


RESEARCH ARTICLE

Open Access



Genome annotation with long RNA reads reveals new patterns of gene expression and improves single-cell analyses in an ant brain

Emily J. Shields^{1,2}, Masato Sorida¹, Lihong Sheng¹, Bogdan Sieriebriennikov^{3,4}, Long Ding³ and Roberto Bonasio^{1*} 

Abstract

Background: Functional genomic analyses rely on high-quality genome assemblies and annotations. Highly contiguous genome assemblies have become available for a variety of species, but accurate and complete annotation of gene models, inclusive of alternative splice isoforms and transcription start and termination sites, remains difficult with traditional approaches.

Results: Here, we utilized full-length isoform sequencing (Iso-Seq), a long-read RNA sequencing technology, to obtain a comprehensive annotation of the transcriptome of the ant *Harpegnathos saltator*. The improved genome annotations include additional splice isoforms and extended 3' untranslated regions for more than 4000 genes. Reanalysis of RNA-seq experiments using these annotations revealed several genes with caste-specific differential expression and tissue- or caste-specific splicing patterns that were missed in previous analyses. The extended 3' untranslated regions afforded great improvements in the analysis of existing single-cell RNA-seq data, resulting in the recovery of the transcriptomes of 18% more cells. The deeper single-cell transcriptomes obtained with these new annotations allowed us to identify additional markers for several cell types in the ant brain, as well as genes differentially expressed across castes in specific cell types.

Conclusions: Our results demonstrate that Iso-Seq is an efficient and effective approach to improve genome annotations and maximize the amount of information that can be obtained from existing and future genomic datasets in *Harpegnathos* and other organisms.

Keywords: Iso-Seq, Long-read RNA-seq, *Harpegnathos saltator*, Ants, Genome annotation, 3' UTR annotation, Single-cell sequencing, Alternative splicing

Background

Improved sequencing technologies have enabled studies in previously inaccessible organisms, but annotations remain the bottleneck to thorough genomic and

epigenomic analyses. Specifically, gene annotations of many non-model organisms suffer from the limitations imposed by their reliance on traditional, short-read RNA-seq coupled with gene prediction software [1–3]. This approach can identify splice junctions but cannot capture complex combinations of exons that define full transcript isoforms. Furthermore, local fluctuations of RNA sequencing coverage can make it difficult to

* Correspondence: roberto@bonasiolab.org

¹Epigenetics Institute and Department of Cell and Developmental Biology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

identify the 5' and 3' untranslated regions (UTRs), resulting in inaccurate transcription start sites (TSSs) and transcription termination sites (TTSs) [4]. Inaccurate annotation of the 3' end of genes is especially problematic for the analysis of droplet-based single-cell RNA sequencing data, such as those obtained with the widely used 10x Genomics 3' gene expression platform [5] or Drop-seq [6], both of which have a strong 3' bias due to the capture of transcripts using an oligo-dT sequence [7].

Just as long-read sequencing of DNA has led to better genome assemblies [8, 9], long-read sequencing of RNA molecules can be used to address the limitations of short-read-based annotation. PacBio Single Molecule Real-Time (SMRT) sequencing for RNA, called Iso-Seq, produces full-length transcript sequences, resolving issues of isoform reconstruction and accurate identification of the end of transcripts. Iso-Seq has been used in many settings and organisms to reveal alternative polyadenylation sites [10], provide insight into alternative splicing [11], and identify tissue-specific transcriptional isoforms [12]. The genome of the ant *Harpegnathos saltator* was first sequenced in 2010 [13], was improved using Pacific Biosciences (PacBio) long-read DNA sequencing in 2018 [9], and was re-annotated by NCBI (NCBI Release 102, released in 2018 [14];). While long-read DNA sequencing technology was utilized to great effect to improve the reference *Harpegnathos* genome assembly, existing gene annotations still suffered from the shortcomings listed above, imposed by their reliance on traditional, short-read RNA-seq coupled with gene prediction software.

Harpegnathos is a promising model system to study the epigenetic regulation of brain and behavioral plasticity. Similar to colonies of other social insects, *Harpegnathos* colonies are founded by a mated reproductive female ("queen") and contain many non-reproductive individuals ("workers") that carry out all tasks necessary for colony survival. As in most social insect species [15], *Harpegnathos* queens and workers differ greatly in reproductive physiology, social status and behavior, and lifespan, despite possessing the same genomic instructions. In addition, *Harpegnathos* ants display a rare form of phenotypic plasticity: workers retain the ability to convert to reproductive individuals called "gamergates" throughout their adult life [16, 17]. In the absence of a dominant reproductive, *Harpegnathos* workers participate in a ritual dueling tournament, whereby winners become gamergates that activate their ovaries and acquire a queen-like social status [18, 19]. Workers that become gamergates lay eggs, cease activities associated with the worker caste [20], and acquire a longer lifespan [21]. Previous works identified changes in brain transcriptomes [20] and cell type proportions [22] following

this behavioral transition, establishing *Harpegnathos* as a powerful model for studying the epigenetic regulation of brain and behavioral plasticity.

Here, we use Iso-Seq long-read RNA sequencing to further improve the genomic infrastructure for genomic and epigenomic studies in *Harpegnathos* by generating more comprehensive annotations of splice isoforms and gene boundaries. These new annotations resulted in greatly improved analyses of bulk and single-cell RNA-seq, which revealed new caste-specific genes and splicing events and extended the reach of our single-cell atlas of the *Harpegnathos* brain.

Results

Using Iso-Seq to update *Harpegnathos* gene annotation

We previously generated a single-cell RNA-seq atlas of the *Harpegnathos* brain during the worker–gamergate transition and discovered extensive changes in cell type composition in glia and neurons [22]. While inspecting these sequencing data [23], we noticed that in many cases, even when using the latest NCBI annotation (NCBI Release 102; hereafter referred to as HSAL50), the single-cell RNA-seq reads mapped outside gene model boundaries, typically downstream of the annotated TTS, resulting in decreased coverage and information loss. As an example we show the case of the *Ref1* gene (Fig. 1A, red box), resulting in decreased coverage and information loss. Motivated by examples such as this and by a desire to obtain a more comprehensive catalog of splicing isoforms, we sought to improve the *Harpegnathos* gene annotations using PacBio Iso-Seq long RNA reads.

To maximize library complexity, we sequenced two separate polyA+ Iso-Seq libraries: one from a pool of *Harpegnathos* brains from different castes, and one from a mixture of ovary and fat body tissues. After processing the raw PacBio subreads (Additional File 1: Fig. S1A and Fig. 1B), we obtained 34,867 and 33,520 full-length "polished" reads with median length of 2034 bp and 2137 bp for the brain sample and the fat body/ovary sample, respectively (Fig. 1C). After aligning the polished Iso-Seq reads to the *Harpegnathos* genome, we compared gene coverage with that of previously obtained short-read RNA-seq in matching tissues [9, 24]. More than half of the genes detectable (RPKM_s > 0.5) in our collection of deep short-read RNA-seq were also covered by Iso-Seq reads (Fig. 1D). As expected, genes detected by Iso-Seq tended to be more highly expressed (Additional File 1: Fig. S1B). From the mapped reads, we collapsed redundant transcript models, generated predicted isoforms for each of the Iso-Seq libraries, and used these isoforms to refine the existing HSAL50 annotation. We further improved these models by manually adding 89 genes and reviewing all merged genes (Additional File 1:

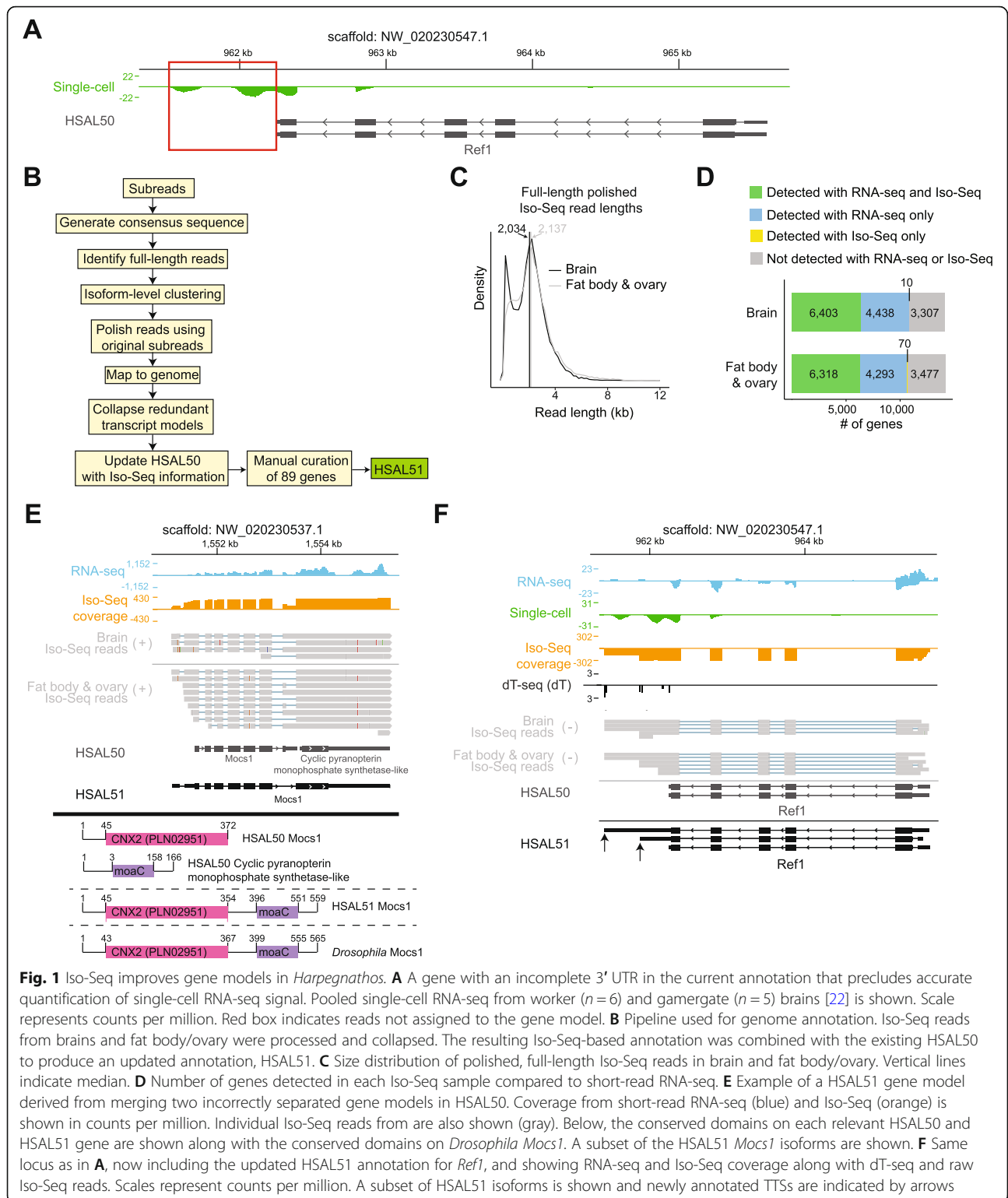


Fig. 1 Iso-Seq improves gene models in *Harpegnathos*. **A** A gene with an incomplete 3' UTR in the current annotation that precludes accurate quantification of single-cell RNA-seq signal. Pooled single-cell RNA-seq from worker ($n = 6$) and gamergate ($n = 5$) brains [22] is shown. Scale represents counts per million. Red box indicates reads not assigned to the gene model. **B** Pipeline used for genome annotation. Iso-Seq reads from brains and fat body/ovary were processed and collapsed. The resulting Iso-Seq-based annotation was combined with the existing HSAL50 to produce an updated annotation, HSAL51. **C** Size distribution of polished, full-length Iso-Seq reads in brain and fat body/ovary. Vertical lines indicate median. **D** Number of genes detected in each Iso-Seq sample compared to short-read RNA-seq. **E** Example of a HSAL51 gene model derived from merging two incorrectly separated gene models in HSAL50. Coverage from short-read RNA-seq (blue) and Iso-Seq (orange) is shown in counts per million. Individual Iso-Seq reads from are also shown (gray). Below, the conserved domains on each relevant HSAL50 and HSAL51 gene are shown along with the conserved domains on *Drosophila Mocs1*. A subset of the HSAL51 *Mocs1* isoforms are shown. **F** Same locus as in **A**, now including the updated HSAL51 annotation for *Ref1*, and showing RNA-seq and Iso-Seq coverage along with dT-seq and raw Iso-Seq reads. Scales represent counts per million. A subset of HSAL51 isoforms is shown and newly annotated TTSs are indicated by arrows

Fig. S1C and Additional File 2: Table S1–3) and designated this upgraded annotation as HSAL51.

