DNA Research, 2018, 25(4), 375–382
doi: 10.1093/dnares/dsy010
Advance Access Publication Date: 2 April 2018
Full Paper

OXFORD

Full Paper

A new approach for comprehensively describing heterogametic sex chromosomes

Shenglong Li[1], Masahiro Ajimura[1], Zhiwei Chen[1], Jianqiu Liu[1], Enxiang Chen[1], Huizhen Guo[1], Vidya Tadapatri[2], Chilakala Gangi Reddy[2], Jiwei Zhang[1], Hirohisa Kishino[3], Hiroaki Abe[3], Qingyou Xia[1], Kallare P. Arunkumar[2], and Kazuei Mita[1,*]

[1]State Key Laboratory of Silkworm Genome Biology, Southwest University, Chongqing 400716, China, [2]Laboratory of Molecular Genetics, Centre for DNA Fingerprinting and Diagnostics, Hyderabad 500001, Telengana, India, and [3]Department of Biological Production, Faculty of Agriculture, Tokyo University of Agriculture and Technology, Fuchu, Tokyo 183-8509, Japan

*To whom correspondence should be addressed. Tel. +861 77 8317 3040. Fax. +860 23 6825 0892.
Email: mitakazuei@gmail.com

Edited by Dr. Toshihiko Shiroishi

Received 29 November 2017; Editorial decision 7 March 2018; Accepted 9 March 2018

Abstract

Notwithstanding the rapid developments in sequencing techniques, Y and W sex chromosomes have still been mostly excluded from whole genome sequencing projects due to their high repetitive DNA content. Therefore, Y and W chromosomes are poorly described in most species despite their biological importance. Several methods were developed for identifying Y or W-linked sequences among unmapped scaffolds. However, it is not enough to discover functional regions from short unmapped scaffolds. Here, we provide a new and simple strategy based on k-mer comparison for comprehensive analysis of the W chromosome in *Bombyx mori*. Using this novel method, we effectively assembled *de novo* 1281 W-derived genome contigs (totaling 1.9 Mbp), and identified 156 W-linked transcript RNAs and 345 W-linked small RNAs. This method will help in the elucidation of mechanisms of sexual development and exploration of W chromosome biological functions, and provide insights into the evolution of sex chromosomes. Moreover, we showed this method can be employed in identifying heterogametic sex chromosomes (W and Y chromosomes) in many other species where genomic information is still scarce.

Key words: heterogametic sex chromosome, silkworm, non-coding RNA

1. Introduction

In most animal species, sex-determination is governed by chromosomal differences involving two common sex determination systems: XX/XY sex chromosomes determine female/male, e.g. in *Homo sapiens*, *Drosophila melanogaster* and *Anopheles gambiae* and other species and WZ/ZZ sex chromosomes determine female/male, e.g. in *Bombyx mori*, *Gallus gallus*. Previously people believed a sex determining gene would be localized in a unique sequence domain of a heterogametic sex chromosome; thus many studies were focused on fishing or positioning unique sequence domains on W- or Y-chromosomes. In this way, *DMY* of the medaka fish was identified as a sex-determining gene.[1] However, Y and W chromosomes present special

challenges for whole genome shotgun assembly,[3–5] because they are enriched with interspersed repeat elements and segmental duplications.[6,7] Even though some unique domains exist on Y or W, many transposable elements (TEs) interrupt these unique sequences into short pieces, making it difficult to find such unique domains. Thus, whole genome shotgun methods, even the latest third generation sequencing techniques (e.g. PacBIO) which can produce very long reads, cannot effectively assemble high quality contigs or scaffolds of Y and W chromosomes at present stage,[8–10] which obstructs studies of sex-determination and sexual development and reproduction.

