region for which a high-quality reference is available, running and re-running the software and comparing the resulting predictions against the reference. Community groups providing annotation for species- or clade-specific genomes typically release updated annotations following the initial release. Affordable transcriptome sequencing provides individual labs with data to specifically improve annotations for particular genes of interest, for example with respect to alternative splicing. In each of these scenarios, multiple annotations associated with a common set of genomic sequences

require comparative assessment.

A variety of comparison methods exist, but none can fully address the growing needs of the community (see Table 1). Manual comparison approaches can trivially be ruled out as slow, tedious, error prone, and hopelessly unscalable. Although genome browsers have had a huge impact by making gene annotations accessible to a wide variety of scientists, they likewise do little to provide the automation and precision needed in whole-genome annotation comparisons. Large genome sequencing projects and centers have certainly developed in-house scripts and pipelines over the years to address this need. However, these pipelines are typically not standardized, not openly shared, and do not migrate well.

Tools such as the Eval package [23] and the GFPE program [24] represent some of the earliest efforts to provide a reusable, easy-to-use annotation comparison tool to the community. Eval in particular stands out based on the amount of detail provided by its reported comparison statistics and by the ability to visualize the distributions of these statistics. Eval takes as input annotation files in Gene Transfer Format (GTF) and calculates a rich set of descriptive statistics summarizing the differences between the annotations. Because whole-genome annotations typically include thousands (or tens of thousands) of genes, these statistics are intended to condense the information into a comprehensive yet concise summary (at the resolution of entire sequences or sets of sequences), facilitating targeted improvement of gene prediction software. Unfortunately, this condensing process discards large amounts of valuable information at the resolution of individual gene loci, making the tool unsuitable for analyses that target a particular gene, sets of genes, or gene loci with characteristics of interest from within a larger set of genes. Such locus-resolution comparisons are useful not only to software developers and annotation producers who need to know whether their software has distinct advantages or disadvantages, e.g., favoring long over shorter gene models on average, or failing in untranslated region (UTR) prediction, but they are of primary interest for specialists

concerned with a particular gene family or pathway.

Motivated by a need for genome-scale evaluations with locus-scale detail, we developed ParsEval, a program for comparing and analyzing distinct sets of gene structure annotations for the same input sequences. The program is designed to incorporate all of the benefits of existing methods while addressing their shortcomings. ParsEval identifies differences in exon/intron assignments and in coding sequence (CDS) and UTR designations, at both feature-level (exon, CDS segment, UTR segment) and nucleotide-level resolution. The output consists of a set of commonly used statistics that provide quantitative measures of agreement when comparing predicted gene structures against a standard reference [28, 29, 6]. This output is presented in a detailed report for each gene locus, supplemented with genome browser styled graphics to enable additional visual assessment and analysis of the annotations. The statistics are also presented in a single summary report that aggregates the statistics across all loci, providing a condensed high-level view of the similarity between the two sets of annotations. For gene loci that include alternatively spliced genes or overlapping genes (or both), ParsEval determines the optimal matching of reference transcripts to prediction transcripts, and additionally reports any novel transcript predictions that have been identified.

Implementation

Overview

ParsEval is a gene annotation comparison and analysis tool, designed with a focus on speed, resource efficiency, and portability. The program takes as input a pair of gene structure annotations corresponding to the same sequence (in GFF3 format [30]), analogous to two separate annotation tracks one might see in a genome browser. For comparison purposes, the first set of annotations is treated as the *reference* while the other is treated as the *prediction*, although ParsEval makes no assumptions regarding the

respective quality of the two annotation sets. The output of the program is a set of reports containing common comparison statistics intended to highlight relevant similarities and differences between the two sources of annotation.

ParsEval first loads the annotation data into memory, identifies start and end coordinates for gene loci, and associates each gene annotation with a single locus. Next, the program does a comparative assessment of the gene annotations for each locus, calculating and storing a variety of informative similarity statistics. Finally, ParsEval generates reports providing a detailed readout of these statistics.

Implemented in ANSI C, ParsEval is fast, memory efficient, and portable, designed to run on all POSIX-compliant UNIX systems (Linux, Mac OS X, Cygwin, Solaris, etc.). Most of the analysis code is implemented with shared memory parallelization, providing additional performance gains when running on multicore processors that are becoming increasingly common in commodity hardware. ParsEval's only external dependency is the GenomeTools library [31], which provides an API for generating annotation graphics with AnnotationSketch [32], as well as implementations of a variety of data parsers and dynamic data structures.