Overall, HSAL51 contained 13,957 gene models that corresponded unambiguously to HSAL50 genes, 392

new genes either predicted from Iso-Seq signal or added manually (Additional File 2: Table S1–2), and 57 gene models that merged two or more HSAL50 gene models into a single one in HSAL51 (Additional File 1: Fig. S1D

and Additional File 2: Table S3). An example of a merged gene was the combination of two adjacent HSAL50 genes, each of which contained one of the two known *Drosophila melanogaster* MOCS1 protein domains; the resulting merged gene model for *Mocs1* in HSAL51 encodes a protein with a domain structure identical to its ortholog in *Drosophila* (Fig. 1E). Returning to the example of *Ref1* (Fig. 1A), Iso-Seq reads indicated the existence of at least two isoforms with TTS downstream of the one annotated in HSAL50, which captured the single-cell sequencing signal missed with the old annotation (Fig. 1F, arrows). For additional verification of the TTSs predicted in our new annotation, we devised a custom RNA-seq protocol that compares short reads obtained with an anchored oligo-dT primer with random hexamers to remove signal from internal A-stretches and identifies the position of the polyA tail on mature mRNAs (“dT-seq,” Additional File 1: Fig. S1E–F, see methods). In the case of *Ref1*, dT-seq signal analyses confirmed the existence of the new termination sites (Fig. 1F).

Thus, using long-read sequencing, we updated the *Harpegnathos* gene annotation and recovered gene models that were split incorrectly in the HSAL50 assembly or that did not have a correctly annotated 3' UTRs.

Comprehensive annotation of transcriptional isoforms with Iso-Seq

Since its development in 2013 [25], Iso-Seq has been performed on a genome-wide scale in a range of plants and animal species to improve the annotation of transcriptional isoforms [11, 26–28]. The ability of Iso-Seq to sequence RNA molecules in their entirety confers an advantage in detecting splicing patterns compared to the typical short-read annotation strategy of relying on reads that cover a limited span across splice junctions. Indeed, HSAL51 contained a greater number of annotated transcripts with distinct splicing patterns (i.e., beyond simple extension of 5' or 3' UTRs) (Fig. 2A). In addition, gene models in HSAL51 exhibited more instances of all seven types of alternative splicing [29]: skipped exon (SE), mutually exclusive exons (ME), alternative 5' splice site (A5), alternative 3' splice site (A3), retained introns (RI), alternative first exon (AF), and alternative last exon (AL) (Fig. 2B). Examples of genes with newly annotated alternative splicing events are presented in Additional File 3: Fig. S2A–C.

Next, we identified genes whose relative transcript expression varies between tissues (also called differential transcript usage, or DTU) [30] using previously published bulk RNA-seq data from 6 *Harpegnathos* tissues: non-visual brain, ovary, fat body, retina, optic lobe, and antenna [9, 31]. Most genes (80%) exhibiting DTU in HSAL50 between at least two tissues were also detected

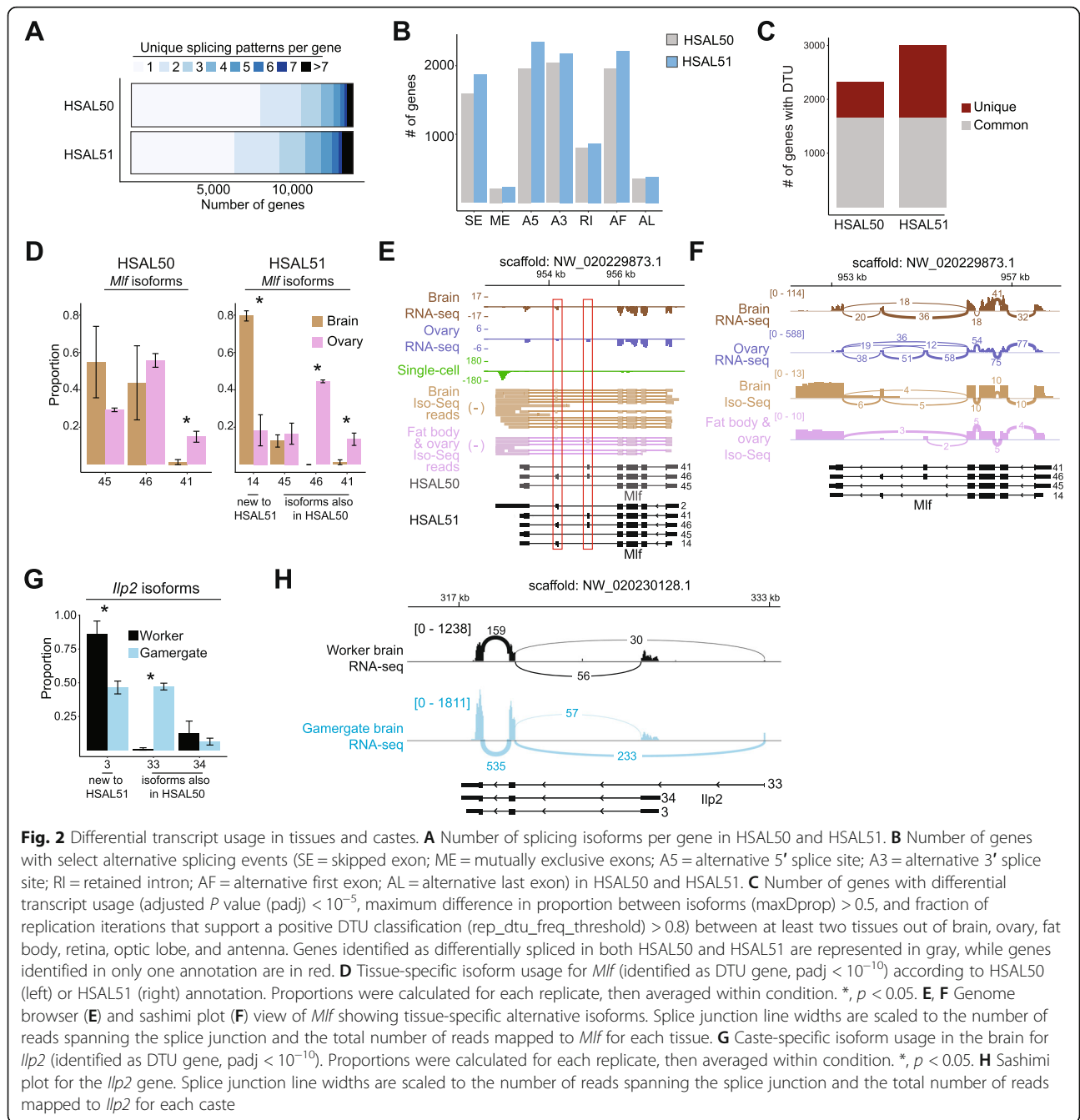
in HSAL51, but the number of genes displaying DTU increased by 681 in the new annotation (Fig. 2C). For example, a newly annotated isoform (isoform 14) of the *myeloid leukemia factor* (*Mlf*) gene accounted for 80% of transcripts produced in the brain (Fig. 2D). This isoform contained exon 6 but not exon 5 of *Mlf* (Fig. 2E, F) and was not identified with short reads alone, highlighting the power of Iso-Seq in untangling complicated exon structures, especially in cases of mutually exclusive exons. Once annotated, short reads spanning the alternative splice junctions could be properly assigned to this isoform resulting in the identification of brain-specific DTU.

In addition to genes with DTU between tissues, we identified several genes with caste-specific transcript usage in the brain using previously published RNA-seq from *Harpegnathos* [24]. One notable gene with caste-biased isoforms between workers and gamergates was *insulin-like-peptide 2* (*Ilp2*; *LOC105188195*), a gene similar to canonical insulin whose absolute transcript levels are higher in the brains or heads of reproductive individuals compared to those of non-reproductives in many ant species [32]. In general, insulin signaling has been identified as a key pathway regulating caste identity in social insects [33]. In addition to the higher gene-level *Ilp2* expression in brains of *Harpegnathos* gamergates compared to workers (Additional File 3: Fig. S2D), isoforms 3 and 33 were used at different levels between the castes (Fig. 2G), with the upstream first exon being used more frequently in gamergates compared to workers (Fig. 2H). While these alternative splicing events appear to only affect the 5' UTR, they might still have important consequences on insulin signaling, for example by regulating translation of the resulting mRNA, as observed in mammals [34, 35]. Upon reanalysis of published data [32, 36–38], two other ant species, the carpenter ant *Camponotus planatus* (Formicinae) and the giant ant *Dinoponera quadricaps* (Ponerinae), also displayed caste-biased selection of the first exon in brains (Additional File 3: Fig. S2E–F). In both species, as in *Harpegnathos*, the reproductive caste was more likely than the non-reproductive caste to use the upstream first exon, suggesting that alternative splicing of *Ilp2* mRNA might be an evolutionary conserved mechanism for the caste-specific regulation of the insulin pathway.

Together, these analyses demonstrate that an Iso-Seq-enriched genomic annotation captures a greater complexity in the transcriptome which, in turn, can provide a comprehensive view of alternative splicing events between different biological samples—in this case tissues and castes.

Extended 5' and 3' gene boundaries increase sensitivity of bulk RNA-seq

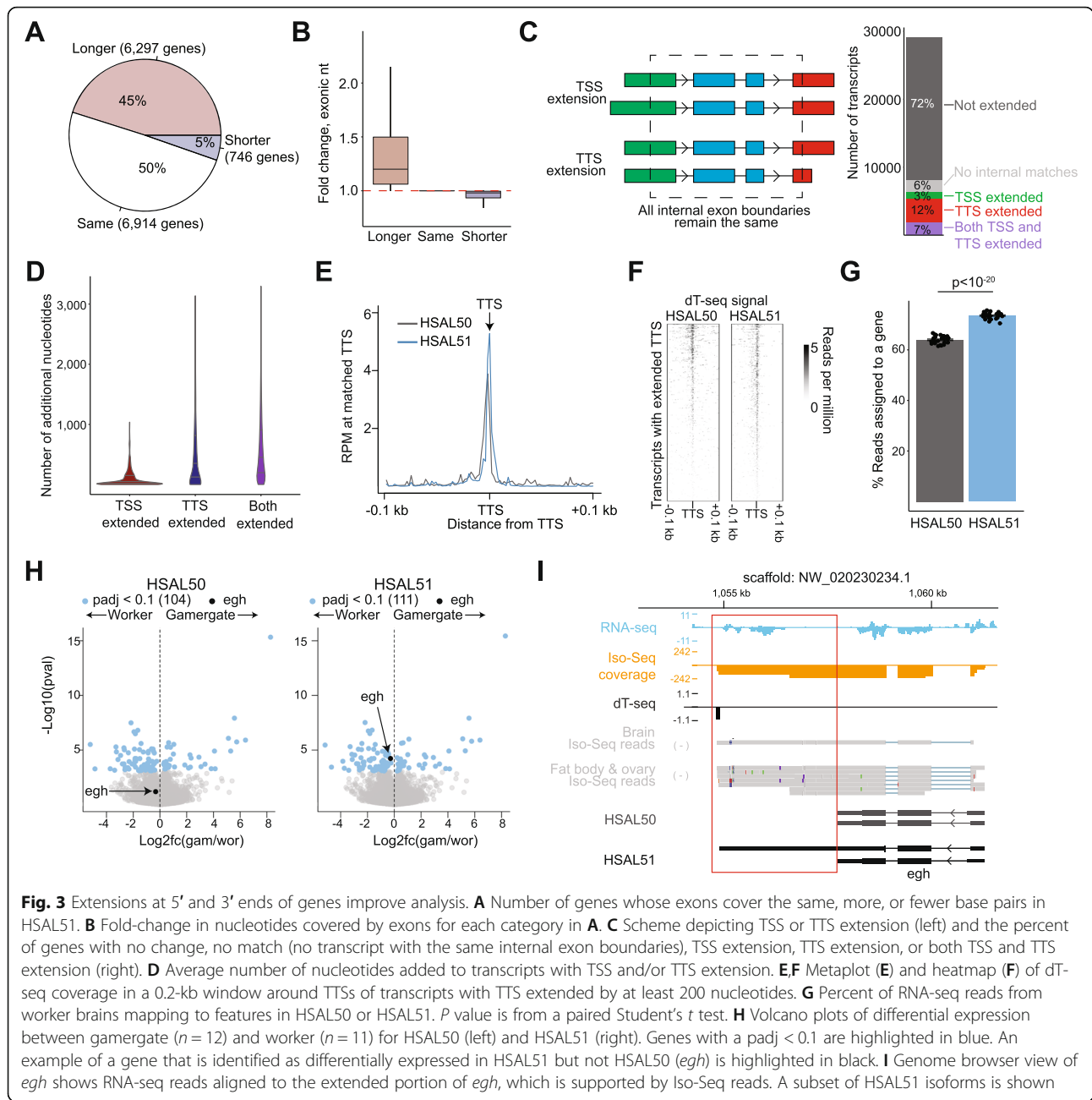
In addition to a more comprehensive view of transcriptional isoforms originating from alternative splicing, the



