Ontano *et al. BMC Genomics* (2026) 27:228
https://doi.org/10.1186/s12864-025-12512-x

BMC Genomics

RESEARCH Open Access

Independent centromeric expansions define giant hornet genomes

Andrew Ontano[1*], Sheina B. Sim[2], Jerry Jenkins[3], Scott M. Geib[2], Anna K. Childers[4], Brian Scheffler[5] and Justin Vaughn[1,3,6]

Abstract

Background The *Vespa* lineage of hornets demonstrate the potential to displace native species and cause significant damage to US apiculture through predation. In spite of introductions in recent years to North America, eradication efforts have prevented the Northern Giant Hornet, *Vespa mandarinia*, from establishing. Improved genomic resources could offer insight into the traits that define this lineage: large body size variance, adaptability to new environments, and high potential for invasiveness.

Results We sequenced and assembled genomes of two lineages of the northern giant hornet, *Vespa mandarinia*, and one of the European hornet, *Vespa crabro*, using HiFi long read sequencing technology. We found centromeric and pericentric satellite repeats account for nearly half the total DNA of the hornet genomes and their identities were largely unique across species, indicating active, independent expansion. The intraspecific northern giant hornet genomes exhibit asymmetrical expansion across homologous chromosomal regions localized with Hi-C scaffolding. We leveraged pangenomic alignments of the hornet genomes to identify derived mutations, particularly those that affect repeat content and transposable elements (TEs). We found that TEs do not contribute to the bulk repeat content and show no differential expansion in genic space.

Conclusions Large tandemly repetitive DNA account for large structural variations across the *Vespa* and between *V. mandarinia*. Localization within the genomes necessitated a suite of tools to support the assembly and identification of those elements. Typical repercussions of long-term reductions in population size – namely, reduced diversity and TE expansion – are not present. The degree to which alternative explanations, such as cell and body size selection or centromeric drive, cause massive, localized repeat amplification will require more extensive sampling.

Keywords Evolution, Centromere, Invasive species, Repeat expansion, Repetitive elements, Pangenome

*Correspondence:
Andrew Ontano
andrew.ontano@uga.edu
[1]Department of Crop and Soil Sciences, University of Georgia, Athens, GA 30602, USA
[2]Agricultural Research Service, Tropical Pest Genetics and Molecular Biology Research Unit, USDA, Daniel K. Inouye U.S. Pacific Basin Agricultural Research Center, 64 Nowelo Street, Hilo, HI 96720, USA
[3]Hudson Alpha Institute for Biotechnology, Huntsville, AL, USA
[4]Agricultural Research Service, Bee Research Laboratory, Beltsville Agricultural Research Center, USDA, Beltsville, MD 20705, USA
[5]Jamie Whitten Delta States Research Center, Genetics and Bioinformatics Research Unit, USDA, Agricultural Research Service, 141 Experiment Station Rd, Stoneville, MS 38776, USA
[6]Former affiliation: Jamie Whitten Delta States Research Center, Genetics and Bioinformatics Research Unit, USDA, Agricultural Research Service, 141 Experiment Station Rd, Stoneville MS 38776, USA

Background

Hornets of the genus *Vespa* are large, eusocial wasps with extant ranges across Europe and Asia. The 22 species of *Vespa* hornets are highly predacious on other insects, including species of agricultural importance like the European Honeybees, *Apis mellifera*, and the Eastern honeybee, *Apis cerana*. Several species have been introduced to North America including the European hornet, *Vespa crabro*, which first arrived then became established in the eastern USA in the 1800s and more recently was collected in Guatemala in 2010 [1], the northern giant hornet, *Vespa mandarinia*, which was detected in British Columbia, Canada and Washington state, USA and British Columbia of Canada [2], and most recently the yellow-legged hornet, *Vespa velutina*, was detected in the eastern USA [3]. Prior phylogenetic analyses on the mitogenomes of *V. mandarinia* revealed support for two distinct introduction events into North America with origins likely from Japan and South Korea [4]. *V. mandarinia* is the largest and most aggressive of the *Vespa* hornets and attacks honeybee hives in groups to kill workers and bring larvae back to the nest as a food source [5]. The introduction of *V. mandarinia* proffers unchecked predation that the North American honeybee breeding lines are not equipped to defend against. Precautions to detect and exterminate individuals and nests have been taken to prevent establishment [2].

Competition between these introduced populations and native hornets could cause the displacement of native species as well as negatively impacting their prey food sources, thus leading these populations to become invasive to North America, but efforts thus far have prevented this outcome. *Vespa* hornets are an excellent evolutionary model to explore the interface between body size, genome size, and invasiveness. Multiple lineages have undergone expansions in their habitable range and these expansions have been facilitated by anthropocentric factors such as urbanization [6]. Various species of hornets that have been introduced or expanded their natural ranges are in direct competition with other *Vespa* hornets as they overlap in nesting habits and structure as well as competition for their prey food sources. This displacement imposes a threat to agriculturally relevant prey like the European and Asian honeybees that are impacted more heavily by these hornets, but also offers the model for comparison between their life histories and genetics.

Repetitive DNA comprises large portions of the genomes of Eukaryota. Commonly repeats are classified into several groups, including DNA transposons that act as "selfish" elements able to proliferate and relocate across the genome with both positive and negative repercussions to the organisms [7]. In contrast, tandemly repetitive DNA are short motifs that expand on end for hundreds to millions of times to form a repeat satellite. Motifs can differ greatly in complexity where some are comprised of smaller 3 bp couplets with high similarity to one another, whereas other motifs may span hundreds to thousands of basepairs with less internal repetition. These regions impose a challenge during genome sequencing where the size of satellites can be masked due to relatively low complexity and high sequence identity between the underlying sequencing reads [8]. A common architecture of the centromeres among animals are satellites stretching upwards of millions of base pairs, but resolution of these regions may suffer from sequencing limitations.

Few genomic resources in the *Vespa* clade existed until quite recently. The sequencing of several notable genomes of *Vespa* hornets, including one offered from this study for early release, has provided insight into gene content and differential gene regulation across the caste system [9]. In other insects, employing long-read sequencing technology has helped to elucidate the under-studied regions of insect genomes and improve their contiguity towards a telomere-to-telomere scale [10]. Pangenomic graphs have emerged as a powerful tool for assessing variation across closely related organismal genomes [11]. Comparison of structural variation between multiple genomes can be used to facilitate investigation of population dynamics, assessing heterozygosity among founder populations, or finding genomic regions that are under greater selection than others. These tools set a framework for comparative analyses to identify and localize variation between these hornets as more genomic data is generated.

The centromeres of insects remain a challenging region of study due in part to variations such as multiple lineage-specific shifts to holocentricity [12]. Some centromeric forms also appear to act independently of the histone H3 centromere-specific protein CENP-A [13]. Improved sequencing and assembly methodologies offer greater resolution of the tandemly repetitive DNA elements that surround centromeres. Centromere size broadly tends to correlate with genome size among Eukaryota [14]. Thus, improving measurement of one will improve estimate of the other and vice versa. Broader and deeper sampling can also provide phylogenetic information to elucidate the origin of these centromeric innovations [15]. Comparing the variations across *Vespa* hornets requires integrating all these factors to understand their evolutionary histories.

In this study, we dissect the underlying mutations causing genomic divergence between unrelated *Vespa mandarinia* introductions to North America as well as comparisons between the two largest-bodied *Vespa* hornets in North America. All chromosomes express variability in repeat expansion, although some reflect extreme polymorphisms. We consider such divergent loci critical for future study of the interplay between

centromeric structure and genome size in evolution. Our study also provides additional genomic resources for three wasp species, including a pangenomic graph of two major North American hornet introductions.