Gene locus identification

Comparative analysis of two sets of gene annotations requires determining how annotations from one set correspond to annotations from the other, as well as the genomic coordinates (the *gene locus*) that should be considered in each comparison. For rare cases in which a single reference annotation and a single prediction annotation line up perfectly, determining the gene locus and the corresponding genes is trivial. However, in most cases this task is complicated a variety of factors. For example, a single gene prediction workflow may annotate multiple genes at a single location, so one must determine how to associate these annotations with corresponding annotations from an alternative source. Furthermore, when one or more gene annotations from one source overlap with

multiple annotations from another source, one must determine how to compare these gene annotations and which coordinates to include in the comparison.

One common approach involves designating one set of annotations as the *reference* set and then using the coordinates of each reference gene annotation to define a distinct gene locus to serve as the basis for subsequent comparison (see Figure 1). However, this approach is unfavorable for several related reasons. First, reference gene annotations that overlap are handled separately, when it makes more sense to associate them with the same locus and handle them together. Second, it forces a quality judgment between the two sets of annotations when their relative quality is often unknown. The two sets of annotations likely include complementary information, and unless there is a clear distinction in quality between the two, choosing one as a reference discards clearly related information from the other. Third, relevant information from predicted gene models that extend beyond the boundaries of the corresponding reference annotation is ignored.

Although ParsEval uses the terms *reference* and *prediction* to distinguish between the two sets of annotations, both are considered equally when identifying gene loci. Each gene annotation corresponds to a node in an interval graph G. There is an edge between two nodes G_i and G_j if the corresponding gene annotations overlap (see Figure 2). Each connected component in G then corresponds to a distinct gene locus, which we define as the smallest genomic region containing every gene annotation associated with the corresponding subgraph. Defining a gene locus in this way makes no assumptions as to the relative quality of the two sets of annotations, and ensures that no potentially relevant data are discarded. Furthermore, according to this definition each gene locus is independent, enabling the subsequent comparative analysis tasks to run in parallel.

Gene structure representation

To facilitate analysis at each gene locus, ParsEval converts GFF3 annotations for each gene into a character string representing the annotated gene structure (a *model*

vector). This model vector is similar to a sequence in Fasta format, except instead of using the alphabet $\{A, C, G, T\}$ to represent chemical composition at each nucleotide, the alphabet $\{C, F, G, I, T\}$ representing gene structure is used: C for coding sequence, F for 5'-UTR, T for 3'-UTR, I for introns, and G for intergenic sequence. Using this alphabet, each transcript can be represented by a single model vector. ParsEval uses these model vectors when comparing reference and prediction gene annotations.

In many cases, a single pair of model vectors (one for the reference, one for the prediction) is sufficient to fully represent annotated gene structure at a given locus. This is certainly true when both the reference and the prediction annotate a single gene with a single mRNA product at the locus. But even if the reference (or the prediction) annotates multiple genes or transcripts, non-overlapping annotations can be encoded in the same model vector and compared simultaneously with corresponding annotations from the other data set. However, if either the reference or the prediction contains annotations for overlapping transcripts, either because of alternative splicing or because of overlapping gene models, a single pair of model vectors is insufficient to represent the complete annotated gene structure at that locus. In these more complicated cases, the reference or the prediction or both will be associated with multiple model vectors. Thus, the algorithmic requirement is to represent all annotated transcript structures in the locus using the smallest number of model vectors.

This problem reduces to a common problem in graph theory known as the *maximal clique enumeration problem* [33]. We treat each transcript as a node in an undirected graph and place an edge between two nodes if the corresponding transcripts do not overlap (unlike the locus identification step, reference annotations and prediction annotations are handled separately in this step). Each maximal clique (maximal fully-connected subgraph) in this graph corresponds to a set of transcripts that do not overlap and can therefore be collapsed into a single model vector. ParsEval uses the Bron-Kerbosch algorithm [33] to enumerate all maximal transcript cliques, first for the reference and then

for the prediction. A model vector is generated for each clique, after which ParsEval compares all reference model vectors with all prediction model vectors.

Comparative analysis of annotations

Given a pair of equal-length model vectors representing a pair of gene structure annotations at a given locus, ParsEval computes a variety of comparison statistics to measure the level of agreement between the pair of annotations. Calculated at different levels of resolution, these statistics provide a detailed assessment of similarity between the reference and the prediction. At the resolution of distinct annotation features, ParsEval calculates the sensitivity and specificity as described in [28], the F1 score as described in [29], and the annotation edit distance as described in [6, 34]. These statistics are calculated for exons, CDS segments, and UTR segments. Note that for a prediction feature to be considered a true positive, ParsEval requires both the start and end coordinates to match the reference perfectly.

