Published online 20 January 2021

doi: 10.1093/nargab/lqaa113

The regulatory genome of the malaria vector *Anopheles gambiae*: integrating chromatin accessibility and gene expression

José L. Ruiz[1], Lisa C. Ranford-Cartwright[2] and Elena Gómez-Díaz[1,*]

[1]Instituto de Parasitología y Biomedicina López-Neyra (IPBLN), Consejo Superior de Investigaciones Científicas, 18016 Granada, Spain and [2]Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Science, University of Glasgow, Glasgow G12 8QQ, UK

Received September 21, 2020; Revised December 15, 2020; Editorial Decision December 22, 2020; Accepted December 26, 2020

ABSTRACT

***Anopheles gambiae* mosquitoes are primary human malaria vectors, but we know very little about their mechanisms of transcriptional regulation. We profiled chromatin accessibility by the assay for transposase-accessible chromatin by sequencing (ATAC-seq) in laboratory-reared *A. gambiae* mosquitoes experimentally infected with the human malaria parasite *Plasmodium falciparum*. By integrating ATAC-seq, RNA-seq and ChIP-seq data, we showed a positive correlation between accessibility at promoters and introns, gene expression and active histone marks. By comparing expression and chromatin structure patterns in different tissues, we were able to infer *cis*-regulatory elements controlling tissue-specific gene expression and to predict the *in vivo* binding sites of relevant transcription factors. The ATAC-seq assay also allowed the precise mapping of active regulatory regions, including novel transcription start sites and enhancers that were annotated to mosquito immune-related genes. Not only is this study important for advancing our understanding of mechanisms of transcriptional regulation in the mosquito vector of human malaria, but the information we produced also has great potential for developing new mosquito-control and anti-malaria strategies.**

INTRODUCTION

Chromatin structure is the basal element determining dynamic regulatory landscapes, i.e. the set of regulatory sequences, and the proteins binding to them, that control the definition of phenotypes during development, and in response to the external environment, in metazoa (1). *Cis*-regulatory elements (CREs) are regions of non-coding DNAs capable of regulating transcription. For example, accessibility at promoters allows for the interaction of transcription factors (TFs) with their cognate motifs, recruiting other co-factors involved in chromatin remodeling and transcriptional activation, which enables the spatiotemporal control of gene expression (2–4). Additionally, transcriptional enhancers work in concert with the core promoter in regulating gene expression, acting as a scaffold for the recruitment of TFs and chromatin-modifying enzymes (5–8). Other relevant regulatory regions are insulators that typically work in long-range distances and contain binding sites for specific TFs, such as the CCCTC-binding factor (CTCF) (9). Chromatin structure and accessibility at these regulatory regions can also influence alternative splicing (10). The regulation of this process often involves intronic or exonic CREs that are bound by DNA-binding proteins that interfere with RNA polymerase II transcriptional elongation or associate with enhancers or silencers in a time- and tissue-specific manner (11–14).

Based on these fundamental principles, variable levels of chromatin accessibility at regulatory regions are expected to reflect the level of transcriptional activity at a given tissue or condition and time point. It is also expected that these active regulatory sites display a typical pattern of histone post-translational modifications (hPTMs) characteristic of active chromatin (15). As a consequence, chromatin accessibility can be used as a proxy to globally identify active promoters and enhancers, and to predict gene activity (16,17). By profiling genome-wide chromatin accessibility in several model organisms, such as the fruit fly *Drosophila melanogaster*, recent studies have mapped thousands of *cis*- and *trans*-regulatory elements, defining their functional roles in the regulation of genes involved in processes such as development, physiology and disease (18,19). For example, there are more than 40 000 Drosophila enhancers and target genes described in the EnhancerAtlas

[*]To whom correspondence should be addressed. Tel: +34 958 181 621; Fax: +34 958 181 633; Email: elena.gomez@csic.es
Present address: Instituto de Parasitología y Biomedicina López-Neyra (IPBLN), Consejo Superior de Investigaciones Científicas, 18016 Granada, Spain.

database (19). These enhancers, in general, have been shown to modulate the transcription levels of several target genes, regardless of orientation or distance to the target promoters (5,20,21), and they can be located between and within genes, i.e. within introns or exons (6,18,22).

Compared to the knowledge accumulated on transcriptional regulation in the fruit fly, little is known about the regulatory genome of other insects, such as mosquitoes. This is despite the major role of these arthropod vectors in the transmission of important human infectious diseases. Among mosquito-borne diseases, malaria is the deadliest and the one with the highest global health and economic burden (23). Human malaria is transmitted by *Anopheles* mosquitoes, with members of the species complex *A. gambiae* recognized as the main vectors in Africa (24). Controlling and targeting vector populations is key in the ongoing efforts to fight malaria, but further progress toward alternative molecular-based approaches has been hampered by the lack of epigenetic and functional genomic studies in mosquitoes (25,26). Indeed, the regulatory genome of Anopheles remains practically unexplored and the regulatory networks of most genes in the *A. gambiae* genome are unknown, including the genes involved in important biological processes such as mosquito insecticide resistance and immunity. The vast majority of *cis*-regulatory sequences reported to date in mosquitoes are computational predictions without experimental verification (27–34). For example, less than a dozen enhancer sequences have been experimentally validated in *A. gambiae* (30,35), a negligible number compared with the functionally annotated enhancers that are publicly available in several *Drosophila* databases (18,19,36,37). Other aspects of the genomics of mosquitoes that remain understudied from a mechanistic perspective are the *trans*-acting factors that bind unknown CREs, including, for example, insulator elements like CTCF.

To fill this important gap, here we used the assay for transposase-accessible chromatin by sequencing (ATAC-seq) (38,39) in *A. gambiae* to further characterize mosquito gene regulatory networks *in vivo*. This analysis of chromatin accessibility was necessary for the genome-wide identification of promoter regions and enhancers, as well as the prediction of TF binding events (2). In the research presented here, we performed ATAC-seq and RNA-seq analyses of both *A. gambiae* midguts (MGs) and salivary glands (SGs) infected with *Plasmodium falciparum*, and integrated such datasets with ChIP-seq data for various histone modifications (H3K9ac, H3K4me3, H3K27ac and H3K9m3) (34). We report a genome-wide association between chromatin accessibility, epigenetic states and tissue-specific regulation of transcription. Analyses of DNA-binding motif enrichment of active regulatory regions allowed us to predict binding sites similar to several Drosophila TFs, many of which have been functionally validated. Furthermore, we provide a comprehensive map of enhancer-, transcription start site (TSS)- and CTCF-like novel regulatory sequences, which are conserved with those previously characterized in the fruit fly, and appear to be active in the mosquito. Our results provide a more complete annotation of the regulatory genome of the major vector of human malaria, and add new insights into mechanisms controlling mosquito functional gene expression.

MATERIALS AND METHODS

Mosquito rearing and experimental infections

Five-day-old female *A. gambiae* s.s. Kisumu mosquitoes from a genetically outbred laboratory colony in the University of Glasgow were used for experimental infections. Mosquitoes were maintained under standard insectary conditions (27 ± 2°C, 80 ± 5% relative humidity, 12:12 LD). Females were fed through membrane feeders on blood containing gametocytes of *P. falciparum* clone 3D7, prepared according to standard protocols (40). Thereafter the mosquitoes were given access to a solution of 5% glucose/0.05% 4-aminobenzoic acid *ad libitum*. Three independent experimental infections (Infections 1, 2 and 3) were carried out. Prevalence (percentage of infected mosquitoes) and intensity of infection (median number of oocysts) are described in Supplementary Table S1. We performed dissection of MGs on adults at 7 days post-infection and of SGs at 14 days post-blood meal.

ATAC-seq library preparation and sequencing

We performed the ATAC-seq protocol using fresh MGs and SGs from ∼20 individual mosquitoes from two independent infections (Supplementary Table S1). Mosquito tissues were resuspended in lysis buffer (38,39) to permeabilize membranes. The nuclei pellet was resuspended in the transposition reaction mix (25 μl of 2 × TD Buffer, 1.25 μl of Tn5 Transposase and 23.75 μl of nuclease free water), and incubated for 30 min at 37°C. All samples were purified using the Qiagen MiniElute Kit. Library amplification was carried out with 2× KAPA HiFi mix and 1.25 μM of Nextera primers (38,39). The optimal cycle number was determined by quantitative polymerase chain reaction with conditions as originally described in (38,39). ATAC-seq libraries were sequenced at BGI (China) with an Illumina HiSeq4000 sequencer to obtain 25–37 M of 2 × 50 bp paired-end reads (Supplementary Table S2).

RNA isolation, RNA-seq library preparation and sequencing

We prepared RNA-seq libraries from *P. falciparum*-infected MGs and SGs obtained in two independent experimental infections (Infections 2 and 3; Supplementary Table S1). Immediately after dissection, tissues were stored in TRIzol (Invitrogen) and frozen at −80°C until subsequent processing. Total RNA was extracted from a pool of ∼30 MGs and a pool of ∼60 SGs from 30 mosquitoes using the TRIzol manufacturer protocol. RNA concentration was quantified using a Qubit® 2.0 Fluorometer, and RNA integrity was determined with an Agilent 2100 Bioanalyzer. We used the Ovation® Universal RNA-seq System (Nugen Technologies) for strand-specific RNA-seq library construction following the manufacturer instructions. Custom primers specific to mosquito ribosomal sequences were designed to reduce the percentage of ribosomal reads in the sample and the ribo-depletion step was incorporated into the standard workflow (Supplementary Table S3). Libraries were sequenced at Cabimer (Spain) using an Illumina NextSeq500 for both 2 × 150 bp paired-end and 1 × 75 bp single-end reads.

ATAC-seq data processing and analyses

We conducted ATAC-seq data analyses according to the recommendations of the ENCODE Pipeline (https://www.encodeproject.org/atac-seq/, (41)). First, raw reads were trimmed 20 bases from the 3′ end of each read (−3 20), and aligned to the *A. gambiae* PEST genome (AgamP4) with Bowtie2 (42) (v2.4.1) using default parameters, except for –no-unal –no-mixed -X 2000. We then applied a MAPQ score threshold of 10 and sorted and deduplicated the reads using SAMtools (43) (v1.10). To adjust the known bias and ensure the mapping of Tn5 cutting sites, we shifted aligned reads +4 bp for + strands and −5 bp for − strands with ATACSeqQC (44) (v1.10). We removed not properly paired reads and extracted nucleosome-free fragments with a size threshold of 130 bp (SAMtools). We performed peak-calling on nucleosome-free reads with MACS2 (45) (v.2.1.2) *callpeak* module and the following parameters: -f BAMPE -g 273109044 -q 0.01 -B –keep-dup all –nomodel –nolambda. We refer to ATAC-seq peaks as Tn5 hypersensitive sites (THSs). To identify THSs unique or common across samples, we used BEDTools *intersect* (46) requiring a minimum overlap of 51% (-f/-F 0.51). We annotated THSs to genomic features combining HOMER (47) (v4.11) and ChIPseeker (48) (v1.22) (see Supplementary Methods). We used the AgamP4.12 gene set (49) and considered the first coordinate of the 5′ UTRs as the TSSs. For genes without an annotated 5′ UTR we took the translation start site (ATG codon) as the reference point. In either case, the 1 Kb upstream window was considered as the putative promoter region. The mean length of the annotated 5′ UTRs was 253 bp, and the higher density in the distribution of values is ∼100 bp, so even if some genes do not have annotated 5′ UTRs, we expect a 1 Kb window to capture the whole promoter (Supplementary Figure S5A). Indeed, by using BEDTools *intersect* we observed that 52.2% of the THSs categorized as promoters (18 448 out of 35 323) overlap with chromatin states that are enriched in H3K4me3, which is distinctive of active promoters (34) (Supplementary Table S4). Nevertheless, we cannot rule out the possibility that a small portion of THSs might be misannotated, in particular for genes without annotated 5′ UTRs. Metadata for the annotated genes was obtained from the Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Pfam databases using DAVID (50) and PANTHER (51), and the WikiPathways (52) and ImmunoDB (53) databases (Supplementary Table S4). To quantify the ATAC-seq nucleosome-free signal enrichment at genomic regions of interest, we used BEDTools *intersect* -c. The read counts were normalized (RPKM) and we added a pseudocount (0.1) when required to get finite values. We categorized genes into high, medium or low groups based on promoter chromatin accessibility with threshold values determined by dividing the signal in three quantile groups according to their means (*Hmisc::cut2* R function). We extracted mononucleosomal reads applying a 171–256 bp size threshold (SAMtools) and predicted nucleosome dyads using NucleoATAC (54) (v0.3.4) (see Supplementary Methods).

