A draft genome sequence of an invasive mosquito: an Italian *Aedes albopictus*

Vicky Dritsou^{*1}, Pantelis Topalis^{*2}, Nikolai Windbichler³, Alekos Simoni³, Ann Hall³, Daniel Lawson⁴, Malcolm Hinsley⁴, Daniel Hughes⁴, Valerio Napolioni¹, Francesca Crucianelli¹, Elena Deligianni², Giuliano Gasperi⁵, Ludvik M. Gomulski⁵, Grazia Savini⁵, Mosè Manni⁵, Francesca Scolari⁵, Anna R. Malacrida⁵, Bruno Arcà⁶, José M. Ribeiro⁷, Fabrizio Lombardo⁶, Giuseppe Saccone⁸, Marco Salvemini⁸, Riccardo Moretti⁹, Giuseppe Aprea⁹, Maurizio Calvitti⁹, Matteo Picciolini¹, Philippos Aris Papathanos^{3,10}, Roberta Spaccapelo¹⁰, Guido Favia¹¹, Andrea Crisanti^{1,3,10}, Christos Louis^{2,10}

¹Polo d'Innovazione di Genomica, Genetica e Biologia (Polo GGB), Loc. S. Andrea delle Fratte, Perugia, Italy, ²Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas, Heraklion, Crete, Greece, ³Imperial College London, Department of Life Sciences, South Kensington Campus, UK, ⁴European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genomes Campus, Cambridge, Cambridgeshire, UK, ⁵Department of Biology and Biotechnology, University of Pavia, Italy, ⁶Department of Public Health and Infectious Diseases – Division of Parasitology, Sapienza University of Rome, Italy, ⁷National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA, ⁸Department of Biology, University of Naples Federico II, Italy, ⁹ENEA – Italian National Agency for New Technologies, Energy and Sustainable Economic Development, CR Casaccia, Rome, Italy, ¹⁰Department of Experimental Medicine, University of Perugia, Italy, ¹¹Scuola di Bioscienze e Medicina Veterinaria, University of Camerino, Italy

The draft genome sequence of Italian specimens of the Asian tiger mosquito *Aedes (Stegomyia) albopictus* (Diptera: Culicidae) was determined using a standard NGS (next generation sequencing) approach. The size of the assembled genome is comparable to that of *Aedes aegypti*; the two mosquitoes are also similar as far as the high content of repetitive DNA is concerned, most of which is made up of transposable elements. Although, based on BUSCO (Benchmarking Universal Single-Copy Orthologues) analysis, the genome assembly reported here contains more than 99% of protein-coding genes, several of those are expected to be represented in the assembly in a fragmented state. We also present here the annotation of several families of genes (tRNA genes, miRNA genes, the sialome, genes involved in chromatin condensation, sex determination genes, odorant binding proteins and odorant receptors). These analyses confirm that the assembly can be used for the study of the biology of this invasive vector of disease.

Keywords: NGS, WGS, BUSCO, Repetitive DNA, Transposable elements, Invasive species, Disease vector, Dengue fever, Chikungunya

Introduction

Aedes albopictus, the Asian tiger mosquito native in Southeast Asia, depicts a very aggressive host seeking and biting behaviour that causes a high degree of nuisance and, most significantly, it also acts as a vector for several important diseases such as arboviral (e.g. dengue virus, West Nile virus, chikungunya

virus) and other parasitic (e.g. dirofilaria) infections.¹ Ever since its presence was recorded in Albania² and Italy³ in the mid 1970s and late 1980s, respectively, there have been increased voices asking for intensified monitoring in order to prevent a potential re-emergence of dengue fever in Europe. Indeed, a major dengue fever epidemic in Greece occurred about 90 years ago: two back-to-back dengue fever epidemics, transmitted by *Aedes aegypti* (*Stegomyia fasciata*), resulted in almost one million cases and, officially, 1553 deaths mostly in Athens.⁴⁻⁶ Since then, vectors of dengue fever were only occasionally reported in Europe. *A. aegypti* for example, which

Correspondence to: Christos Louis, IMBB-FORTH, N. Plastira 100, Vassilikia Vouton, GR-700 13 Heraklion, Crete, Greece. Email: louis@imbb.forth.gr

* Should be considered as equal first authors

was first recorded in Macedonia^{7,8} and was then found throughout Greece,⁹ was never again collected there, even in systematic searches, following the DDT-based malaria eradication campaign that took place in the 1950s.¹⁰ In spite of the absence of dengue fever's classical vector, the recent continent-wide invasion of the Asian tiger mosquito¹¹ has prominently included this disease in the agenda of European control agencies.

The spread of *A. albopictus* in Europe, and in particular in Italy, seems to have originated from the Americas, whose invasion is assumed to have occurred less than a decade earlier.^{12,13} Most, if not all movements of this species across continents are assumed to be the result of worldwide trade, especially, that of used car tyres.¹⁴ By now, wild Asian tiger mosquito specimens have been isolated in a large number of countries in all continents with the exception of the Antarctic (see¹⁵ for a map of its current global distribution).

Although European disease control agencies feared the imminent re-emergence of dengue fever, it was, first, a small epidemic caused by a different arbovirus, chikungunya virus (CHIKV) that was attributed to transmission by *A. albopictus*. In the summer of 2007, a few hundred people near the coast of Emilia-Romagna, Italy, were infected with the CHIKV virus (reviewed by Tomasello and Schlagenhauf¹⁶). This was followed, indeed, by several isolated cases of dengue fever in different locations in Europe, and an epidemic outbreak of the disease did then occur in Madeira.¹⁷ It should be noted, of course, that although it belongs to Portugal, Madeira is located about 900 km southwest of continental Europe.

The present situation is certainly not dramatic from the public health point of view, at least not in terms of the actual cases of diseases transmitted by *A. albopictus*. It is clear, though, that to keep it from becoming worse, measures have to be taken, ideally in the sense of a potential eradication of *A. albopictus* from Europe the soonest possible. This implies a boosted investment in the research on the biology of the Asian tiger mosquito, especially in its European environment.¹⁸ There has been an increased emphasis in the study of 'European' tiger mosquito 'strains', mostly referring to its population biology and ecology.¹⁵ As members of the European Infravec Consortium,¹⁹ we decided to shift gears and to determine the sequence of specimens of *A. albopictus* isolated from the wild in Italy. This would provide an additional advantage, namely the possibility to compare its 'post-invasion' genome to the genome of Asian specimens once this is publicly available (Xiaoguang Chen, Anthony A. James, *et al.*, personal communication) as well as the genome of a strain isolated in the South Indian Ocean island of La Réunion.²⁰

Although the time frame of the establishment of the species in Italy is extremely short in an evolutionary scale (probably <30 years), it is a fact that the Asian tiger mosquitoes in this country have gone through one, and possibly two major bottlenecks. These bottlenecks occurred during their probable initial transfer from East Asia to the United States, followed a few years later by the second transfer to Italy. Furthermore, the mosquitoes may have been crossed to insects that arrived in Italy from other geographic locations such as, for example, Albania. This could have happened both in the field and, possibly, even in the laboratory, during the process of establishing the strain *Fellini*, which was utilised as the source of biological material in the study reported here.²¹

We report here the determination of the whole genome sequence of the *Fellini* strain of *A. albopictus*, as well as its preliminary characterisation. We discuss the assembly obtained, present several examples of annotated genetic features (gene families and repetitive elements) and also discuss the reasons for the difficulties in obtaining a higher quality assembly.

Materials and Methods

Biological material and preparation of genomic DNA

The strain *Fellini* (a.k.a. Rimini) was used for the determination of the genome sequence. This was derived from 500 eggs collected in 2004 with ovitraps in an urban environment from the northern Italian city of Rimini (44°03'24" N 12°33'52" E, Romeo Bellini, personal communication). At generation 40, 10 000 eggs were transferred to Imperial College. Multiple rounds of isofemale selection to reduce heterozygosity of the colony were carried out. At generation 73 about 3800 larvae (stage L4) were collected and total genomic DNA was extracted in batches of 15–25 using either the Wizard® Genomic DNA Purification Kit (Promega) according to the manufacturer's instructions, or by Phenol:Chloroform:Isoamyl alcohol (25:24:1, Sigma-Aldrich, St. Louis, MO, USA) and purified by EtOH. The extracted genomic DNA (8465 µg) was resuspended in 10 mM Tris-1 mM EDTA buffer.