At the nucleotide-level resolution, ParsEval also calculates the sensitivity, specificity, F1 score, and annotation edit distance, as well as the simple matching coefficient and the correlation coefficient as described in [28]. These statistics are calculated for coding nucleotides (CDS) and untranslated exonic nucleotides (UTR). Overall identity at the nucleotide level, of which the simple matching coefficient is a generalization, is also computed.

For complex loci requiring multiple comparisons, the locus report includes an aggregate summary of the similarity statistics at the locus level in addition to the reports for each individual comparison. This locus-level summary also includes the splice complexity statistic [6], which ParsEval computes and reports for both the reference and the prediction at the locus level.

Based on the computed statistics, each comparison is classified in terms of similarity. A comparison is classified as a *perfect match* if the model vectors (and by implication the

annotated gene structures) are identical. A comparison is classified as a *CDS structure match* if the comparison is not a perfect match, but there is perfect agreement in terms of CDS structure. A comparison is classified as an *exon structure match* if there are differences in the coding sequence that nevertheless preserve exon structure (as resulting from different start and/or stop codons). A comparison is classified as a *UTR structure match* if there are differences in CDS and exon structure, but the UTR structures are identical. All other comparisons are classified as *non-matches*.

Note that, as with feature-level statistics, match classifications require perfect agreement. For instance, a pair of annotations may have very similar CDS structures, and this will be reflected in the nucleotide-level CDS statistics. However, if the CDS structures are not precisely identical, the comparison will not be classified as a *CDS structure match*.

As comparison statistics are computed on a locus-by-locus basis, ParsEval also maintains a running total of all comparison counts (such as true positives and false positives) from which the statistics are computed. When all loci have been considered, each comparison statistic is then recomputed using these running totals to provide an overall assessment of similarity.

Reporting comparison scores

For each gene locus, comparison statistics are calculated for each corresponding pair of reference and prediction model vectors. If multiple comparisons are required at a locus, however, statistics are not reported for each comparison. The comparisons are ranked using the previously described similarity statistics and are reported so as to ensure each transcript (or transcript clique) is considered at most one time. In cases where there is an unequal number of reference and prediction transcripts (or transcript cliques) associated with a particular locus, some will be labeled as novel or unmatched transcripts, and corresponding statistics are not included in ParsEval's reports.

ParsEval presents the comparison statistics in a collection of reports. The first is a single summary report providing the aggregated statistics for a high-level assessment of similarity, as is standard for tools of this kind. Additionally, ParsEval produces a dedicated comparison report for each individual locus. The detail provided by these locus-level reports is extremely valuable, and ParsEval is the only tool of its kind that preserves and reports comparisons at this level. By default, ParsEval generates these reports in an easy-to-parse and easy-to-read text format. However, ParsEval can also generate the reports as hyperlinked HTML files to facilitate browsing and network-based distribution. Furthermore, ParsEval can supplement HTML reports with embedded PNG graphics providing a genome-browser-like view of each locus' genomic context and enabling visual assessment of the annotations.

If more targeted reporting is desired, ParsEval also provides some filtering features. Using a simple optional configuration file, the user can exclude some gene loci from the reports based on a variety of features: locus length, number of genes, number of transcripts, number of transcripts per gene, number of exons, and CDS length. No comparisons are performed for loci that are filtered out, and thus do not contribute to the reported aggregate summary statistics and comparison classifications.

To facilitate integration of comparison reports with popular genome browsers such as GBrowse [35] and PlantGDB [36], ParsEval can generate an additional output file (in GFF3 format) containing the coordinates of each gene locus. These genome browsers commonly allow users to anonymously create private custom tracks with uploaded data, which provides the quickest mechanism for integration. Once a track is populated with the uploaded locus data, the user can configure the track configuration so that each locus feature in the track is hyperlinked to the corresponding ParsEval report stored, for example, on that user's local machine (see Figure 3). Alternatively, if a more permanent and public solution is desired, a user with administrative privileges for the genome browser can follow standard procedures for populating a new track with the GFF3 data, and

then configure the track so that locus features are linked to network-accessible ParsEval reports.

Results and Discussion

We present several use cases to demonstrate ParsEval's capabilities, benchmark its performance, and compare its utility relative to existing methods. The input data for these demonstrations were obtained from a variety of public databases with different respective formatting conventions. Accordingly, all data files were processed and converted to a uniform format before analysis. A detailed description of this conversion process, along with all code and commands used, are provided in the Supplemental Data as well as in ParsEval's source code distribution.