RNA-seq data processing and analyses

We trimmed adapters from the raw reads using BBDuk (v38.79) with -tbo -tpe -minlength = 35 parameters and removed rRNA contamination using SortMeRNA (55) (v4.2) with default parameters. Apart from the default rRNA databases, we used additional large and small subunit *A. gambiae* and *P. falciparum* rRNA sequences from the SILVA database (56,57) (LSU r132/SSU r138). Cleaned directional RNA-seq reads were mapped against the AgamP4 v2.00 reference genome using STAR (58) (v2.7.3a). Two different sets of reads were available, 2 × 150 and 1 × 75 bp, which were processed in parallel until we combined the raw counts. Raw counts at the gene level were obtained using CoCo (59) (v.0.2.2) and then provided to DESeq2 (60) (v1.26) for normalization (see Supplementary Methods). Correlation and PCA plots by deepTools2 (61) (v3.4.1) showed higher similarity between infections than between tissues and no clustering based on the sequencing approach (Supplementary Figure S5B and C), which is consistent with results using summed counts. Normalized counts were comparable between infections and higher in SGs (Supplementary Figure S5D). We also observed high correlation (∼75%) between the RNA-seq data in this study and data from our previous study (34) (Supplementary Figure S5E). We categorized normalized counts in high, medium or low groups as described above (*Hmisc::cut2* R function). Normalized RNA-seq counts (DESeq2) for each gene are included in Supplementary Table S5.

Integration of ATAC-seq, RNA-seq and ChIP-seq data

We used ChromHMM (62) (v1.2) to compute genome-wide chromatin state predictions on the *A. gambiae* genome based on ATAC-seq nucleosome-free signal enrichment levels and hPTMs data from our previous study (34) (see Supplementary Methods). This analysis was performed only on MGs for which hPTMs data was available.

For the correlation and integrative analyses of ATAC-seq, ChiP-seq and RNA-seq data, we restricted the analysis to a subset of 8245 genes that harbor high-confidence regulatory regions. We first discarded genes with very close adjacent genes encoded on opposing strands (with upstream promoters likely overlapping) and genes with the gene bodies or putative promoters overlapping or embedded into many other gene bodies or promoters (>2). Following their categorization into high, medium and low accessibility groups (see above) we identified genes displaying different patterns of regulation: activating when the change in accessibility is in the same direction, and potential repressor events when the promoter is accessible but the gene is weakly expressed or silent.

Differential ATAC-seq and RNA-seq analyses

We used the DiffBind R package (63) (v2.14) to assess differential chromatin accessibility at given locations between *P. falciparum*-infected mosquito MGs and SGs. As input for DiffBind, we used the ATAC-seq nucleosome-free reads and the THSs. Infection 1 and Infection 2 were used as biological replicates. Differential gene expression analyses between infected mosquito MGs and SGs were conducted using the DESeq2, edgeR (64) (v3.28.1) and DREAMSeq (65) (v1.0.4) R packages to identify differentially expressed genes (DEGs) (Supplementary Table S6). We used the IsoformSwitchAnalyzeR R package (66) (v1.8) to analyze differential gene isoforms expression between MGs and SGs. See Supplementary Methods for more information and details on these analyses.

Characterization of novel regulatory elements

To map active regulatory sites, we used *A. gambiae* enhancers predicted computationally by others from Drosophila enhancers ($N = 1628$), or from Drosophila enhancer motifs ($N = 51$) (29,30), as well a few Anopheles enhancers identified previously by STARR-seq ($N = 6$) (35) (see Supplementary Methods). Next, to generate a set of novel candidates, we downloaded *D. melanogaster* collections of enhancers (19), including some activity-based enhancer-target gene assignments (18). We then used the UCSC LiftOver tool webserver (https://genome.ucsc.edu/cgi-bin/hgLiftOver, (67)), which uses homology data between species, to transfer coordinates to *A. gambiae* (see Supplementary Methods). In both sets of enhancers, we checked the overlap of the novel enhancer-like elements with our THSs using BEDTools *intersect* and incorporated the annotation of the THSs (see above and Supplementary Methods). Of these, we considered to be active enhancers those that, apart from being accessible, overlapped with chromatin states enriched in H3K27ac. The target genes for the previous enhancers were obtained from the previous studies (see above) and we obtained *A. gambiae* orthologs when needed using FlyBase (68), VectorBase (49) and OrthoMCL (69). We considered proximal enhancer-like regions those annotated by our approach to the same genes than the original targets in Anopheles, or any corresponding ortholog in Drosophila. To correct our annotation based on the nearest neighboring gene, we included the target genes identified by others when we could obtain an unambiguous single target gene from the published datasets. We considered a regulatory region to be potentially distal if located >2 Kb away from the promoter of the target gene. To explore the relationship between the chromatin accessibility at these regions and expression of the annotated genes, we quantified the ATAC-seq signal at the THSs-overlapping enhancer-like regions as described above.

To discover novel TSS-like sites, Drosophila TSSs were downloaded from the Eukaryotic Promoter Database (Release 128–005, (70)) and as we did for enhancers, we used a homology-based approach to transfer coordinates to the *A. gambiae* genome. We then checked annotation to genomic features (HOMER), overlap with chromatin states characteristic of TSSs/promoters (H3K4me3 enrichment), overlap with THSs (BEDTools *intersect*) and whether the annotated genes already have 5′ UTRs/TSS annotated, in order to classify these elements as novel.

To characterize CTCF binding sites, we downloaded binding sites for Drosophila CTCF from the ChIP-Atlas (71). As described above, we transferred the coordinates from *D. melanogaster* to the *A. gambiae* genome based on homology.

Motif enrichment analysis

We performed *de novo* motif analysis using HOMER. We applied this pipeline to different sets of THSs: activating or repressor, depending on the pattern of promoter accessibility and gene expression of the annotated gene. This analysis was conducted separately on the set of DiffBind regulatory regions at different locations and on the set of enhancer-like elements (see ‘Results’ section). We used the *findMotifsGenome.pl* module considering the THSs summit and searched for motifs in 100 bp in each direction (−size −100,100). See Supplementary Methods for additional details on this analysis.

RESULTS

Chromatin accessibility correlates with active transcription and epigenetic states

ATAC-seq libraries were produced from adult female *A. gambiae* s.s. mosquitoes infected with the *P. falciparum* parasite clone 3D7. Mosquito midguts (MG) were dissected at 7 days post-infection, and salivary glands (SGs) were dissected at 14 days post-infection, in two independent experimental infections (see ‘Materials and Methods’ section; Supplementary Table S1). These post-infection dates coincide with the presence of *P. falciparum* oocysts (7 days) and sporozoites (14 days) in the mosquito MGs and SGs, respectively. After sequencing, the quality of the ATAC-seq data obtained was high and comparable between tissues and experimental infections (Figure 1A). Reproducibility analyses revealed a higher correlation of ATAC-seq between infections of the same tissue than between tissues (PCA, Supplementary Figure S1A and B). Other quality control measurements, such as the fraction of reads mapping to the mitochondria, or the library complexity coefficients, also indicated the high quality of the ATAC-seq data according to ENCODE standards (Supplementary Table S2) (41). The fragment length distributions for both tissues conformed to previous observations (38,39); most insert sizes corresponded to nucleosome-free regions of less than 130 bp, and a second peak of fragment sizes represented mononucleosomes (Figure 1A). The distribution of the ATAC-seq nucleosome-free signal also matched the typical profiles of higher eukaryotes (44,72,73), with a higher density of insertions at the TSSs and two smaller peaks marking the spacing between adjacent nucleosomes (Figure 1B and Supplementary Figure S1C). By contrast, the highest density of ATAC-seq mononucleosome signal localized to positions flanking the TSSs at the ±1 nucleosome positions (Figure 1B).

Once we had validated the ATAC-seq approach, we used the MACS2 peak-calling software (45) to identify regions significantly enriched in ATAC-seq nucleosome-free signal, which we denote THSs. The computed THSs for all samples are listed in Supplementary Table S4 and include a total of 111 586 unique accessible regulatory regions (43 010 in MGs and 68 576 in SGs) that were present in both experimental infections. The total number of THSs differed