long reads of Iso-Seq are expected to contain more complete UTRs in 3' and, to some extent, 5', resulting in more accurate annotation of TTSs and TSSs, respectively, extending the mappable regions of the gene models. Consistent with these expectations, for 45% of all gene models, the exons annotated in HSA51 covered a larger (median extension, + 20%) sequence than in HSA50, whereas only 5% resulted in smaller gene models and even those were minimally impacted (median contraction, - 2%) (Fig. 3A, B). This remarkable growth in annotated exonic space was largely due to the

extension of 3' UTR using newly annotated TTSs (5338 transcripts among 4269 genes) and, to a smaller degree, to the extension of 5' UTRs using newly annotated TSSs (2878 transcripts) (Fig. 3C). Transcripts were typically extended by more base pairs at the TTS compared to the TSS (Fig. 3D), with median extension length of 251 nt and 31 nt, respectively.

To confirm that these 3' UTR extensions originated from the annotation of bona fide TTSs, we analyzed the position of polyA tails, as determined by dT-seq (see Additional File 1: Fig. S1E), relative to the HSA50 and



HSAL51 gene models. As expected, the dT-seq signal accumulated at TTSs in both annotations, and it was stronger at the TTSs of HSAL51 gene models as compared to HSAL50 (Additional File 4: Fig. S3A), including in a comparison of gene models with extended TTSS in HSAL51 (Fig. 3E, F), demonstrating the accuracy of the newly annotated downstream TTSs.

To evaluate the effect of the new annotation on RNA-seq mapping rates, we calculated the mapping rate of aligned reads using a newly generated dataset of *Harpegnathos* worker brains, which were not used in the construction of the HSAL50 (or HSAL51) annotations.

Significantly more reads mapped to annotated exons in HSAL51 (Fig. 3G). This improvement in RNA-seq mapping rates had tangible benefits on the biological interpretation of sequencing datasets, such as, for example, the identification of additional differentially expressed genes in pairwise comparisons. Reanalyzing the worker vs. gamergate transcriptomes [20, 24] with the new annotation identified the *egh* gene as significantly upregulated in workers (Fig. 3H). The *Drosophila* homolog of this gene is involved in the sex-peptide response and is implicated in the regulation of mating and egg-laying [39], suggesting a biological explanation for its caste-

specific expression in *Harpegnathos*. Iso-Seq reads indicated the existence of a longer gene model for *egh* (Fig. 3I), which increased the number of reads mapping to this gene (Additional File 4: Fig. S3B) and resulted in its confident identification as caste-specific.

These results show that the addition of Iso-Seq information extends the annotated gene models, allowing for the extraction of more information from RNA-seq experiments and resulting in higher sensitivity for differentially expressed genes.

Iso-Seq-based annotation improves single-cell analyses

Given the 3' bias of the most widely used techniques for droplet-based single-cell sequencing, we reasoned that these analyses would be improved by the more accurate 3' UTR annotations found in HSAL51. Indeed, the mapping rate of 10x Genomics single-cell RNA-seq reads from our previous worker–gamergates comparison [22, 23] increased in average by 44% (Fig. 4A). This resulted in substantially increased counts for a large majority of annotated genes, including several with important functions in the brain (Fig. 4B, below diagonal). In all 11 samples analyzed, increased mapping rates resulted in improvements for the total number of cells identified, as well as the average unique molecular identifiers (UMIs) and genes detected per cell (Fig. 4C). Overall, the total number of cells passing quality thresholds increased by 18% from 20,729 using the HSAL50 annotation to 24,560 using HSAL51 (Fig. 4D, Additional File 5: Fig. S4A), showing the importance of accurate 3' UTR annotations to extract the maximum amount of information from single-cell RNA-seq data.

Markers for three major neuronal classes based on neurotransmitter usage, *VAcHT* (cholinergic), *VGlut* (glutamatergic), and *Gad1* (GABAergic), were among the genes that benefitted from the increased mapping rates (Fig. 4B), resulting in an increased number of cells with observed expression of these genes (Fig. 4E, pie charts), and thus easier identification of clusters with specific neurotransmitter expression, as visualized on UMAP projection of the HSAL51 data (Fig. 4E). These improvements were not confined to genes associated with neurotransmitter usage; using HSAL51, we recovered in total 288 previously undetected marker genes with restricted, cell type-specific expression (Additional File 5: Fig. S4B), likely missed in HSAL50 due to reads that were not assigned to the incomplete old gene models (Additional File 5: Fig. S4C).

In addition, we recovered 12 new markers for mushroom body neurons (Additional File 5: Fig. S4D), which are key to learning and memory in insects [40–42]. Some of these markers were biased for mushroom body cells in HSAL50 but did not pass statistical thresholds due to overall low expression. Of these 12 newly identified mushroom body markers in *Harpegnathos*, 10 were previously

described as mushroom body-specific genes in *Drosophila* [43, 44] or honeybees [45] (Additional File 5: Fig. S4D). In particular, two *Harpegnathos* genes with homology to known mushroom body markers *GluR1B* and twin of eyeless (*toy*) [44, 46–48] were barely detectable in HSAL50 but clearly mapped to mushroom body clusters in HSAL51 (Fig. 4F and Additional File 5: Fig. S4D).

We previously showed that neuroprotective ensheathing glia cells are expanded during the worker–gamergate transition and lost at a faster rate in workers than in gamergates during aging [22]. Despite the in-depth investigation of this cell type in our previous study, the updated HSAL51 annotation allowed us to discover a new marker gene, *CG9259* (Fig. 4G), that was previously missed because the near-entirety of the single-cell RNA-seq reads fell within an extended 3' UTR not annotated in HSAL50 (Additional File 5: Fig. S4E). In *Drosophila*, *CG9259* encodes an ecdysteroid kinase-like protein [49], suggesting that *Harpegnathos* ensheathing glia that express this gene might play an important role in the caste-specific regulation of the key developmental hormone ecdysone.

Finally, the increased single-cell transcriptome depth afforded by HSAL51 allowed us to identify differential gene expression between workers and gamergates within specific cell types, with more genes overall classified as caste-specific between single-cell clusters in the HSAL51 analysis (Additional File 5: Fig. S4F). As with the newly detected marker genes, many of the 250 newly called differential genes had increased counts in HSAL51 compared to HSAL50, suggesting their new detection resulted from the increased mapping rates of the 3'-biased single-cell RNA-seq reads (Additional File 5: Fig. S4G). For example, we identified *CycG* as a gene preferentially expressed in specific subtypes of gamergate cells as compared to their counterparts in workers (Fig. 4H). This observation is in agreement with previous studies reporting upregulation of *CycG* in the reproductive caste of various ant species [50], including *Harpegnathos* gamergates (Additional File 5: Fig. S4H). *CycG* regulates the insulin pathway [51, 52], which, as mentioned above, is an important player in caste determination in social insects [32, 33]. The identification of the cell types that are the strongest drivers of *CycG* caste-biased expression will help inform future studies of this gene.

Thus, single-cell RNA-seq analyses were greatly improved by the increased accuracy of 3' UTR annotations in HSAL51, resulting in 18% more single cells identified computationally, a clustering of transcriptional types more reflective of biological function, recovery of additional cell-type markers, and higher sensitivity for differentially expressed genes.

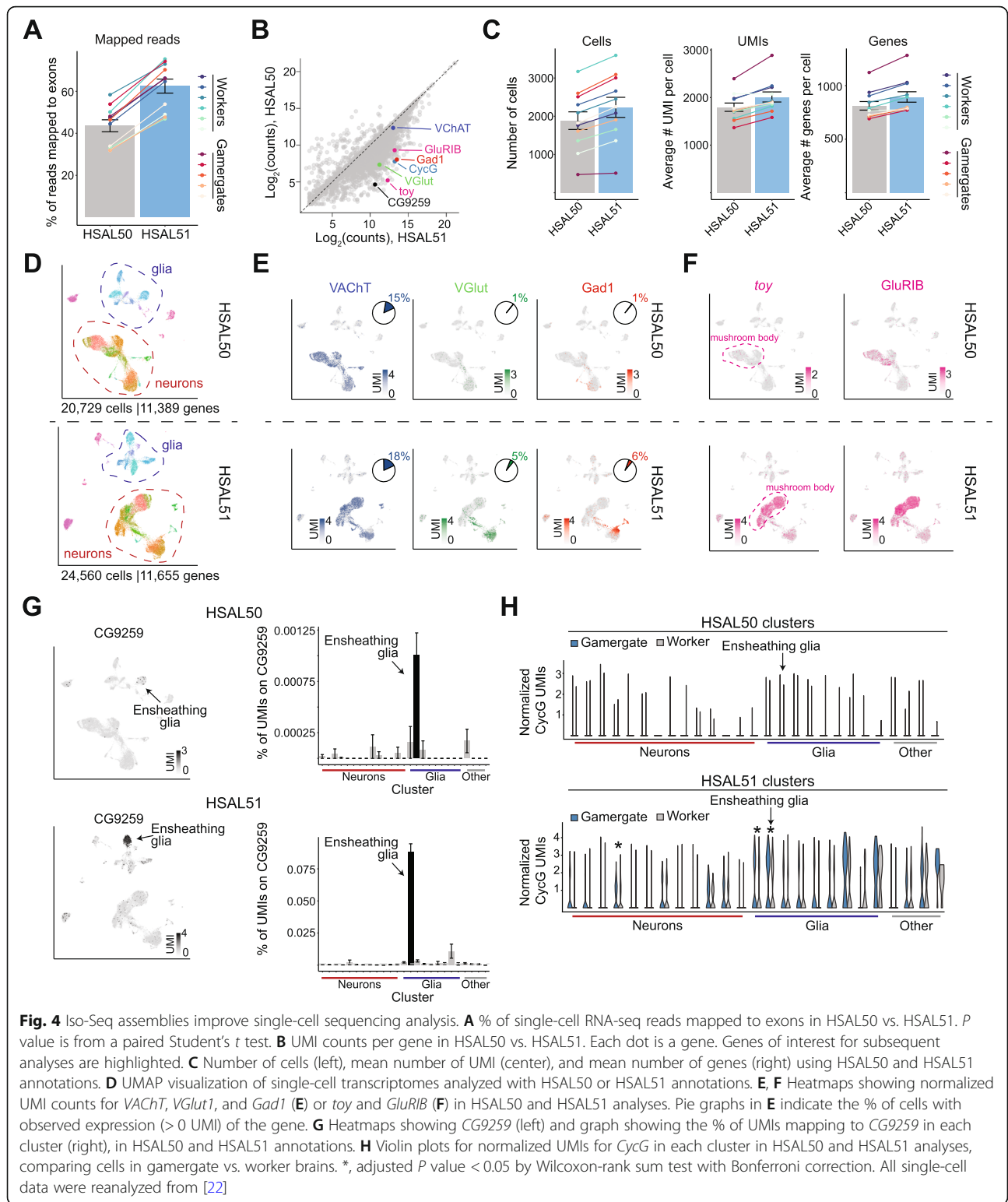


Fig. 4 Iso-Seq assemblies improve single-cell sequencing analysis. **A** % of single-cell RNA-seq reads mapped to exons in HSAL50 vs. HSAL51. *P* value is from a paired Student's *t* test. **B** UMI counts per gene in HSAL50 vs. HSAL51. Each dot is a gene. Genes of interest for subsequent analyses are highlighted. **C** Number of cells (left), mean number of UMI (center), and mean number of genes (right) using HSAL50 and HSAL51 annotations. **D** UMAP visualization of single-cell transcriptomes analyzed with HSAL50 or HSAL51 annotations. **E, F** Heatmaps showing normalized UMI counts for *VChT*, *VGlut1*, and *Gad1* (**E**) or *toy* and *GluRIB* (**F**) in HSAL50 and HSAL51 analyses. Pie graphs in **E** indicate the % of cells with observed expression (> 0 UMI) of the gene. **G** Heatmaps showing *CG9259* (left) and graph showing the % of UMIs mapping to *CG9259* in each cluster (right), in HSAL50 and HSAL51 annotations. **H** Violin plots for normalized UMIs for *CycG* in each cluster in HSAL50 and HSAL51 analyses, comparing cells in gamergate vs. worker brains. *, adjusted *P* value < 0.05 by Wilcoxon-rank sum test with Bonferroni correction. All single-cell data were reanalyzed from [22]