Unless otherwise noted, all use cases and benchmarks described herein were run on a fairly modest desktop computer: a Mac Pro with two 2.8 GHx quad-core Intel Xeon processors and 4 GB of RAM. ParsEval's performance for these demonstrations should therefore be fairly representative of the performance one might expect when running on commodity laboratory or personal hardware.

Use case: predictions vs. gold standard

High-quality gene structure annotations derived from a combination of computational and experimental evidence, and possibly improved with expert manual curation, are indispensably used as "gold standards" for measuring the accuracy of a novel gene prediction method or entire new annotation workflows. Identifying differences between the new method's predictions and such gold standard reference can help identify areas in which the novel method provides or needs improvement. Reports from ParsEval are effective for quickly and clearly identifying such differences.

To demonstrate ParsEval in this context, we reproduced a comparison that was orig-

inally published to assess the performance of the AUGUSTUS gene prediction program [37]. In the original study, AUGUSTUS was tested on the *h178* data set [38], a set of 178 human genomic sequences, each containing a single gene, for which annotations were available from the EMBL database release 50 [39]. Gene predictions from AUGUSTUS were compared the annotations from EMBL, and sensitivity and specificity scores were calculated at the nucleotide level, the exon level, and the gene level.

We obtained the *h178* data set (sequences and EMBL r50 annotations) from [40]. We then used the latest version of AUGUSTUS (2.5.5) to generate gene predictions for the 178 sequences. The data files were reformatted and then compared using ParsEval. Running on a desktop computer, ParsEval generated graphical reports in less than a minute. The summary report provided immediate access to a variety of similarity metrics, including those reported in the original assessment. The sensitivity and specificity values reported by ParsEval are comparable to those reported in the original AUGUSTUS manuscript (see Table 2). Differences in the comparison metrics can likely be explained by improvements to the AUGUSTUS program since publication, although the exact reason is elusive since the original AUGUSTUS software is no longer accessible.

Use case: two sets of annotations

When working with genome annotations, there is an increasing variety of cases in which no gold standard is available for comparison. For example, gene annotations for many model species are available from a variety of sources (i.e., UCSC versus Ensembl). The respective quality of these different annotation sets is not always clear, but comparison is still a necessary and fundamental task. Another example relates to genome projects that typically offer multiple releases of gene annotations between each major genome assembly release. Although newer releases may offer marginal improvements over the older ones, neither one can truly be considered a high-quality standard reference for comparison. An additional example relates to the increased affordability of

genome sequencing and the number of new and exotic species for which genome sequence is available. Gene annotation software is based on complex statistical models containing many parameters, and it is not always initially clear which parameter values to use up front. Therefore, when annotating a newly sequenced genome, it is common to extract a subset of the genome on which to perform repeated optimization runs to determine the parameter values that should be used subsequently to annotate the entire genome.

In each of these scenarios, multiple annotation sets must be compared, despite having no intuition as to the relative quality of the respective annotations. ParsEval was designed precisely for this type of analysis. Reports from ParsEval provide both an overall summary and locus-level detail, enabling the user to make informed decisions about annotations for individual loci, as well as for annotation sets as a whole.

As a demonstration of ParsEval's capability in this context, we downloaded two recent gene annotation releases (releases 64 and 65) for *Mus musculus* from the Ensembl database [41]. We compared these annotations using ParsEval, which required approximately 3 minutes of runtime on a desktop computer. A brief review of ParsEval's summary report shows that a total of 20,362 gene loci were identified using these annotations (see Table 3 for a complete breakdown). Of these gene loci, 6,725 had only annotations from release 64.

23,590 comparisons were performed by ParsEval, of which 22,333 (94.7%) were perfect matches between releases 64 and 65. A small number (83, 0.4%) of comparisons were classified as UTR structure matches. For the remaining 1,174 comparisons (5.0%) that were classified as non-matches, transcripts from release 64 contained an average of 16.47 exons, whereas transcripts from release 65 contained an average of 8.11 exons. A brief review of a handful of selected loci showed that many long transcripts (with many exons) that had been present in release 64 were absent in release 65.

This use case is an ideal demonstration of ParsEval's capabilities. Although the authors have no prior experience working with these particular data sets, a cursory

examination ParsEval's reports clearly draw attention to an important fact—between release 64 and 65, changes to Ensembl's annotation pipeline (perhaps different values for parameters that influence joining/splitting annotations, or implementation of stricter filters for gene length) affected approximately 5% of the gene annotations. Not only does ParsEval provide this information in a summarized form, it also provides detailed locus reports enabling users to scrutinize the results on a gene-by-gene basis. This breadth and detail of information is of great benefit to a wide variety of scientists and will empower them to more fully understand the available data and make informed decisions regarding alternative sources of annotation.