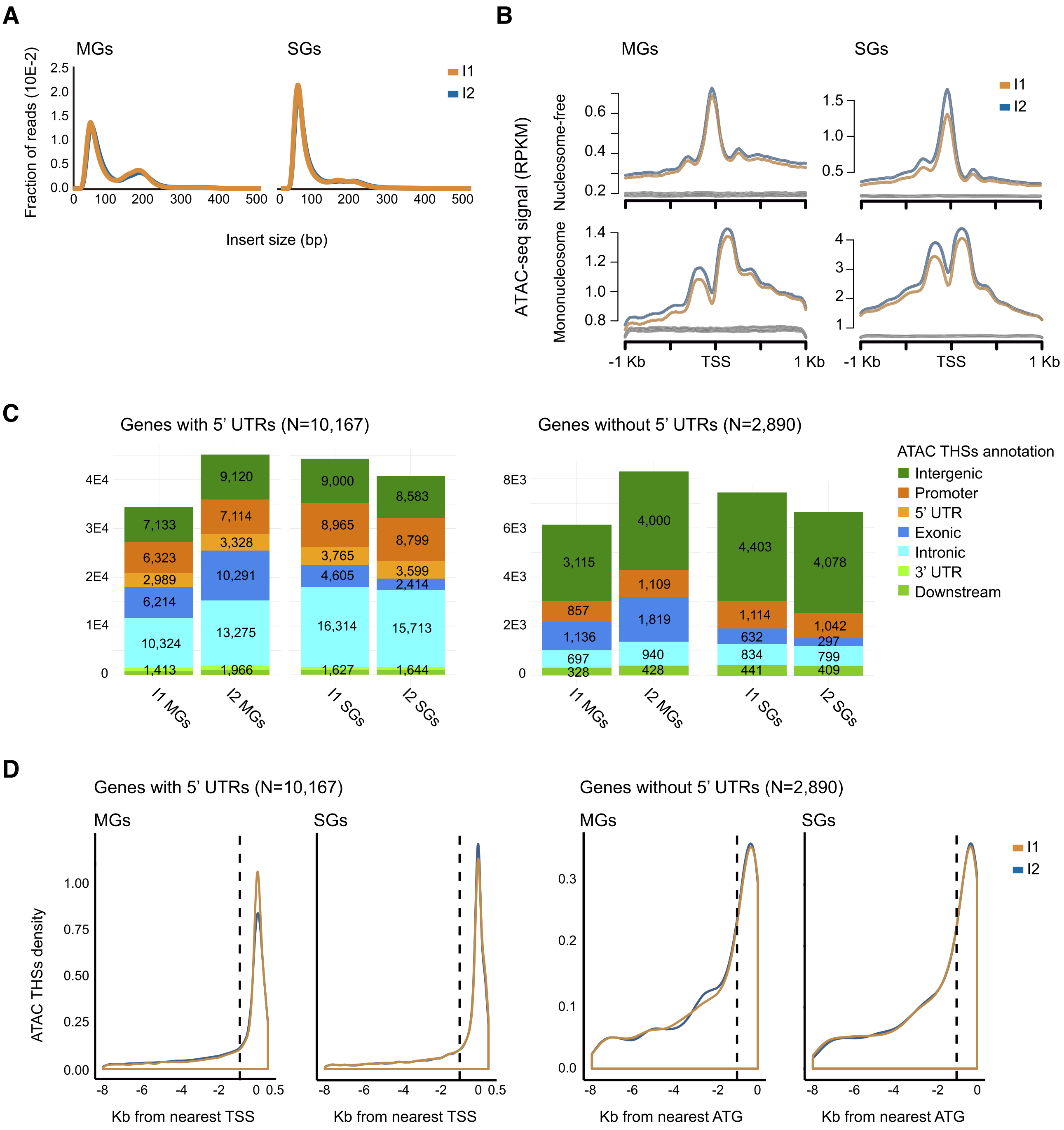

Figure 1. ATAC-seq allows for the genome-wide profiling of chromatin accessibility and nucleosome occupancy in *Anopheles gambiae*. (**A**) ATAC-seq fragment size distribution corresponding to *Plasmodium falciparum*-infected *A. gambiae* MGs and SGs. I1 and I2 are biological replicates (independent infections). A large proportion of reads are <100 bp, which represents the nucleosome-free region. The plot also shows a clear periodicity, which is indicative of nucleosome occupancy. To filter THSs (nucleosome-free regions) and mononucleosomes, we selected reads in the ranges of 50–130 and 171–256 bp, respectively. (**B**) Average profile plots of normalized (RPKM) ATAC-seq nucleosome-free and mononucleosomal reads surrounding *A. gambiae* annotated TSSs (±1 Kb). Higher mononucleosomal signals flank the nucleosome-free region at TSSs. Profiles in gray represent read density at random genomic coordinates. (**C**) Annotation of THSs to features genome-wide: intergenic regions, promoters, 5′ UTRs, exons, introns, 3′ UTRs and downstream regions. Most THSs were annotated to introns. (**D**) Density plot showing the position of THSs with respect to the TSSs (or ATG start codons for genes without annotated 5′ UTRs). Higher densities of THSs occur within 1 Kb upstream the TSSs or ATG sites. The dashed lines indicate the putative promoter region located 1 Kb upstream.

slightly between tissues (see 'Materials and Methods' section), but we observed that average ATAC-seq enrichment levels at THSs were similar across samples (Supplementary Figure S1D and E). These THSs annotated to >10 000 genes (of a total of 13 057 genes in *A. gambiae*). Approximately 20% of the total *A. gambiae* genes do not have 5′ UTRs/TSS annotated, and the promoter for those genes is assumed to lie in the 1 Kb region upstream of the ATG start codon (see 'Materials and Methods' section). The annotation of the THSs to genomic features showed a preferential location within introns (30.5%), followed by promoters (18.3%), exons (14.2%) and 5′ UTRs (7.1%) (Figure 1C and Supplementary Table S4). The concentration of accessible sites within introns has been previously reported, for example in *D. melanogaster*, with around 50% of ATAC-seq THSs found within introns (74). The density of THSs at promoters diminishes with distance from the TSS. For genes without annotated 5′ UTRs, THSs density also decreases with distance from the ATG start codon of the gene (Figure 1D).

In addition to THSs, ATAC-seq data is also informative about patterns of nucleosome positioning. The pattern for metazoans is such that the 5′ promoter appears to be nucleosome-free and the transcribed regions are occupied by a more periodic array of positioned nucleosomes (75,76). Similarly, in our study we observed a larger proportion of mononucleosome signal at exons and introns compared to promoters and 5′ UTRs (Supplementary Figure S1F). To validate this more systematically, we used the NucleoATAC nucleosome calling software (54) to predict nucleosome dyads positions (symmetry axis of the nucleosomal DNA), and we found dyads within 100 bp windows of around 90% of the THSs (Supplementary Table S7).

Certain hPTMs are known to play crucial roles in the remodeling of chromatin structure and transcriptional regulation, with *a priori* well-established activating (H3K9ac, H3K27ac and H3K4me3) or repressor (H3K9me3) functions (77,78). Activating hPTMs are expected to be enriched at mononucleosomes flanking active and accessible regulatory regions, like TSSs, whereas repressor histone marks are associated with non-accessible and nucleosome-occupied heterochromatin. To test such a relationship between chromatin accessibility and epigenetic states, we examined ATAC-seq signal relative to ChIP-seq peaks for various hPTMs that we obtained from our previous study on *P. falciparum*-infected and non-infected *A. gambiae* MGs (34). ATAC-seq nucleosome-free signals appeared more enriched (compared to a random distribution) at peaks of H3K9ac, H3K27ac and to a lesser extent H3K4me3, and depleted at H3K9me3 peaks (Supplementary Figure S1G). Next, we applied the segmentation algorithm ChromHMM (62), that integrates the ATAC-seq and ChIP-seq data, to partition the genome. The purpose was to detect recurring epigenetic/accessibility patterns genome-wide and then assign a state to each region in the genome. This analysis resulted in eight chromatin states (Supplementary Figure S2A and Table S8). Most of the genome (∼65%) appeared to be in a depleted state without ATAC-seq or ChIP-seq signal. Intergenic regions were largely depleted of ATAC-seq nucleosome-free signal and enriched in H3K4me3 and H3K9me3. By contrast, 5′ UTRs and promoters displayed states of open-chromatin, and activating hPTMs (H3K9ac, H3K27ac and/or H3K4me3). This was also the case for introns and exons, which also appeared in an open state (Figure 2A). Nearly half of the intronic THSs coincided with H3K27ac-enriched chromatin states (states 5/7; Supplementary Figure S2A) and only 7.5% with the chromatin state H3K9ac/H3K27ac/H3K4me3 (state 6; Supplementary Figure S2A and Table S4). The observation of chromatin accessibility spanning into exons is also in agreement with the pattern previously reported in Drosophila (17,79).

The integration of ATAC-seq and RNA-seq data can be used to infer clusters of co-regulated genes and common regulatory mechanisms. Looking at the association between nucleosome occupancy and gene expression, mononucleosomes were positioned more frequently at the promoters of more highly expressed genes (Supplementary Figure S2B and Table S5). Our results also showed that nucleosome-free ATAC-seq enrichment at promoters was positively associated with gene expression levels of the corresponding genes (Spearman test: rho 0.39, $P < 0.001$ (MGs); rho 0.47, $P < 0.001$ (SGs); Figure 2B). To further investigate this pattern in a more quantitative manner, we filtered non-overlapping genes with THSs located at promoters or 5′ UTRs, and categorized them by their expression and chromatin accessibility levels at promoters (see 'Materials and Methods' section). In ∼90% of cases, the high/medium expressed genes appeared in the high/medium promoter accessibility groups and conversely, medium/low expressed genes appeared in the medium/low promoter accessibility categories (Figure 2C; Supplementary Figure S2C and Table S5). This suggests that the function of these regulatory regions is gene activation. Notably, around 10% of the THSs-annotated genes displayed opposite profiles for chromatin accessibility and gene expression levels (high promoter accessibility/low expression or low promoter accessibility/high expression; Spearman test: rho −0.71, $P < 0.001$ (MGs); rho −0.65, $P < 0.001$ (SGs); Supplementary Figure S2D and Table S5), suggesting that these accessible THSs correspond to binding sites for repressor TFs. Finally, by integrating ATAC-seq and RNA-seq data with ChIP-seq data, we observed that higher accessibility and expression correlated positively with H3K9ac/H3K27ac enrichment in the promoter, and negatively with H3K9me3 in MGs (Spearman test: rho 0.14, $P < 0.001$ (H3K9ac); rho 0.45, $P < 0.001$ (H3K27ac); rho −0.23, $P < 0.001$ (H3K9me3); Figure 2C).

A large proportion of THSs (30.5%) were located within introns, suggesting that accessibility in these regions may have a role in the regulation of gene expression. Indeed, 49.3% (29 055 out of 58 896) displayed H3K4me3 enrichment, suggesting that they could correspond to alternative TSSs (Supplementary Table S4). Consistent with this hypothesis, genes that are more intronic relative to the gene length (intron-rich genes, see 'Materials and Methods' section) were found to have a higher number of THSs that accumulate within introns. Conversely, for intron-poor genes, the THSs tended to be located at promoters (Figure 2D). This pattern is independent on whether the genes have 5′ UTRs annotated or not (genes with annotated 5′ UTRs: intron-rich 49.3% intronic/12.5% promoter THSs, intron-poor 2.3% intronic/31.8% promoter THSs; genes without annotated 5′ UTRs: intron-rich 31.2% intronic/8.9% pro-