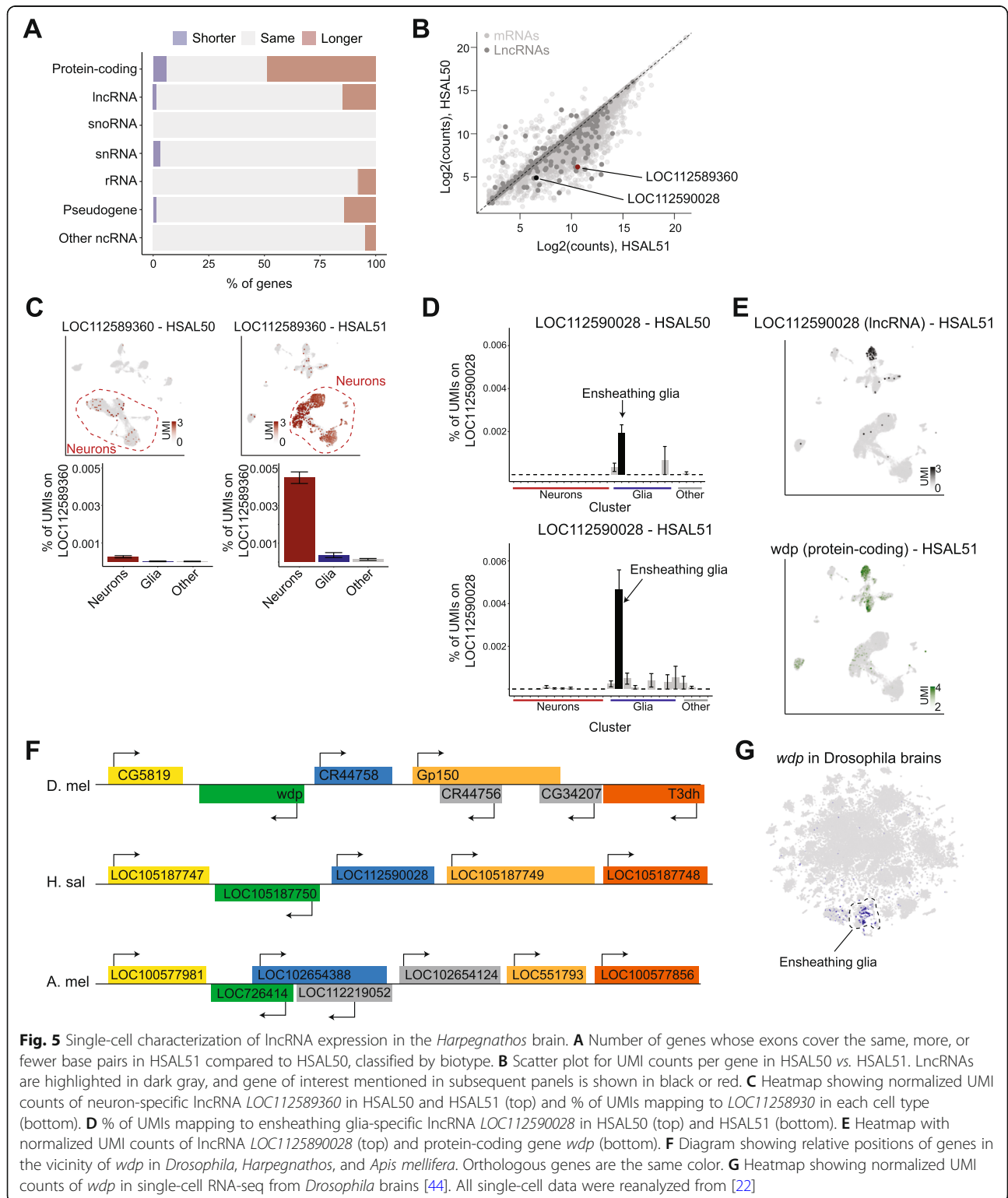
Long noncoding RNAs in single-cell sequencing analysis revealed by Iso-Seq

Protein-coding genes are often the focus of transcriptomic studies, but many genes are transcribed into

noncoding RNAs with important regulatory roles [53–55]. Similar to the case of protein-coding transcripts, several gene models for various types of noncoding RNAs, and in particular long noncoding RNAs

(lncRNAs), were also extended in HSAL51 compared to HSAL50 (Fig. 5A, Additional File 6: Fig. S5A), although not to the same extent, possibly due to their overall lower expression level.

Single-cell analyses using the updated gene models in HSAL51 revealed 130 lncRNAs with more UMIs compared to HSAL50 (Fig. 5B), suggesting that the new annotation might provide additional insight on the



patterns of lncRNA expression in the *Harpegnathos* brain. We recovered a set of lncRNAs with neuronal-specific expression profiles, some of which could not be detected using HSAL50 gene models (Additional File 6: Fig. S5B). One example was LOC112589360, which had over 20 times more mapping reads in HSAL51 compared to HSAL50 (Fig. 5B), with a corresponding increase in its calculated expression levels in neurons (Fig. 5C), as well as the fraction of neurons where this lncRNA could be detected, from 0.6% in HSAL50 to 10.1% in HSAL51 (Additional File 6: Fig. S5B).

In addition to these neuronal lncRNAs, we identified several cell type-specific lncRNAs in ensheathing glia. LOC112588339.LOC112588340 was merged from two adjacent lncRNAs annotated in HSAL50, with Iso-Seq reads clearly supporting the HSAL51 gene model (Additional File 6: Fig. S5C). The new merged gene model had negative coding potential as assessed by CPC and PhyloCSF [56, 57] and was one of the strongest markers of ensheathing glia (Additional File 6: Fig. S5D). Another updated lncRNA, LOC112590028, was specific to ensheathing glia, but missed by previous analyses due to low mapping rates in HSAL50 (Fig. 5D, E). The protein-coding gene adjacent to this lncRNA, *windpipe* (*wdp*), was also preferentially expressed in ensheathing glia (Fig. 5E and Additional File 6: Fig. S5E), suggesting potential co-regulation of the coding and noncoding transcript in *cis* as previously reported for other lncRNAs-mRNA pairs [58]. *Drosophila wdp* encodes a transmembrane protein with known functions in the wing disc [59] and the trachea [60], and it has also been implicated in synaptic target recognition [61] and learning [62]. While the sequence of the lncRNA LOC112590028 itself is not conserved, *Drosophila* has a lncRNA, *CR44758* in the same position as LOC112590028, between *wdp* and *Gp150* (Fig. 5F), suggesting that synteny of this locus, and potentially its molecular regulation, have been maintained over 350 million years of divergent evolution (Fig. 5F). In fact, *wdp* is also expressed specifically in *Drosophila* ensheathing glia (Fig. 5G) [44], further supporting a conserved regulation of this locus across distantly related insect species.

Thus, similar to protein-coding genes, lncRNA annotations were also improved by the addition of Iso-Seq data and this resulted in increased visibility of these regulatory transcripts in single-cell analyses.

Discussion

Genomic resources are becoming increasingly common for a wide range of species, beyond traditional model organisms, facilitating molecular analyses of an ever-growing variety of biological phenomena. However, in addition to high-quality genome assemblies, accurate gene annotations are indispensable for genome-wide

studies. While we are not the first group to leverage long reads to improve gene models, especially at the crucial 3' region [26, 63], we report here that our new *Harpegnathos* gene annotations, upgraded using Iso-Seq, resulted in more accurate detection of alternative splicing events, increased sensitivity of differential gene expression analyses, and, importantly, deeper single-cell transcriptomes.

Long RNA reads for more complete gene annotations

Often, genome annotations are constructed de novo from short-read RNA-seq, with or without the guide of an existing genome assembly [1–3]. While RNA-seq-based annotations are typically sufficient to identify protein-coding mRNAs and characterize their expression levels in bulk RNA samples, they are limited in their ability to fully annotate complete transcript isoforms, extended UTRs, and lncRNAs. Because of its ability to sequence long RNA molecules in a single read, Iso-Seq overcomes these limitations and has already been used to identify novel transcriptional isoforms in various genomes [12, 27] including the very well annotated human genome [25]. Ideally, these more comprehensive isoform maps can be used to detect genes with differential splicing between biological conditions. In human cells, targeted long-read sequencing was used to examine splicing of neurexins [64], leading to the association of aberrant splicing of *NRXN1* with psychosis disorders [65].

Incomplete annotations can impede proper analysis of massively high-throughput single-cell RNA-seq, most of which is heavily biased towards the 3' end of the gene due to oligo-dT capture by beads [7]. Thus, having a correct annotation of the 3' UTR and the TTS of genes becomes crucial. Annotation of the 3' ends of genes is hampered by widespread occurrences of multiple polyadenylation and cleavage sites [66, 67] and intrinsic limitations of existing annotation methods [68]. Even in very high-quality reference genomes, such as the human genome, reads from single-cell RNA-seq can fall past the annotated 3' UTR, resulting in information loss [69]. Notably, there is no reliable strategy that we are aware of to integrate single-cell RNA-seq reads from widely used droplet-based technologies into existing models to improve 3' UTR annotations. As these reads profile only a short region at the most terminal 3' end of a gene, there is often no way to definitively link the reads to their source gene, especially if the reads map far outside an annotated termination site.

Previous studies have employed computational strategies to improve the assignment of these reads to gene models, including extending every 3' UTR by 2 kb [70] or mapping to non-overlapping windows on the genome and assigning each window to a gene based on the

proximity to an annotated TTS [71]. Although methods exist that model 3' UTR annotations based on deviation in RNA-seq coverage [68, 72], here we employed an empirical approach, based on long-read sequencing, to improve the annotation of 3' UTRs genome-wide (Fig. 1).

Splicing analysis with additional Iso-Seq information

Differential alternative splicing occurs between tissues and biological conditions [73–75]. Specific isoform usage, potentially mediated by varying RNA-binding protein expression, is widespread and affects protein-protein interaction networks [67, 76, 77]. Identification of novel isoforms with Iso-Seq revealed genes with differential splicing patterns between tissues, including one gene with a newly annotated mutually exclusive exon (Fig. 2).

Caste-specific alternative splicing has been observed in social insects. Doublesex (*dsx*) is differentially spliced between queens and workers in the ants *V. emeryi*, *S. invicta*, and *W. auropunctata* in addition to its differential splicing in non-social and social insects between males and females [78–81]. While we did not find splicing changes between worker and gamergate in *dsx*, we identified caste-specific isoform usage in the brain for *Ilp2*, a known factor in caste determination that is also differentially expressed on the gene level in the brains of many ant species including *Harpegnathos* [32]. Other ants also seem to have caste-biased splicing of the first exon, similar to *Harpegnathos*, suggesting a possible conserved mechanism that will require more investigation to understand.

Bulk and single-cell RNA-seq analysis with refined gene models

The integration of Iso-Seq into the annotation resulted in improved gene models mostly due to an extension of first and last exon, corresponding to the 5' and 3' UTR, respectively (Fig. 3A–F). Many 3' UTR extensions were confirmed by the presence of dT-seq signal, designed to capture the location of non-templated polyA tails. The new annotation captured more information from RNA-seq, with a median of 15% more reads. Increased mapping rates had immediate tangible effects on discovery, as a reanalysis of existing RNA-seq data from the worker–gamergate transition led to the identification of 7 new caste-specific genes in the brain. One of these, *egh*, was previously implicated in reproductive behavior in *Drosophila* [39], and therefore represents a high value candidate for the dissection of the molecular regulation of social behavior in ants.