Results

Extreme repeat content differs across the *Vespa*

We aggregated and aligned highly conserved proteins from the chromosome-scale genomes of *Vespa* hornets and close relatives from the genus *Vespula*. Using these alignments, maximum-likelihood phylogenetic analyses of the hornets and their relatives show minimal intraspecific differences among the genomes of *Vespa crabro* and the genomes of *Vespa mandarinia* respectively (Fig. 1a). These hornets form a clade together with *Vespa velutina* placed as sister to this clade. As evidenced by self-by-self dotplots and corroborated by randomly selected raw reads, the branches consisting of *V. crabro* and *V. mandarinia* exhibits extensive enrichment of very simple tandem repeats, which consisted of 6–12 bp motifs upon investigation (Fig. 1b).

Satellite repeats account for nearly half of total DNA in *Vespa*

To understand the nature of the repeat enrichment observed above, k-mer profiles were generated for all four assemblies and raw HiFi read data. The 1,000 most abundant canonical k-mers (K = 21) from each assembly overlapped across assemblies but the extent of overlap ranged by relatedness. For example, the female Canadian *Vespa mandarinia* assembly shared 879 of the top 1,000 k-mers with the male Washington state *Vespa mandarinia* assembly, 343 with the *Vespa crabro* assembly, and 115 with the *Vespula pensylvatica* assembly. Simple repetitive motifs were the most dominant k-mers among each assembly. Three k-mers represented shifts in an (AAT)n motif. Top k-mers were taken from each of the four libraries and paired with matching k-mers from the corresponding scaffolded assemblies and were normalized by the single-copy sequencing coverage estimated from k-mer profile (Fig. 2). K-mers in the *Vespa* assemblies have a bimodal representation centered around 4k-5k copies. *Vespula pensylvatica* copy number was far lower, at ~ 3.2k, and the bimodal distribution was absent.

We defined fold-enrichment, relative the single copy coverage, of the top 1,000 canonical k-mers counted in the raw sequencing reads (Fig. 2). This distribution shifts substantially in the scaffolded genome. The elements underlying these k-mers were collapsed during the assembly process at different rates across species. The lower-copy peak of the bimodal distributions among the scaffolded *Vespa* genomes are more disparate than their raw counterparts. These elements of the raw HiFi data are dissimilar enough to form unique contigs, but high

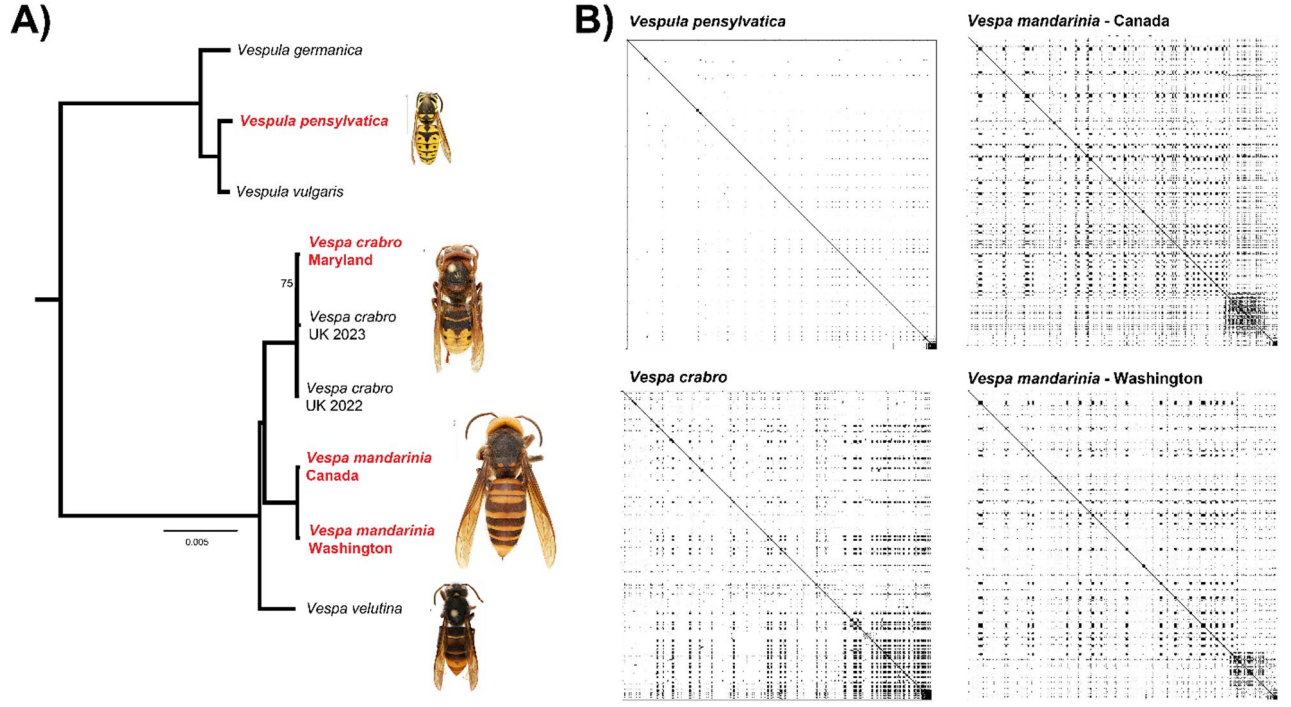

Fig. 1 Phylogeny of hornets and the repetitive DNA inherent to hornets. **A** Maximum likelihood tree of genomic resources of *Vespa* and *Vespula* assessed from a concatenated dataset of 500 BUSCOs from the hymenoptera_odb10 dataset. Nodes had 100% support unless specified with a support level. Genomes assembled for this study are in bold and have a representative photo. The specimen scale bars represent 1 cm. **B** Self-by-self dotplots of random samples of raw HiFi reads using a word length of 30 base pairs for the sequenced data

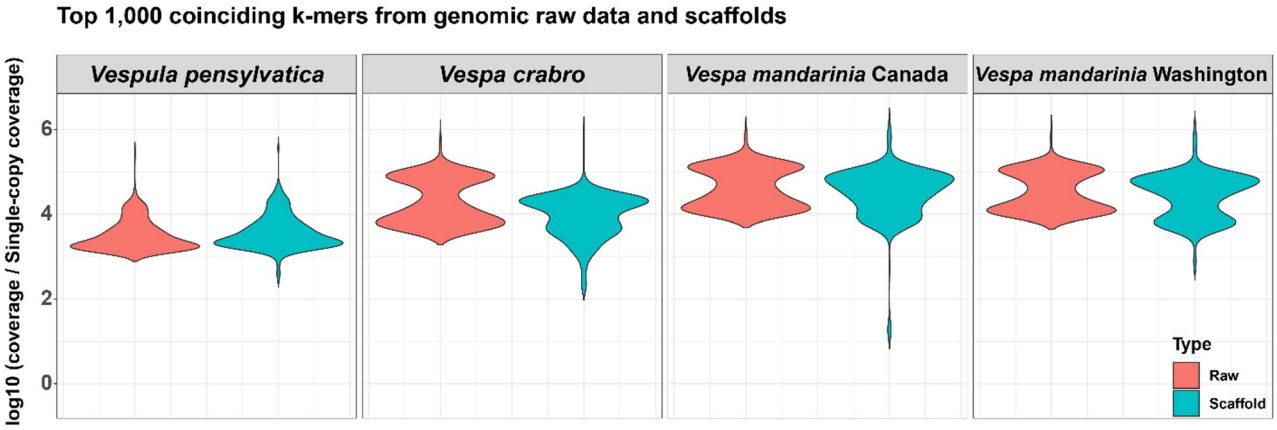

Fig. 2 Comparison of k-mers in raw data and scaffolded assemblies. Violin plots showing the abundance of k-mers (k = 21). The top 1,000 most prevalent canonical k-mers from each dataset of raw HiFi reads were counted in both raw reads (red) and assembled genomes (blue). K-mer counts for the raw HiFi reads were adjusted by the haploid coverage peak of each dataset