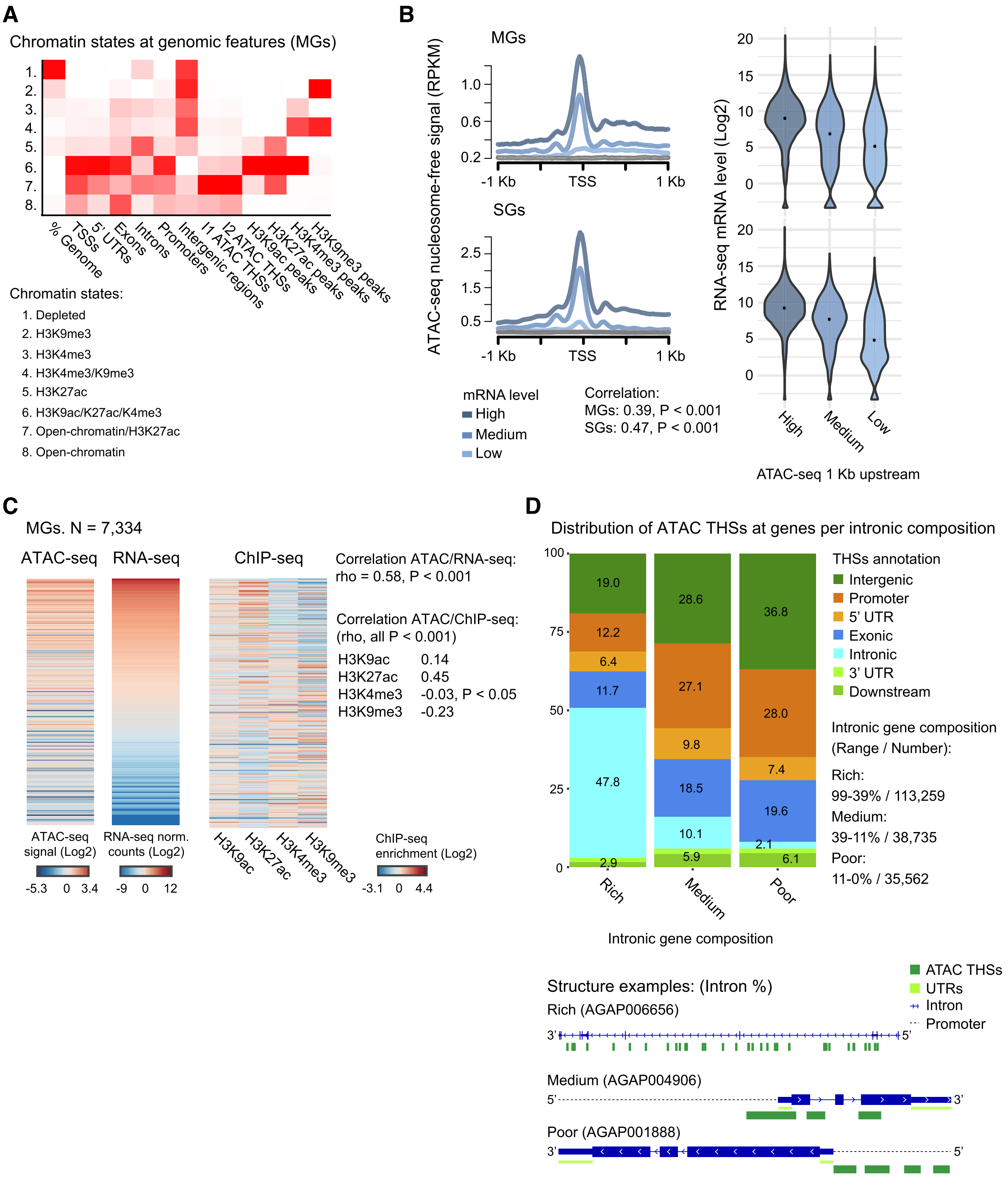

Figure 2. Chromatin accessibility by ATAC-seq is predictive of active hPTMs and tissue-specific gene expression. (**A**) Heatmap showing the overlap of various genomic features with chromatin states inferred using ChromHMM. Darker red in the first column indicates higher percentage of the genome overlapping with a particular state. In all other columns the red indicates the likelihood of finding a chromatin state compared to the random expectation. Most of the genome is in a depleted state. Introns and promoters display a typical state of open-chromatin and activating hPTMs. (**B**) Correlation between ATAC-seq nucleosome-free signal at TSSs and promoters and gene expression. Profile plots (left) show changes in ATAC-seq nucleosome-free signal enrichment at each tissue with respect to the TSSs (±1 Kb). Genes are divided into groups and ranked by their mRNA levels (high, medium or low). Violin plots (right) show mRNA levels for genes grouped by their level of ATAC-seq nucleosome-free signal at promoters: high, medium and low chromatin accessibility. Plot width accounts for the density of repeated values in the range. Median values are marked with a black dot. (**C**) Heatmap showing ATAC-seq nucleosome-free enrichment, gene expression levels of the annotated gene and hPTMs enrichment at promoters. Data correspond to a subset of non-overlapping genes with a THSs annotated in which there is a positive relationship between accessibility and gene expression, i.e. transcriptional activation (see 'Materials and Methods' section). The plot is for MGs. Genes are ordered by mRNA levels. ATAC-seq and ChIP-seq enrichments at promoters are normalized (RPKM) and the ChIP-seq input-corrected. Data are mean-centered. (**D**) Frequency of THSs annotated to various regions for genes grouped by their intronic composition: intron-rich, intron-medium and intron-poor. THSs tend to localise at introns, rather than at the promoter, in high intronic genes and the opposite is true for low intronic genes. Diagrams show the archetypal structure for the three categories of genes based on their intronic composition.

moter THSs, intron-poor 1.3% intronic/16.3% promoter THSs). Indeed, genes without annotated 5′ UTRs did not seem to accumulate more THSs within introns when compared to genes with annotated 5′ UTRs (Mann–Whitney U test: $P > 0.05$; Figure 1C). The mean intronic accessibility was positively correlated with chromatin accessibility at promoters (Spearman test: rho 0.33, $P < 0.001$ (MGs); rho 0.41, $P < 0.001$ (SGs)). There was also a positive and significant correlation between mean intronic accessibility and gene expression (Spearman test: rho 0.20, $P < 0.001$ (MGs); rho 0.27, $P < 0.001$ (SGs)), which was higher for intron-rich genes (rho 0.25, $P < 0.001$ (MGs); rho 0.33, $P < 0.001$ (SGs)), and when considering the first intron only (rho 0.32, $P < 0.001$ (MGs); rho 0.45, $P < 0.001$ (SGs)). Quantitatively, we also observed that intron-rich genes tend to be more expressed (Spearman test: rho 0.16, $P < 0.001$ (MGs); rho 0.30, $P < 0.001$ (SGs); Supplementary Figure S2E)). Overall, these observations suggest that chromatin accessibility within introns could be functionally involved in gene expression regulation and is dependent on the gene architecture.

Among the target genes with a THS annotated, we found 241 important immune-related genes such as *stat2*, *rel2* and members of the *TEP* family (53) (Supplementary Table S4), as well as 671 genes that we reported in an earlier study to be malaria-related by comparing *P. falciparum*-infected and non-infected *A. gambiae* MGs (34). These include, for example, genes encoding CLIP serine proteases, argonaute 1 and the defensin DEF1 (Supplementary Table S4). Regulatory sites at these genes are located predominantly within introns (39.4%), but also at 5′ UTRs (7.2%) and promoters (11.5%). Lastly, for this set of genes we also report here a significant positive correlation between chromatin accessibility and gene expression levels (MG, Spearman test: rho 0.31 $P < 0.001$ (promoters); rho 0.10, $P < 0.05$ (introns); Supplementary Figure S2F).

Overall, our results show a genome-wide association between chromatin accessibility at regulatory regions (i.e. promoters and introns) and the gene expression levels for each tissue assayed. We also confirm that these regulatory sites display the typical pattern of hPTMs characteristic of active chromatin.

Tissue-specific chromatin accessibility correlates with differential gene expression

We found evidence for tissue-specific chromatin accessibility profiles through differential chromatin accessibility analyses at THSs between *A. gambiae* MGs and SGs. A higher proportion of differentially accessible regions (DiffBind regions) were more accessible in SGs (85.1%, $n = 21\ 243$) compared to MGs (Supplementary Figure S3A and Table S6), and the majority were located at promoters (26.9%) or within introns (30.2%) (Supplementary Figure S3B and Table S6). The majority of the DiffBind regions corresponded to changes in the level of accessibility between tissues, rather than a presence/absence of a THS in one of the tissues (Supplementary Figure S3C). In total, 21 400 DiffBind regions coincided with a THS present in both biological replicates, and we used this high-confidence set for downstream analysis (Supplementary Table S6).

The integration of the ATAC-seq and RNA-seq data for genes that appeared differentially expressed and accessible between the two tissues (1920 out of a total of 3584 DEGs; Supplementary Figure S3D and Table S6), revealed a significant correlation between the level of accessibility at the regulatory sites and the levels of gene expression in the same tissue (5382 Diffbind regions at 5′ UTRs, promoters or introns of 1920 DEGs; Spearman test, rho 0.25, $P < 0.001$ (MGs); rho 0.12, $P < 0.001$ (SGs); Figure 3A). Additionally, there was also a positive correlation between the magnitude of the change in accessibility at the DiffBind regions and the change in gene expression of the DEGs between tissues (Spearman test, promoters: rho 0.37, $P < 0.001$; introns: rho 0.19, $P < 0.001$; Figure 3B). Full details of these analyses are given in Supplementary Data.

The GO over-representation functional analyses, performed on the DiffBind regions at 5′ UTRs, promoters or introns of DEGs, showed different processes to be enriched in each tissue, correlating with different functional activities such as digestion and peptide catabolism in the MGs, or amino acid metabolism and glycolysis in the SGs (Figure 3C; Supplementary Figure S3E and Table S6). These results conform with the enriched processes expected for each tissue. For example, genes related to digestion, such as trypsins, are known to be specifically upregulated in a tissue-specific manner in MGs after a blood meal (80). Previous studies have also shown the upregulation of glycolytic processes in mosquito SGs (81,82) and of amino acid metabolism in Anopheles SGs in response to Plasmodium infection (83). In addition, we found more expression linked with more accessible promoters in the same tissue for immune-related genes, such as proteases in MGs and serpins in SGs (Supplementary Figure S3E). These included, for example, serpins involved in the Toll pathway (SRPN2 and SRPN6), and which have been linked to the *A. gambiae* immune response to *Plasmodium berghei* sporozoites in the SGs (84). Other examples of mosquito genes in which a functional link was established are TF-encoding genes, such as *rel2*, which modulates anti-Plasmodium factors, and other immune-related genes including the C-type lectin *ctl4* and the defensin *def1* (Supplementary Table S6). The gene encoding the gambicin antimicrobial peptide *gam1* (AGAP008645) is a good case example. This gene was more expressed in SGs, and this upregulation coincides with higher chromatin accessibility at several regions located at the TSS/promoter and spanning into exons and introns (Figure 3D).

A high proportion of the THSs and DiffBind regions were found within introns (44.7%) or exons (36.6%), so they could be also implicated in regulation of expression at the isoform rather than the gene level. Our analysis revealed evidence of isoform switching at 176 *A. gambiae* genes, with 346 isoforms changing between the two infected tissues (Supplementary Figure S3F and Table S9). The majority (90%) of these genes contained DiffBind regions, mostly within introns (61.7%) or at promoters (24.6%) (Supplementary Figure S3G). These results suggest a functional link between chromatin accessibility dynamics at regulatory regions (mainly introns) and gene isoform switching.