The improvements for single-cell RNA-seq (Fig. 4) were even more remarkable. Mapping existing single-cell RNA-seq reads from gamergate and worker brains to the new annotation improved the median percent of

reads mapped to exons from 47 to 66%. The increased mapping resulted in 18% more cells passing the minimum UMI and gene thresholds and the addition of 266 new cell type-specific genes to our single-cell atlas of the *Harpegnathos* brain. The extraction of more information from this existing dataset provided several new insights. A number of biologically relevant genes had increased UMI mapping in the new HSAL51 annotation that translated to improved identification and visualization of cells expressing these genes, including established mushroom body markers *toy* and *GluIRB*, and the neurotransmitter-associated markers *VAcHT*, *VGlut*, and *Gad1*. Using the new annotation, more genes were classified as specific to the mushroom body or to a specific cell type. We identified *CG9259* as a new protein-coding gene specifically expressed in *Harpegnathos* ensheathing glia, a cell type previously linked to caste regulation and aging [22]. Overall, our results indicate that leveraging long reads to annotate 3' UTRs with more precision will increase the depth of existing and future single-cell RNA-seq datasets, an important consideration for many model systems, especially those with less complete annotations.

The new annotation also improved analysis of lncRNAs in the single-cell data set (Fig. 5) revealing a lncRNA that is a marker of these ensheathing glia and is expressed in a similar set of cells as its adjacent protein-coding gene *wdp*. In single-cell analysis of the *Drosophila* brain, *wdp* expression is largely restricted to ensheathing glia, although the lncRNA syntenic to the *Harpegnathos* ensheathing glia marker LOC112590028 is not detected in this data set. Further work would be required to confirm the link between LOC112590028 and *wdp* expression in *Harpegnathos*, but *wdp* expression patterns in *Drosophila* suggest a conservation in the regulation and, possibly, function of this gene.

Conclusions

The new *Harpegnathos* annotation, bolstered by Iso-Seq, allowed us to uncover new patterns of differential alternative splicing, differentially expressed genes, and markers of cell types in single-cell sequencing. The identification of new candidate genes involved in tissue-specific and caste-specific regulation of gene expression highlights the advantages of a more complete gene annotation. Future genomic studies in *Harpegnathos* will undoubtedly benefit from these improved annotations, which will help obtain new insights on the molecular regulation of their remarkable phenotypic plasticity.

Beyond the implications for *Harpegnathos* genomics, our analyses show that Iso-Seq is an effective strategy for improving incomplete gene annotations, maximizing the amount of information garnered from genome-wide sequencing data.

Methods

Ant colonies and husbandry

As previously described [22], *Harpegnathos* ants were descended from a gamergate colony collected in Karnataka, India, in 1999 and bred in various laboratories. Ant colonies were housed in plaster nests in a clean, temperature- and humidity-controlled ant facility on a 12-h light/dark cycle. Ants were fed three times per week with live crickets.

PacBio Iso-Seq

Non-visual brains and combined samples of fat body and ovary were dissected from *Harpegnathos* ants. Each sample contained tissues from ~ 8 ants from a variety of ages (5 days, 30 days, or ~ 120 days old) from both worker and gamergate castes. Brains were homogenized in TRIzol by pipetting up and down. Chloroform was added and tubes were shaken by hand before centrifugation at 12,000g for 15 min at 4 °C. The aqueous phase was transferred to a new tube and the pellet was precipitated using an equal volume of isopropanol and 1 µl glycoblue (Ambion #AM9516), with successive washes using 70% and 80% ethanol. Liquid was removed, the pellet was suspended in BTE, and DNA was removed by incubating for 30 min at 37 °C with Turbo DNase (Thermo Fisher). After the incubation, 1 mL of TRIzol was added and all purification steps described above were repeated before final suspension of the RNA in BTE.

Total RNA was submitted to the University of Washington PacBio Sequencing Services (Seattle, WA) and prepared for sequencing according to the Iso-Seq method. Briefly, the samples were enriched for polyA+ RNA via pulldown (Dynabeads mRNA Purification Kit, Thermo Fisher Scientific). The polyA+ RNA was converted to cDNA (SMARTer PCR cDNA Synthesis Kit, Clontech) and amplified with 14 (brain) or 12 (fat body/ovary) cycles of PCR. The resulting cDNAs were converted to SMRTbell libraries with the Template Prep Kit version 1 (PacBio) which involves DNA damage repair, end repair, ligation of barcoded hairpin adapters, and exonuclease digestion of imperfect templates. The material was then fractionated by size using AMPure XP beads (PacBio): a 0.4× volume of beads was added to the library to bind longer fragments to magnetic beads, then the supernatant was removed and an additional 0.6× volume of AMPure XP beads (for an effective bead buffer ratio of 1.0×) added to bind the shorter material. Each set of beads was washed twice with ethanol (80% v/v in water) and the DNA was eluted into PacBio EB. After quantitation (Qubit, Thermo Fisher Scientific) and measurement of size (Bioanalyzer, Agilent), the four components (brain and fat body/ovary, each with fractions 0.4× and 1.0×) were pooled in an equimolar

fashion for sequencing. The pool was sequenced on the Sequel II platform (PacBio) on one SMRT Cell 8 M using chemistry version 1.0 and a 30-h movie time.

RNA-seq of *Harpegnathos* worker brains (Fig. 3G)

Brains from transitioning *Harpegnathos* ants were dissected in phosphate-buffered saline, placed in TRIzol (Invitrogen #15596026) and stored at - 70 °C until RNA extraction. Each brain was processed separately. To extract RNA, thawed brains were homogenized in TRIzol by pestle, and frozen and thawed again. Chloroform was added, followed by vigorous vortexing and centrifugation at 21,000g for 10–15 min at 4 °C. The aqueous phase was purified using RNA Clean and Concentrator kit (Zymo Research #R1013) following the manufacturer's instructions. Extracted RNA was quantified using NanoDrop 2000 (Thermo Scientific) and RNA integrity was checked using High Sensitivity RNA ScreenTape (Agilent #5067-5579, 5067-5580, 5067-5581) on a TapeStation 2200 or 4200 (Agilent). 250–500 ng extracted RNA was used to prepare libraries with NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs #E7760S) following the manufacturer's instructions. Libraries were quantified using Qubit dsDNA High Sensitivity Assay Kit (Invitrogen #Q32854) on a Qubit 2.0 fluorometer (Invitrogen) and fragment size distribution was checked using High Sensitivity D1000 ScreenTape (Agilent #5067-5584, 5067-5585, 5067-5587) on a TapeStation 2200 or 4200 (Agilent). Libraries were combined in two pools and sequenced in two runs of a NextSeq 500 machine (Illumina) in High-Output mode with a 2 × 150 bp configuration.

Iso-Seq data processing and annotation construction

Full-length Iso-Seq reads were classified from circular consensus sequences using lima (Pacific Biosciences; <https://github.com/PacificBiosciences/barcoding>) and refined with isoseq3 (Pacific Biosciences; <https://github.com/PacificBiosciences/IsoSeq>) with --require-polya to filter for reads ending in a polyA tail. Full-length reads (FLNC) were clustered using isoseq3 cluster and polished using isoseq3 polish. FASTQ files output from polishing step were mapped using the STARlong module of STAR [82] to the *Harpegnathos* genome (GCF_003227715.1) with parameters suggested for mapping Iso-Seq reads to a genome provide by cDNA_Cupcake (https://github.com/Magdoll/cDNA_Cupcake/wiki/Best-practice-for-aligning-Iso-Seq-to-reference-genome:-minimap2,-deSALT,-GMAP,-STAR,-BLAT). Redundant transcript models were collapsed using TAMA Collapse [83]. Transcript models were generated separately for brain and fat body/ovary tissues and were merged together with the existing annotation produced by NCBI (GCF_003227715.1) using TAMA Merge [83], with the

no_cap option and prioritizing the brain Iso-Seq models followed by the fat body/ovary Iso-Seq models and then the existing annotation.

Several automatic and manual processing steps were performed to refine this annotation (see Additional File 1: Fig. S1C for an overview). The mitochondrial scaffold and odorant receptor were annotated manually due to the challenges of annotating these genes computationally (see below, “Manual annotation of mitochondrial scaffold and odorant receptor genes”). A list of the odorant genes added and the genes from the previous annotation that were replaced can be found in Additional File 2: Table S1. Genes “merged” in the Iso-Seq annotation (evidence suggests that two gene models should be one gene model; see example in Fig. 1E) were manually reviewed, with evidence from Iso-Seq, bulk RNA-seq, and homology of the genes in question taken into account to ensure that genes were not spuriously merged.

Due to suspicions of transcript models with retained introns representing pre-mRNAs sequenced by Iso-Seq, new transcripts with retained introns found by SUPPA [29] were subjected to further filtering. BLASTn was used to find homologs for transcript models with and without the retained intron, using a BLAST index created from all transcripts from *Drosophila melanogaster* (GCF_000001214.4), *Apis mellifera* (GCF_003254395.2), *Nasonia vitripennis* (GCF_009193385.2), and *Bombyx mori* (GCF_014905235.1) NCBI annotations. The query cover for the pairs of transcripts with or without retained introns was compared. If the transcript model without the retained intron had higher query coverage, the transcript with the retained intron was removed from the annotation. Performing the homology search with BLASTX or with additional Hymenoptera species did not substantially affect these results.

From these final gene and transcript models, Transdecoder [84] was run to find coding sequences. The longest ORFs were BLASTed (BLASTp) to a reference of proteins from *Drosophila melanogaster*, *Apis mellifera*, and *Homo sapiens* (GCF_000001405.39) with an e-value cutoff of 10^{-5} .

Manual annotation of mitochondrial scaffold and odorant receptor genes

The scaffold NW_020230424.1 was identified as the mitochondrial scaffold, as it contained the best BLAST hits of *Drosophila* mitochondrial genes. Several single-copy *Drosophila* genes had multiple hits on this scaffold and many genes were annotated as pseudogenes, necessitating a manual reannotation of these loci. We performed BLAST of the NW_020230424.1 sequence against itself and established that the positive strand of the 3' end of the scaffold aligns with the positive strand of its 5' end. This alignment pattern and the double

BLAST hits of the *D. melanogaster* genes are consistent with NW_020230424.1 being an erroneous linear assembly of ~2 iterations of the circular mtDNA sequence. Furthermore, visual examination of aligned RNA-seq reads revealed an abnormally high number of base mismatches and small indels in this scaffold. This suggests that NW_020230424.1 may have not been optimally polished when the genome was being assembled and explains why most genes contain frame shifts and/or premature stop codons and are annotated as pseudogenes.

Considering these assembly and annotation issues, we removed existing gene models in NW_020230424.1 and manually re-annotated coding genes in its 3' half. We identified regions that exhibited contiguous coverage in bulk RNA-seq, looked for predicted ORFs in such regions, and performed BLAST of their translations against *Drosophila* mitochondrial proteins. We also took into account gene order, as the mitochondrial genomes of *Drosophila* and *Harpegnathos* are completely syntenic [85]. In the instances where putative assembly errors caused frame shifts and/or premature stops, we split the genes into several fragments and added an alphabetical index to the fragments' names. For example, *Drosophila mt:Cyt-b* corresponds to *mt:Cyt-ba*, *mt:Cyt-bb*, and *mt:Cyt-bc* in *Harpegnathos*.