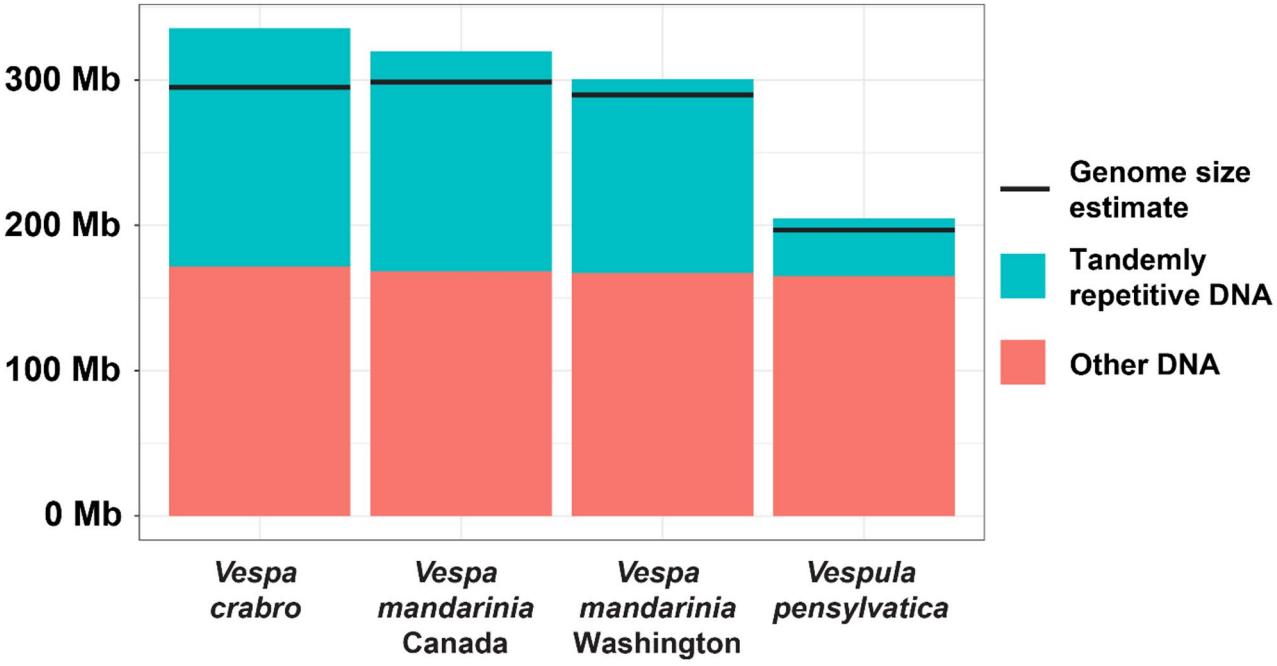

Fig. 3 Comparison between non-repetitive DNA and tandemly repetitive DNA. Quantification of assembled genomes by the total base pairs reported as tandemly repetitive DNA (blue) and base pairs not reported as tandemly repetitive (red). Black bars represent the estimated genome size based on k-mer analyses

identity prevents scaffolding and causes some underrepresentation in the final assembly. Repeat absence appears to only affect a small set of chromosomes (see below).

Repeat enriched regions of the assembled genome sequence were identified with a suite of software for the identification and localization of repetitive elements. Low complexity repetitive DNA, or DNA with a high similarity and a repeated motif shorter than 400 base pairs in length, accounted for 44.4–48.9% of the total genome size of scaffolded *Vespa* assemblies and 19.5% of the *Vespula* assembly (Fig. 3). K-mer based genome size estimates showed a minimal excess of repetitive DNA in the assemblies with *Vespa crabro* having the most over-enriched genome. The tandemly repetitive DNA was predominantly comprised of the centromeres. The core centromeric satellite of the *Vespa mandarinia* genomes was identified as a 341 bp motif spanning up to 4 million base pairs with an AT-rich consensus sequence and 87–97% similarity along the satellites. The core centromeric satellite of *Vespa crabro* had a similar architecture, but the motif undergoes periodic shifts from 341 base pairs to motifs in a range of consensus sequence lengths from 337 to 340 base pairs spanning up to 2.5 million base pairs. The core centromeric satellite of *Vespula*

pensylvatica was comprised of a 166 bp motif spanning up to 2.5 million base pairs that ranged between 92 and 98% similarity along the satellites. Immediately flanking the core centromeric satellites across all *Vespa* genomes were long stretches of short, degenerate AT-rich tandem repeats. The predominant motifs were 6 or 9 base pairs in length and incorporated derivations on a core (AAT)n motif spanning up to 2.5 million base pairs which differed in complexity and consensus sequences across species, chromosomes, chromosomal arms, and domains along the pericentric regions. A satellite repeat spanning 3.2 million base pairs with a consensus pattern deriving from a (AATGATGAT)n motif was identified on homologous chromosomes of both *Vespa mandarinia* genomes that was adjacent to the pericentric region but separated by a region of interspersed repetitive elements.

Tandemly repetitive DNA elements were the predominant quantity of repetitive elements in the genomes over interspersed repeats and RNA elements, which accounted for 6.3–7.7% of the *Vespa* genomes and 9.1% of the *Vespula* genome. Identified elements were split evenly between DNA transposons and retroelements (Supplementary Tables S1 & S2). Long interspersed nuclear elements (LINE) of the Jockey family were found across all four genomes but were twice as prevalent in the *Vespa* genomes over *Vespula pensylvatica*. These elements were localized predominantly in unplaced contigs and pericentric regions of genomic scaffolds. DNA transposons from various families were found throughout genic space of each genome. Adjacent to the pericentric satellite repeats were regions of interspersed LINE type Cin4, PiggyBac DNA transposons, long terminal repeat (LTR) type Gypsy, and many other less abundant elements. The centromeric landscape of *Vespa* hornets is dominated by pervasive repetitive elements, and regions dense with tandemly repetitive DNA elements had many interspersed transposable elements. Expansions in the centromeric and pericentric regions were not detected in *V. pensylvatica*, but interspersed repetitive elements accounted for more DNA in non-centromeric regions than in the *Vespa*. Most of the TE activity may be ancestral to these *Vespa* lineages allowing for the decay of the element's sequences, but this activity is largely independent of TE activity in *Vespula*.

Anchored repeat expansions is at root of intraspecific variance

High-fidelity long-reads enabled the localization of many of *Vespa*'s high repetitively tracks, allowing some centromeric borders to be placed with their respective chromosomes and compared at orthologous sites. While the genomes of *Vespa mandarinia* are comparable in size, the localization and sizes of tandemly repetitive elements are often different (Fig. 4a). Homologous chromosomes from the female Canadian and male Washington assemblies have similarly high contiguity through genic space but variable contiguity through the centromeric regions. The identities of repetitive elements are similar, and pericentric 9 bp motifs were localized to homologous chromosomes. Twenty-three centromeric satellite repeats were localized to chromosomes on the female Canadian assembly whereas nineteen were localized to chromosomes of the male Washington state assembly. The differences in size of centromeric satellites, where localized, ranged from 2.5 million base pairs larger in the Canadian assembly to 2.8 million base pairs larger in the Washington state assembly. On average the centromeric satellites of the Canadian assembly were larger by 252,000 base pairs, and accounted for ~1.9% of the total genome(Fig. 4b).

Given the difficulty of assembling large repetitive regions comprised of short repetitive patterns, some centromeric elements could have been misplaced during scaffolding; however, pairwise differences when satellite elements are simply ordered by size suggest a high degree of plasticity in centromeric sizes between the two assemblies. Pericentric tandemly repetitive DNA shows similar plasticity in sizes and the elements in the Washington state assembly were 120,000 base pairs longer on average than the Canadian assembly across chromosomes with all comparable elements present. Contiguity through these regions reveals abrupt transitions along the chromosomes where the core centromeric satellite is immediately adjacent to a 9 bp pericentric satellite. HiFi long read alignments to the genome corroborate with contiguous assembly of the large repetitive regions. Similar transitions can be found in the assembly of *Vespa crabro* where core centromeric satellites contained abrupt transition into pericentric satellites.