Determination of genome size

The genome size of the *Fellini* strain was estimated at 0.94 Gb using a real-time PCR-based method.²² A series of ten-fold dilutions of known concentration of a linearised plasmid containing the amplicon sequence corresponding to part of the *A. albopictus* G6PDH gene identified from RNA-seq data (Genbank/EMBL Accession Nr KT279821) was monitored in real-time using a MiniOpticon (Bio-Rad). As the length of the amplicon was known, the concentrations of the dilutions could be

calculated as copies per microlitre. Comparison of the calibration curve of the C_T values derived from the standard dilution series with those derived from the same G6PDH amplicon in genomic *A. albopictus* DNA samples of known concentration, permitted the absolute number of target gene copies to be calculated, and hence the genome size.

Library preparation and next generation sequencing

Mate-Pair (MP) and Paired-End (PE) libraries were prepared using Nextera Mate-Pair and Nextera DNA Sample Preparation kits (Illumina, San Diego, CA, USA), respectively, starting from one microgram of gDNA for MP and 50 ng of gDNA for the PE libraries; the manufacturer's instructions were used. A total of 14 libraries were generated (seven for each type). Mate-Pair and PE libraries were then separately pooled and clustered on eight different lanes of one HiSeq PE Flow Cell v3 (Illumina, San Diego, CA, USA). Library quality control was performed on a 2100 Bioanalyzer using a High Sensitivity DNA chip (Agilent Technologies). The libraries were quantified using Qubit dsDNA BR Assay (Life Technologies) and by qPCR using the Illumina library quantification kit (Kapa Biosystem). The sequencing was performed on Illumina's HiSeq1500 with a 2×93 cycling. The whole sequencing run generated 252 Gb (94% $\geq Q30$). The paired-end library had a $60 \times$ coverage and an insert size equal to 400 bp.

To determine the average insert size of the MP library to be used for scaffolding, we mapped all 438 867 831 reads included in the contigs of the Mi and So assemblies (see *Assembly*) with a size of 10 Kb or larger (256 and 218 contigs, respectively) using Seqmap²³ allowing no mismatches. We then identified the respective mate pairs that could be mapped within the same contig and calculated the distance between the two reads.

Small RNA sequencing

The protocol described by Friedländer *et al.*²⁴ was followed. Essentially, total RNA was prepared from 30 whole 4- to 5-day-old mosquitoes using TRIzol (Invitrogen) according to manufacturer's protocol. Males, sugar-fed females and blood-fed females (1 day post feeding) were sampled. Three independent biological replicates for sugar fed and two for blood-fed mosquitoes were used for the small RNA library preparation. Strand-specific cDNA libraries with different barcodes (six base index) were generated using a TruSeq Small RNA kit (Illumina). Small RNA library was validated with 2100 Bioanalyzer (Agilent) and sequenced with MiSeq (Illumina). Adaptor and index sequences were extracted from the Miseq raw data.

Assembly

For the assembly of the reads, we used a 64-core cluster with a total of 512 Gb of RAM. After having tested a number of software/assemblers (e.g. Velvet,²⁵ ABySS²⁶), we decided to proceed with SOAPdenovo,²⁷ which produced the most qualitative result. Due to hardware limitations, we divided the PE reads into two groups named Mi and So, each one containing half of the initial reads (i.e. each one containing ~ 625 million reads or a $30 \times$ genome coverage each). We then assembled them individually and subsequently used the MP library for the scaffolding procedure. Following this, the two assemblies were merged, a procedure that included the removal of redundant contigs as well as all contigs that were shorter than 260 bp (see Results and Discussion section). The assembly was concluded by building scaffolds using, again, SOAPdenovo. Assembly 1.0 (see Results and Discussion) was submitted to the NCBI with the BioProject ID number PRJNA289460.

Given the relatively poor statistics of the assembly obtained, we assessed its overall inclusiveness/completeness by performing a BUSCO analysis.²⁸

Annotation

The automatic annotation of the genome was performed using the MAKER pipeline.²⁹ In an attempt to provide more evidence to the pipeline for the annotation process, we have used, within MAKER, evidence from publicly available *A. aegypti* ESTs as well as from RNA-seq data of male and female antennae (Fellini strain, Gomulski, Gasperi *et al.*, manuscript in preparation), and female antennae and palps, whole females and male heads isolated from a strain of *A. albopictus* established in 2012 from wild mosquitoes collected in Rome (Arcà *et al.*, manuscript in preparation). In addition to the automatic annotation, a group of collaborators working on different research areas are currently working towards the manual annotation of the genome. A summary analysis of some of these data is reported here in order to provide an assessment of the quality of the genome assembly. Further details will be published in a series of manuscripts that will follow. For the manual annotation of the following gene families the specifics are as follows.

Manual annotation of specific genes

Masking and identification of repetitive sequence elements

To mask the repeats present in the assembly, we used the MAKER pipeline²⁹ due to several advantages that this offers: First, it consists of two different masking pieces of software, RepeatMasker and RepeatRunner that both accept and use external libraries of repeated elements. In our case, we used the version 20.0.1 of RepBase plus the default RepeatRunner library,³⁰

in addition to the *A. aegypti* repeated elements reported in VectorBase [https://www.vectorbase.org/download/info/aedes-aegypti-liverpoolrepeatslib]. Second, the MAKER pipeline can take advantage of computers with multiple cores rendering the analysis more rapid and, third, it processes each scaffold/contig separately allowing the whole process to be resumed in case of involuntary interruptions during the run.

tRNA genes

To identify tRNA genes, we used the tRNA-SE³¹ software on the previously masked sequences. Again, this software can be used from within the MAKER pipeline for the reasons stated above.

miRNA-encoding genes

For the detection of reads representing miRNAs, miRDeep2 was used using default parameters.²⁴ The miRDeep2 core algorithm requires as inputs two reference files: one of mature and one of precursors miRNAs of related organisms in addition to a reference file of mature miRNAs of similar organisms. For the related species, mature and precursor miRNAs sequences from the *A. albopictus* cell line C710 were used³¹ while for the similar species, we used all the mature miRNA sequences of metazoan found in miRBase v.21 (http://www.mirbase.org)

Genes involved in chromatin condensation

Various genes involved in chromatin compaction³² have been selected based on their differential expression associated to the presence of the endosymbiont *Wolbachia*³³ while trying to identify factors of paternal origin, which contribute to the Cytoplasmic Incompatibility (CI) phenomenon.³⁴ Specifically, three RNA-seq samples from *A. albopictus* testes belonging to three lines of *A. albopictus* characterised by a different *Wolbachia* infection pattern (A: wild-type infection, i.e. wAlbA and wAlbB *Wolbachia* strains; 0: no infection, obtained by antibiotic treatment; P: artificial infection with wPip *Wolbachia* strain, obtained by transinfection)³⁵ have been used for a differential gene expression analysis. Reads were aligned to the *A. albopictus* transcriptome³⁶ using TopHat v2.0.13,^{37–39} with default options except for the following: read mismatch set to 3. Cuffdiff v2.2.1^{40,41} was then used to identify differentially expressed genes. Cufflinks was run, followed by Cuffmerge and Cuffdiff, in all cases with default options. Significance was scored at a cut-off false discovery rate (FDR) of 0.05. After that, genes known to be involved in chromatin condensation and showing a significantly different expression related to *Wolbachia* infection pattern were mapped to the *A. albopictus* genome with the TBLASTN tool v2.2.28+,^{42,43} with the following non default option: max intron length set to 5000, estimated

according to Nene *et al.*⁴⁴ In addition, a set of 10 key genes from *Drosophila melanogaster* (infected by wMel *Wolbachia* strain) known to be implicated in chromatin regulation were also tentatively mapped to the *A. albopictus* genome to ascertain the presence of orthologues.

Odorant binding protein and Odorant receptor genes

TBLASTN searches were performed using as queries protein sequences of the 111 Odorant binding proteins (OBPs) and 112 Odorant receptors (ORs) of *A. aegypti*.^{45–47} The contigs and scaffolds that produced hits ($e < 1E-10$) were used to interrogate, using BLASTX, local protein databases of the *Ae aegypti* OBPs and ORs. Multiple contigs and scaffolds with hits to an OBP or OR were assembled using CAP3.⁴⁸ Gene-Wise⁴⁹ was used to obtain gene model predictions based on homology with *A. aegypti* OBP and OR proteins.

Sex-determination genes

A TBLASTN search of the present assembly was performed using, as queries, the protein sequences of 14 putative sex determining genes of *A. aegypti*⁴⁴ and the *Ceratitis capitata* TRA protein sequence (GenBank acc. num.: AF434936.1⁵⁰). For the *A. aegypti* genes, we utilised the longest *protein sequence*, in case of multiple isoforms for a given gene, except for the *dsx* and *fru* orthologues for which we utilised the full protein sequence of each sex-specific isoforms.