In most genome projects, Y and W chromosome sequences are fragmented into a large number of small, unmapped scaffolds.[11] Recently a few researchers have applied the CQ (Chromosome Quotient) method[12] or YGS (Y chromosome Genome Scan) method[11] to identify Y-linked sequences in these unmapped scaffolds. The CQ method uses the number of alignments from male and female sequence data to determine whether a sequence is Y-linked.[12] The YGS method, on the other hand, is based on a comparison of the assembled genome with female short-reads.[11,13] We tried to apply both methods in *B. mori* to find W-derived contigs. However, both methods can only identify limited number of contigs. In CQ method all repeat sequences must be masked and the female to male ratio is not always precise, especially in short sequences (length $<=2$ kb) with few reads aligned. As YGS method is based on genome sequences, implementing it in species without reference genome is difficult. Consequently, both methods cannot work very effectively in *B. mori*, especially for short sequences. Hence, neither of them provides a reliable strategy for *de novo* assembly of Y or W chromosomes, which restricted the application of the two methods in many other species.

Just like the previously mentioned two methods, many studies concentrated on how to eliminate repetitive sequences from the sequence data. However, the majority of Y or W chromosomes are composed of repetitive sequences and many functional domains may be inserted into TEs, which means very limited unique contigs can be found if all repetitive sequences are removed. Nevertheless, many mutations accumulate on Y or W repetitive sequences and Y/W chromosome have no recombination with X/Z chromosome, which may generate some specific SNPs that can be used as molecular markers to find Y or W linked sequences. So we can utilize these repetitive sequences, instead of masking them, for identifying Y or W linked sequences.[11] Based on this concept, we applied a novel *K*-mer Quotient (KQ) method as a more easy, effective and accurate way for comprehensive analysis of the *B. mori* W chromosome. It can provide a strategy for *de novo* assembling of contigs of Y or W chromosomes, unlike the CQ or YGS method. Moreover, in addition to *B. mori*, we successfully applied this method to two diptera, *D. melanogaster* and *An. gambiae*, as XX/XY sex determination species. As a complete genome sequence is not necessary and removal of repeat sequences is not needed for the successful implementation of our method, it can also be applied to species where genomic information is scarce. Additionally, the KQ method can be extended for the identification of Y or W-linked transcript RNAs and small RNAs, thus providing an innovative approach to study the functions of W or Y chromosomes without complete sequence information.

2. Materials and methods

2.1. Short reads

All sequencing data produced in the present study have been deposited in the NCBI Short Read Archive and can be accessed under the SRA accession number PRJNA393916. *B. mori* female and male HiSeq reads were produced on an Illumina HiSeq 2000 sequencer (101-bp pair-end using pools of larval posterior silk gland which yielded ~73-fold coverage of genome for each sex). *B. mori* female MiSeq reads (only used for assisting *de novo* assembly of W contigs) were produced on an Illumina MiSeq sequencer (250-bp pair-end; pools of female larvae; ~20-fold genome coverage). All reads were from Dazao, the inbred strain used in the genome project.[14] The next generation sequencing data for male and female *D. melanogaster* pooled 5 day old mated adults was downloaded from the NCBI Sequence Read Archive [SRA: SRP007888]; the next generation sequence data for male and female *An. gambiae* pooled wild type strain was downloaded from NCBI Sequence Read Archive [SRA: SRP014756].[12,15] Preferably, to ensure the accuracy of the KQ method, male and female sequence data should be from highly inbred populations and data for both sexes should be of the same amount. Since the coverage of the next-generation sequence data can affect the calculation of the KQ value, from our experience, we suggest that more than 50X coverage is an appropriate choice, which can produce a more precise high-frequency W-15-mer. *B. mori* RNA-Seq data of mixed sex embryos from 0–24 hpo were produced using an Illumina HiSeq 2000 sequencer, yielding 101-bp pair-end and 24 G clean data for each of the different periods (0, 2, 4, 8, 16 and 24 hpo). Small RNA sequences have been deposited in GenBank/EMBL/DDBJ under accession numbers from AB386191–AB424683.