Diversity in haplotypes supports independent introductions of *Vespa mandarinia*

Given the threat of *Vespa* species to honeybee and human health, significant efforts have been made to detect and eradicate any introduced hornets. Thus, population genetic data on these introduction events is quite rare. The female sample of one of the *Vespa mandarinia* in this study allowed us to contrast haplotype diversity with intra-specific diversity. To fully exploit this data, we created a chromosome scale alignment as a pangenomic graph. The pangenome graph produced for the *Vespa* hornets and *Vespula pensylvatica* represents six genomes across four species. Biallelic variants, where nucleotide polymorphic alleles were called for each genome, accounted for 7,554,462 variants. 527,349 variants were uniquely homozygous in *V. mandarinia* with respect to all others, and 651,452 variants among only the *Vespa*. A similar number of variants were uniquely

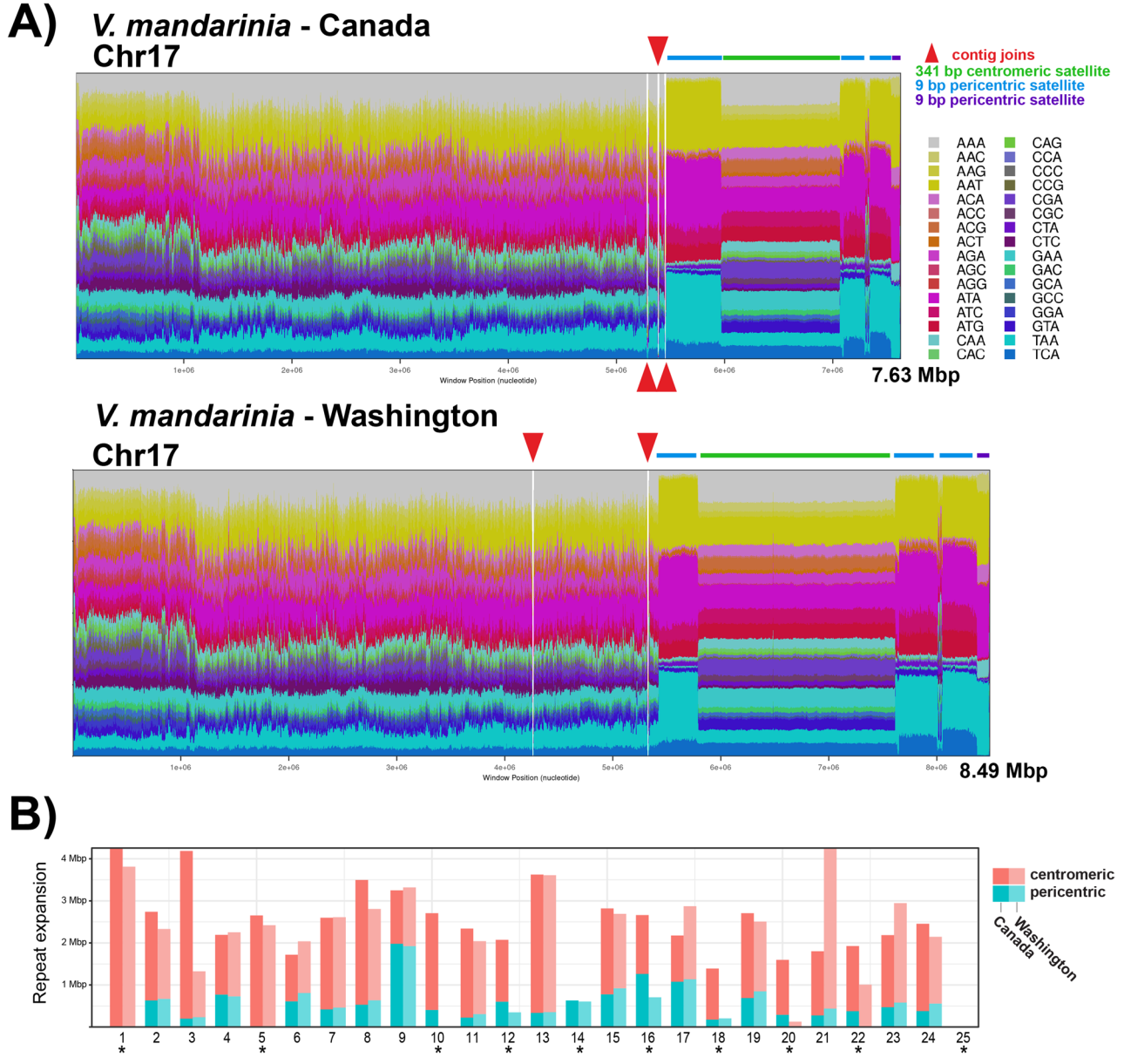

Fig. 4 Analysis of anchored centromeres in *Vespa mandarinia*. **A** Distribution of each canonical k-mer (K = 3) along 3,000 bp windows across the homologous chromosomes Chr17 in the Canada assembly and the Washington assembly using Spectra. Regions with consistent k-mer ratios correlate with large tandem repeats of varying motifs. Arrowheads indicate location of contig joins in the scaffolding. Colored bars above indicate comparable repetitive elements. **B** Size of centromeric (red) and pericentric (blue) repetitive elements for the 25 homologous chromosomes. Left bars indicate the Canadian assembly, right bars indicate the Washington state assembly. Asterisks indicate where any elements were not clearly identified in the genomes

homozygous among *Vespa crabro* with 542,054 variants with respect to all others and 667,506 among only *Vespa*. Variation within species accounted for a total of 40,864 heterozygous variants in *V. mandarinia* and 21,894 in *V. crabro*. The genomes of *Vespa crabro* represent two distinct populations from a native European source and an introduced North American source. Prior analysis of the mitochondrial genomes of *Vespa mandarinia* supported two distinct maternal lineages that were introduced to British Columbia, Canada from Japan and introduced to Washington, USA from South Korea [4]. A population genomic study of individuals from the native ranges as well as individuals from both introductions validated these origins, but found a high degree of genetic diversity of native South Korean hornets and introduced Washington hornets [16]. Variation between the *V. mandarinia* samples was higher than the variation between the *V. crabro* samples (Fig. 5). The level of variation between the haplotypes of the Canadian *V. mandarinia* specimen represents a high baseline of heterozygosity that is