Salivary gland genes

Salivary proteins/cDNAs previously identified in *A. albopictus*⁵¹ and *A. aegypti*⁵² were mapped to the genome using the BLAT tool,⁵³ and compared to other databases using BLASTX or RPS BLAST.⁵⁴ Results are displayed in a hyperlinked spreadsheet as previously described.⁵⁵ Manual annotation of salivary genes was done with the tool Artemis.⁵⁶

Results and Discussion

Assembly

We obtained sequencing reads on an Illumina's HiSeq1500 automatic sequencer to a final coverage of about 60-fold from a paired-end library and about 60-fold from a Mate Pair library. We faced a clear failure of most assembly software used to achieve an assembly that would have an 'acceptable' N50 value for contigs. The most obvious explanation for this is the high number of repetitive sequences present in the genome of the Asian tiger mosquito⁵⁷ combined with the short length of the NGS sequence reads. After repeated attempts, we decided to settle on an assembly that was based on the merger of two sub-assemblies (Mi and So), each performed with one half of the 'short' paired end reads obtained that yielded the best results, and then try to improve it by further 'cleaning', i.e. removal of duplicate contigs (see Table 1 for assembly statistics).

We proceeded with the elimination of singletons with a length <100 nt (~52% of the total). Then all remaining contigs were aligned to each other using the nucmer application of the MUMmer package.⁵⁸ The contigs of Mi and So were then divided into four categories: (i) Contigs identical between Mi and So, (ii) Contigs in So fully contained within Mi contigs (identity 99%), (iii) Contigs in So which contained a Mi contig (identity 99%) and finally (iv) contigs that showed no or only partial overlap. We then removed from the assembly all Mi contigs that were identical to So contigs as well all Mi contigs that were contained within So contigs, a total of 3 152 748. Similarly, we also removed from the assembly 2 445 970 So contigs contained in Mi contigs, making sure that we did not remove contigs identical to previously removed Mi contigs from the first two steps. We therefore verified this and re-added those to the assembly (1 169 778 contigs). At the end 5 326 714 contigs remained, which were then used as input to SOAPdeNovo for scaffolding.

For scaffolding, the average insert size determined for the MP library was used. We calculated the distance between all the mate pairs we identified within contigs longer than 10 Kb and we determined an average distance between them of 3150 bp with a median value 2930 bp. Based on that, we used an insert size of 3000 bp for the scaffolding of the contigs to produce the pre-final assembly that we called MiSo 1.0. After analysing the MiSo 1.0 assembly for the presence of repetitive DNA segments, we finally removed from the assembly all contigs that had a length between 100 and 260 bp (see the BUSCO analysis, below), and finally called this assembly MiSo 1.01, which is the assembly that we used for annotation purposes. The relevant overall statistics of the assembly are shown in Table 1.

At the time of writing this paper, we repeated the sizing of the MP library, this time using all contigs from the MiSo assembly that had a size of 3.5 Kb or larger (3120 contigs) and this led to an estimated average insert size of 2032 bp with a mean value of 2186 bp.

Evaluation of the fragmentation and completeness of the assembly

Given the computed low contig N50 number obtained for our assembly, we used an alternative approach for the evaluation of its completeness. This consisted of a BUSCO analysis, which is based

on the determination of the presence, in a given assembled genome of a set of single copy orthologues.²⁸ In our case, we used BUSCOs from both *A. gambiae* and *D. melanogaster*, the closest species to *A. albopictus* for which such sets were available. The results are shown in Table 2.

The fruit fly based analysis yielded results that were very similar to those obtained with the mosquito orthologues (less than 5% divergence) verifying the solidity of the methodology. The overall analysis with 2685 *A. gambiae* BUSCOS showed that 2662 (99.14%) of them were found in our assembly while 23 were missing. Of these positive BUSCOs, 1747 (or 65.63% of all present in the assembly) were found to produce single hits, while 925 (the remaining 34.37%) yielded multiple hits. As BUSCOS are ideally expected to produce single hits, the multiple-hit BUSCOs can only represent genes that are all present in the assembly, though exhibiting varying degrees of fragmentation.

Concerning the BUSCOs that yield no hit, one explanation is that they are missing from the assembly, but another acceptable explanation is that a missing BUSCO does not represent *a priori*, a problem in the assembly. For example, one *A. gambiae* BUSCO is also missing from the complete, version 6 assembly of the *D. melanogaster* genome, while in the opposite analysis, 21 fruit fly BUSCOs (0.68%) are missing from the *A. gambiae* assembled genome (not shown) that is based on a much more complete assembly than the one reported here. In addition to a given BUSCO being absent from the assembly, other reasons for such missing BUSCOs may be an extreme fragmentation of the gene, a gene that has undergone substantial sequence changes in evolution such that, although present in the genome, it no longer fulfils the criteria set for a positive result in the analysis; finally, a gene can always be missing even from an evolutionarily close relative. This latter explanation can be best exemplified by the fact that although <1% of the *A. gambiae* BUSCO orthologues are absent from the assembly, the percentage of missing BUSCOs goes up to 2.6 when the fruit fly orthologues are analysed (not shown). Finally, as stated, around one-third of all BUSCOs found in the assembled Asian tiger mosquito genome are found in more than one contig/scaffold, as a result of fragmentation of the genes in question. This fragmentation is

Table 1 The table shows the summary statistics of the assemblies described in this paper

	Size with N, in bp	Size without N, in bp	No of scaffolds + contigs	No of scaffolds	No of singletons	Scaffolds N50	Contigs N50
MiSoClean (Ass. 1.0)	2 432 868 255	1 589 519 495	4 901 513	189 306	4 712 207	1105	341
MiSoClean with contigs >260 (Ass. 1.01)	1 965 518 921	1 122 114 752	2 141 557	189 141	1 952 416	3255	516

Table 2 The table shows the summary statistics of the BUSCO analysis described in this paper

	No hit DMELA	Single hits	Multiple hits	%	No hit AGAMB	Single hits	Multiple hits	%
<i>Drosophila melanogaster</i>	–				1	3	2681	99.96
<i>A. gambiae</i>	21	2922	129	99.32	–			
All scaffolds and contigs	80	1988	1004	97.4	23	1747	915	99.14
Only scaffolds	881	1867	324	71.32	653	1707	325	75.68
All contigs only	410	1703	959	86.66	232	1540	913	91.36
All contigs ≥ 2000 bp	2650	372	50	13.74	2299	347	39	14.38
260 $\leq$ contigs < 2000	525	1634	913	82.91	318	1486	881	88.16
Contigs ≥ 260	438	1711	923	85.74	246	1555	884	90.84

DMELA and AGAMB refer to the *D. melanogaster* and *A. gambiae* BUSCOS, respectively. The first two rows show the number of hits detected when using the two sets of BUSCOS on the current assemblies of the fruit fly and the *A. gambiae*, while the next rows show the respective analysis performed on the *A. albopictus* assembly.

certainly smaller for the BUSCOs that are present in larger contigs and scaffolds (10–15%).

We also performed the BUSCO analysis on subsets of the assembly, namely on scaffolds alone, on all contigs, and on all contigs longer than 2000 bp and those with a size ranging from 260 to 2000 bp. These data are also shown in Table 2. Here, two interesting observations are worth mentioning. First, the higher frequency at which BUSCOs are found in large contigs, not unexpectedly, is $\sim 10\times$ higher than the total average and, second, the rate of single to multiple hits is also highest for large contigs followed by BUSCOs found in scaffolds.

The results of the BUSCO analysis, taken together, show that the assembled *A. albopictus* genome can be used for the isolation of most protein-coding genes that are to be studied, although due to fragmentation, this may require some additional steps such as PCR-based isolation and additional sequencing steps.

Finally, the BUSCO analysis provides another significant piece of information. With only one exception, all *A. gambiae* BUSCOs present in the assembled genome are found in contigs and scaffolds that are longer than 260 bp, strongly suggesting that the high number of short contigs present in the assembly are comprised of sequences that do not code for proteins. Based on that and additional pieces of evidence such as the repeat masking and further computations (not shown) we decided, as briefly mentioned earlier, to include in the ‘final’ assembly only contigs that are longer than 260 bp as well as all scaffolds that were created.

We also performed a masking of repeats present in the assembly; the results are described in Table 3. We have separated the total assembly in four categories, namely scaffolds, contigs between 261 and 2000 bp in length and contigs longer than 2,000 bp (the longest contig in the assembly has a length of 39 933 bp) as well as, naturally, the total assembly.