Benchmarks

To demonstrate its speed, scalability, and efficiency, we benchmarked ParsEval by analyzing pairs of whole-genome gene structure annotations for four common model organisms representing a wide range of eukaryotic diversity: *Arabidopsis thaliana* (thale cress), *Drosophila melanogaster* (fruit fly), *Glycine max* (soybean), and *Homo sapiens* (human) (see Table 4). To give a detailed demonstration of its performance, ParsEval was run 24 times for each species—3 technical replicates while varying the output mode (text and HTML/PNG) and the number of dedicated processors (1, 2, 4, and 8). Reported runtimes were obtained by taking the mean of the 3 corresponding replicates.

Performance in text output mode

ParsEval demonstrated optimal performance when running in text output mode, with runtimes ranging between about 30 seconds to about 4 minutes. Running ParsEval in parallel on multiple processors provided noticeable improvement in runtime for *Drosophila* and human, although no improvement was seen for *Arabidopsis* and soybean. It is likely that for loci with relatively small and simple gene structures, ParsEval's runtime is bound more by serial I/O related tasks than by actual analytical computations,

which would explain why no improvement was observed for the plant species.

Performance in HTML output mode with PNG graphics

Running ParsEval in HTML/PNG output mode increased the runtimes by an order of magnitude, although parallel processing kept these runtimes within a reasonable range (about a half hour for the most intensive comparison) with observed speedup factors ranging from 3 to 5 when using all 8 processors. Because these improvements in runtime were observed for all species, it is likely that ParsEval's runtime is bound primarily by computationally intensive graphics generation tasks when running in HTML/PNG output mode.

Notes on benchmark results

The results of the *A. thaliana* benchmark were not surprising. Perfect matches and CDS matches account for 97.5% of the comparisons, which makes sense considering that TAIR10 represents minor cumulative updates to TAIR9 (in contrast, perfect matches and CDS matches account for only 4.2% of comparisons between TAIR6 and TAIR10). There were even fewer differences between FlyBase and Ensembl annotations for the *D. melanogaster* benchmark ($\approx 0.1\%$ of loci), suggesting perhaps that these differences may be the consequence of technical artifacts in one data set or the other.

The results of the other two benchmarks, for *G. max* and *H. sapiens*, were somewhat surprising. In each case, approximately 10% of the comparisons reflected perfect matches between the two annotations (6.4% for soybean and 15.3% for human), while approximately 50% of the comparisons reflected CDS matches (45.1% for soybean and 54.9% for human). Therefore, for the remaining approximate 30% of human genes and 50% of soybean genes, the annotated coding sequence (and the associated polypeptide) is different depending on the data source. These differences are likely the result of different annotation strategies between the alternative sources of annotation. Regardless, this is

an important point of consideration both for consumers and producers of gene structure annotations, and we hope that the ParsEval tool will be a useful asset to a wide variety of scientists that rely on reliable gene annotations for their research.

Performance evaluation in comparison to Eval software

To evaluate ParsEval's performance in comparison to existing methods, we used the Eval tool [23] to repeat one of the previously described use cases. Gene annotations for *Mus musculus* were retrieved from releases 64 and 65 of the Ensembl database, and subsequently analyzed using both Eval and ParsEval. Some small differences were observed in the similarity statistics computed by the two programs, although this was not unexpected as Eval uses a different approach than ParsEval for matching reference annotations to prediction annotations. Also, the two programs provide a different breakdown of the similarity statistics, making a rigorous comparison between the Eval results and the ParsEval results impractical.

Running Eval on the complete data sets exhausted the desktop computer's memory resources after several minutes, so comparison of Eval and ParsEval was only possible after restricting the data sets to annotations for *M. musculus* chromosomes 1 through 10. To analyze these reduced data sets, Eval required an average of 12 minutes 13 seconds and consumed all available memory. On the other hand, ParsEval, running on a single processor, required an average of 1 minute 44 seconds, with memory consumption peaking at approximately 0.5 GB. When run on 4 processors, ParsEval's performance margin increased with an average runtime of 47 seconds.

To ensure that Eval's performance was not being severely affected by the desktop's limited system memory, the comparison was also performed in a high-performance computing environment in which memory could not have been a limiting factor. ParsEval continued to demonstrate superior performance in this environment as well, although by a slightly less drastic margin. The Eval program required an average of 7 minutes 18

seconds of runtime, while ParsEval required an average of 1 minute 19 seconds using a single processor, or 37 seconds using 4 processors.