We performed motif enrichment analysis on the regions with differential chromatin accessibility to predict the sets

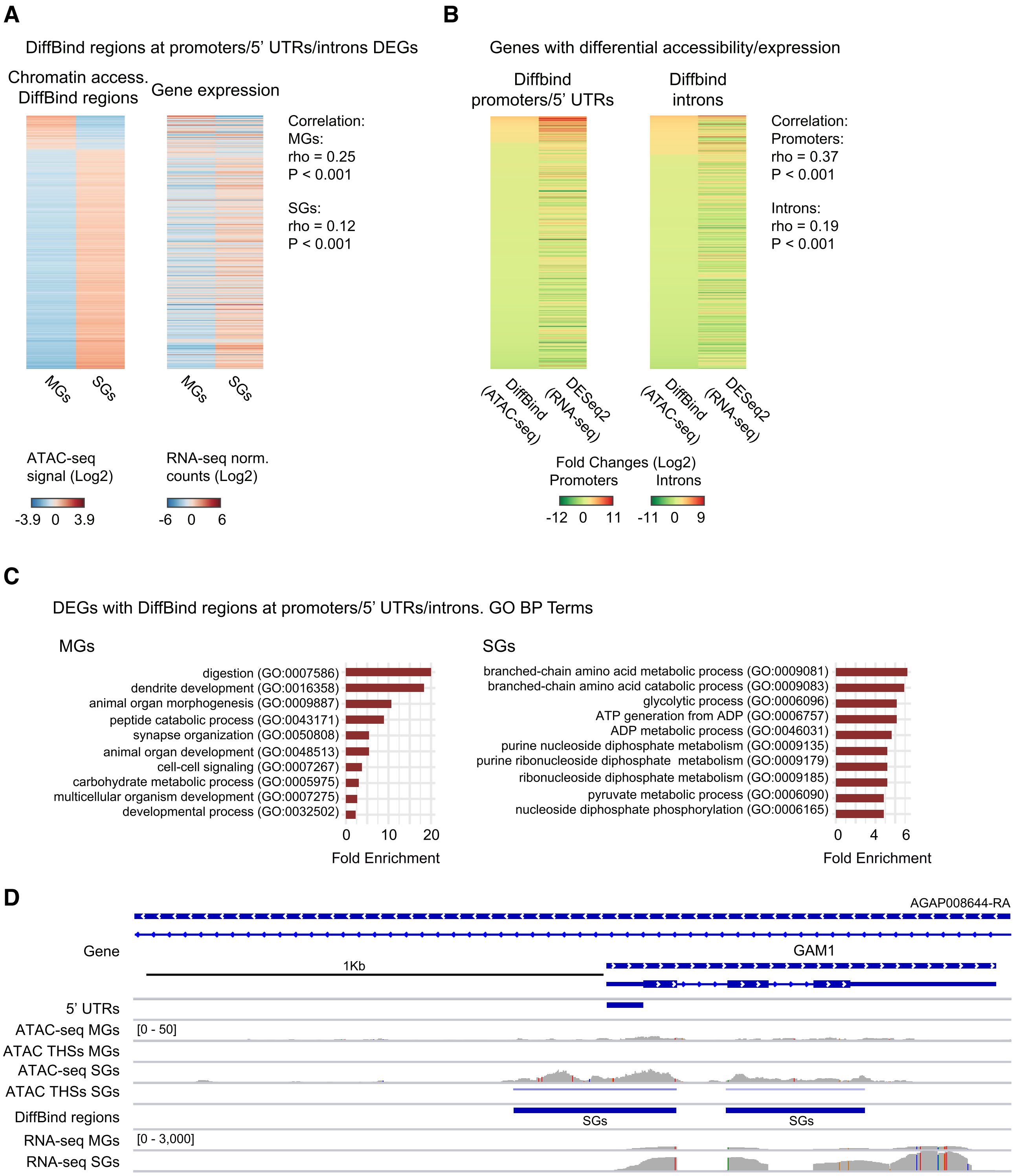

Figure 3. Differential chromatin accessibility between tissues correlates with changes in gene expression. (**A**) Heatmap showing ATAC-seq nucleosome-free enrichment at DiffBind regions located at 5′ UTRs, promoters or introns of DEGs and their expression levels. There is a positive and significant correlation between chromatin accessibility at these regions and gene expression. Genes are ordered by normalized ATAC-seq enrichment (RPKM). Data are mean-centered and for infection 2. (**B**) Heatmap showing chromatin accessibility and gene expression fold changes between tissues for DEGs that display a DiffBind region at the promoters and/or 5′ UTRs and at the introns. Changes occur in most cases in the same direction and there is a positive and significant correlation between the magnitude of changes in accessibility and expression. (**C**) Top GO biological processes terms over-represented in the set of DEGs with DiffBind regions located at promoters, 5′ UTRs and/or introns for each tissue. (**D**) Chromatin accessibility and gene expression profiles in the region containing the antimicrobial peptide GAM1-encoding gene (AGAP008645) which is differentially expressed and differentially accessible between tissues. The tracks displayed are for MGs and SGs from infection 2. The location of 5′ UTRs, THSs, and the regions of differential accessibility (DiffBind) are indicated by colored bars. All tracks are shown at equal scale.

of TFs that may be involved in tissue-specific functional gene expression. By comparing the tissue in which the regulatory sites were more accessible and the annotated genes more expressed, we could infer the functions of the TFs in transcriptional activation or repression (i.e. higher accessibility in a tissue corresponded to higher or lower expression in the same tissue; Supplementary Table S6). We first focused on the set of DiffBind regions annotated to 5′ UTRs or promoters for DEGs specific to MGs and SGs (Supplementary Table S10). For the subset of accessible regions that annotate to active genes, we found *de novo* motifs that match consensus sequences (binding sites) for TFs that are known Drosophila activators with MG and SG-specific functions such as serpent or odd-paired (MGs) and homothorax or broad (SGs). The subset of regions in which the accessibility change was linked to silencing, were enriched in motifs similar to binding sites of known Drosophila repressors, such as even-skipped or pleiohomeotic for MGs, and forkhead or hunchback specific to SGs. Next, provided that the majority of DiffBind regions were located within introns (see above) we performed analogous analyses to predict TFs at these regions (Supplementary Table S10). In the majority of cases, we found enriched motifs similar to the ones in promoters and 5′ UTRs (see above). We also predicted TFs that appeared to be particular to introns, such as bric a brac 1 or schnurri.

Altogether, chromatin accessibility and gene expression differential analyses between tissues allowed us to identify mosquito genes in which the switch between the open/closed chromatin is associated with transcriptional activation in a tissue-specific manner. We also provide some evidence that differential accessibility at intronic regulatory regions could be related with changes in isoform rather than gene expression. Lastly, the motif enrichment analyses allowed us to predict binding sites for potential activator or repressor TFs with tissue-specific functions.

Chromatin accessibility allows for the identification of novel *A. gambiae cis*-regulatory elements

ATAC-seq has been shown to capture regulatory sequences with high precision, and therefore is an ideal assay to characterize novel CREs such as enhancers and TSSs. Of the 1685 *A. gambiae* enhancers suggested from previous studies (29,30,35), mostly based on computational predictions, 42% (708) were identified as THSs (Supplementary Table S11), and therefore may correspond to active enhancers in the tissues assayed here. Furthermore, using a homology approach and *D. melanogaster* enhancer maps such as the EnhancerAtlas (19) (see 'Materials and Methods' section and Supplementary Data), we predicted 1122 novel enhancer-like regions in *A. gambiae*. Around 10% of these, (93 regions) were found to overlap with THSs by ATAC-seq (Supplementary Table S11). As a consequence, the final database comprises a total of 811 accessible enhancer-like regions, overlapping with 2272 THSs annotated to 563 genes. The majority are located within introns (43.3%) or promoters (21%) (Figure 4A and Supplementary Table S11). Of these 811 accessible enhancer-like regions, 293 displayed signatures typical of active enhancers: chromatin accessibility and H3K27ac enrichment (Figure 4B and Supplementary Table S11). For around 80% (633) of the accessible enhancer-like regions, the annotated gene coincides with the target of the enhancer reported previously by others (18,30,35) and therefore these are most likely to be proximal enhancer elements. There was a significant association between chromatin accessibility at the regulatory sites and the expression of the annotated target genes for proximal enhancers (Spearman test: rho 0.35, $P < 0.001$ (MGs); rho 0.15, $P < 0.05$ (SGs); Figure 4C), but not for distal ($P = 0.66$ for MGs and $P = 0.32$ for SGs). The remainder (178) may be distal enhancers (see 'Materials and Methods' section and Supplementary Table S11).

The integration of the differential gene expression data between the tissues and differential chromatin accessibility at enhancer-like regions was used as a proxy to validate functions. Fifty-one regions were identified both as proximal enhancer-like regions and DiffBind regions, and annotated to DEGs between the two tissues. Of these, the most frequent role was activating, that is, in 37 cases the enhancer region was more accessible in the tissue where the gene was more highly expressed. However, in 14 cases the relationship was the opposite, pointing to a repressor function. There was also a significant positive correlation between the changes in accessibility and the differential expression of the annotated genes (Spearman test: rho 0.60, $P < 0.001$) (Figure 4D). The *lrim1* gene (AGAP006348), that encodes for a leucine-rich immune protein, is a good example of a malaria-related gene (34), that has an experimentally validated enhancer and that was differentially accessible between tissues. Here, we observed the gene to be more highly expressed in SGs, and this was associated with a DiffBind region upstream of the promoter, which was more accessible in SGs, and that coincided with the enhancer experimentally validated by others (35) (Figure 4E). Apart from differences in gene expression, these regulatory elements may be also involved in differences in gene isoform expression. Indeed, there are 38 enhancer-like regions annotated to 19 genes with isoforms that switch expression. In addition, 20 out of these 38 enhancer-like regions contained instances of motifs that have been shown to be characteristic of Drosophila intronic-splicing enhancers, such as CTCTCT and TTATAA (85).

Finally, on the set of 2272 THSs that overlap *A. gambiae* enhancer-like regions, we performed motif enrichment analyses to predict enriched *de novo* motifs similar to known binding sites of Drosophila TFs that could be acting through binding of enhancers. Among the top hits, we predicted TFs involved in processes such as nucleosome organization, reproduction and regulation of immune response (Supplementary Table S12). These included regulators such as trithorax-like, pleiohomeotic, zeste and the deformed epidermal autoregulatory factor-1, which had motifs that annotate to 306 genes.

ATAC-seq peaks can also be used to support TSS prediction and discovery. The majority of *A. gambiae* genes displayed THSs located at the promoter or the 5′ UTRs. In 46.5% of those, the THSs coincided with the annotated TSSs (i.e. the 5′ coordinate of the annotated UTR). Around 20% of genes in the current *A. gambiae* genome annotation (2890) do not have annotated 5′ UTRs. We observed that 35% of these (1009) displayed THSs located at pro-