For the odorant receptors, we converted annotations manually generated for previous versions of the *Harpegnathos* genome [86] to the current genome assembly. We mapped the mRNAs of these old predictions to the current genome assembly using exonerate v. 2.2.0 and visually compared them to the HSAL50 annotations of the corresponding loci [87]. Where necessary, we manually updated HSAL51 predictions taking into account the exon-intron structure of the mapped annotations of Zhou et al. [86] and bulk RNA-seq coverage.

dT-sequencing

Using previously isolated RNA [9] from worker tissues (fat body, ovary, non-visual brain, and optic lobe) and gamergate tissues (fat body, ovary, and non-visual brain), a version of RNA-seq was performed to specifically capture 3' ends of transcripts with a polyA tail, here called “dT-seq” (see Additional File 1: Fig. S1E for overview). RNA was fragmented by adding 4 μ L of 5 \times 1st strand buffer from SSIII RT kit (Invitrogen catalog #18080-044) to RNA and incubating at 94 °C for 16 min, followed by a ramp down to 4 °C. PolyA⁺ selection was performed with OligodT25 dynabeads (Invitrogen catalog #610-02) with 3 washes with Oligo-dT washing buffer and eluted with BTE (10 mM bis-tris pH 6.7, 1 mM EDTA) containing either oligo-dT primer or random hexamer primers. Library construction was completed following established protocols [9] in parallel for the oligo-dT-primed

and the random hexamer-primed samples. Libraries were sequenced in paired-end mode on a NextSeq500.

dT-sequencing analysis

FASTQs from oligo-dT primed samples (“dT”) were filtered to keep only reads with at least 5 Ts with one mismatch at the 5′ end of the read, then trimmed using prinseq [88] to remove tails from reads. These reads and random hexamer reads (“hex”) were aligned to the *Harpegnathos* genome using STAR with default parameters, except `--alignIntronMax 50000`.

Hex samples were used to filter out reads coming from polyA tracts within a transcript (see Additional File 1: Fig. S1E, bottom). Genomic coverage of the first read (pair closest to polyA tail) was computed for both dT and hex samples using GenomicRanges [89]. For each “peak,” defined as a contiguous region of coverage with at least one read, tentative summits were defined as any position with coverage at least 90% of the highest read total within the peak. The “summit” of each peak was defined as the most downstream tentative summit. dT peaks were sorted into three categories: (1) dT peaks that did not overlap with a hex peak, (2) dT peaks whose peak was upstream of a hex peak—see Additional File 1: Fig. S1E, “discarded” box (3) dT peaks whose summit was downstream or equal to the hex summit—see non-discarded peaks in Additional File 1: Fig. S1E. Peaks in categories (1) and (2) were discarded, leaving a list of dT peaks that did not have hex signal downstream, indicating dT reads coming from polyA tails instead of internal polyA stretches. Reads overlapping these peaks were retained and used for further analysis and for any genome browser snapshots showing dT signal.

To verify that the retained dT reads were from the end of transcripts (see Additional File 1: Fig. S1F), aligned dT and hex reads were reduced to the first base at the 5′ end of the read using GenomicRanges. dT “peaks” were again detected using the strategy above and categorized as 3′ UTR peaks or CDS peaks based on their position within HSAL51 transcripts. Read coverage for dT and hex reads were computed at each of these peak sets.

Splicing

Isoform-level counts from *Harpegnathos* samples were generated using kallisto [90] with HSAL50 or HSAL51 annotations with any single-exon transcripts removed. For gamergate and worker brains [20] (single-end sequencing), kallisto quant was run with the parameters `-b 30 --rf-stranded--single -l 200 -s 1`. For tissue samples (non-visual brain, ovary, fat body, antenna, retina, optic lobe; paired-end) [9], kallisto quant was run with the parameters `-b 30 --rf-stranded`. Differential transcriptional usage was tested with RATS [91] with the parameters `p_`

`thresh = 0.05`, `dprop_thresh = 0.2`, `abund_thres = 5` (for tissues) or `p_thres = 0.01`, `dprop_thresh = 0.1`, `abund_thresh = 1` (for gamergate/worker brains). The number of genes with DTU using each annotation (Fig. 2C) was the number of gene with an adjusted *P* value (`padj`) $< 10^{-5}$, maximum difference in proportion between isoforms (`maxDprop`) > 0.5 , and fraction of replication iterations that support a positive DTU classification (`rep_dtu_freq_threshold`) > 0.8 between any two tissues. Proportions of transcript usage were calculated for each replicate using TPMs computed with kallisto. Proportions for each isoform were calculated by averaging replicates.

Analysis of extended transcripts and genes

Genes with additional exonic nts were identified by comparing the total number of nts covered by exons of the gene in HSAL50 and HSAL51.

Genes with extensions of their transcription start or termination sites were identified by first finding pairs of HSAL50 transcripts with a corresponding HSAL51 transcript with the same internal exon structures in the two annotations (see Fig. 3C, left), defined by transcripts with all exon boundaries the same except for the start and/or termination sites. A small number of HSAL50 transcripts had no matching transcript in the HSAL51 annotation (light gray, “no internal matches” in Fig. 3C). The start and termination sites of the remaining HSAL50 transcripts were compared to their paired HSAL51 transcripts and sorted into categories of not extended, TSS extended, TTS extended, and both TSS/TTS extended.

Bulk RNA-seq analysis

Bulk RNA-seq data for *Harpegnathos* from RNA newly sequenced from worker brains (see above) or previously published worker/gamergate brains [20, 24] or tissues [9, 31], was aligned to the genome using STAR with default parameters except `--alignIntronMax = 50000`. Previously published RNA-seq from brains of *Camponotus planatus* (PRJNA472392) [32, 37] and *Dinoponera quadriciceps* (PRJNA255520) [36, 38] was aligned with the same parameters to the *Camponotus floridanus* (GCF_003227725.1) and *Dinoponera quadriciceps* (GCF_001313825.1) genomes and annotations, respectively, as no genome assembly or annotation has been published for *Camponotus planatus*. Read counts or TPM (Additional File 1: Fig. S1B, Additional File 3: Fig. S2D and Additional File 5: Fig. S4H) were produced by an in-house script using GenomicRanges summarizeOverlaps (counting mode = union) [89] that counts the number of reads overlapping each gene model. Differential expression analysis was performed using DESeq2 [92].

Single-cell tracks shown as examples were produced by aligning previously published single-cell RNA sequencing from *Harpegnathos* workers and gamergates [22,

23] to the *Harpegnathos* genome aligned using STAR with default parameters except `--alignIntronMax = 50000` [82].

Genome browser screenshots

All genome browser screenshots were produced using IGV v2.8.6. Bigwig tracks for visualization were produced using DeepTools [93]. Sashimi plots were produced using IGV v2.8.6, with a custom scaling of splicing lines (line widths scaled to the total number of reads mapped to the locus, with a constant scaling factor used between sequencing from different castes for each technology).

Single-cell analysis

Single-cell RNA sequencing from brains of *Harpegnathos* workers and gamergates previously generated with 10x Genomics [22, 23] was reanalyzed using Cell Ranger [5] with default parameters and either the HSAL50 or HSAL51 annotations provided. Cell Ranger was used to produce digital gene expression matrices for all samples for both annotations. These matrices were processed by Seurat v3 [94]. Cells with at least 200 genes and 500 UMIs were retained, with genes required to be expressed in at least 3 cells. UMIs were log-normalized with a scale factor of 10,000, the top 2000 variable features detected using the “vst” selection method, and data were scaled so the mean of each gene across cells was 0 and variance was 1. As the samples were produced in three separate batches, the experiment was regressed out during this step. Cells were clustered using the variable features previously detected.

Clustering and visualization were performed with Seurat’s principal component analysis followed by JackStraw to detect significant principal components. All components were selected until a component had a $P > 0.05$. Cells were clustered with a resolution of 1 and clusters were visualized using UMAP. Cell type for each cluster was performed using previously established markers [22].

For cluster-level pseudobulk expression analyses (“% UMIs on gene”), the number of UMI for each gene from all cells in each cluster for each sample were added together and normalized by the total number of UMI detected in that sample and cluster.

To compare clusters from HSAL50 and HSAL51, cluster groupings were produced by comparing the cells present in each cluster, which largely stayed constant between analyses using the two annotations. In all cases, one cluster from HSAL51 corresponded to one or two clusters from HSAL50 (in two cases, one HSAL51 cluster was split into two clusters in HSAL50).

Marker genes for each of these cluster groups (Additional File 5: Fig. S4B and C) were defined using Seurat

FindAllMarkers with the parameters `only.pos = TRUE`, `min.pct = 0.25`, `logfc.threshold = 1`. Only markers with a `padj` less than 0.05 were retained. Differentially expressed genes between castes within each cluster were identified using Seurat FindMarkers, run within each cluster with worker cells and gamergate cells provided as the two identities and default parameters. Only genes with a `padj` less than 0.01 were retained.

Mushroom body markers

For Additional File 5: Fig. S4D, mushroom markers in the HSAL50 and HSAL51 analysis were found using FindMarkers with the mushroom body clusters (defined by *mub*) as one group and all other cells as another group, with the parameters `only.pos = T`, `min.pct = 0.25`, and `logfc.threshold = 1.5`. Markers detected in HSAL51 and not HSAL50 were classified as new mushroom body markers and heatmaps of their expression in cells in each cluster was plotted using Seurat DoHeatmap. Lists of mushroom body-enriched genes found by comparing *Drosophila* head to mushroom body transcriptomes (“*D. mel* MB vs head”) [43] and mushroom body markers from single-cell sequencing in *Drosophila* [44] and *Apis mellifera* [45] were taken from supplemental information of published work; for *Apis mellifera*, mushroom body clusters were defined as the clusters with *mub* as a marker gene.

Drosophila single-cell sequencing

For Fig. 5G, we utilized published single-cell RNA sequencing data from *Drosophila* [44, 95]. The x and y positions of each cell on the tSNE were specified by this object, and the position of ensheathing glia indicated was informed by the cluster identities defined in Davie et al. The expression level of *wdp* was computed by normalizing the expression levels provided; taking the \log_2 of UMIs normalized for the total UMI in each cell and multiplied by a scaling factor of 10,000.

Synteny analysis of *wdp* locus

Corresponding genes between *Drosophila*, *Apis mellifera*, and *Harpegnathos* were found using a BLASTp search; each *Harpegnathos* gene, excluding the lncRNA LOC112590028, had a best match to the gene indicated in the other two species.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-021-01188-w>.

Additional file 1: Figure S1. Statistics and methods used to create and evaluate the new *Harpegnathos* annotation. (A) Length distribution of all raw Iso-Seq subreads. (B) Transcripts per million (TPM) from short-read RNA-seq of genes with and without Iso-Seq coverage in brain and fat body/ovary. (C) Pipeline for manual annotation following combination of

Iso-Seq-based and RNA-seq-based annotations. (D) Relationship between gene models in HSAL50 and HSAL51. (E) Schematic of the dT-seq approach. RNA was chemically fragmented. PolyA+ molecules were purified and split into two reverse transcription reactions, one primed with an anchored oligo-dT primer and one with random hexamers. The resulting cDNA was assembled into libraries and sequenced. The scheme at the bottom shows that the expected read distribution in dT- and hexamer-primed reactions differs for true polyA tails and internal A-stretches. This information was used to discard peaks that did not correspond to *bona fide* TTSs (red square). (F) Expected (top) and observed (bottom) signal at dT peaks found in the CDS (let) or 3' UTR (right) from oligo-dT primed libraries ("dT", top) and random hexamer primed libraries ("hexamers", bottom).

Additional file 2: Table S1. Genes manually added to HSAL51. **Table S2.** Genes computationally added to HSAL51. **Table S3.** Genes merged in HSAL51.

Additional file 3: Figure S2. More comparisons of alternative splicing in HSAL50 and HSAL51. (A–C) Examples of a transcript with newly identified alternative splicing patterns of (A) mutually exclusive exons, (B) a skipped exon, and (C) an alternative first exon. Boxes indicate regions of the gene that is alternatively spliced. A subset of HSAL51 isoforms is shown. (D) TPM of *llp2* (*LOC105188195*) in worker ($n = 11$) and gamergate ($n = 12$) brains. Padj is from DESeq2 differential expression analysis. (E) Sashimi plot for the *llp2* gene (*LOC105257206*) in *Camponotus planatus* (RNA-seq from [32]; using *Camponotus floridanus* genome and annotation) for worker ($n = 5$) and queen ($n = 5$) brains. Splice junction line widths are scaled to the number of reads spanning the splice junction and the total number of reads mapped to *llp2* for each caste. Red boxes indicate positions of first exon for each isoform. (F) Sashimi plot for the *llp2* gene (*LOC106750697*) in *Dinoponera quadricaps* (RNA-seq from [36]) for worker ($n = 6$) and gamergate ($n = 6$) brains. Splice junction line widths are scaled to the number of reads spanning the splice junction and the total number of reads mapped to *llp2* for each caste. Red boxes indicate positions of first exon for the two major isoforms.