These tests conclusively demonstrate two important points regarding the performance of ParsEval relative to Eval: not only is ParsEval markedly faster, but its resource efficiency also makes it much better equipped to run whole-genome comparisons on the laptop or desktop computers one might expect to see in the typical biology lab. The initial runtimes reported herein should be fairly representative of what users can expect to observe when running ParsEval on commodity hardware.

Conclusions

The accessibility of genome annotation tools to an increasingly wider variety of scientists will soon be accompanied by an increased demand for supplementary tools to manage and analyze genome annotations. We address this need with ParsEval, a tool for fulfilling a common, fundamental analytical need for which existing software is lacking. ParsEval is a portable, easy-to-install, and efficient program for comparing gene structure annotations, and facilitates a wide variety of downstream comparative analyses. We demonstrate the speed and scalability of ParsEval, even when working with large eukaryotic genomes. Furthermore, we highlight the capability of the detailed comparison statistics in ParsEval reports to highlight relevant biological trends in the data. We anticipate that ParsEval will enable a wide variety of biologists to more fully take advantage of the vast genome annotation data resources accumulating in their individual labs and in the community at large.

Availability and requirements

Source code for ParsEval is available at http://parseval.sourceforge.net under an ISC license. ParsEval is implemented in ANSI C and is designed to run on all POSIX-

compliant UNIX systems (Linux, Mac OS X, Cygwin, Solaris, etc.). Aside from a C compiler with OpenMP support (such as GCC 4.2 or higher), ParsEval's only external dependency is the GenomeTools library [31].

Authors contributions

DS designed and implemented the software and drafted the manuscript. VB supervised the project and provided design and feature suggestions. Both authors conceived the project, edited the manuscript, and approved the final version.

Acknowledgements

The authors would like to thank the developers of the GenomeTools software for helpful feedback regarding integration of AnnotationSketch, and our colleagues Carolyn Lawrence and Amy Toth as well as anonymous reviewers whose suggestions were a valuable contribution to this manuscript.

Funding: This work was supported in part by the U.S.A. National Science Foundation Plant Genome Research Program grant ISO#1126267 to V.B..

CHAPTER 4. ILOCI: SCALABLE GENOME ANNOTATION FOR PROVISIONAL GENOME ASSEMBLIES

A paper to be submitted to *BMC Genomics*.

Standage DS, Brendel VP

Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus ac feugiat mauris. Nunc sed felis a purus finibus cursus in eu ligula. Nam cursus iaculis augue eget rutrum. Curabitur sed lorem posuere, ultricies nisl ac, dictum est. Praesent accumsan urna turpis, nec tristique nulla rutrum sollicitudin. Vivamus eu sapien id risus fringilla faucibus. Vestibulum euismod, nibh nec rutrum interdum, ex urna vehicula lorem, et fermentum tortor nunc at nisi. Curabitur urna metus, suscipit a ipsum ac, consectetur pharetra augue. Sed sit amet turpis vel risus vehicula dapibus. Duis mattis metus tellus, sit amet placerat lacus tincidunt ut. Nullam dictum lacus magna, in porttitor elit malesuada nec. Quisque quis massa luctus dui tincidunt hendrerit vel eget nisl.

Suspendisse et massa dolor. Cras cursus finibus enim in dapibus. Morbi aliquet placerat arcu, sed tristique ante pulvinar nec. Proin non metus non felis imperdiet tristique tristique vel augue. Cras posuere condimentum purus, vitae tempus tellus. Sed nibh velit, scelerisque vitae felis sit amet, dignissim sollicitudin tellus. Nam eget lacus

vitae dolor fermentum fermentum id id magna. Donec auctor euismod porta. Cras in ante scelerisque, placerat enim eu, dictum nisi. Integer nunc eros, elementum tempor arcu sed, tristique hendrerit leo.

As can be seen in Table 4.1 it is truly obvious what I am saying is true.

Table 4.1 This table shows a standard empty table

Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Cras tincidunt vehicula mi in ultrices. Proin vitae mauris aliquam, rutrum mi non, maximus libero. Cras sit amet metus sit amet nisi posuere eleifend. Nam et sapien odio. In ultrices elit nibh, sit amet commodo purus lobortis vitae. Quisque ac felis interdum, ornare nulla fringilla, posuere augue. Nullam dictum et arcu non ornare. In faucibus hendrerit nibh nec mollis. Nam eu dolor sodales mauris fermentum ornare ac ac ante. Etiam non odio sed odio faucibus luctus sed sed nulla. Aliquam sit amet est bibendum, lacinia velit eget, ornare mi. Nam eros neque, scelerisque quis cursus egestas, placerat eu nisl. Vivamus scelerisque odio at ipsum faucibus faucibus. Mauris consequat eu felis nec vulputate.