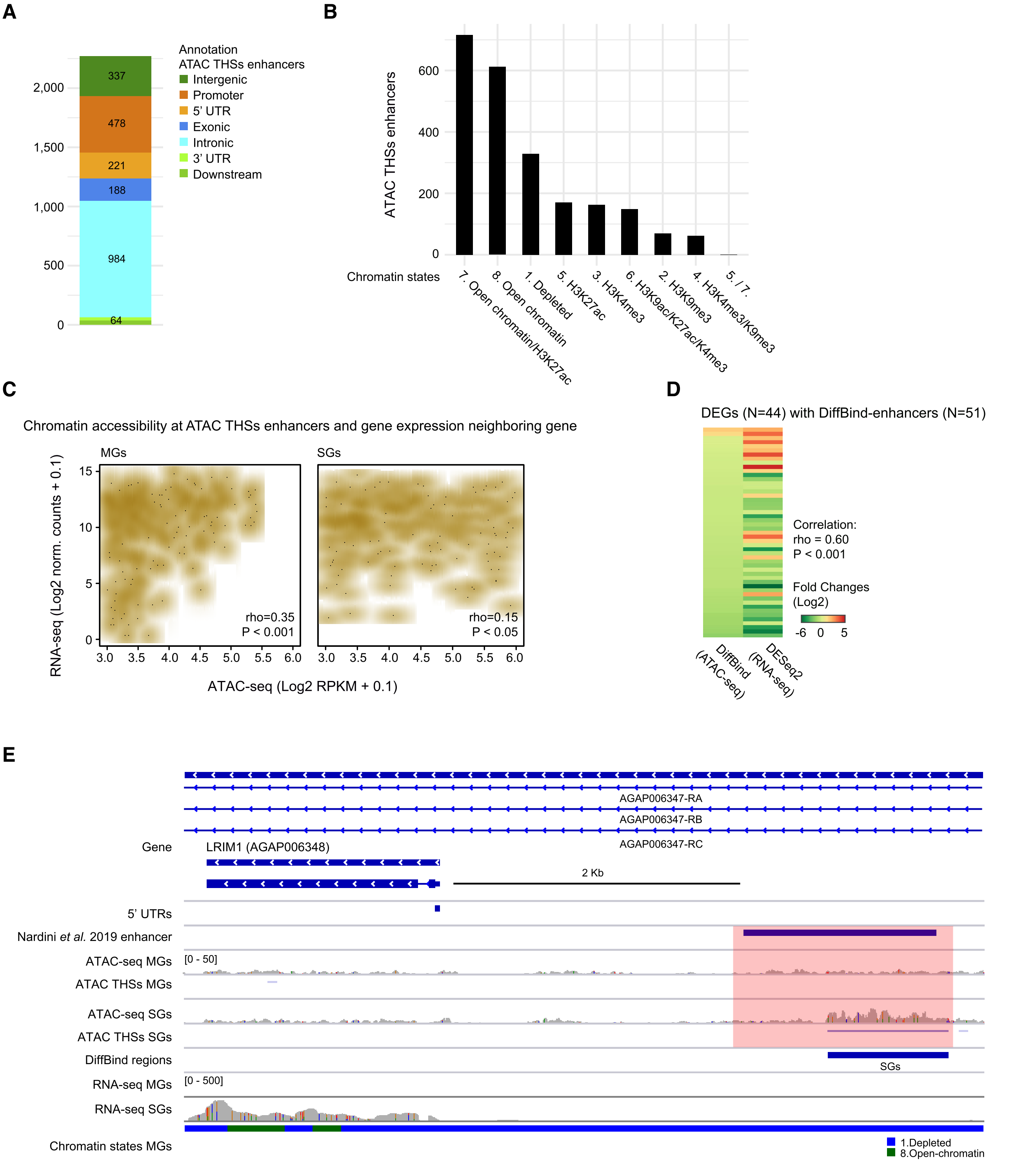

Figure 4. Genome-wide *in vivo* mapping and functional characterization of *Anopheles gambiae* enhancers. (**A**) Annotation of accessible enhancers to various genomic features: intergenic regions, promoters, 5′ UTRs, exons, introns, 3′ UTRs and downstream regions. The majority of enhancers locate at introns or promoters. (**B**) Chromatin states at the accessible enhancers. As expected, these regulatory regions appear to be H3K27ac-enriched. (**C**) Scatter plots showing a positive correlation between chromatin accessibility for a subset of proximal enhancers (see 'Results' section), and gene expression of the nearest target gene. Data are for MGs (left) and SGs (right). (**D**) Heatmap displaying chromatin accessibility and gene expression fold changes between tissues for DEGs that show a DiffBind region coinciding with a proximal enhancer element. Changes occur in most cases in the same direction and there is a positive correlation in the magnitude of the change. (**E**) Chromatin accessibility and gene expression profiles in the region containing the LRIM1-encoding gene (AGAP006348), a Plasmodium-responsive gene based on our previous study (34). Here, this gene is differentially expressed and displays a differentially accessible enhancer between tissues. The tracks shown are from MGs and SGs from infection 2. All tracks are shown at equal scale. The location of various genomic features: 5′ UTRs, THSs, DiffBind regions and chromatin states, are indicated by colored bars. The enhancer element as predicted by others (35) that is coinciding with the differential accessibility region is highlighted as the pink box.

moters, which could be novel TSSs (Supplementary Table S4), and in agreement, 51% of them displayed chromatin states characteristic of TSSs (i.e. open-chromatin, H3K9ac, H3K27ac and/or H3K4me3 enrichment) (Supplementary Table S11). To validate novel TSSs, and based on the assumption that TSSs tend to be conserved (86,87), we used a Drosophila dataset (70) and identified 917 homolog TSS-like sites in *A. gambiae* annotated to 819 genes (Supplementary Table S11). Integrating this set of homolog TSS sites with the mosquito ATAC-seq data, we found that the 28.5% of the transferred TSSs (217 out of 917) overlapped with our THSs at promoters or 5′ UTRs (Supplementary Figure S4A), and displayed enrichment in active hPTMs characteristic of promoter regions (i.e. H3K9ac, H3K27ac and H3K4me3; Supplementary Figure S4B). As a proof of principle, 79.7% (173) of homolog TSSs validated mosquito annotated TSSs, and for the rest, the THSs did not coincide with the annotated TSSs and thus could be considered as alternative TSSs. Finally, we report 14 potentially novel TSSs that annotated to genes without 5′ UTRs (Supplementary Table S11). Future studies in *A. gambiae*, applying techniques such as CAGE, will be needed to confirm these novel TSSs.

A third type of *cis*-regulatory sequences are insulators, specific DNA sequences that play an important role in regulating gene expression (88,89). The CCCTC binding factor (CTCF) is a TF known to bind insulators and domain boundaries in vertebrates and Drosophila. It contributes to long-range chromatin interactions, including enhancer-promoters, and organization of chromatin architecture. In *A. gambiae* the CTCF-like gene (AGAP005555) is the ortholog to CTCF in Drosophila and appears to be expressed in the tissues assayed here. Based on the assumption that the binding sites for CTCF as determined by ChIP-seq in Drosophila (71) should be highly conserved in Anopheles (90), we could identify 2516 homolog *A. gambiae* CTCF sites, which we propose could function as mosquito insulator elements. 30.4% of the transferred CTCF sites (764 out of 2516) overlapped with our ATAC-seq THSs (Supplementary Table S11). For 51.3% of these (392 out of 764), the annotated gene was ortholog to the nearest neighboring gene to the CTCF binding sites in Drosophila. We also report a fraction of potential CTCF binding sites that coincide with differentially accessible regions between MGs and SGs (177 out of 2519). Of these, 29 were annotated to DEGs such as the ortholog gene in *A. gambiae* of the abdominal A (abd-A) gene in Drosophila (Supplementary Figure S4C), which is part of the Bithorax Complex (89).

DISCUSSION

The aims of this study were to investigate mechanisms underlying tissue-specific regulation of gene expression, and to map genome-wide enhancer- and TSS-like activity in *A. gambiae in vivo*. In a previous study, we characterized different post-translational modifications of histone tails (hPTMs), and the transcriptional profiles of *P. falciparum*-infected and non-infected *A. gambiae* MGs, which allowed us to identify changes in the epigenomic landscape of the mosquito linked to malaria infection (34). However, this information was insufficient to capture with high precision the location and function of tissue-specific mosquito CREs. To this end, in this study we performed the first genome-wide chromatin accessibility profiling by ATAC-seq, together with gene expression analysis by RNA-seq, in *A. gambiae* tissues infected with *P. falciparum*.

We report thousands of accessible regulatory sequences (THSs) involved in tissue-specific transcriptional regulation, which were distributed genome-wide, particularly at introns and promoters (regions 1 Kb upstream of genes). This pattern is in agreement with the distribution of accessible regulatory sites reported in Drosophila by ATAC-seq (17,74), and in *Aedes aegypti* and *A. gambiae* mosquitoes by FAIRE-seq (31–33). Chromatin accessibility at regulatory sites is generally considered a good predictor of gene activity (41,91), which our results support, showing a positive correlation between open-chromatin at regulatory regions and gene expression. By integrating our ATAC-seq and ChIP-seq data for various hPTMs (34), we also describe a relationship between accessibility and epigenetic states at these regulatory sites: the enrichment of hPTMs with *a priori* activating (H3K9ac, H3K27ac) or repressor (H3K9me3) roles that relate to gene function.

In *D. melanogaster*, introns are known to harbour regulatory sequences and to have an important role in the spatio-temporal control of gene expression (92,93). In *A. gambiae*, we demonstrate a relationship between accessibility at introns, gene expression, and H3K27ac enrichment at the active site, suggesting that intronic THSs are involved in functional regulation of gene expression. Another important observation is that gene architecture influences the proportion of open CREs at introns respective to promoters, and *vice versa*: genes with higher intronic content contain more intronic THSs, and moreover, chromatin accessibility at these regions also correlates with higher accessibility of the cognate promoters and higher gene expression. Our results agree with previous hypotheses that longer introns would be more efficient in transcriptional enhancement (94), and is also in agreement with observations that introns can influence transcription by looping or coupling with promoters, transcribing small RNAs, or accumulating regulatory sequences (8,93). This is a poorly understood phenomenon that has been termed intron-mediated enhancement and that seems to be conserved among eukaryotes, including Drosophila (95,96).

Our comparative analyses of mosquito MGs and SGs allowed us to unveil tissue-specific regulatory elements that may underlie functional differences between tissues. We identified thousands of THSs displaying differential chromatin accessibility between tissues, annotated to DEGs. A major proportion of differentially accessible regions were more accessible in SGs, and were located at promoters or introns. A higher fraction of ATAC-seq nucleosome-free reads were seen at SGs when compared to MGs, which could reflect both that the ATAC-seq assay worked better in this tissue, or that SGs display a more accessible regulatory landscape. Notably, the integration of our differential datasets revealed a correlation between accessibility and the transcriptional state, i.e. genes tended to be more expressed in the tissue where the regulatory sites were more accessible. Moreover, these genes appeared to have tissue-specific functions, such as digestion in MGs and amino acid metabolism in SGs, which has also been shown to be a pathway affected by Plasmodium infection (83). The next step was to predict the regulatory proteins involved in these functional responses. Here we report that the CREs with tissue-specific accessibility appear enriched in motifs resembling the consensus binding sites of tissue-specific Drosophila TFs, including TFs with immune functions in MGs and SGs such as serpent and relish. The functions of most of these predicted Drosophila TFs are likely conserved in mosquitoes (97,98). Indeed, many of the TFs predicted in this study, such as serpent, deformed epidermal autoregulatory factor-1 or trithorax-like, agree with those predicted by previous studies that employed FAIRE-seq in different *A. aegypti* and *A. gambiae* tissues (31,32). Among the differentially expressed genes that seem to be regulated by differentially accessible CREs and that contain the above motifs, we found examples of immune-related genes (53), such as *srpn6* or *gam1*, and genes that we identified in our previous study as Plasmodium-responsive (34), such as the defensin *def1*. The regulation of these genes is likely to be crucial in determining mosquito infection phenotypes that impact traits such as vector competence, immune response, longevity or reproduction. Until now, the regulatory elements of most immune-related and Plasmodium-responsive *A. gambiae* genes remain uncharacterized.