2.2. Criteria for filtrating valid W-specific *k*-mer

The genome sequence of males and females differ only by the W chromosome, which produces W-specific *k*-mers. These k-mers are present only in the female data, but absent from the male data. Before doing a female *vs*. male *k*-mer comparison, removal of sequencing errors in the reads is important in order to achieve high resolution. This was done using two criteria: (i) masking low quality bases and (ii) filtering *k*-mers with rare frequency. Specifically, for the *B. mori* reads, the sequencing errors were filtered by masking bases with phred score <20 and removing 15-mers that were present fewer than 5X in reads. Counting *k*-mers for separate female and male data and removal of sequencing errors can be done using Jellyfish.[16,17] We set up the KQ parameter, for a given *k*-mer *K*i, $KQ(ki) = M(ki)/F(ki)$, where $M(ki)$ is the frequency of *K*i in male reads, $F(ki)$ is the frequency of *K*i in female reads. To obtain reliable W-derived *k*-mers, we used the strict criterion of $KQ = 0$. To reduce the interference of individual polymorphism in sequencing reads, we filtered out these *k*-mers ($KQ = 0$) with low frequency (<15, Supplementary Fig. S1) in female reads.

2.3. W-derived reads and *de novo* assembly strategy

Reads containing W-15-mers could be W-derived and therefore called W-reads. We screened the W-reads from *B. mori* Illumina female reads with Bowtie.[18] Since the reads were paired ends, if a single read sequence was W-derived (contained W-15-mer), its corresponding pair end read could be also W-derived even without a W-15-mer. Based on this assumption, we screened W-derived paired-end reads and assembled them using the MaSuRCA assembler. To prevent incorporation of invalid contigs originating from misassembly, a criterion was followed to filter out invalid contigs. We aligned all the W-15mers to all contigs with 0 mismatch using Bowtie, and calculated the amount of W-15-mers per Kb sequence (AWK) for each contig. From the scatter distribution map (Supplementary

Fig. S2), we found that when the AWK value was less than 10, the scatters deviated from the main trend line, indicating that such contigs with fewer W-15-mers were possibly not W-derived. Thus, we set a threshold of AWK = 10 for filtering invalid contigs. Whereas in the case of very short contigs (less than 1 kb), we suggested that at least 10 W-15-mers were required to be aligned to such contigs.

2.4. AWK threshold in transcripts and small RNAs

For identification of W derived contigs from the already assembled genome sequences, it is also appropriate to set the same threshold with AWK = 10 as followed previously. However, for identification of transcripts and small RNA sequences we must follow different criteria. Considering that *de novo* assembled RNA transcripts were composed of different exons and free from introns, we could increase the threshold to AWK = 50 for W-derived transcripts. Thus, we suggested that it is better to choose a number from 10 to 50, since the higher AWK value gave more reliable results but fewer W-derived transcript RNA candidates and the lower AWK values gave more candidate sequences but less reliable results. As for the small RNAs, we counted the number of aligned W-15-mers for each small RNA and set a threshold with no less than three aligned W-15-mers. Similarly, this threshold is also flexible and can be increased for higher confidence results, or can be decreased for a higher number of W-linked small RNA candidates.

2.5. Application and limitation of the KQ method

Different species have different genome sizes and therefore the *k* value of *k*-mer needs to be changed according to the genome size. The *k*-mer size should be large enough so that identical *k*-mers seldom occur from sequencing errors.[19,20] On the other hand, run time and memory usage would be sharply increased with larger *k* values. Thus, $k = 15$ for *Drosophila*, and $k = 18$ for humans are typical values used for filtering errors in short reads from these genomes.[19,20] Taking the above information and results from the YGS method into consideration, we recommend initial values of $k = 15$ for insect-like genomes (approximately size 100–500 Mb) and $k = 18$ for vertebrate-like genomes (approximately size ~2 Gb).