Fusce finibus erat nulla, eget vestibulum diam tristique ac. Fusce nisi diam, finibus vitae fermentum nec, placerat sodales tellus. Praesent et accumsan nunc. Pellentesque quam orci, rutrum quis ultricies quis, facilisis a ante. Curabitur felis ex, efficitur ut blandit eu, luctus quis enim. Aliquam ac lacinia massa. Quisque aliquam, quam at aliquam venenatis, magna ante auctor purus, nec pharetra turpis urna et diam. Duis eu

lectus eget risus ultrices lacinia. Quisque tincidunt purus ac nunc ornare, at pharetra erat rutrum. Sed massa sem, iaculis at vestibulum eget, accumsan eu nibh.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Sed lacus augue, euismod sed lacinia at, rutrum eget lectus. Donec egestas massa ac risus finibus mattis id eu velit. Mauris fermentum ligula vel tempor mattis. Etiam vehicula arcu a venenatis elementum. Ut cursus molestie ex eget auctor. Morbi eget risus a purus sodales semper. Quisque efficitur laoreet nunc, interdum volutpat orci. Sed quam ligula, dignissim sed dui vel, finibus ultricies elit. Praesent ipsum lectus, finibus sit amet mattis vitae, tempus non turpis. Nulla facilisi. Curabitur et dignissim nibh.

Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

This can also be seen in Figure 4.1 that the rest is obvious.

Figure 4.1 This table shows a standard empty figure

CHAPTER 5. GENEANNOLOGY: SCALABLE AND REPRODUCIBLE GENOME ANALYSIS WITH GENE ANNOTATION VERSION CONTROL

A manuscript to be submitted to *BMC Bioinformatics*.

Standage DS, Brendel VP

Introduction

Here initial concepts and conditions are explained and several hypothesis are mentioned in brief.

Of course, data on this as seen in Table 5.1 is few and far between.

Table 5.1 Moon Data

Element	Control	Experimental
Moon Rings	1.23	3.38
Moon Tides	2.26	3.12
Moon Walk	3.33	9.29

Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

Or graphically as seen in Figure 5.1 it is certain that my hypothesis is true.

Figure 5.1 Durham Centre

Parts of the hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

Second Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

Parts of the second hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

Criteria Review

Here certain criteria are explained thus eventually leading to a foregone conclusion.

CHAPTER 6. SUMMARY AND DISCUSSION

Introduction

Here initial concepts and conditions are explained and several hypothesis are mentioned in brief.

Or graphically as seen in Figure 6.1 it is certain that my hypothesis is true.

Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

As can be seen in Table 6.1 it is truly obvious what I am saying is true.

Parts of the hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

Second Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

Table 6.1 This table shows almost nothing but is a sideways table and takes up a whole page by itself

Element	**Control**	**Experimental**
Moon Rings	1.23	3.38
Moon Tides	2.26	3.12
Moon Walk	3.33	9.29

Parts of the second hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

Criteria Review

Here certain criteria are explained thus eventually leading to a foregone conclusion.

Figure 6.1 Durham Centre— Another View

APPENDIX A. ADDITIONAL MATERIAL

This is now the same as any other chapter except that all sectioning levels below the chapter level must begin with the *-form of a sectioning command.

More stuff

Supplemental material.

APPENDIX B. STATISTICAL RESULTS

This is now the same as any other chapter except that all sectioning levels below the chapter level must begin with the *-form of a sectioning command.

Supplemental Statistics

More stuff.

BIBLIOGRAPHY

[1] Earl D et al. (2011) Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research* 21(12):2224–2241.

[2] Bradnam KR et al. (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2(1):1–31.

[3] Salzberg SL et al. (2012) Gage: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* 22(3):557–567.

[4] Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) Quast: quality assessment tool for genome assemblies. *Bioinformatics* 29(8):1072–1075.

[5] Guig R et al. (2006) Egasp: the human encode genome annotation assessment project. *Genome biology* 7 Suppl 1:S2.131.

[6] Eilbeck K, Moore B, Holt C, Yandell M (2009) Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics* 10(1):67.

[7] Denton JF et al. (2014) Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol* 10(12):1–9.