Another motivation of this study was the genome-wide mapping, discovery and validation of functional enhancers and TSSs in *A. gambiae;* several enhancer elements have previously been predicted bioinformatically, but their *in vivo* characterization and functions remain poorly explored. Relatively few mosquito candidate enhancer sequences have been defined, compared to the fourty thousand enhancers described in Drosophila (EnhancerAtlas database, (19)), and only a very small number have been experimentally validated in Anopheles (30,35). Using our ATAC-seq data, we mapped *in vivo* 42% (708) of the 1685 *A. gambiae* enhancers predicted bioinformatically (29,30,35). In addition, we found 1122 potentially novel *A. gambiae* regulatory regions that are homolog to known *D. melanogaster* enhancers (18), and which do not coincide with previously predicted mosquito regulatory sequences. Of these 1122 enhancer-like regions, around 10% were found to be accessible by ATAC-seq analysis. The remaining 1019 regions homolog to Drosophila enhancers were not accessible according to ATAC-seq analysis. This could be due to the fact that these enhancers were originally identified in *D. melanogaster* under different experimental conditions and tissues. In total we report 811 enhancer-like regions accessible according to ATAC-seq analysis, that are located throughout the genome, mainly in introns and exons. Our results also show that the majority of enhancer sites (around 80%) would regulate the neighboring gene. This distribution and pattern is in agreement with previous observations in other model organisms, which suggest enhancers are mainly proximal to or intragenic of target genes (18,99).

In support of the functional role of active enhancer sites in *A. gambiae*, a positive correlation was seen for the majority of enhancer sites between chromatin accessibility at proximal regulatory sites and the gene expression of the target genes. A small number of potentially distal regulatory networks, which are known to play important roles during development and differentiation processes in Drosophila, were also identified in Anopheles, based on the location of ortholog genes in Drosophila.

Finally, a powerful tool to validate the enhancer function is the combined analysis of differential chromatin accessibility at enhancers and gene expression changes between tissues. We report 51 enhancer-like regions with differential chromatin accessibility in MGs and SGs that annotate to differentially expressed genes. In 37 of these, the enhancer was more accessible in the same tissue in which the gene was more expressed, and this association was also quantitative. Importantly, we observed that 141 of the accessible enhancers were annotated to genes that we previously reported to be Plasmodium-responsive (34). One example is the leucine-rich repeat protein 1 (LRIM1) encoding gene that harbors an enhancer that switches chromatin states between tissues, and thus is implicated in tissue-specific gene expression regulation. Future studies in *A. gambiae*, applying approaches such as STARR-seq, transgenesis or Hi-C, are now needed to validate the function of this and other mosquito enhancers.

In summary, we applied for the first time high-throughput genome-wide chromatin accessibility profiling by ATAC-seq in Plasmodium-infected *A. gambiae*, adding new evidence on the mechanisms of transcriptional regulation in mosquitoes. The integrative analyses of ATAC-seq, RNA-seq and ChIP-seq allowed us to link chromatin accessibility and structure with function, as well as to characterize tissue-specific CREs potentially involved in mosquito immune responses to Plasmodium. We also show ATAC-seq is a powerful tool for the *in vivo* discovery and characterization of functionally active enhancers as well as insulator sequences, which are still poorly understood in mosquitoes. Such a detailed map of the regulatory genome of the main human malaria vector *A. gambiae* fills an important gap in the field, and it is essential for designing new strategies of disease control based on the genetic manipulation of mosquitoes.

DATA AVAILABILITY

The ATAC-seq and RNA-seq data generated and analyzed during the current study are available in the GEO repository under accession number GSE152924 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152924). The datasets supporting the conclusions of this article are included within the article and its additional files.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors wish to dedicate this work to the memory of Prof. José Luis Gómez-Skarmeta (1966–2020) for his role as inspiring mentor and colleague, and help with the application of the ATAC-seq technique in mosquitoes. We thank Rodrigo G. Arzate-Mejía for suggestions on the analyses and useful discussion. We are grateful to A. Hidalgo-Galiana for her contribution in the technical ATAC-seq work, and to Dorothy Armstrong and Elizabeth Peat for their maintenance of the mosquito colonies and support for parasite cultures at the University of Glasgow.

FUNDING

Spanish Ministry of Economy and Competitiveness Grant [BFU2015-65000-R to E.G.-D.]; Severo Ochoa Fellowship [BES-2016-076276 to J.L.R.].

Conflict of interest statement. None declared.

REFERENCES

1. Li,B., Carey,M. and Workman,J.L. (2007) The role of chromatin during transcription. *Cell*, **128**, 707–719.
2. Li,X.Y., Thomas,S., Sabo,P.J., Eisen,M.B., Stamatoyannopoulos,J.A. and Biggin,M.D. (2011) The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biol.*, **12**, R34.
3. Voss,T.C. and Hager,G.L. (2014) Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet.*, **15**, 69–81.
4. Reiter,F., Wienerroither,S. and Stark,A. (2017) Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.*, **43**, 73–81.
5. Ong,C.-T. and Corces,V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, **12**, 283–293.
6. Pennacchio,L.A., Bickmore,W., Dean,A., Nobrega,M.A. and Bejerano,G. (2013) Enhancers: five essential questions. *Nat. Rev. Genet.*, **14**, 288–295.
7. Shlyueva,D., Stelzer,C., Gerlach,D., Yanez-Cuna,J.O., Rath,M., Boryn,L.M., Arnold,C.D. and Stark,A. (2014) Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Mol. Cell*, **54**, 180–192.
8. Shlyueva,D., Stampfel,G. and Stark,A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–286.
9. Arzate-Mejia,R.G., Recillas-Targa,F. and Corces,V.G. (2018) Developing in 3D: the role of CTCF in cell differentiation. *Development*, **145**, dev137729.
10. Nieto Moreno,N., Giono,L.E., Cambindo Botto,A.E., Munoz,M.J. and Kornblihtt,A.R. (2015) Chromatin, DNA structure and alternative splicing. *FEBS Lett.*, **589**, 3370–3378.
11. Wang,Z. and Burge,C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, **14**, 802–813.
12. Das,D., Clark,T.A., Schweitzer,A., Yamamoto,M., Marr,H., Arribere,J., Minovitsky,S., Poliakov,A., Dubchak,I., Blume,J.E. *et al.* (2007) A correlation with exon expression approach to identify *cis*-regulatory elements for tissue-specific alternative splicing. *Nucleic Acids Res.*, **35**, 4845–4857.
13. Rambout,X., Dequiedt,F. and Maquat,L.E. (2018) Beyond transcription: roles of transcription factors in pre-mRNA splicing. *Chem. Rev.*, **118**, 4339–4364.
14. Urbanski,L.M., Leclair,N. and Anczuków,O. (2018) Alternative-splicing defects in cancer: splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. *Wiley Interdiscip. Rev. RNA*, **9**, e1476.
15. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C. and Ching,K.A. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
16. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
17. Davie,K., Jacobs,J., Atkins,M., Potier,D., Christiaens,V., Halder,G. and Aerts,S. (2015) Discovery of transcription factors and regulatory regions driving *in vivo* tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. *PLos Genet.*, **11**, e1004994.
18. Kvon,E.Z., Kazmar,T., Stampfel,G., Yanez-Cuna,J.O., Pagani,M., Schernhuber,K., Dickson,B.J. and Stark,A. (2014) Genome-scale functional characterization of Drosophila developmental enhancers *in vivo*. *Nature*, **512**, 91–95.
19. Gao,T. and Qian,J. (2020) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.*, **48**, D58–D64.
20. Zabidi,M.A. and Stark,A. (2016) Regulatory enhancer-core-promoter communication via transcription factors and cofactors. *Trends Genet.*, **32**, 801–814.
21. Chen,H., Levo,M., Barinov,L., Fujioka,M., Jaynes,J.B. and Gregor,T. (2018) Dynamic interplay between enhancer–promoter topology and gene activity. *Nat. Genet.*, **50**, 1296–1303.
22. Bulger,M. and Groudine,M. (2010) Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.*, **339**, 250–257.
23. World Health Organization (2017) Global vector control response 2017–2030. Licence: CC BY-NC-SA 3.0 IGO, Geneva.
24. World Health Organization (2019) *World Malaria Report 2019*. World Health Organization, Geneva, Switzerland.
25. Catteruccia,F. (2007) Malaria vector control in the third millennium: progress and perspectives of molecular approaches. *Pest. Manag. Sci.*, **63**, 634–640.
26. Compton,A., Sharakhov,I.V. and Tu,Z. (2020) Recent advances and future perspectives in vector-omics. *Curr. Opin. Insect. Sci.*, **40**, 94–103.
27. Sieglaff,D.H., Dunn,W.A., Xie,X.S., Megy,K., Marinotti,O. and James,A.A. (2009) Comparative genomics allows the discovery of *cis*-regulatory elements in mosquitoes. *Proc. Natl Acad. Sci. U.S.A.*, **106**, 3053–3058.
28. O'Brochta,D.A., Pilitt,K.L., Harrell,R.A. 2nd, Aluvihare,C. and Alford,R.T. (2012) Gal4-based enhancer-trapping in the malaria mosquito *Anopheles stephensi*. *G3 (Bethesda)*, **2**, 1305–1315.
29. Ahanger,S.H., Srinivasan,A., Vasanthi,D., Shouche,Y.S. and Mishra,R.K. (2013) Conserved boundary elements from the Hox complex of mosquito, *Anopheles gambiae*. *Nucleic Acids Res.*, **41**, 804–816.
30. Kazemian,M., Suryamohan,K., Chen,J.Y., Zhang,Y., Samee,M.A., Halfon,M.S. and Sinha,S. (2014) Evidence for deep regulatory similarities in early developmental programs across highly diverged insects. *Genome Biol. Evol.*, **6**, 2301–2320.
31. Behura,S.K., Sarro,J., Li,P., Mysore,K., Severson,D.W., Emrich,S.J. and Duman-Scheel,M. (2016) High-throughput *cis*-regulatory element discovery in the vector mosquito Aedes aegypti. *BMC Genomics*, **17**, 341.
32. Perez-Zamorano,B., Rosas-Madrigal,S., Lozano,O.A.M., Castillo Mendez,M. and Valverde-Garduno,V. (2017) Identification of *cis*-regulatory sequences reveals potential participation of lola and Deaf1 transcription factors in *Anopheles gambiae* innate immune response. *PLoS One*, **12**, e0186435.
33. Mysore,K., Li,P. and Duman-Scheel,M. (2018) Identification of Aedes aegypti *cis*-regulatory elements that promote gene expression in olfactory receptor neurons of distantly related dipteran insects. *Parasit. Vectors*, **11**, 406.
34. Ruiz,J.L., Yerbanga,R.S., Lefevre,T., Ouedraogo,J.B., Corces,V.G. and Gomez-Diaz,E. (2019) Chromatin changes in *Anopheles gambiae* induced by *Plasmodium falciparum* infection. *Epigenet. Chromatin*, **12**, 5.
35. Nardini,L., Holm,I., Pain,A., Bischoff,E., Gohl,D.M., Zongo,S., Guelbeogo,W.M., Sagnon,N.F., Vernick,K.D. and Riehle,M.M. (2019) Influence of genetic polymorphism on transcriptional enhancer activity in the malaria vector *Anopheles coluzzii*. *Sci. Rep.*, **9**, 15275.
36. Gallo,S.M., Gerrard,D.T., Miner,D., Simich,M., Des Soye,B., Bergman,C.M. and Halfon,M.S. (2010) REDfly v3. 0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res.*, **39**, D118–D123.
37. Rivera,J., Keranen,S.V.E., Gallo,S.M. and Halfon,M.S. (2019) REDfly: the transcriptional regulatory element database for Drosophila. *Nucleic Acids Res.*, **47**, D828–D834.
38. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.