Assembly 1.01 has a total length of ~ 1.12 Gb (before scaffolding) and ~ 1.97 after scaffolding using an average of 3,000 bp for the length of the insert of the MP library. The former number is

relatively close to genome size determined by quantitative PCR (i.e. 0.94 Gb), while the latter is clearly much higher. Obviously, using the recalculated average distance of the Mate Pairs, the length would become approximately 0.2 Gb shorter. The masking resulted in $\sim 34.12\%$ of the total length being tagged as representing repetitive sequences. Although the percentage of masked sequences is only 3% higher when the shorter contigs are examined, we have found that the percentage increases dramatically for the shortest contigs with a length of < 260 bp (not shown); this reaffirms our decision to exclude them from the assembly 1.01 (see above). We stress, of course, that the masking of the repetitive segments of the assembly is certainly not complete. The software used as well as the libraries used by the software can certainly not be comprehensive, thus several such repetitive elements may have remained undetected. This will naturally affect the size of the computed genome size that will also be affected by all contigs still present in the assembly that are not unique.

Annotation

Repetitive sequence segments

One of the problematic features of the project anticipated since its beginning was the large number of repeated sequences present in the genome of *A. albopictus*. The fact that older publications reported extremely varying sizes of the genome for wild specimens collected at different places and times made it probable that this was due to the massive expansion or loss of non-coding DNA, very possibly representing transposable elements.^{57,59} Since the differences of DNA content could between several populations be the result of the invasion process, we decided to address the issue in more detail. We approached the matter using the data obtained by the repeat masking of the assembly. A total of 27 501 repetitive segments were identified in the assembly that were non-identical; of those 797 were not further recognised through the databases used for the masking. These numbers represent a minimum of elements since if two or more copies of a

Table 3 Detailed statistics performed of the *A. albopictus* assembly, performed on fractions or the total assembly

	Fellini scaffolds	Fellini small contigs (260–000 bps)	Fellini large contigs (> 2000 bps 2001–39 933)	Total
Number #	189 306	1 946 237	6030	2 141 573
Total bases	1 151 063 905	798 233 382	16 259 831	1 965 557 118
Scaffolding Ns	843 348 760	0	0	843 348 760
% Scaffolding Ns	73.27	0	0	42.90
Remaining bases	307 715 145	798 233 382	16 259 831	1 122 208 358
Masked bases	101 286 315	276 390 922	5 164 936	382 842 173
% Masked bases	32.92	34.63	31.77	34.12
Unique bases	206 428 830	521 842 460	11 094 895	739 366 185
% unique bases	67.08	65.37	68.23	65.88
(Scaffolding Ns not included)				
% Unique bases (Scaffolding Ns included)	17.93	65.37	68.23	37.61

The numbers indicate bp except for the cases in which percentages are shown.

given sequence are present in the assembly, these will be counted as 1. The remaining hits contained several sequences derived from transposable elements known from other species, but also of sequences representing satellite DNA ($97 \times$ or 5.4% of the assembly) and simple repeats ($12\,496 \times$ or 1.12% of the assembly). Table 4 shows the percentage of the assembly taken up by different classes/families. Interestingly, 22 contigs in the assembly were found to contain rDNA sequences, a family of genes that is known to be extremely difficult to be assembled correctly in WGS (Whole Genome Sequence) projects.⁶⁰ The most abundant elements of the ‘Other’ class of Table 4 includes (but is not limited to) the fruit fly elements *17.6*, *S*, *Doc*, *Bari1*, *piwi*, *F*, *TART* and *Asterix* as well as the rice gypsy-type retrotransposon; these are listed in decreasing order of frequency identified, which, in no case, was higher than 0.2%.

A recent publication reported the assembly of the ‘repeatome’ of *A. albopictus*.²⁰ In this case, the strain sequenced was from the island of La Réunion in the Indian Ocean, where the Asian tiger mosquito has also recently been established. In addition to the high number of repeats, a direct comparison is rendered difficult by the fact that the two studies used different software utilising different algorithms and databases for the identification of repeated sequences. Still, in the 12 families of repeated sequences that can be directly compared between the two strains, one can only distinguish a few differences that could be considered to be significant (see Table 4). With the possible exception of the LINE/LOA family whose members are more frequently encountered in the assembly of the strain from La Réunion all other differences in the frequency of transposable elements are not significant enough to pinpoint to a noteworthy variance. Finally, although the frequency of the satellite sequences in the two assemblies is clearly different (6.5 times higher in the Italian strain), we believe that this only represents a dissimilarity that has ‘technical’ reasons (i.e. software used) and does not reflect real biological differences.

Automatic annotation pipeline

The *ab initio* analysis led to the ‘identification of 719 636 ORFs as well as a number of “gene models”’. Given the relatively low N50 value for the assembly as well as the previously concluded BUSCO analysis, we were not able to determine a somewhat more accurate value for the number of ‘genes’ present in the *A. albopictus* genome. To remedy this, we checked all sequences longer than 50 bps reported in the gff output (6 154 839 out of a total 77 066 684) and subjected them to both BLASTN and TBLASTX analysis querying the set of transcripts of *A. aegypti* obtained from Vectorbase. The results are shown in Table 5. The nucleotide searches indicate that the annotated genome contains 238 898 unique high scoring pairs (HSP) that are similar to 14 581 different *A. aegypti* transcripts with an identity that is higher than 80%. This value represents $\sim 77.34\%$ of the total of *A. aegypti* genes. The value obtained from the TBLASTX analysis is $\sim 99.87\%$ (18 815 with 14 822 to produce high scoring pairs longer than 50 bps). The fact that we miss approximately 0.13% of the genes in our assembly, based on the automatic annotation, could be interpreted as being due to a variety of reasons. The most obvious one, with a potentially biological significance, would be a divergence of the sequences due to changes that occurred throughout evolution, but technical reasons such as extreme fragmentation could also be blamed. Nevertheless, the values obtained at this level of analysis also suggest that our assembly is useful in identifying genes of *A. albopictus*. Significantly, this result also points to the high conservation between the two *Aedes* species on the level of the primary structure of the proteins encoded by the two genomes.

Annotation of specific genes: a potpourri of examples

tRNA genes

The number of tRNA genes in a given genome varies tremendously from organism to organism. This is true not only when comparing species that are distantly related to each other but it is also the case

Table 4 Percentage of the assembly taken up by different classes/families of repeated sequences

Family	Average size of family members	% of assembly	'French' data
Satellite	306	5.40	0.83%
LINE/RTE-BovB	251	3.25	3.41%
LINE/R1	256	1.45	1.26%
Simple_repeat	53	1.12	
LTR/Gypsy	200	1.00	
LTR/Pao	262	0.94	
LINE/LOA	231	0.94	1.93%
DNA	136	0.86	
DNA/Sola	141	0.54	0.25%
LTR/Copia	199	0.49	
LINE/CR1	222	0.49	0.14%
LINE/I	160	0.40	0.57%
SINE/tRNA	141	0.38	0.23%
DNA/TcMar?	167	0.21	
DNA/CMC-Chapaev	121	0.20	0.33%
LINE/RTE	158	0.19	
DNA/CMC-EnSpm	117	0.19	0.41%
LINE/Penelope	130	0.18	
LINE/L2	213	0.17	
DNA/hAT-hATm	144	0.16	0.11%
SINE/tRNA-I	115	0.14	0.13%
DNA/CMC-Chapaev-3	144	0.14	
DNA/CMC-Transib	156	0.14	
LINE/Jockey	221	0.11	
DNA/Zator	104	0.10	
LINE/Dong-R4	280	0.10	
DNA/Kolobok-Hydra	124	0.09	
Unknown	116	0.08	
RC/Helitron	161	0.07	
Low_complexity	53	0.07	
DNA/TcMar-Tc1	103	0.07	
LINE/L1	171	0.07	
DNA/Crypton	131	0.06	
DNA/TcMar-Fot1	98	0.05	
DNA?/Crypton?	104	0.04	
Unspecified	122	0.04	
DNA/MULE-MuDR	131	0.04	
LINE/L1-Tx1	222	0.03	
DNA/hAT-Charlie	77	0.03	
tRNA	66	0.03	
DNA/hAT-Tip100	107	0.02	
DNA/hAT-Ac	102	0.01	
DNA/hAT	101	0.01	
DNA/PiggyBac	103	0.01	
DNA/hAT-hATx	88	0.01	
LTR/ERV1	81	0.01	
DNA/P	103	0.00	
DNA/Maverick	68	0.00	
DNA/PIF-Harbinger	106	0.00	
DNA/TcMar-Tigger	83	0.00	
DNA/Dada	119	0.00	
DNA/hAT-Blackjack	88	0.00	
Other	80	13.99	

for species that are part of the same genus or even represent sibling species as was shown in comparisons of the genomes of species in the genera *Drosophila*⁶¹ and *Anopheles*.⁵⁷ In *A. Aegypti*, a total of 906 tRNA genes, including pseudogenes, have been reported for its annotated genome;⁶² the corresponding number is indicated as 984 in VectorBase (see Table 6). In the present analysis, whose results are also shown in Table 6, we identified 5852 tRNA genes, including 4094 pseudogenes. This extremely high number of genes includes 422 genes that encode tRNA-Gly, 432 tRNA-Ala and 276 tRNA-Met genes. Significantly, in these three cases, 88.6,