Additional file 4: Figure S3. Transcript extensions and RNA-seq analysis. (A) dT-seq coverage (see Additional File 1: Fig. S1E and methods) coverage at all unique TTSs in HSAL50 (gray) and HSAL51 (blue). (B) Reads mapping to *egh* in HSAL50 and HSAL51. P-value is from a paired Student's t-test.

Additional file 5: Figure S4. Additional single-cell analyses using HSAL50 and HSAL51 annotations. (A) Heatmaps of markers for neurons (*nSyb*), glia (*bd1*), mushroom body neurons (*mub*), and ensheathing glia (*Tsft1*) in HSAL51 single-cell clustering. (B) Number of marker genes (padj < 0.05, LFC > 1) for each cluster. Marker genes common to HSAL50 and HSAL51 analyses are shown in black, while markers unique to HSAL50 are in gray and markers unique to HSAL51 are in blue. (C) Scatter plot for UMI counts in HSAL50 vs. HSAL51 with marker genes highlighted according to (B). (D) Heatmap of newly identified mushroom body markers in HSAL51 (padj < 0.05 and logFC > 1.5). Arrows denote mushroom body clusters, as determined by *mub* expression (top row of heatmap). Classification of each new marker in other data sets (fly MB vs head, [43]; fly single-cell, [44]; honeybee single-cell, [45]) is indicated in heatmap to right, with black boxes indicating marker was identified as mushroom body-enriched (see methods). (E) Genome browser view showing CG9259 with RNA-seq, Iso-Seq, and single-cell coverage along with dT-seq and raw Iso-Seq reads. Scales represents counts per million. A subset of HSAL51 isoforms is shown. (F) Number of genes differentially expressed (DE) within each cluster (padj < 0.01). Differentially expressed genes common to HSAL50 and HSAL51 analyses are shown in black, while genes unique to HSAL50 are in gray and genes unique to HSAL51 are in blue. (G) Scatter plot for UMI counts in HSAL50 vs. HSAL51 with differentially expressed genes highlighted according to (F). (H) TPM of *CycG* in worker ($n = 11$) and gamergate ($n = 12$) brains. Padj is from DESeq2.

Additional file 6: Figure S5. Additional single-cell analyses of lncRNA expression. (A) Fold-change in nucleotides covered by exons of lncRNAs for each category in Fig. 5A. (B) Examples of neuronal lncRNAs detected in HSAL51 showing % of neurons, glia, and other cells expressing the indicated genes in HSAL50 and HSAL51. (C) Genome browser view showing RNA-seq signal, Iso-Seq signal, and raw Iso-Seq reads of a locus

containing two ensheathing glia marking lncRNAs which were merged into one gene model in HSAL51. Scales for RNA-seq and Iso-Seq represent counts per million. (D) Heatmap showing normalized UMI counts of the new merged lncRNA *LOC112588339.LOC112588340* in HSAL51. (E) % of UMIs mapping to *wdp* in HSAL51.

Acknowledgements

The authors thank Katy Munson and Alexandra Mackenzie at the UW PacBio Sequencing service for performing Iso-Seq; Jakub Mlejnek for technical support; Janko Gospocic for helpful discussion; Claude Desplan and Danny Reinberg for support and encouragement.

Authors' contributions

E.J.S. and R.B. designed the study. M.S. performed dT-sequencing. L.S. performed RNA extractions for Iso-Seq. B.S. and L.D. provided short-read RNA-seq of worker brains. E.J.S. performed all bioinformatic analyses with help from B.S., E.J.S. and R.B. wrote the manuscript with help from all authors. All authors read and approved the final manuscript.

Funding

This work in the Bonasio Lab was supported in part by the NIH (R21MH123841, R01AG071818) and the Max Planck-von Humboldt Research Award 2020. M.S. was supported by the Naito Foundation Grant for Studying Overseas. L.D. and B.S. were supported by NIH (R01AG058762); B.S. was supported by a Long-Term Fellowship LT000010/2020-L from the Human Frontier Science Program.

Availability of data and materials

Next-generation sequencing data generated for this study, along with the annotation in GTF format and associated transcript and peptide files, have been deposited in the NCBI GEO with accession number GSE172309 [96].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Epigenetics Institute and Department of Cell and Developmental Biology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. ²Department of Urology and Institute of Neuropathology, Medical Center-University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. ³Department of Biology, New York University, New York, NY, USA. ⁴Department of Biochemistry and Molecular Pharmacology, NYU Grossman School of Medicine, New York, NY, USA.

Received: 3 June 2021 Accepted: 10 November 2021

Published online: 27 November 2021

References

- Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* 2013;8:1494–512. <https://doi.org/10.1038/nprot.2013.084>.
- Haas BJ, Zody MC. Advancing RNA-Seq analysis. *Nat Biotechnol.* 2010;28:421–3. <https://doi.org/10.1038/nbt0510-421>.
- Martin JA, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet.* 2011;12:671–82. <https://doi.org/10.1038/nrg3068>.
- Steijger T, Abril JF, Engstrom PG, Kocokinski F, Consortium R, Hubbard TJ, et al. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods.* 2013;10:1177–84. <https://doi.org/10.1038/nmeth.2714>.
- Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun.* 2017;8:14049. <https://doi.org/10.1038/ncomms14049>.

6. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. 2015;161:1202–14. <https://doi.org/10.1016/j.cell.2015.05.002>.
7. Zhang X, Li T, Liu F, Chen Y, Yao J, Li Z, et al. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Mol Cell*. 2019;73:130–42 e5. <https://doi.org/10.1016/j.molcel.2018.10.020>.
8. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017;27:722–36. <https://doi.org/10.1101/gr.215087.116>.
9. Shields EJ, Sheng L, Weiner AK, Garcia BA, Bonasio R. High-quality genome assemblies reveal long non-coding RNAs expressed in ant brains. *Cell Rep*. 2018;23:3078–90. <https://doi.org/10.1016/j.celrep.2018.05.014>.
10. Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One*. 2015;10:e0132628. <https://doi.org/10.1371/journal.pone.0132628>.
11. Feng S, Xu M, Liu F, Cui C, Zhou B. Reconstruction of the full-length transcriptome atlas using PacBio Iso-Seq provides insight into the alternative splicing in *Gossypium australe*. *BMC Plant Biol*. 2019;19:365. <https://doi.org/10.1186/s12870-019-1968-7>.
12. Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat Commun*. 2016;7:11708. <https://doi.org/10.1038/ncomms11708>.
13. Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin N, et al. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science*. 2010;329:1068–71. <https://doi.org/10.1126/science.1192428>.
14. Thibaud-Nissen F, Souvorov A, Murphy T, DiCuccio M, Kitts P. Eukaryotic Genome Annotation Pipeline. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2013.
15. Keller L, Genoud M. Extraordinary lifespans in ants: a test of evolutionary theories of ageing. *Nature*. 1997;389:958–60. <https://doi.org/10.1038/40130>.
16. Peeters C. The occurrence of sexual reproduction among ant workers. *Biological Journal of the Linnean Society*. 1991;44:141–52. <https://doi.org/10.1111/j.1095-8312.1991.tb00612.x>.
17. Bonasio R. Emerging topics in epigenetics: ants, brains, and noncoding RNAs. *Ann N Y Acad Sci*. 2012;1260:14–23. <https://doi.org/10.1111/j.1749-6632.2011.06363.x>.
18. Peeters C, Holldobler B. Reproductive cooperation between queens and their mated workers: the complex life history of an ant with a valuable nest. *Proc Natl Acad Sci U S A*. 1995;92:10977–9. <https://doi.org/10.1073/pnas.92.24.10977>.
19. Peeters C, Liebig J, Hölldobler B. Sexual reproduction by both queens and workers in the ponerine ant *Harpegnathos saltator*. *Insectes Sociaux*. 2000;47:325–32. <https://doi.org/10.1007/pl00001724>.
20. Gospocic J, Shields EJ, Glastad KM, Lin Y, Penick CA, Yan H, et al. The neuropeptide corazonin controls social behavior and caste identity in ants. *Cell*. 2017;170(748-59):e12. <https://doi.org/10.1016/j.cell.2017.07.014>.
21. Ghaninia M, Haight K, Berger SL, Reinberg D, Zwiebel LJ, Ray A, et al. Chemosensory sensitivity reflects reproductive status in the ant *Harpegnathos saltator*. *Sci Rep*. 2017;7:3732. <https://doi.org/10.1038/s41598-017-03964-7>.
22. Sheng L, Shields EJ, Gospocic J, Glastad KM, Ratchasanmuang P, Berger SL, et al. Social reprogramming in ants induces longevity-associated glia remodeling. *Sci Adv*. 2020;6:eaba9869. <https://doi.org/10.1126/sciadv.aba9869>.
23. Bonasio R, Shields EJ. Single-cell RNA sequencing of *Harpegnathos saltator* brains. *GEO* <https://identifiers.org/geo:GSE135513>. 2020.
24. Bonasio R, Gospocic J, Shields EJ. Brain RNA-seq 120 days after worker-gamergate transitions in *Harpegnathos* ants. *GEO* <https://identifiers.org/geo:GSE83798>. 2017.
25. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol*. 2013;31:1009–14. <https://doi.org/10.1038/nbt.2705>.
26. Beiki H, Liu H, Huang J, Manchanda N, Nonneman D, Smith TPL, et al. Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. *BMC Genomics*. 2019;20:344. <https://doi.org/10.1186/s12864-019-5709-y>.
27. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, et al. A survey of the sorghum transcriptome using single-molecule long reads. *Nat Commun*. 2016;7:11706. <https://doi.org/10.1038/ncomms11706>.
28. Minio A, Massonnet M, Figueroa-Balderas R, Vondras AM, Blanco-Ulate B, Cantu D. Iso-Seq allows genome-independent transcriptome profiling of Grape Berry Development. *G3 (Bethesda)*. 2019;9:755–67. <https://doi.org/10.1534/g3.118.201008>.
29. Alamancos GP, Pages A, Trincado JL, Bellora N, Eyra E. Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA*. 2015;21:1521–31. <https://doi.org/10.1261/ma.051557.115>.
30. Van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, Love MI, et al. RNA sequencing data: hitchhiker's guide to expression analysis. *Annual Review of Biomedical Data Science*. 2019;2:139–73. <https://doi.org/10.1146/annurev-biodatasci-072018-021255>.
31. Shields EJ, Bonasio R. RNA-seq from *Harpegnathos* tissues and brain regions. *GEO* <https://identifiers.org/geo:GSE112843>. 2018.
32. Chandra V, Fetter-Prunedo I, Oxley PR, Ritger AL, McKenzie SK, Libbrecht R, et al. Social regulation of insulin signaling and the evolution of eusociality in ants. *Science*. 2018;361:398–402. <https://doi.org/10.1126/science.aar5723>.
33. Toth AL, Robinson GE. Evo-devo and the evolution of social behavior. *Trends Genet*. 2007;23:334–41. <https://doi.org/10.1016/j.tig.2007.05.001>.
34. Minn AH, Lan H, Rabaglia ME, Harlan DM, Peculis BA, Attie AD, et al. Increased insulin translation from an insulin splice-variant overexpressed in diabetes, obesity, and insulin resistance. *Mol Endocrinol*. 2005;19:794–803. <https://doi.org/10.1210/me.2004-0119>.
35. Shalev A, Blair PJ, Hoffmann SC, Hirshberg B, Peculis BA, Harlan DM. A proinsulin gene splice variant with increased translation efficiency is expressed in human pancreatic islets. *Endocrinology*. 2002;143:2541–7. <https://doi.org/10.1210/endo.143.7.8920>.
36. Patalano S, Vlasova A, Wyatt C, Ewels P, Camara F, Ferreira PG, et al. Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. *Proc Natl Acad Sci U S A*. 2015;112:13970–5. <https://doi.org/10.1073/pnas.1515937112>.
37. The Rockefeller University. RNA-Seq on the brains of queens and workers (or reproductives and non-reproductives) from multiple ant species. *BioProject*. <https://identifiers.org/bioproject:PRJNA472392>. 2018.
38. Patalano S, Vlasova A, Wyatt C, Ewels P, Camara F, Ferreira PG, et al. Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. *GEO* <https://identifiers.org/geo:GSE59525>. 2015;112(45):13970–5.
39. Soller M, Haussmann IU, Hollmann M, Choffat Y, White K, Kubli E, et al. Sex-peptide-regulated female sexual behavior requires a subset of ascending ventral nerve cord neurons. *Curr Biol*. 2006;16:1771–82. <https://doi.org/10.1016/j.cub.2006.07.055>.
40. Fahrbach SE. Structure of the mushroom bodies of the insect brain. *Annu Rev Entomol*. 2006;51:209–32. <https://doi.org/10.1146/annurev.ento.51.110104.150954>.
41. Strausfeld NJ, Hansen L, Li Y, Gomez RS, Ito K. Evolution, discovery, and interpretations of arthropod mushroom bodies. *Learn Mem*. 1998;5:11–37. <https://www.ncbi.nlm.nih.gov/pubmed/10454370>.
42. Zars T. Behavioral functions of the insect mushroom bodies. *Curr Opin Neurobiol*. 2000;10:790–5. [https://doi.org/10.1016/s0959-4388\(00\)00147-1](https://doi.org/10.1016/s0959-4388(00)00147-1).
43. Jones SG, Nixon KCJ, Chubak MC, Kramer JM. Mushroom body specific transcriptome analysis reveals dynamic regulation of learning and memory genes after acquisition of long-term courtship memory in *Drosophila*. *G3 (Bethesda)*. 2018;8:3433–46. <https://doi.org/10.1534/g3.118.200560>.
44. Davie K, Janssens J, Koldere D, De Waegeneer M, Pech U, Kreft L, et al. A single-cell transcriptome atlas of the aging *Drosophila* brain. *Cell*. 2018;174(982-98):e20. <https://doi.org/10.1016/j.cell.2018.05.057>.
45. Traniello IM, Bukhari SA, Kevill J, Ahmed AC, Hamilton AR, Naeger NL, et al. Meta-analysis of honey bee neurogenomic response links deformed wing virus type A to precocious behavioral maturation. *Sci Rep*. 2020;10:3101. <https://doi.org/10.1038/s41598-020-59808-4>.
46. Crocker A, Guan XJ, Murphy CT, Murthy M. Cell-type-specific transcriptome analysis in the *Drosophila* mushroom body reveals memory-related changes in gene expression. *Cell Rep*. 2016;15:1580–96. <https://doi.org/10.1016/j.celrep.2016.04.046>.
47. Kurusu M, Nagao T, Walldorf U, Flister S, Gehring WJ, Furukubo-Tokunaga K. Genetic control of development of the mushroom bodies, the associative learning centers in the *Drosophila* brain, by the eyeless, twin of eyeless, and Dachshund genes. *Proc Natl Acad Sci U S A*. 2000;97:2140–4. <https://doi.org/10.1073/pnas.040564497>.
48. Noveen A, Daniel A, Hartenstein V. Early development of the *Drosophila* mushroom body: the roles of eyeless and dachshund. *Development*. 2000;127:3475–88. <https://www.ncbi.nlm.nih.gov/pubmed/10903173>.