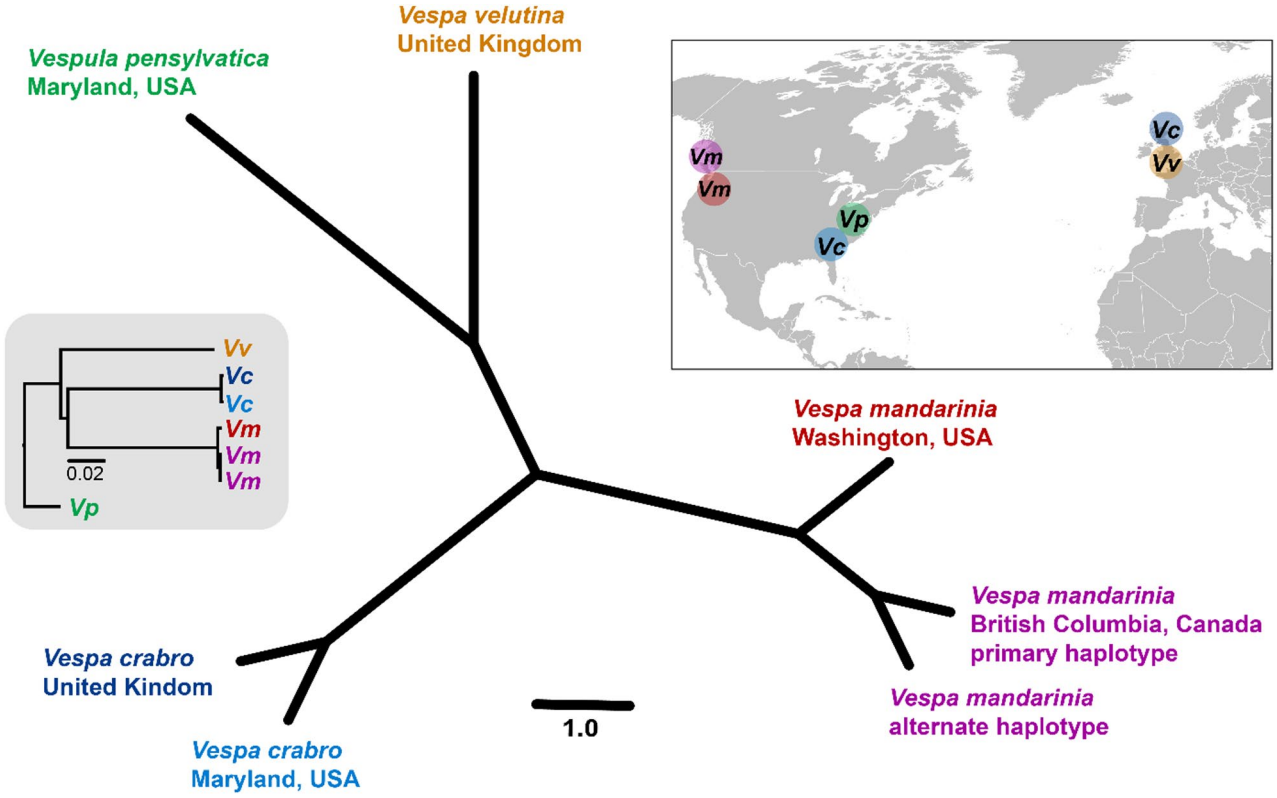

Fig. 5 *Vespa mandarinia* has more variation than *Vespa crabro*. Unrooted neighbor-joining tree assessed from biallelic nucleotide polymorphisms. Confidence intervals were 100% support unless deonoted. Branch lengths were normalized on a $\log_{10}$ scale for visibility, and non-normalized branch lengths are as shown in inset rooted tree. Genomes are color labeled corresponding to their collection locality on the map

independent of the specimen from Washington state (Fig. 5).

Insertion events relative to *Vespula pensylvatica* accounted for approximately 800,000 base pairs of each *Vespa* assembly's size. Repetitive DNA accounted for 33.3–34.6% of mutations in the *Vespa* genomes (Supplementary Tables S3 & S4). The majority of repeats were comprised of interspersed repetitive DNA such as transposons and retroelements, accounting for 187,859 – 188,143 base pairs in *Vespa mandarinia*, 182,017–183,100 base pairs in *Vespa crabro*, and 190,668 in *Vespa velutina*. The numbers of specific elements were consistent within species and tended to differ by no more than one to four, and inserted elements never differed by more than 500 base pairs in total occupied length.

Selection generally results in aberrant diversity profiles at particular genomic loci. We investigated the two most dynamically shifted loci in our results – the centromeric regions of Chr 3 and Chr 21 of the Washington *Vespa mandarinia* (Fig. 6; identified in Fig. 4b). When compared with genome-wide distributions on variant diversity within *V. mandarinia*, these two regions both fell well below the genome-wide average of our sliding-window diversity index (Fig. 6). After normalizing by extant *V. mandarinia* diversity, heterozygosity in the female is marginally lower for these loci (Supplementary Figure S1).

Discussion

We have presented herein extensive analyses of repetitive DNA in the genomes of *Vespa* hornets. Long read sequencing technology offers greater insight into the assembly of challenging genomes. While full telomere-to-telomere resolution was not yet possible, long read sequence data allowed for repetitive DNA to be incorporated with high contiguity (Fig. 4a). Characterizing the accumulation and the localization of various centromeric and pericentric repetitive elements is facilitated by a multi-technology sequencing approach utilizing both HiFi and Hi-C data as well as identifying repeat homology through Spectra (Fig. 4b). Large stretches of tandemly repetitive DNA accounted for large portions of HiFi long read sequencing data and nearly half of the scaffolded genome assemblies in contrast to limited repetitive DNA in *Vespula pensylvatica* (Fig. 3). Intraspecific radiations among the *Vespa* have permitted the accumulation of repetitive DNA in similar in structure but with unique underlying repeat motifs.

Small effective population size (N_e) during species introduction can create a permissive environment for

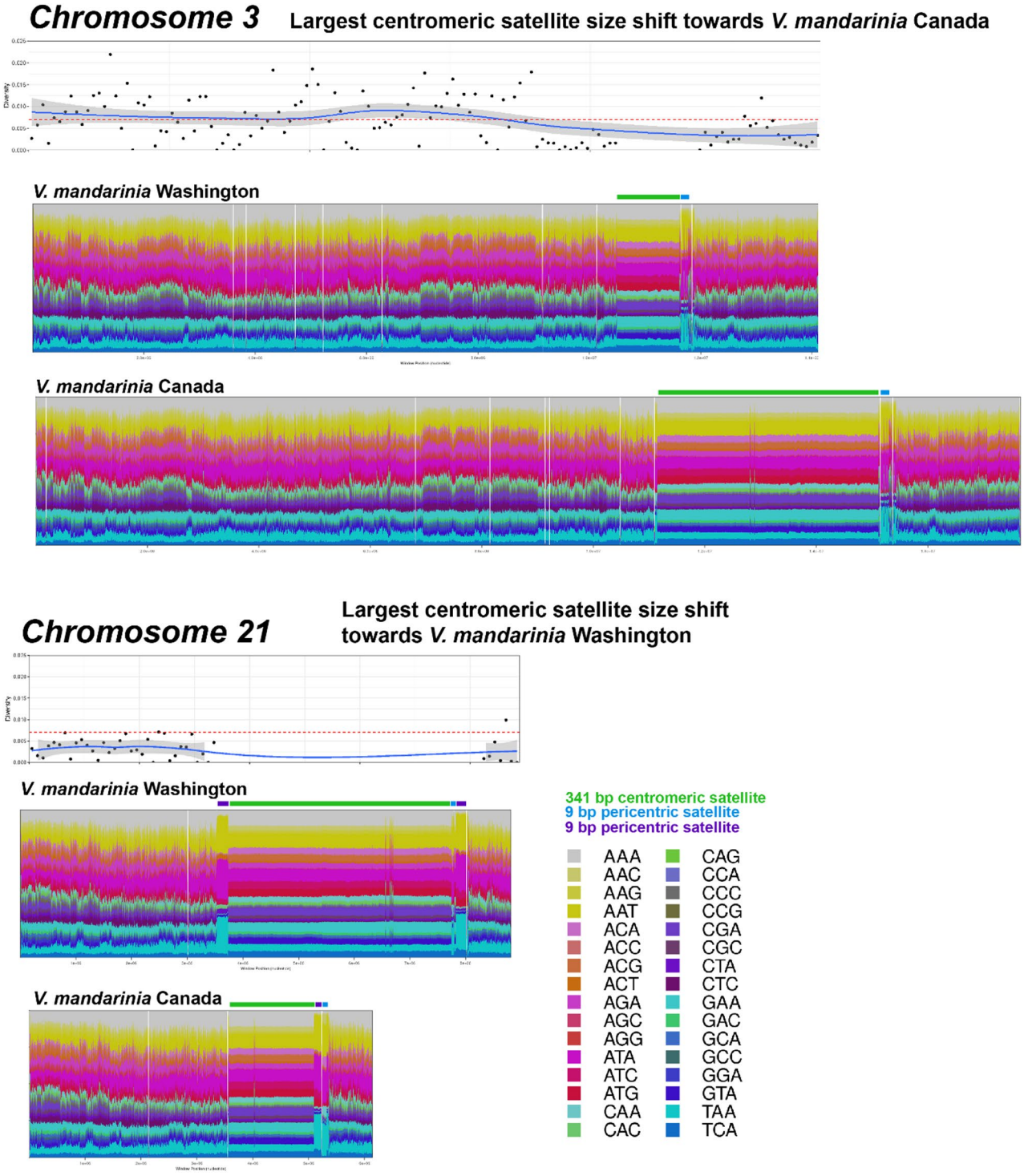

Fig. 6 Diversity of *Vespa mandarinia* alleles along chromosomes 3 and 21. Dots represent the number of variants segregating in *V. mandarinia* divided by all *Vespa* variants, computed in 100,000 bp windows; blue line and shaded region denotes LOESS regression; red dashed line denotes genome-wide average variation across windows (top). Distribution of each canonical k-mer (K = 3) along 3,000 bp windows. Regions with consistent k-mer ratios correlate with large tandem repeats of varying motifs. Colored bars above indicate repeats with high similarity in motif patterns(bottom)