76.9, and 73.9%, respectively, are found in small contigs with a length between 260 and 2000 bp. Since our overall data indicated that this part of the assembly contains a higher number of repetitive sequences and protein-coding genes are found at a lower frequency, we would like to speculate that these three classes of tRNA genes may be associated to specific families of repetitive sequences, which were not masked, for example transposable elements not present in the libraries used for masking. We should note that a disproportionately high number of tRNA-Ala genes are also found in the genome of *A. aegypti*.⁶² Given the draft assembly, gene numbers

Table 5 Results of querying the *ab initio* gene predictions of the assembly using the set of transcripts of *A. aegypti* with BLASTN and TBLASTX

Identity% and alignments length (bps)	No. of genes similar to <i>A. aegypti</i> based on blastn	% (Blastn)	No. of genes similar to <i>A. aegypti</i> based on tlastx	% (tblastx)
90% and alignment length >90 bps	9005	47.80	9033	47.95
90% and alignment length >80 bps	9324	49.49	9930	52.71
90% and alignment length >70 bps	10075	53.48	10905	57.88
90% and alignment length >60 bps	10311	54.73	11874	63.03
90% and alignment length >50 bps	10496	55.71	12749	67.67
90% and alignment length >40 bps	10680	56.69	13456	71.42
90% and any alignment length	10913	57.92	18545	98.43
80% and alignment length >90 bps	13542	71.88	11091	58.87
80% and alignment length >80 bps	13754	73.00	12032	63.86
80% and alignment length >70 bps	14296	75.88	13022	69.12
80% and alignment length >60 bps	14394	76.40	13999	74.30
80% and alignment length >50 bps	14443	76.66	14822	78.67
80% and alignment length >40 bps	14497	76.95	15423	81.86
80% and any alignment length	14571	77.34	18815	99.87
70% and alignment length >90 bps	13844	73.48	12080	64.12
70% and alignment length >80 bps	14040	74.52	13072	69.38
70% and alignment length >70 bps	14574	77.36	14062	74.64
70% and alignment length >60 bps	14664	77.83	14998	79.61
70% and alignment length >50 bps	14705	78.05	15711	83.39
70% and alignment length >40 bps	14751	78.30	16249	86.25
70% and any alignment length	14821	78.67	18839	99.99
60% and alignment length >90 bps	13846	73.49	12699	67.40
60% and alignment length >80 bps	14042	74.53	13646	72.43
60% and alignment length >70 bps	14576	77.37	14582	77.40
60% and alignment length >60 bps	14666	77.85	15486	82.20
60% and alignment length >50 bps	14707	78.06	16132	85.63
60% and alignment length >40 bps	14753	78.31	16595	88.08
60% and any alignment length	14823	78.68	18840	100.00
50% and alignment length >90 bps	13846	73.49	13018	69.10
50% and alignment length >80 bps	14042	74.53	13938	73.98
50% and alignment length >70 bps	14576	77.37	14898	79.08
50% and alignment length >60 bps	14666	77.85	15733	83.51
50% and alignment length >50 bps	14707	78.06	16327	86.66
50% and alignment length >40 bps	14753	78.31	16768	89.00
50% and any alignment length	14823	78.68	18840	100.00

are likely overestimates and will change as the assembly is improved.

Odorant binding protein and Odorant receptor genes

A. albopictus, like most other mosquito species, relies on olfactory cues for host-seeking, blood-feeding and oviposition site detection. Insect olfaction is the result of a signal transduction cascade involving odorant binding proteins (OBPs) and odorant receptors (ORs), amongst others. The water-soluble OBPs bind odorant molecules that enter the pores of the sensilla and transport them through the lymph to activate the membrane-bound ORs.⁶³ In *A. aegypti*, 111 OBPs and 112 ORs have been identified.^{45–47} Previous studies have identified a small number of OBPs^{64–66} and only one OR⁶⁷ in *A. albopictus*.

Based on searches using the *A. aegypti* OBP and OR gene families, orthologues were identified for 110 and 98 *A. aegypti* OBPs and ORs, respectively (Tables S1 and S2). It was not possible to determine the actual number of OBPs and ORs present in the genome due to the extensive fragmentation of many of the genes. Fifty-four of the *A. albopictus* OBP sequences were complete, although the sequence

often spanned more than one contig. Partial sequences were obtained for 56 OBP orthologues, for 25 of which, it was not possible to identify the region corresponding to the signal peptide, perhaps due to the higher variability of these regions compared to the rest of the gene.

Thirty-five of the OR sequences were complete (although two contained frame-shifts), again frequently spanning several different contigs; six of the complete ORs were identified in single contigs, whereas 12 were identified in single scaffolds. Twenty-one of the incomplete OR sequences lacked one or both termini and encoded conceptual polypeptides between 305 and 411 amino acids in length. Orthologues of 14 *A. aegypti* ORs (OR3, 5, 14, 16, 17, 28, 65, 96, 98, 103, 107, 114, 124 and 127) appear to have been lost in *A. albopictus*, however, a number of gene lineage expansions are clearly evident, for example, AegOR60, AegOR61 and AegOR63 have at least two homologues in *A. albopictus*, whereas AegOR70 has at least three.

Further analyses, currently in progress including transcriptome data, will help complete the *A. albopictus* OBP and OR repertoires and will open

Table 6 tRNA genes identified in the assembly

Genes	Scaffolds	Small contigs (260-2000 bps)	Large contigs (2001-9993)	Total	<i>A. aegypti</i> tRNA genes in VectorBase
Total tRNAs	1172	4616	64	5852	984
Pseudo tRNAs	856	3202	36	4094	0
Remaining tRNAs	316	1414	28	1758	984
Gly	78	640	4	722	47
Val	11	34	2	47	48
Ala	90	332	10	432	327
Leu	8	35	0	43	58
Ile	7	8	2	17	5
Cys	3	4	0	7	14
Met	66	204	6	276	35
Asp	5	20	4	29	54
Asn	1	2	0	3	23
Glu	6	8	0	14	49
Gln	5	12	0	17	25
His	0	2	0	2	37
Arg	1	10	0	11	42
Lys	0	4	0	4	54
Phe	1	0	0	1	19
Tyr	1	0	0	1	28
Trp	1	4	0	5	14
Ser	13	26	0	39	38
Thr	12	32	0	44	24
Pro	3	6	0	9	38
SeC	1	5	0	6	1
Undetermined	3	20	0	23	4
Sup-TTA	0	6	0	6	0
Sup-CTA	0	0	0	0	0

the way for the functional characterisation of these proteins with the aim of developing novel and effective attractants and repellents for monitoring and control.

Genes involved in chromatin condensation

Cytoplasmic Incompatibility is a phenomenon of conditional sterility observed in various insect species, which is caused by the symbiotic bacterium *Wolbachia*; CI is initiated in an embryo when male and female pronuclei originated from parents that exhibited a different *Wolbachia* infection pattern. Specifically, if females are uninfected or infected by a non-compatible *Wolbachia* strain, karyogamy fails and an embryo does not develop. CI can be exploited to produce functionally sterile males to be released against wild target populations as suppressive tool.³³ In addition, it has also been suggested as a method to spread genes of interest or useful physiological features into wild populations (as an example, resistance to pathogens).⁶⁸ Therefore, understanding the genetic basis of this phenomenon could have a profound impact on the exploitation of *Wolbachia* for genetic control of insect pests in both agriculture and medicine. *Wolbachia* is not present in mature sperm but it is known that male pronuclei are indirectly affected by its presence during spermatogenesis.⁶⁹ Although various studies have aimed at determining all factors implicated in CI,⁷⁰⁻⁷² no model that can fully describe the phenomenon has yet been developed. What is certainly known is that the embryonic death resulting from CI crosses

is a consequence of improper paternal chromatin condensation leading to an abnormal migration of paternal chromosomes. This epigenetic modification has to be induced by *Wolbachia* before spermatid stage, when these bacteria are completely removed. Thus, we selected and analysed a series of *A. aegypti* genes known to be involved in chromatin condensation and whose orthologues were found to be differently expressed when *Wolbachia* is present or not in *A. albopictus* testes (Moretti *et al.*, unpublished data).