49. Aradska J, Bulat T, Sialana FJ, Birner-Gruenberger R, Erich B, Lubec G. Gel-free mass spectrometry analysis of *Drosophila melanogaster* heads. *Proteomics*. 2015;15:3356–60. <https://doi.org/10.1002/pmic.201500092>.
50. Nagel M, Qiu B, Brandenburg LE, Larsen RS, Ning D, Boomsma JJ, et al. The gene expression network regulating queen brain remodeling after insemination and its parallel use in ants with reproductive workers. *Sci Adv*. 2020;6(38):eaaz5772. <https://doi.org/10.1126/sciadv.aaz5772>.
51. Fischer P, La Rosa MK, Schulz A, Preiss A, Nagel AC. Cyclin G functions as a positive regulator of growth and metabolism in *Drosophila*. *PLoS Genet*. 2015;11:e1005440. <https://doi.org/10.1371/journal.pgen.1005440>.
52. Fischer P, Preiss A, Nagel AC. A triangular connection between Cyclin G, PP2A and Akt1 in the regulation of growth and metabolism in *Drosophila*. *Fly (Austin)*. 2016;10:11–8. <https://doi.org/10.1080/19336934.2016.1162362>.
53. Shields EJ, Petracovici AF, Bonasio R. IncRedibly versatile: biochemical and biological functions of long noncoding RNAs. *Biochem J*. 2019;476:1083–104. <https://doi.org/10.1042/BJC20180440>.
54. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. *Annual review of biochemistry*. 2012;81:145–66. <https://doi.org/10.1146/annurev-biochem-051410-092902>.
55. Bonasio R, Shiekhattar R. Regulation of transcription by long noncoding RNAs. *Annual review of genetics*. 2014;48:433–55. <https://doi.org/10.1146/aannurev-genet-120213-092323>.
56. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*. 2007;35:W345–9. <https://doi.org/10.1093/nar/gkm391>.
57. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*. 2011;27:i275–82. <https://doi.org/10.1093/bioinformatics/btr209>.
58. Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*. 2016;539:452–5. <https://doi.org/10.1038/nature20149>.
59. Takemura M, Noborn F, Nilsson J, Bowden N, Nakato E, Baker S, et al. Chondroitin sulfate proteoglycan Windpipe modulates Hedgehog signaling in *Drosophila*. *Mol Biol Cell*. 2020;31:813–24. <https://doi.org/10.1091/mbc.E19-06-0327>.
60. Huff JL, Kingsley KL, Miller JM, Hoshizaki DK. *Drosophila* windpipe codes for a leucine-rich repeat protein expressed in the developing trachea. *Mech Dev*. 2002;111:173–6. [https://doi.org/10.1016/S0925-4773\(01\)00609-8](https://doi.org/10.1016/S0925-4773(01)00609-8).
61. Kurusu M, Cording A, Taniguchi M, Menon K, Suzuki E, Zinn K. A screen of cell-surface molecules identifies leucine-rich repeat proteins as key mediators of synaptic target selection. *Neuron*. 2008;59:972–85. <https://doi.org/10.1016/j.neuron.2008.07.037>.
62. Williams-Simon PA, Posey C, Mitchell S, Ng'oma E, Mrkvicka JA, Zars T, et al. Multiple genetic loci affect place learning and memory performance in *Drosophila melanogaster*. *Genes Brain Behav*. 2019;18:e12581. <https://doi.org/10.1111/gbb.12581>.
63. Wang K, Wang D, Zheng X, Qin A, Zhou J, Guo B, et al. Multi-strategic RNA-seq analysis reveals a high-resolution transcriptional landscape in cotton. *Nat Commun*. 2019;10:4714. <https://doi.org/10.1038/s41467-019-12575-x>.
64. Treutlein B, Gokce O, Quake SR, Sudhof TC. Cartography of neuroligin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proc Natl Acad Sci U S A*. 2014;111:E1291–9. <https://doi.org/10.1073/pnas.1403244111>.
65. Flaherty E, Zhu S, Barretto N, Cheng E, Deans PJM, Fernando MB, et al. Neuronal impact of patient-specific aberrant NRXN1alpha splicing. *Nat Genet*. 2019;51:1679–90. <https://doi.org/10.1038/s41588-019-0539-z>.
66. Beaulieu E, Gautheret D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res*. 2001;11:1520–6. <https://doi.org/10.1101/gr.190501>.
67. Wang ET, Sandberg R, Luo S, Khrebukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456:470–6. <https://doi.org/10.1038/nature07509>.
68. Shenker S, Miura P, Sanfilippo P, Lai EC. IsoSCM: improved and alternative 3' UTR annotation using multiple change-point inference. *RNA*. 2015;21:14–27. <https://doi.org/10.1261/rna.046037.114>.
69. Ntranos V, Yi L, Melsted P, Pachter L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat Methods*. 2019;16:163–6. <https://doi.org/10.1038/s41592-018-0303-9>.
70. Sebe-Pedros A, Saudemont B, Chomsky E, Plessier F, Mailhe MP, Renno J, et al. Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-Seq. *Cell*. 2018;173:1520–34 e20. <https://doi.org/10.1016/j.cell.2018.05.019>.
71. Sa JM, Cannon MV, Caleon RL, Welles TE, Serre D. Single-cell transcription analysis of *Plasmodium vivax* blood-stage parasites identifies stage- and species-specific profiles of expression. *PLoS Biol*. 2020;18:e3000711. <https://doi.org/10.1371/journal.pbio.3000711>.
72. Huang Z, Teeling EC. ExUTR: a novel pipeline for large-scale prediction of 3'-UTR sequences from NGS data. *BMC Genomics*. 2017;18:847. <https://doi.org/10.1186/s12864-017-4241-1>.
73. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol*. 2017;18:437–51. <https://doi.org/10.1038/nrm.2017.27>.
74. Streuli M, Saito H. Regulation of tissue-specific alternative splicing: exon-specific cis-elements govern the splicing of leukocyte common antigen pre-mRNA. *EMBO J*. 1989;8:787–96. <https://www.ncbi.nlm.nih.gov/pubmed/2524382>.
75. Xu Q, Modrek B, Lee C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res*. 2002;30:3754–66. <https://doi.org/10.1093/nar/gkf492>.
76. Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, et al. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell*. 2012;46:871–83. <https://doi.org/10.1016/j.molcel.2012.05.039>.
77. Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell*. 2012;46:884–92. <https://doi.org/10.1016/j.molcel.2012.05.037>.
78. Burtis KC, Baker BS. *Drosophila* doublesex gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides. *Cell*. 1989;56:997–1010. [https://doi.org/10.1016/0092-8674\(89\)90633-8](https://doi.org/10.1016/0092-8674(89)90633-8).
79. Cho S, Huang ZY, Zhang J. Sex-specific splicing of the honeybee doublesex gene reveals 300 million years of evolution at the bottom of the insect sex-determination pathway. *Genetics*. 2007;177:1733–41. <https://doi.org/10.1534/genetics.107.078980>.
80. Mine S, Sumitani M, Aoki F, Hatakeyama M, Suzuki MG. Identification and functional characterization of the sex-determining gene doublesex in the sawfly, *Athalia rosae* (Hymenoptera: Tenthredinidae). *Appl Entomol Zool*. 2017;52:497–509. <https://doi.org/10.1007/s13355-017-0502-3>.
81. Miyakawa MO, Tsuchida K, Miyakawa H. The doublesex gene integrates multi-locus complementary sex determination signals in the Japanese ant, *Vollenhovia emeryi*. *Insect Biochem Mol Biol*. 2018;94:42–9. <https://doi.org/10.1016/j.ibmb.2018.01.006>.
82. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
83. Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, et al. Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics*. 2020;21:751. <https://doi.org/10.1186/s12864-020-07123-7>.
84. Haas B, Papanicolaou A. Transdecoder. <http://transdecoder.github.io>.
85. Garesse R. *Drosophila melanogaster* mitochondrial DNA: gene organization and evolutionary considerations. *Genetics*. 1988;118:649–63. <https://doi.org/10.1093/genetics/118.4.649>.
86. Zhou X, Slone JD, Rokas A, Berger SL, Liebig J, Ray A, et al. Phylogenetic and transcriptomic analysis of chemosensory receptors in a pair of divergent ant species reveals sex-specific signatures of odor coding. *PLoS Genet*. 2012;8:e1002930. <https://doi.org/10.1371/journal.pgen.1002930>.
87. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;6:31. <https://doi.org/10.1186/1471-2105-6-31>.
88. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011;27:863–4. <https://doi.org/10.1093/bioinformatics/btr026>.
89. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. *PLoS Comput Biol*. 2013;9:e1003118. <https://doi.org/10.1371/journal.pcbi.1003118>.
90. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34:525–7. <https://doi.org/10.1038/nbt.3519>.
91. Froussios K, Mourao K, Simpson G, Barton G, Schurch N. Relative Abundance of Transcripts (RATs): identifying differential isoform abundance from RNA-seq. *F1000Res*. 2019;8:213. <https://doi.org/10.12688/f1000research.17916.1>.

92. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550. <https://doi.org/10.1186/s13059-014-0550-8>.
93. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 2016;44:W160–5. <https://doi.org/10.1093/nar/gkw257>.
94. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol.* 2018;36:411–20. <https://doi.org/10.1038/nbt.4096>.
95. Davie K, Janssens J, Koldere D, Aerts S. A single-cell transcriptome atlas of the ageing *Drosophila* brain. GEO <https://identifiers.org/geo:GSE107451>. 2018.
96. Shields EJ, Bonasio R. Harpegnathos saltator RNA-seq and Iso-Seq. GEO <https://identifiers.org/geo:GSE172309>. 2021.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