"selfish" or "junk" DNA to expand in the genome, which may lead to the activity of transposons across the genome [17]. The activity of DNA transposons in the *Vepsa* genomes did not show evidence of transposon expansion and heightened activity (Supplementary Tables S3 & S4), indicating that long-term reductions in effective population size are not driving genome expansion. Supporting this conclusion, the heterozygosity we observe in the female sample is a sizable fraction – 45% - of the diversity observed within the species, and the species diversity is also substantial relative to *Vespa* as a whole (Fig. 5). A broader study of the hornet genomes would help elucidate how simple-repeat dynamics relate to genotypic diversity.

With regard to "selfish" DNA, selection at multiple levels of biological organization may favor the same mutation type. Microtubule asymmetry during female egg development can select for expansion of centromere satellite DNA due to competition for these microtubules [18]. To our knowledge, such asymmetry has not been confirmed in wasps, but variation in the sizes of centromeric satellites among the *Vespa* species (Fig. 4b) could be attributed to "centromere drive" [19]. Rapid changes in the identity and size of centromeric satellites may be exacerbated by the haplo-diploid reproductive cycle of hymenopterans [20]. Though evolution at the intragenomic level may be present, this does not preclude selection from exploiting the mutations at an organismal level as well.

Nucleotypes, i.e. bulk DNA, could be acting as a driver for phenotypic variation [21]. Though historically treated as an aggregate feature of the genome, nucleotypes can manifest as individual genomic loci when genotypic variants are so large that they represent phenotypically relevant contributions to the total DNA content. Long-read assemblies are now a viable method to accurately characterize such large events. Larger cell size via an increase in bulk DNA is one mechanism to rapidly increase tissue and body size [22], although tissue and organism developmental control can substantially decouple these features [23]. Tentative results from our sparse but high-quality genetic diversity data suggest that diversity has been reduced around the most dynamically divergent centromeric loci (Fig. 6). Barring the coincidence of advantageous genotypic mutations in linkage with these loci, it appears they have been under selection and yet remain dynamic in the population (Fig. 4b). It is likely that phenotypically related genotypic mutations - in cell-cycle genes, for example - will be far rarer and potentially far too extreme to improve fitness. Because selection takes advantage of the optimal variant at hand, expanded repeat loci could play a substantial role in *Vespa* evolution. With the increasing ease of generating long-read sequence data, further work could enhance the identification of specific genomic regions that are operating in a nucleotypic capacity.

The haplo-diploid genetics of male/female determination is also significant to the divergent centromeric analysis of *Vespa*. Because males and females are comparable in size in spite of 1x and 2x copies of the genome, nucleotypic theory predicts that females will have fewer, larger cells or males will compensate with endoduplication methods. Indeed, in other hymenoptera, endoduplication has been observed to be more intensive in males as a mechanism to retain comparable body size [24]. Ultimately this study highlights the need for a wider breadth of sampling paired with cytological and body size information which would aid in uncovering the potential link between nucleotypically driven lineages and the diversity in body sizes of the hornets.

Conclusion

The genomic architecture of Hymenoptera varies greatly among their many lineages, and repetitive DNA comprises a large portion of that variability. However, repetitive DNA impedes sequencing and telomere-to-telomere assembly of chromosomes. We have shown that large stretches of tandemly repetitive DNA can be sequenced and assembled with the aid of long-read technology. Such resolution allows us to begin addressing very challenging questions about the relationship between genome structure and evolution at multiple levels of biological organization.

Methods

Specimen collection

A *Vespa mandarinia* specimen was turned in to local authorities on Vancouver Island, in British Columbia, in May of 2019, and an additional hornet was found in August 2019 [4]. In September 2019, local beekeepers in Nanaimo, BC, tracked down and eradicated the colony, located at 49.152809 N, 123.943125 W [4]. All insects were transferred into 70% ethanol. Once dead, the insects were removed from the alcohol and stored at -20˚C until they could be transferred to the University of British Columbia, where they were subsequently stored at -80˚C. The female *V. mandarinia* specimen utilized for sequencing herein was stored at -20˚C for approximately 2 weeks.

Washington State Department of Agriculture verified multiple reports of single *V. mandarinia* hornets in northwest Washington state starting in the Fall of 2019 [25]. On October 22, 2020, the first US nest was located in Blaine, Washington (30 miles south from Vancouver). The nest was eradicated on October 24. The male specimen utilized for sequencing herein was collected alive, flash frozen with liquid N and stored at -80˚C.

Male *Vespula pensylvatica* (western yellowjacket) and *Vespa crabro* specimens were live collected in a USDA,

ARS Bee Research Laboratory honey bee apiary in Beltsville, MD on October 24, 2019 and November 8, 2019, respectively. The specimens were flash frozen with liquid N and stored at -80˚C.

DNA extraction, library preparation, and sequencing

High molecular weight (HMW) DNA extraction was performed using the Qiagen MagAttract HMW DNA Kit (Qiagen, Hilden Germany) following the protocol for fresh or frozen tissue. To improve the purity of the extracted DNA prior to shearing, a 1.75x bead clean-up was performed by adding 175 µl of room temperature polyethylene glycol containing solid-phase reversible immobilization beads solution [26] to each sample. DNA was sheared to a mean size of 20 kb using the Diagenode Megaruptor 2 (Denville, New Jersey, USA). SMRTbell libraries were prepared using a SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, California, USA). The prepared libraries were sequenced at the USDA-ARS Tropical Pest Genetics and Molecular Biology Research Unit in Hilo, Hawaii, USA on a Sequel IIe system (Pacific Biosciences, Menlo Park, California, USA). One 8 M SMRT cell was used for each of the libraries generated for *Vespa crabro* and *Vespula pensylvatica*, two cells were used for the Canadian female *Vespa mandarinia*, and 3 cells were used for the Washington male *V. mandarinia*. Sequencing began with a 2-hour pre-extension followed by a 30-hour movie collection time. Raw subreads were converted to HiFi data by processing with Circular Consensus Sequencing (CCS) to call a single high quality consensus sequence for each molecule, using a 99.5% consensus accuracy cutoff.

The mitochondrial genome of *Vespa mandarinia* was assembled utilizing the data from the female Canadian specimen following a general strategy described previously [27]. Mitochondrial reads were selected by alignment of the full HiFi read dataset against the *Vespa mandarinia* mitogenome reference (NCBI accession NC_027172) using minimap2 [28]. IPA was run on the mitochondrial read set, followed by manual end overlap trimming and start-point rotation (to the NCBI reference start point) of the resulting mitochondrial genome contigs. Multiple contigs were obtained due to the presence of an extended variable number tandem repeat (VNTR) region corresponding to the control region (this mitogenome portion is absent in the NCBI reference), with different copy numbers (ranging from 5 to 9) of an 823 bp repeat unit. Remapping of the mitochondrial read set provided estimates of the relative abundance of the mitogenome VNTR variants in the hornet individual (Supplementary Table S5). The most abundant mitogenome variant (6 repeat copies) was designated as the mitogenome sequence, and is deposited with the Canadian *Vespa mandarinia* under project PRJNA1088093. Mitochondrial sequences were identified and removed prior to assembly of the nuclear genome.