Based on a TBLASTN analysis, summarised in Table S3, *A. aegypti* orthologues mostly showed a significant sequence conservation in the genome assembly. The amino acid sequence identity referred to the single BLAST hits peaked in some cases at 100%. However, orthologues are generally scattered among different scaffolds/contigs (ranging 6-185). The best results regard two isoforms of the nucleosome assembly protein gene (*Nap1*) that are located in seven and six scaffolds/contigs, respectively, of the Fellini genome assembly. In the above cases, single scaffolds account for 63 and 99% of query coverage respectively with 32% of sequence identity in the former and 90% in the latter. In addition, regarding the first case, a contig was found to host a hit showing 95% similarity and representing 58% of the query coverage. Two chromatin accessibility complex (CHRAC) gene isoforms are also represented in a relatively low number of scaffolds/contigs of the genome. The analysis of the

first gave seven hits, among which the best scoring showed 62% of query coverage and 97% of max identity. Eight hits characterise the second CHRAC gene isoform with the best hit showing just 24% query coverage and 87% maximum identity.

The analysis of a second set of genes from *D. melanogaster*, known to be involved in chromatin regulation, also gave interesting results even though fewer hits were obtained generally with lower levels of query coverage and sequence identity (Table S4). The histone-binding protein Caf1 is significantly conserved in *A. albopictus*. A hit, out of 10, is located on a scaffold and showed 52% of query coverage and 97% of max sequence identity. An orthologue of the heterochromatin protein (HP1) was also found with 63% query coverage and 39% max identity. Again, the Nap1 protein mapped to the assembly, although with lower similarity compared to the two isoforms from *A. aegypti* (53% query coverage and 45% max identity referring to the best scored hit). Finally, three isoforms of a HIRA protein homologue were located in the assembly. The expression of this gene is already known to be down-regulated by *Wolbachia*, possibly determining some of the cytological effects characterising the CI phenomenon.^{70,71} Based on the genes/proteins analysed in this work, future research will aim at identifying the main factors determining the sperm modifications responsible for the CI.

Sex-determination genes

As is the case in other culicine mosquitoes, *A. albopictus* lacks heteromorphic sex chromosomes.^{73,74} In this species the primary signal of sex determination seems to be a dominant male determining factor linked to the chromosome 1, according to *A. aegypti*–*A. albopictus* linkage maps.^{75,76} We initiated the search for the *A. albopictus* genes involved in sex determination using the fruit fly homologues as starting blocks.

The results of this study are shown in Table S5. A TBLASTN analysis revealed sequence conservation between *Drosophila* orthologues and *A. albopictus* putative genes in the genome assembly, amino acid sequence identity ranging from 44 to 98%. Only two genes (the homologues of *dpm* and *vir*) are located in a single scaffold/contig of the present genome assembly. The other orthologues are distributed among scattered scaffolds/contigs. However, for each gene we identified, although scattered, all the *A. aegypti* orthologous exons. The name, length and ORF strand of the scaffolds/contigs containing each of the *A. albopictus* orthologues, as well as links to *Drosophila* and *A. aegypti* orthologue annotations (Flybase/Ensembl) and scaffolds/contigs statistics, are reported in Table S5. Interestingly, we have found that the male-specific and the female-specific *dsx* isoforms, similarly to *A. aegypti*, are conserved in *A. albopictus* (80 and 90%, respectively); moreover, *A. aegypti* and *A.*

albopictus possess FRU orthologues sharing 90% identity. As previously described for *A. aegypti* and other mosquitoes, a *tra* orthologue seems to be absent in the *A. albopictus* genome, leaving open the question on the molecular nature of the upstream splicing regulator/regulators of *dsx* and *fru* genes in this mosquito species. The *dsx* and *fru* genomic regions of *A. albopictus* will provide a basis for an in-depth comparative search of these and other potential *cis* elements eventually conserved in *Aedes* mosquitoes. Finally, a very recent report⁷⁷ showed that the *A. aegypti* *nix* gene encoding a putative splicing factor related to *Drosophila* *tra-2* seems to correspond to a Male determining factor only expressed in males. *Aedes nix* controls the *dsx* and *fru* male-specific splicing and interestingly it seems to be conserved in *A. albopictus* as a male-specific gene.

It is obviously conceivable that sex-specifically expressed genes such as *dsx* and *fru*, could be utilised to devise and apply biotechnological approaches for vector population suppression, proposed for other mosquito species.^{78–80} Their usage in approaches such as SIT and similar techniques in which sex separation is involved obviously jumps to mind.⁸¹

Salivary gland genes

Mosquito saliva plays an important physiological role in blood feeding through its anti-platelet, anti-clotting and vasodilatory activities, however it also affects pathogen transmission by virtue of its immuno-modulatory and anti-inflammatory properties.^{82,83} Salivary genes of blood feeding arthropods, perhaps also as a consequence of the host immune pressure, exhibit an accelerated evolutionary rate and this explains, at least in part, both the elevated diversification observed when comparing the salivary repertoires of anopheline to culicine mosquitoes and the presence of a substantial number of genus-specific salivary genes and gene families.^{84,85}

A total of 68 putative secreted products, mostly found in adult female saliva or expressed in the salivary glands of both sexes, were identified in a previous *A. albopictus* sialotranscriptome analysis.⁵¹ We selected 58 putatively secreted salivary proteins and searched the *A. albopictus* genome to proceed to manual annotation and obtain corresponding gene models. Genomic sequences encoding 35 full-length salivary proteins, and including among others serine protease inhibitors, D7 and Antigen five family members, 34 kDa proteins, a salivary lysozyme and putative antimicrobials of the HHH family, were retrieved. For other 19 proteins only partial sequences could be obtained, whereas a short D7 family member (gi 56417443), two putative 13 kDa proteins (gi 56417419, gi 56417421) and a 7.6 kDa protein (gi 56417477) were found to be most likely alleles or alternative splicing products

(not shown). The availability of the complete *A. albopictus* genome also allowed searching for orthologues of *A. aegypti* salivary proteins⁵² not previously identified in the tiger mosquito. This way 52 novel putative secreted salivary proteins were identified in *A. albopictus* and among these 20 represent full-length genes (Table S6). The availability of the complete salivary repertoire of *A. albopictus* will allow further functional studies and may help in developing reliable serological tools to evaluate human exposure to *Aedes* arboviral vectors, as previously done for *Anopheles malaria* vectors.^{86,87}

miRNA-encoding genes

The class of microRNAs (miRNAs) contains small, 22–23 nt long molecules that have been found to play crucial roles in regulating gene expression both in plants and metazoans. Several papers have dealt with the characterisation of miRNAs in mosquitoes, including *A. albopictus*.^{88,89} In the first referenced case a total of 104 miRNA species were identified while in the second referenced study 64 species were found; it should be stressed, though, that the latter study concerned the analysis of a cell line of the Asian tiger mosquito. In our analysis of the Assembly 1.01, we were able to discover a total of 93 miRNAs, of which 30 were characterised as novel. From those novel miRNAs, 26 have the same seed with a miRNA found in miRBase while four share no homology or similarity with any known miRNA stored in miRBase. The allocation of the miRNAs to three groups (male adult mosquitoes, sugar- and blood-fed adult females) was 84, 51 and 65, respectively. Naturally, some of the miRNAs were shared by the three groups; thus, 47 were common between male and sugar fed females, 47 between sugar fed and blood females while 57 between males and blood fed females. The miRNA genes identified are listed in Table S7. Whether our analysis has identified all miRNA genes present in the genome studied cannot be assessed, but due to the small size of the genes we are confident that, no genes will remain undetermined due to the fragmentation reported earlier.

Conclusions

We reported here the summary of the whole genome sequencing project of *A. albopictus*. This assembled genome does not represent an optimal result; indeed, as was the case for *A. aegypti*, the genome of *albopictus* is comprised of a large proportion of repetitive DNA segments, including many transposable elements, a fact that makes an optimal assembly extremely tedious, if not impossible. In addition to that, of course, in contrast to the published *A. aegypti* genome the *A. albopictus* genome was based entirely on NGS. Using this approach, the assembly into contigs of short sequences that are often repeated hundreds and thousands of times is

an extremely difficult undertaking. Given the fact that the assembly, as it stands today, can be used to isolate and study most protein-coding genes it can, thus, be considered a milestone to be used to ‘find one’s way’. This is especially true when one keeps in mind the cost-benefit of such a project.

The analysis of the genome did not provide any final clues as to the reasons of the dramatic differences seen in the size of the genomes of different strains of *A. albopictus*. It is more than probable that the high number of repetitive sequences (among which a large number of transposable elements) could be the cause for the differing genome size determined for this species. Whether the diversity only concerns, in this magnitude, specimens that have been collected in different parts of the globe, or whether it is also true for individuals that form part of the same population cannot be answered with the present data. We hope to be able to provide more precise answers when the three genomes that are currently near closure will be compared to each other. Finally, since our annotation had used the assembly 1.01, we reported these data in the present report, but we plan to release, in the near future, an updated assembly onto which the features presented here will be re-mapped, and which will be based on updated data such as, for example, the newly determined average insert size of the MP library.