[8] Warr A et al. (2015) Identification of low-confidence regions in the pig reference genome (sscrofa10.2). *Frontiers in Genetics* 6(338).

[9] MacManes MD (2014) On the optimal trimming of high-throughput mrna sequence data. *Frontiers in Genetics* 5(13).

[10] Williams CR, Baccarella A, Parrish JZ, Kim CC (2016) Trimming of sequence reads alters rna-seq gene expression estimates. *BMC Bioinformatics* 17(1):1–13.

[11] Jandt JM, Toth AL (2015) in *Genomics, Physiology and Behaviour of Social Insects*, Advances in Insect Physiology, eds. Zayed A, Kent CF. (Academic Press) Vol. 48, pp. 95 – 130.

[12] Chen X et al. (2012) Transcriptome comparison between honey bee queen- and worker-destined larvae. *Insect Biochemistry and Molecular Biology* 42(9):665 – 673.

[13] GROZINGER CM, FAN Y, HOOVER SER, WINSTON ML (2007) Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (apis mellifera). *Molecular Ecology* 16(22):4837–4848.

[14] Whitfield CW, Cziko AM, Robinson GE (2003) Gene expression profiles in the brain predict behavior in individual honey bees. *Science* 302(5643):296–299.

[15] Ometto L, Shoemaker D, Ross KG, Keller L (2011) Evolution of gene expression in fire ants: The effects of developmental stage, caste, and species. *Molecular Biology and Evolution* 28(4):1381–1392.

[16] Simola DF et al. (2013) Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Research*.

[17] Harrison MC, Hammond RL, Mallon EB (2015) Reproductive workers show queen-like gene expression in an intermediately eusocial insect, the buff-tailed bumble bee bombus terrestris. *Molecular Ecology* 24(12):3043–3063.

[18] Ferreira PG et al. (2013) Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. *Genome Biology* 14(2):1–15.

[19] Li-Byarlay H et al. (2013) Rna interference knockdown of dna methyl-transferase 3 affects gene alternative splicing in the honey bee. *Proceedings of the National Academy of Sciences* 110(31):12750–12755.

[20] Lyko F, Maleszka R (2011) Insects as innovative models for functional studies of dna methylation. *Trends in Genetics* 27(4):127–131.

[21] Sumner S (2014) The importance of genomic novelty in social evolution. *Molecular Ecology* 23(1):26–28.

[22] Johnson BR, Tsutsui ND (2011) Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. *BMC Genomics* 12(1):1–10.

[23] Keibler E, Brent M (2003) Eval: A software package for analysis of genome annotations. *BMC Bioinformatics* 4(1):50.

[24] Wang J, Kraemer E (2003) GFPE: gene-finding program evaluation. *Bioinformatics* 19(13):1712–1713.

[25] Li Y et al. (2013) Truesight: a new algorithm for splice junction detection using rna-seq. *Nucleic Acids Research* 41(4):e51.

[26] (year?) liftOver. http://genome.ucsc.edu/cgi-bin/hgLiftOver.

[27] Swain MT et al. (2012) A post-assembly genome-improvement toolkit (pagit) to obtain annotated genomes from contigs. *Nat. Protocols* 7(7):1260–1284.

[28] Burset M, Guigó R (1996) Evaluation of gene structure prediction programs. *Genomics* 34(3):353 – 367.

[29] Zhao XM, Wang Y, Chen L, Aihara K (2008) Gene function prediction using labeled and unlabeled data. *BMC Bioinformatics* 9(1):57.

[30] GFF3 Specification, The Sequence Ontology Project http://www.sequenceontology.org/gff3.shtml.

[31] GenomeTools library http://genometools.org.

[32] Steinbiss S, Gremme G, Schrfer C, Mader M, Kurtz S (2009) AnnotationSketch: a genome annotation drawing library. *Bioinformatics* 25(4):533–534.

[33] Bron C, Kerbosch J (1973) Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* 16:575–577.

[34] Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12(1):491.

[35] GBrowse: the generic genome browser http://gmod.org/wiki/GBrowse.

[36] Duvick J et al. (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Research* 36(suppl 1):D959–D965.

[37] Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19(Suppl 2):ii215–ii225.

[38] Guigó R (2000) An assessment of gene prediction accuracy in large dna sequences. *Genome Research* 10(10):1631–1642.

[39] EMBL nucleotide sequence database http://www.ebi.ac.uk/embl/.

[40] Genome Informatics Research Lab, Institut Municipal d'Investigació Mèdica http://genome.imim.es/datasets/gpeval2000/.

[41] Ensembl project http://ensembl.org.