Contig assembly

Resulting circular consensus sequence (CCS) reads were processed by filtering for sequences containing adapter contamination from the raw reads using the HiFiAdapterFilt pipeline [29]. A *de novo* genome assembly was performed from the filtered CCS reads using HiFiASM v 0.14.2 [30] for each organism sampled using default parameters, resulting in a primary assembly as well as an alternate assembly containing alternative haplotypes.

Genome scaffolding

Hi-C Illumina reads from the Washington male and Canadian female *Vespa mandarinia*, were separately aligned to the Canada and Washington contig sets using Juicer [31], and chromosome scale scaffolding was performed with 3D-DNA [32]. No misjoins were identified in either assembly. The contigs were then oriented, ordered, and joined together into 25 chromosomes each for each respective genome using JUICEBOX [33]. A total of 104 joins were applied to the Canadian female assembly, and 117 joins for the Washington state male assembly. Each chromosome join is padded with 10,000 Ns. Contigs terminating in significant telomeric sequence were identified using the (CCCTGACGCAA)n repeat, and care was taken to make sure that they were properly oriented in the production assembly. The final 25 chromosomes for the Canadian female assembly contained 261.0 Mb of sequence, consisting of 129 joined contigs with a contig N50 of 3.4 Mb. 177 contigs representing 59.2 Mb remained unscaffolded. The final 25 chromosomes for the Washington state male assembly contained 252.2 Mb of sequence, consisting of 142 contigs with a contig N50 of 3.0 Mb. 152 contigs representing 48.6 Mb remained unscaffolded.

HiFiASM contigs from *Vespa crabro* were scaffolded against the genomic reference iyVesCrab1.2 and HiFiASM contigs from *Vespula pensylvatica* were separately scaffolded against the reference ASM1446617v1 using RagTag v2.1 [34]. Contig sets were first corrected for assembly errors that resulted in visually identified chimeric contigs using "ragtag.py correct -b 100000" and the resulting contig sets were then scaffolded by the reference sequences using "ragtag.py scaffold -r". A total of 781 joins were applied to reference-corrected contigs of *Vespa crabro* and 999 joins were applied to the reference-corrected contigs of *Vespula pensylvatica*. The final *Vespa crabro* chromosomes contained 237.4 Mb of sequence, consisting of 806 corrected contigs with a contig N50 of 0.6 Mb. The final *Vespula pensylvatica* chromosomes contained 195.9 Mb of sequence, consisting of 1,024 corrected contigs with a contig N50 of 0.4 Mb (Table 1).

Table 1 Genome assembly information

Genome	Total Length (Mbp)	Scaffolds	Scaffold N50 (Mbp)	Busco % Complete (Single / Duplicated) / Fragmented / Missing
Vespa crabro	335.58	1,168	8.91	97.2 (96.9 / 0.3) / 0.8 / 2.0%
Vespa mandarinia Canada	320.25	202	11.42	97.1 (96.5 / 0.6) / 0.9 / 2.0%
Vespa mandarinia Washington	300.76	177	9.29	96.7 (96.4 / 0.3) / 0.9 / 2.4%
Vespula pensylvatica	204.90	178	8.75	97.3 (97.0 / 0.3) / 0.8 / 1.9%

K-mer profiling

K-mer analysis was completed for each set of HiFi CCS reads using Jellyfish v2 [35] utilizing a range of k-mer sizes (K = 17–27) and counts of the canonical k-mers only. Genome sizes were estimated based on the profile of single copy k-mer coverage. Canonical 21-mers were also counted for each scaffolded genome. The 1,000 most prevalent k-mers from each set of HiFi raw reads were normalized against the heterozygous peak of k-mer coverage for that set of reads, which was compared against the counts of matching k-mers in the scaffolded assemblies.

Genome assessment

Completeness of all four assemblies was determined with BUSCO v5.2.2 [36] using the Hymenopteran-specific orthologous gene database v10 (hymenoptera_odb10). Read coverage quality was assessed with minimap2 [28] utilizing predefined settings for HiFi CCS reads and retaining only the best matching primary alignment for each read above a mapping quality threshold of 50.

Phylogeny

A phylogeny was constructed representing the genomic resources we produced as well as genomes available of representatives across the family Vespidae. 4,328 BUSCO genes that were identified with a single copy in each genome along with orthologous genes for 12 publicly available genomes of representatives from the family Vespidae were aligned using mafft v7.505 [37] with a maximum of 1000 iterations, and alignments were trimmed to retain only regions with robust alignments using trimAl [38] with the "-strictplus" flag. In order to maintain the maximal amount of alignment coverage, genes were pruned if an ortholog was absent in any of the lineages. Trimmed alignments were sorted by their length and only the longest were used to construct a maximally occupied dataset. Maximum-likelihood gene trees of the 500 longest amino acid gene alignments were produced using IQ-TREE v2.0.6 [39, 40] each with the best fitting AA model assessed with ModelFinder [41] performed from among a subset of several general matrix models. Trimmed alignments were concatenated and a maximum-likelihood tree was produced with IQ-TREE v2.0.6 utilizing gene-specific model partitions and 1,000 ultrafast bootstrap replicates.

Repeat profiling

Large stretches of tandemly repetitive DNA were identified in HiFi CCS read data and scaffolded genomes of each assembly using TRF v4.10 [42] with the settings "trf FASTA 2 5 7 80 10 50 2000 -l 10 -h". Scaffolded genomes were masked using the "-m" flag. Identity of tandemly repetitive DNA in the scaffolded assemblies was evaluated using Spectra v1.0, a toolkit for Python and R we developed for the visualization of the signature of repetitive DNA in genomic sequences by measuring the prevalence of short k-mers (k = 3) along sequences. K-mer localization was measured in the Canadian and Washington *Vespa mandarinia* genomes along 3,000 bp intervals with canonical k-mers counted as their counterpart using "Spectra.py count -w 3000 -s 3000 -c .". Chromosomes were visualized with scaling across chromosomes maintained using "Spectra-plot.r -k .". Repeat motifs were then categorized as belonging to a centromeric satellite if localization coincided with the inferred centromeric region and if the size and identity of the motif were consistent across TRF identifications. Tandemly repetitive DNA consistently identified adjacent to the centromeric satellite but bearing distinct motif identities and sizes were categorized as pericentric elements. All other instances of tandemly repetitive DNA remained uncategorized.

Identification of *de novo* repetitive elements was performed using RepeatModeler v2 [43] for each genome masked with TRF. Repetitive elements were characterized from each genome incorporating LTR structures and then classified based on homology to RepeatMasker Repeat Protein Database and Dfam. Repeat libraries classified from each genome were combined to create an exhaustive database of repetitive elements for the hornets, and identifiable centromeric elements from each genome were added to the database. Elements in each scaffolded genome were counted and localized with RepeatMasker utilizing the combined database of repetitive elements.

Pangenome construction

Utilizing minigraph-cactus v2.6.8 [44], a pangenomic graph was generated from the sequences above as well as two publicly available *Vespa* genomes. Briefly, genomes with tandemly repetitive regions masked with TRF were aligned with the scaffolded genome of the haploid male *Vespa mandarinia* from Washington state as the reference lineage. Unscaffolded contigs from the assembly of the diploid female *V. mandarinia* from Canada were incorporated as both their primary and alternate haplotypes assembled through HiFiASM to ascertain their level of heterozygosity. Each other genome was supplied as scaffolded. The graph was used to generate variants, which were filtered to contain only biallelic SNPs with a valid allele call in all *Vespa* (including both the Canadian *V. mandarinia* haplotypes). A neighbor joining tree was constructed from a distance matrix calculated in TASSEL 5 [45]. Confidence intervals were measured with 1,000 bootstrap replicates from all SNPs. Diversity of *V. mandarinia* was evaluated by filtering the graph for alleles where one or both haplotypes of the Canadian genome were different from the Washington genome and then measured as a ratio against the total number of variants where no *Vespa* alleles were missing and each allele was not fixed among *Vespa*.