Acknowledgements

We would like to dedicate this paper to our friend, colleague and mentor Fotis Kafatos on occasion of his 75th birthday. We want to thank Romeo Bellini for furnishing the initial stock of *A. albopictus* used for the determination of the genome sequence, Francesca Marini and Angiola Desiderio who helped with the transcriptomic analysis of *A. albopictus* testes aiming at identifying *Wolbachia*-induced sperm modifications.

Disclaimer statements

Contributors

VD and PT were responsible for the assembly and overall annotation of the genome, as well as for the annotation of the tRNA, miRNA genes and the repetitive sequences in the genome, NW, AS, AH, DL, MH, DH, VN, FC were responsible for the acquisition of the primary sequence data, ED assisted the acquisition of the short RNAseq data and was responsible for the annotation of genes encoding miRNAs, GG, LMG, GS, MM, FS, AM were responsible for the annotation of the odorant binding protein and odorant receptor genes, BA, JMR, FL were responsible for the annotation of the salivary gland genes, GS, MS were responsible for the annotation of the sex-determination genes, RM, GA, MC, were responsible for the annotation of the genes

involved in chromatin condensation, MP assisted with the bioinformatics analysis, PAP contributed to the interpretation of data, RS, GF, AC as well as all authors mentioned above helped draft and revise the article, AC coordinated the overall project, CL coordinated the bioinformatics part of the project and drafted the article.

Funding

This work was funded by the Infravec Consortium (FP7 programme of the European Union) and, partially, by funds from the Hellenic Secretariat General for research and Technology. VD and CL were supported, in part, by i-Move fellowships from the EU's Marie Curie programme. JMR was funded by the intramural programme of the NIAID. The sex determination studies were supported by a grant from Campania Region (LG 5/2001; 2008; Biotechnological approaches for the control of the human disease vector *Aedes albopictus*) to GS and by a grant from the University of Naples Federico II and Compagnia di San Paolo, Naples, ITALY, to MS in the frame of Programme STAR (STAR2013_25) from the University of Naples Federico II and Compagnia di San Paolo, Naples, ITALY. ARM was funded by a grant from the Italian Ministry of Health (RF-2010-2318965).

Conflicts of interest

None for any of the authors.

Ethics approval

Not applicable for this paper.

References

- Gratz NG. Critical review of the vector status of *Aedes albopictus*. *Med Vet Entomol*. 2004;18:215–27.
- Adhami J, Murati N. Prani e mushkonjës *Aedes albopictus* në shqipëri. *Rev Mjekësore*. 1987;1:13–16.
- Sabatini A, Raineri V, Trovato G, Coluzzi M. *Aedes albopictus* in Italia e possibile diffusione della specie nell'area mediterranea. *Parasitologia*. 1990;32:301–04.
- Halstead SB, Papaevangelou G. Transmission of dengue 1 and 2 viruses in Greece in 1928. *Am J Trop Med Hyg*. 1980;29(4):635–37.
- Rosen L. Dengue in Greece in 1927 and 1928 and the pathogenesis of dengue hemorrhagic fever: new data and a different conclusion. *Am J Trop Med Hyg*. 1986;35(3):642–53.
- Louis C. Daily newspaper view of dengue fever epidemic, Athens, Greece, 1927–1931. *Emerg Infect Dis*. 2012;18(1):78–82.
- Waterston J. On the mosquitoes of Macedonia. *Bull Entomol Res*. 1918;9:1–12.
- Joyeux C. Notes sur les Culicides de Macedoine. *Bull Soc Pathol Exot*. 1918;11:530–47.
- Samanidou-Voyadjoglou A, Darsie RF Jr. New country records for mosquito species in Greece. *J Am Mosq Control Assoc*. 1993;9(4):465–66.
- Livadas GA. Malaria eradication in Greece. *Riv Malariol*. 1958;37:173–91.
- Medlock JM, Hansford KM, Schaffner F, Versteirt V, Hendrickx G, Zeller H, et al. A review of the invasive mosquitoes in Europe: ecology, public health risks, and control options. *Vector Borne Zoonotic Dis*. 2012;12:435–47.
- Forratini OP. Identificacao de *Aedes (Stegomyia) albopictus* do Brasil. *Rev Saude Publica*. 1986;20(244):245.
- Hawley WA. The biology of *Aedes albopictus*. *J Am Mosq Control Assoc*. 1988;4(Suppl):1–40.
- Knudsen AB. Global distribution and continuing spread of *Aedes albopictus*. *Parasitologia*. 1995;37(2-3):91–97.
- Bonizzoni M, Gasperi G, Chen X, James AA. The invasive mosquito species *Aedes albopictus*: current knowledge and future perspectives. *Trends Parasitol*. 2013;29(9):460–68.
- Tomasello D, Schlagenhauf P. Chikungunya and dengue autochthonous cases in Europe, 2007–2012. *Travel Med Infect Dis*. 2013;11(5):274–84.
- Schaffner F, Mathis A. Dengue and dengue vectors in the WHO European region: past, present, and scenarios for the future. *Lancet Infect Dis*. 2014;14(12):1271–80.
- Medlock JM, Hansford KM, Versteirt V, Cull B, Kampen H, Fontenille D, et al. An entomological review of invasive mosquitoes in Europe. *Bull Entomol Res*. 2015;25:1–27.
- Crisanti A. INFRAVEC: research capacity for the implementation of genetic control of mosquitoes. *Pathog Glob Health*. 2013;107:458–62.
- Goubert C, Modolo L, Vieira C, Valiente-Moro C, Mavingui P, Boulesteix M. De novo assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). *Genome Biol Evol*. 2015;7(4):1192–205.
- Bellini R, Calvitti M, Medici A, Carrieri M, Celli G, Maini S. Use of the sterile insect technique against *Aedes albopictus* in Italy: first results of a pilot trial. In: Vreysen MJB, Robinson AS, Hendrichs J, editors. Area-wide control of insect pests: from research to field implementation. Dordrecht, The Netherlands: Springer; 2007; p. 505–15.
- Wilhelm J, Pingoud A, Hahn M. Real-time PCR-based method for the estimation of genome sizes. *Nucleic Acids Res*. 2003;31(10):e56.
- Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*. 2008;24:2395–96.
- Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res*. 2012;40(1):37–52.
- Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 2008;18:821–29.
- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. *Genome Res*. 2009;19(6):1117–23.
- Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*. 2010;20:265–72.
- Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res*. 2013; 41(Database issue):D358–65.
- Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using MAKER and MAKER-P. *Curr Protoc Bioinformatics*. 2014;48:4.11.1–4.11.39.
- Smith CD, Edgar RC, Yandell MD, Smith DR, Celniker SE, Myers EW, et al. Improved repeat identification and masking in dipterans. *Gene*. 2007;389(1):1–9.
- Lowe TM, Eddy S. tRNAScan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 1997;25:955–64.
- Molla-Herman A, Matias NR, Huynh J-R. Chromatin modifications regulate germ cell development and transgenerational information relay. *Curr Opin Insect Sci*. 2014;1:10–18.
- Bourtzis K, Dobson SL, Xi Z, Rasgon JL, Calvitti M, Moreira LA, et al. Harnessing mosquito-*Wolbachia* symbiosis for vectorand disease control. *Acta Trop*. 2014;132(Suppl):S150–63.
- Landman F, Orsi GA, Loppin B, Sullivan W. *Wolbachia*-mediated cytoplasmic incompatibility is associated with impaired histone deposition in the male pronucleus. *PLoS Pathog*. 2009;5(3):e1000343.
- Calvitti M, Moretti R, Lampazzi E, Bellini R, Dobson SL. Characterization of a new *Aedes albopictus* (Diptera: Culicidae)-*Wolbachia pipiens* (Rickettsiales: Rickettsiaceae) symbiotic association generated by artificial transfer of the wPip strain from *Culex pipiens* (Diptera: Culicidae). *J Med Entomol*. 2010;47:179–87.
- Poelchau MF, Reynolds JA, Denlinger DL, Elisk CG, Armbruster PA. A de novo transcriptome of the Asian tiger mosquito, *Aedes albopictus*, to identify candidate transcripts for diapause preparation. *BMC Genomics*. 2011;12:619.
- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105–11.
- Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.