39. Buenrostro,J.D., Wu,B., Chang,H.Y. and Greenleaf,W.J. (2015) ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.*, **109**, 21.29.1–21.29.29.
40. Carter,R., Ranford-Cartwright,L. and Alano,P. (1993) The culture and preparation of gametocytes of *Plasmodium falciparum* for immunochemical, molecular, and mosquito infectivity studies. *Methods Mol. Biol.*, **21**, 67–88.
41. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
42. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357.
43. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
44. Ou,J., Liu,H., Yu,J., Kelliher,M.A., Castilla,L.H., Lawson,N.D. and Zhu,L.J. (2018) ATACseqQC: a bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics*, **19**, 169.
45. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M. and Li,W. (2008) Model-based analysis of ChIP-seq (MACS). *Genome Biol.*, **9**, R137.
46. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
47. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
48. Yu,G., Wang,L.-G. and He,Q.-Y. (2015) ChIPseeker: an R/bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.
49. Giraldo-Calderón,G.I., Emrich,S.J., MacCallum,R.M., Maslen,G., Dialynas,E., Topalis,P., Ho,N., Gesing,S., Consortium,V. and Madey,G. (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.*, **43**, D707–D713.
50. Jiao,X., Sherman,B.T., Huang,D.W., Stephens,R., Baseler,M.W., Lane,H.C. and Lempicki,R.A. (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, **28**, 1805–1806.
51. Mi,H., Muruganujan,A., Ebert,D., Huang,X. and Thomas,P.D. (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.*, **47**, D419–D426.
52. Slenter,D.N., Kutmon,M., Hanspers,K., Riutta,A., Windsor,J., Nunes,N., Mélius,J., Cirillo,E., Coort,S.L. and Digles,D. (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.
53. Waterhouse,R.M., Kriventseva,E.V., Meister,S., Xi,Z., Alvarez,K.S., Bartholomay,L.C., Barillas-Mury,C., Bian,G., Blandin,S., Christensen,B.M. *et al.* (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science*, **316**, 1738–1743.
54. Schep,A.N., Buenrostro,J.D., Denny,S.K., Schwartz,K., Sherlock,G. and Greenleaf,W.J. (2015) Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.*, **25**, 1757–1770.
55. Kopylova,E., Noé,L. and Touzet,H. (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28**, 3211–3217.
56. Yilmaz,P., Parfrey,L.W., Yarza,P., Gerken,J., Pruesse,E., Quast,C., Schweer,T., Peplies,J., Ludwig,W. and Glöckner,F.O. (2014) The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. *Nucleic Acids Res.*, **42**, D643–D648.
57. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
58. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
59. Deschamps-Francoeur,G., Boivin,V., Abou Elela,S. and Scott,M.S. (2019) CoCo: RNA-seq read assignment correction for nested genes and multimapped reads. *Bioinformatics*, **35**, 5039–5047.
60. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
61. Ramírez,F., Ryan,D.P., Grüning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dündar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.
62. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
63. Stark,R. and Brown,G. (2011) DiffBind: differential binding analysis of ChIP-Seq peak data. *R Package Version*, **100**, 4–3.
64. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
65. Gao,Z., Zhao,Z. and Tang,W. (2018) DREAMSeq: an improved method for analyzing differentially expressed genes in RNA-seq data. *Front. Genet.*, **9**, 588.
66. Vitting-Seerup,K. and Sandelin,A. (2019) IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics*, **35**, 4469–4471.
67. Haeussler,M., Zweig,A.S., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Hinrichs,A.S. and Gonzalez,J.N. (2019) The UCSC genome browser database: 2019 update. *Nucleic Acids Res.*, **47**, D853–D858.
68. Thurmond,J., Goodman,J.L., Strelets,V.B., Attrill,H., Gramates,L.S., Marygold,S.J., Matthews,B.B., Millburn,G., Antonazzo,G. and Trovisco,V. (2019) FlyBase 2.0: the next generation. *Nucleic Acids Res.*, **47**, D759–D765.
69. Chen,F., Mackey,A.J., Stoeckert,C.J. Jr and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
70. Dreos,R., Ambrosini,G., Groux,R., Cavin Perier,R. and Bucher,P. (2017) The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Res.*, **45**, D51–D55.
71. Oki,S., Ohta,T., Shioi,G., Hatanaka,H., Ogasawara,O., Okuda,Y., Kawaji,H., Nakaki,R., Sese,J. and Meno,C. (2018) ChIP-Atlas: a data-mining suite powered by full integration of public Ch IP-seq data. *EMBO Rep.*, **19**, e46255.
72. Zuo,Z., Jin,Y., Zhang,W., Lu,Y., Li,B. and Qu,K. (2019) ATAC-pipe: general analysis of genome-wide chromatin accessibility. *Brief. Bioinform.*, **20**, 1934–1943.
73. Yan,F., Powell,D.R., Curtis,D.J. and Wong,N.C. (2020) From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.*, **21**, 22.
74. Vizcaya-Molina,E., Klein,C.C., Serras,F., Mishra,R.K., Guigo,R. and Corominas,M. (2018) Damage-responsive elements in Drosophila regeneration. *Genome Res.*, **28**, 1852–1866.
75. Mavrich,T.N., Jiang,C., Ioshikhes,I.P., Li,X., Venters,B.J., Zanton,S.J., Tomsho,L.P., Qi,J., Glaser,R.L. and Schuster,S.C. (2008) Nucleosome organization in the Drosophila genome. *Nature*, **453**, 358–362.
76. Mavrich,T.N., Ioshikhes,I.P., Venters,B.J., Jiang,C., Tomsho,L.P., Qi,J., Schuster,S.C., Albert,I. and Pugh,B.F. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, **18**, 1073–1083.
77. Janssen,K.A., Sidoli,S. and Garcia,B.A. (2017) Recent achievements in characterizing the histone code and approaches to integrating epigenomics and systems biology. *Methods Enzymol.*, **586**, 359–378.
78. Zhang,W., Zhang,X., Xue,Z., Li,Y., Ma,Q., Ren,X., Zhang,J., Yang,S., Yang,L., Wu,M. *et al.* (2019) Probing the function of metazoan histones with a systematic library of H3 and H4 mutants. *Dev. Cell*, **48**, 406–419.
79. Thomas,S., Li,X.-Y., Sabo,P.J., Sandstrom,R., Thurman,R.E., Canfield,T.K., Giste,E., Fisher,W., Hammonds,A. and Celniker,S.E. (2011) Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol.*, **12**, R43.

80. Müller,H., Crampton,J., Della Torre,A., Sinden,R. and Crisanti,A. (1993) Members of a trypsin gene family in *Anopheles gambiae* are induced in the gut by blood meal. *EMBO J.*, **12**, 2891–2900.
81. Djegbe,I., Cornelie,S., Rossignol,M., Demettre,E., Seveno,M., Remoue,F. and Corbel,V. (2011) Differential expression of salivary proteins between susceptible and insecticide-resistant mosquitoes of Culex quinquefasciatus. *PLoS One*, **6**, e17496.
82. Dhawan,R., Kumar,M., Mohanty,A.K., Dey,G., Advani,J., Prasad,T.K. and Kumar,A. (2017) Mosquito-borne diseases and omics: salivary gland proteome of the female *Aedes aegypti* mosquito. *OMICS*, **21**, 45–54.
83. Pinheiro-Silva,R., Borges,L., Coelho,L.P., Cabezas-Cruz,A., Valdés,J.J., Do Rosario,V., De La Fuente,J. and Domingos,A. (2015) Gene expression changes in the salivary glands of *Anopheles coluzzii* elicited by *Plasmodium berghei* infection. *Parasit. Vectors*, **8**, 485.
84. Pinto,S.B., Kafatos,F.C. and Michel,K. (2008) The parasite invasion marker SRPN6 reduces sporozoite numbers in salivary glands of *Anopheles gambiae*. *Cell. Microbiol.*, **10**, 891–898.
85. Brooks,A.N., Aspden,J.L., Podgornaia,A.I., Rio,D.C. and Brenner,S.E. (2011) Identification and experimental validation of splicing regulatory elements in *Drosophila melanogaster* reveals functionally conserved splicing enhancers in metazoans. *RNA*, **17**, 1884–1894.
86. Wray,G.A., Hahn,M.W., Abouheif,E., Balhoff,J.P., Pizer,M., Rockman,M.V. and Romano,L.A. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, **20**, 1377–1419.
87. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engström,P.G. and Frith,M.C. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
88. Gaszner,M. and Felsenfeld,G. (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.*, **7**, 703–713.
89. Yang,J. and Corces,V.G. (2011) Chromatin insulators: a role in nuclear organization and gene expression. In: *En Advances in Cancer Research*. Academic Press, pp. 43–76.
90. Kim,T.H., Abdullaev,Z.K., Smith,A.D., Ching,K.A., Loukinov,D.I., Green,R.D., Zhang,M.Q., Lobanenkov,V.V. and Ren,B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231–1245.
91. Kharchenko,P.V., Alekseyenko,A.A., Schwartz,Y.B., Minoda,A., Riddle,N.C., Ernst,J., Sabo,P.J., Larschan,E., Gorchakov,A.A., Gu,T. *et al.* (2011) Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature*, **471**, 480–485.
92. Chorev,M. and Carmel,L. (2012) The function of introns. *Front. Genet.*, **3**, 55.
93. Shaul,O. (2017) How introns enhance gene expression. *Int. J. Biochem. Cell Biol.*, **91**, 145–155.
94. Niu,D.-K. and Yang,Y.-F. (2011) Why eukaryotic cells use introns to enhance gene expression: splicing reduces transcription-associated mutagenesis by inhibiting topoisomerase I cutting activity. *Biol. Direct.*, **6**, 24.
95. Rose,A.B., Emami,S., Bradnam,K. and Korf,I. (2011) Evidence for a DNA-based mechanism of intron-mediated enhancement. *Front. Plant Sci.*, **2**, 98.
96. Gallegos,J.E. and Rose,A.B. (2015) The enduring mystery of intron-mediated enhancement. *Plant Sci.*, **237**, 8–15.
97. Zdobnov,E.M., von Mering,C., Letunic,I., Torrents,D., Suyama,M., Copley,R.R., Christophides,G.K., Thomasova,D., Holt,R.A., Subramanian,G.M. *et al.* (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science*, **298**, 149–159.
98. Meireles-Filho,A.C. and Stark,A. (2009) Comparative genomics of gene regulation—conservation and divergence of *cis*-regulatory information. *Curr. Opin. Genet. Dev.*, **19**, 565–570.
99. Kowalczyk,M.S., Hughes,J.R., Garrick,D., Lynch,M.D., Sharpe,J.A., Sloane-Stanley,J.A., McGowan,S.J., De Gobbi,M., Hosseini,M. and Vernimmen,D. (2012) Intragenic enhancers act as alternative promoters. *Mol. Cell*, **45**, 447–458.