Identifying and annotating derived mutations

Repetitive elements among new mutations across the *Vespa* lineages were characterized with respect to *Vespula pensylvatica*. The graph was filtered to contain any insertions relative to *V. pensylvatica* with no missing alleles in any genome that were at least 50 base pairs in length. Fasta files of insertion events were constructed for each *Vespa* genome. Elements in each fasta were counted and localized with RepeatMasker utilizing the database of *de novo* repetitive elements constructed prior.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-12512-x.

Supplementary Material 1.

Acknowledgements

The authors appreciate the project insights provided by Jonas Korlach and his efforts on genome size estimation and assembly of the Vespa mandarinia mitogenome. The authors thank Todd M. Gilligan, Francisco Posada-Florez, Leonard J. Foster, Chris Looney and the Washington State Department of Agriculture team for their assistance with the coordination and collection of specimens.

Authors' contributions

AC and BS conceptualized this work. AO and JV expanded the concept and wrote this manuscript. AO prepared figures. SS, and SG performed specimen extraction and sequencing. AO, SS, and JJ performed sequence assembly and scaffolding. AO performed comparative analyses and repeat description.

Funding

This work was supported by the U.S. Department of Agriculture, Agricultural Research Service (USDA-ARS) funding to CRIS Project Nos 6066-21310-006-000-D and 8042-30500-002-000-D. This work was also funded, in part, by the USDA-Animal and Plant Health Inspection Service (Project Number: 8042-21000-291-046-I) and reagents provided by Pacific Biosciences. The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Data availability

All raw reads used to assemble and scaffold these genomes, along with the assemblies, were deposited at DDBJ/ENA/GenBank within the Umbrella BioProject PRJNA555319 for the USDA Agricultural Research Service's Ag100Pest Initiative and the genome accessions JBMFGJ000000000, JBMFGK000000000, JBMFGN000000000, JBMFGO000000000 were submitted under the BioProject PRJNA1088093. The Spectra data visualization tool is available at [https://github.com/USDA-ARS-GBRU/Spectra].

Declarations

Ethics approval and consent to participate

No specific permits were required for the samples collected for this study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 15 August 2025 / Accepted: 30 December 2025
Published online: 29 January 2026

References

1. Landolt PJ, et al. A new species of *Vespula*, and first record of *Vespa Crabro* L.(Hymenoptera: Vespidae) from Guatemala, central America. Zootaxa. 2010;2629(1):61–8.
2. Looney C, et al. Detection and description of four *Vespa Mandarinia* (Hymenoptera, Vespidae) nests in Western North America. J Hymenoptera Res. 2023;96:1–20.
3. Yellow-Legged Hornet USDA, Aphis. Mar., 7 2019, https://www.aphis.usda.gov/aphis/ourfocus/planthealth/plant-pest-and-disease-programs/honey-bees/yellow-legged-hornet
4. Wilson TM, et al. First reports of *Vespa Mandarinia* (Hymenoptera: Vespidae) in North America represent two separate maternal lineages in Washington state, united States, and British Columbia, Canada. Ann Entomol Soc Am. 2020;113(6):468–72.
5. Matsuura M. Vespa and Provespa. The social biology of wasps. 1991:232–62.
6. Choi M, Bo SJ, Martin. Distribution, spread, and impact of the invasive Hornet *Vespa velutina* in South Korea. J Asia Pac Entomol. 2012;15(3):473–7.
7. Wells JN, Cédric, Feschotte. A field guide to eukaryotic transposable elements. Annu Rev Genet. 2020;54(1):539–61.
8. Tørresen OK, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019;47(21):10994–1006.
9. Favreau E, et al. Putting hornets on the genomic map. Sci Rep. 2023;13(1):6232.
10. Bliznina A, et al. Telomere-to-telomere assembly of the genome of an individual *Oikopleura dioica* from Okinawa using Nanopore-based sequencing. BMC Genomics. 2021;22(1):222.
11. Eizenga JM, et al. Pangenome graphs. Annu Rev Genom Hum Genet. 2020;21(1):139–62.
12. Marques André, Drinnenberg IA. Same but different: centromere regulations in holocentric insects and plants. Curr Opin Cell Biol. 2025;93:102484.
13. Melters DanielP, et al. Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis. Chromosome Res. 2012;20(5):579–93.
14. Plačková Klára. Petr Bureš, and František Zedek. Centromere size scales with genome size across eukaryotes. Sci Rep. 2021;11(1):19811.
15. Suzuki Y, Myers EW, Morishita S. Rapid and ongoing evolution of repetitive sequence structures in human centromeres. Sci Adv. 2020;6:50.

16. Taylor BA, et al. Population genomics of the invasive Northern giant Hornet Vespa Mandarinia in North America and across its native range. Sci Rep. 2024;14(1):10803.
17. Maside X, Assimacopoulos S, Charlesworth B. Rates of movement of transposable elements on the second chromosome of *Drosophila melanogaster*. Genet Res. 2000;75(3):275–84.
18. Pardo-Manuel, De Villena. Fernando, and Carmen Sapienza. Nonrandom segregation during meiosis: the unfairness of females. Mamm Genome. 2001;125:331–9.
19. Rosin LF, Barbara G. Mellone. Centromeres drive a hard bargain. Trends Genet. 2017;33(2):101–17.
20. Huang Y-C, et al. Evolution of long centromeres in fire ants. BMC Evol Biol. 2016;16(1):189.
21. Bennett MD. Nuclear DNA content and minimum generation time in herbaceous plants. Proceedings of the Royal Society of London. Series B. Biological Sciences. 1972;181(1063):109–135.
22. Gregory T, Ryan PDN, Hebert, Kolasa J. Evolutionary implications of the relationship between genome size and body size in flatworms and copepods. Heredity. 2000;84(2):201–8.
23. Gregory T. Ryan. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev. 2001;76(1):65–101.
24. Scholes DR, Suarez AV, Ken N. Paige. Can endopolyploidy explain body size variation within and between castes in ants? Ecol Evol. 2013;3(7):2128–37.
25. Detection and Eradication Data. Wash State Department Agric. 2024. https://agr.wa.gov/departments/insects-pests-and-weeds/insects/hornets/data.
26. DeAngelis MM, Wang DG. Hawkins. Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Res. 1995;23(22):4742.
27. Soorni A, et al. Organelle_PBA, a pipeline for assembling Chloroplast and mitochondrial genomes from PacBio DNA sequencing data. BMC Genomics. 2017;18(1):49.
28. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
29. Sim SB, et al. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. BMC Genomics. 2022;23(1):157.
30. Cheng H, et al. Haplotype-resolved de Novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–5.
31. Durand NC, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3(1):95–8.
32. Dudchenko O, et al. De Novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92–5.
33. Durand NC, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3(1):99–101.
34. Alonge M, et al. Automated assembly scaffolding using ragtag elevates a new tomato system for high-throughput genome editing. Genome Biol. 2022;23(1):258.
35. Marçais G, Kingsford C. JELLYFISH–fast, parallel k-mer counting for DNA. Bioinformatics. 2011;27:764–70.
36. Manni Mosè, et al. BUSCO: assessing genomic data quality and beyond. Curr Protocols. 2021;1(12):e323.
37. Katoh K, Daron M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
38. Capella-Gutiérrez S. Silla-Martínez JoséM, Gabaldón T. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.
39. Nguyen L-T, et al. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
40. Chernomor O. Haeseler, and Bui Quang Minh. Terrace aware data structure for phylogenomic inference from supermatrices. Syst Biol. 2016;65(6):997–1008.
41. Kalyaanamoorthy S, et al. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587–9.
42. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.
43. Flynn JM et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences. 2020;117(17):9451–9457.
44. Hickey G, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol. 2023;42:663–73.
45. Bradbury PJ, et al. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.

Publisher's Note