- 39 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013;14:R36.
- 40 Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010;28:511–15.
- 41 Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol.* 2013;31:46–53.
- 42 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
- 43 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST – a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25:3389–402.
- 44 Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, et al. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science.* 22 June 2007;316(5832):1718–23.
- 45 Manoharan M, Chong MNF, Vaitinadapoulé A, Frumence E, Sowdhamini R, Offmann B. Comparative genomics of odorant binding proteins in *Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus*. *Genome Biol Evol.* 2013;5(1):163–80.
- 46 Bohbot J, Pitts RJ, Kwon H-W, Rützel M, Robertson HM, Zwiebel LJ. Molecular characterization of the *Aedes aegypti* odorant receptor gene family. *Insect Mol Biol.* 2007;16(5):525–37.
- 47 Bohbot JD, Sparks JT, Dickens JC. The maxillary palp of *Aedes aegypti*, a model of multisensory integration. *Insect Biochem Mol Biol.* 2014;48:29–39.
- 48 Huang X, Madan A. CAP3: a DNA sequence assembly program. *Genome Res.* 1999;9(9):868–77.
- 49 Birney E, Durbin R. Using Genewise in the *Drosophila* annotation experiment. *Genome Res.* 2000;10(4):547–48.
- 50 Pane A, Salvemini M, Delli Bovi P, Polito C, Saccone G. The transformer gene in *Ceratitis capitata* provides a genetic basis for selecting and remembering the sexual fate. *Development.* 2002;129(15):3715–25.
- 51 Arca B, Lombardo F, Francischetti IM, Pham VM, Mestres-Simon M, Andersen JF, et al. An insight into the sialome of the adult female mosquito *Aedes albopictus*. *Insect Biochem Mol Biol.* 2007;37:107–27.
- 52 Ribeiro JM, Arca B, Lombardo F, Calvo E, Phan VM, Chandra PK. An annotated catalogue of salivary gland transcripts in the adult female mosquito, *Aedes aegypti*. *BMC Genomics.* 2007;8:6.
- 53 Kent WJ. BLAT – the BLAST-like alignment tool. *Genome Res.* 2002;12:656–64.
- 54 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25:3389–402.
- 55 Ribeiro JM, Topalis P, Louis C. Anoxcel: an *Anopheles gambiae* protein database. *Insect Mol Biol.* 2004;13:449–57.
- 56 Berriman M, Rutherford K. Viewing and annotating sequence data with Artemis. *Brief Bioinform.* 2003;4:124–32.
- 57 McLain DK, Rai KS, Fraser MJ. Intraspecific and interspecific variation in the sequence and abundance of highly repeated DNA among mosquitoes of the *Aedes albopictus* subgroup. *Heredity (Edinb).* 1987;58(Pt 3):373–81.
- 58 Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biol.* 2004;5:R12.
- 59 Rao PN, Rai K. Inter and intraspecific variation in nuclear DNA content in *Aedes* mosquitoes. *Heredity (Edinb).* 1987;59:253–58.
- 60 Dritsou V, Deligianni E, Dialynas E, Allen J, Poulakakis N, Louis C, et al. Non-coding RNA gene families in the genomes of anopheline mosquitoes. *BMC Genomics.* 2014;15:1038.
- 61 Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature.* 2007;450:203–18.
- 62 Behura SK, Severson DW. Coadaptation of isoacceptor tRNA genes and codon usage bias for translation efficiency in *Aedes aegypti* and *Anopheles gambiae*. *Insect Mol Biol.* 2011;20:177–87.
- 63 Leal WS. Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes. *Annu Rev Entomol.* 2013;58:373–91.
- 64 Armbruster P, White S, Dzundza J, Crawford J, Zhao X. Identification of genes encoding atypical odorant-binding proteins in *Aedes albopictus* (Diptera: Culicidae). *J Med Ent.* 2009;46(2):271–80.
- 65 Li C, Yan T, Dong Y, Zhao T. Identification and quantitative analysis of genes encoding odorant binding proteins in *Aedes albopictus* (Diptera: Culicidae). *J Med Entomol.* 2012;49(3):573–80.
- 66 Deng Y, Yan H, Gu J, Xu J, Wu K, Tu Z, et al. Molecular and functional characterization of odorant-binding protein genes in an invasive vector mosquito, *Aedes albopictus*. *PLoS One.* 2013;8(7):e68836.
- 67 Scialò F, Hansson BS, Giordano E, Polito CL, Digilio FA. Molecular and functional characterization of the odorant receptor2 (OR2) in the tiger mosquito *Aedes albopictus*. *PLoS One.* 2012;7(5):e36538.
- 68 Walker T, Johnson PH, Moreira LA, Iturbe-Ormaetxe I, Frentiu FD, McMeniman CJ, et al. The wMel *Wolbachia* strain blocks dengue and invades caged *Aedes aegypti* populations. *Nature.* 2011;476:450–53.
- 69 Presgraves DC. A genetic test of the mechanism of *Wolbachia*-induced cytoplasmic incompatibility in *Drosophila*. *Genetics.* 2000;154:771–76.
- 70 Zheng Y, Ren P-P, Wang J-L, Wang Y-F. *Wolbachia*-induced cytoplasmic incompatibility is associated with decreased hira expression in male *Drosophila*. *PLoS One.* 2011;6(4):e19512.
- 71 Landmann F, Orsi GA, Loppin B, Sullivan W. *Wolbachia*-mediated cytoplasmic incompatibility is associated with impaired histone deposition in the male pronucleus. *PLoS Pathog.* 2009;5(3):e1000343.
- 72 Beckmann JF, Fallon AM. Detection of the *Wolbachia* protein WIP0282 in mosquito spermathecae: implications for cytoplasmic incompatibility. *Insect Biochem Mol Biol.* 2013;43:867–78.
- 73 McClelland G. Sex-linkage in *Aedes aegypti*. *Trans R Soc Trop Med Hyg.* 1962;56:4.
- 74 Craig GB, Hickey WA. Genetics of *Aedes aegypti*. In: Wright JW, Pal R, editors. Genetics of insect vectors of disease. New York: Elsevier; 1967; p. 67–131.
- 75 Severson DW, Mori A, Kassner VA, Christensen BM. Comparative linkage maps for the mosquitoes, *Aedes albopictus* and *A. aegypti*, based on common RFLP loci. *Insect Mol Biol.* 1995;4(1):41–45.
- 76 Sutherland IW, Mori A, Montgomery J, Fleming KL, Anderson JM, Valenzuela JG, et al. A linkage map of the Asian tiger mosquito (*Aedes albopictus*) based on cDNA markers. *J Hered.* 2011;102(1):102–12.
- 77 Hall AB, Basu S, Jiang X, Qi Y, Timoshevskiy VA, Biedler J. A male-determining factor in the mosquito *Aedes aegypti*. *Science.* 2015;348(6240):1268–70.
- 78 Papathanos PA, Bossin HC, Benedict MQ, Catteruccia F, Malcolm CA, Alphey L, et al. Sex separation strategies: past experience and new approaches. *Malar J.* 2009;8(Suppl 2):S5.
- 79 Alphey L, McKemey A, Nimmo D, Neira Oviedo M, Lacroix R, Matzen K, et al. Genetic control of *Aedes* mosquitoes. *Pathog Glob Health.* 2013;107(4):170–79.
- 80 Marinotti O, Jasinskiene N, Fazekas A, Scaife S, Fu G, Mattingly ST, et al. Development of a population suppression strain of the human malaria vector mosquito, *Anopheles stephensi*. *Malar J.* 2013;12:142.
- 81 Koukidou M, Alphey L. Practical applications of insects' sexual development for pest control. *Sex Dev.* 2014;8(1–3):127–36.
- 82 Ribeiro JM, Mans BJ, Arca B. An insight into the sialome of blood-feeding Nematocera. *Insect Biochem Mol Biol.* 2010;40:767–84.
- 83 Ribeiro JMC, Arca B. From sialomes to the sialoverse: an insight into salivary portion of blood-feeding insects. *Adv Insect Physiol.* 2009;37:59–118.
- 84 Arca B, Struchiner CJ, Pham VM, Sferra G, Lombardo F, Pombi M, et al. Positive selection drives accelerated evolution of mosquito salivary genes associated with blood-feeding. *Insect Mol Biol.* 2014;23:122–31.
- 85 Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science.* 2015;347:1258522.
- 86 Rizzo C, Ronca R, Fiorentino G, Verra F, Mangano V, Poinson A. Humoral response to the *Anopheles gambiae* salivary protein gSG6: a serological indicator of exposure to Afrotropical malaria vectors. *PLoS One.* 2011;6:e17980.
- 87 Stone W, Bousema T, Jones S, Gesase S, Hashim R, Gosling R, et al. IgG responses to *Anopheles gambiae* salivary antigen gSG6 detect variation in exposure to malaria vectors and disease risk. *PLoS One.* 2012;7:e40170.
- 88 Skalsky RL, Vanlandingham DL, Scholle F, Higgs S, Cullen BR. Identification of microRNAs expressed in two mosquito vectors, *Aedes albopictus* and *Culex quinquefasciatus*. *BMC Genomics.* 2010;11:119.
- 89 Gu J, Hu W, Wu J, Zheng P, Chen M, James AA, et al. miRNA genes of an invasive vector mosquito, *Aedes albopictus*. *PLoS One.* 2013;8(7):e67638.