2.6. Molecular biology methods

Bombyx mori (Dazao strain) larvae were reared on fresh mulberry leaves in a dedicated silkworm rearing room at ambient temperature (23–27 °C). Genomic DNA samples were separately isolated from virgin female and male individuals with tissue DNA kits (OMEGA, Norcross, GA). In the case of W contigs that had autosomal or Z paralogs, primers for genomic DNA PCR were designed considering the differences between the autosomal or Z paralogs and the W sequence at the 3' end of the primer. PCR was performed with GO Taq DNA Polymerase (Promega, Madison, WI; Supplementary Table S1–S3), RNA was isolated from individual embryos following the traditional Trizol method (Ambion, Carlsbad, CA) using the total RNA isolation protocol according to the manufacturer's instructions. Residual DNA in samples was used to identify the gender of each embryo with the W_seq1 marker (Supplementary Table S1). Total RNA (following DNase treatment) was subjected to reverse transcription using a PrimeScript™ RT Master Mix (Perfect Real Time; TaKaRa, Kusatsu, Japan) in 50 μl reaction volume (2500 ng total RNA) and then diluted 5-fold. One μg of cDNA was used in 10 μl PCR reaction volume. Small RNAs (following DNase treatment) were subjected to reverse transcription using a Mir-X™ miRNA First-Strand Synthesis Kit (TaKaRa, Kusatsu, Japan). qPCR was performed with SYBR Premix Ex (TaKaRa, Kusatsu, Japan).

3. Results and discussion

3.1. The KQ method

Although most of the *B. mori* W chromosome is composed of TEs whose copies are widely distributed on autosomes or on the Z chromosome, still some subtle variations such as single nucleotide polymorphisms (SNPs) can be observed only on W chromosome. Such W-specific variations may produce W-specific short sequences called *k*-mers.

For obtaining these possible W-specific *k*-mers, we tried to find an optimum *k* value (sequence length). $K = 15$, which is a typical value used for filtering errors in short reads,[11,19] is suitable for female-specific (W-specific) *k*-mers in *B. mori*. We counted the frequencies of each 15-mer separately in female and male sequence data and created a parameter called KQ. For a given *k*-mer *K*i, $KQ(Ki) = M(Ki)/F(Ki)$, where $M(Ki)$ is the frequency of *K*i in male reads and $F(Ki)$ denotes the frequency of *K*i in female reads. As females and males share the same complement of autosomes, autosomal 15-mers are present in both female and male sequencing data in roughly the same quantities. Therefore, autosomal 15-mers have KQ values distributed around one (Fig. 1). As males have two Z chromosomes while females have only one, the Z-specific 15-mers have KQ values distributed around two. Since unique W chromosomal 15-mers are present only in female sequencing data, these KQ values localize to zero. We also applied KQ to *D. melanogaster* and *An. gambiae* [Fig. 1, exchange $M(Ki)$ and $F(Ki)$ position in XX/XY species]. The results were similar to *B. mori*, which confirms that KQ may be applied to most sexually heterogametic species.

In our assumption, these W or Y-specific 15-mers (KQ = 0) could be used as labels to identify sequences from W or Y chromosome. To verify such an assumption, we first applied it to *D. melanogaster* Y, which has been relatively well sequenced. We aligned all Y-specific 15-mers (Y-15-mers) of *D. melanogaster* to its genome scaffolds (Genebank assembly accession: GCA_000001215.4), and a parameter named AYK (Amount of Y-15-mers per Kb sequence) was created for each scaffold. Based on the analysis of each scaffold's AYK value and length (Fig. 2A), we could show clearly that the majority of scaffolds are Y-derived (red points) when AYK >=10, while scaffolds of X or autosome (blue and green points) are with the AYK < 10 and some unmapped scaffolds (grey points) with AYK value >=10 are also possible Y-derived. The result confirmed that our KQ method could effectively identify Y-linked sequences. Except KQ method, we also applied previous CQ and YGS method in *D. melanogaster*, and compared the results of three methods to evaluate which method could be most effective (We only selected contigs with AYK >=50 in KQ method to unify the standard of three methods). For 200 already known Y-scaffolds, CQ method could identify only 63 of 200 as Y-linked, whereas YGS identified 106 and our KQ provided 154 Y-scaffolds. Figure 2B showed that for more than 5 kb scaffolds, three methods could identify most Y-linked scaffolds, whereas for less than 5 kb, KQ method worked obviously better than the other two methods. Thus, compared with other two methods, KQ method performed more effectively for short scaffolds (less than 5 kb). As Y and W chromosome sequences are usually fragmented into a large number of short and mostly unmapped scaffolds due to highly repetitive sequences, KQ method would have wider application in

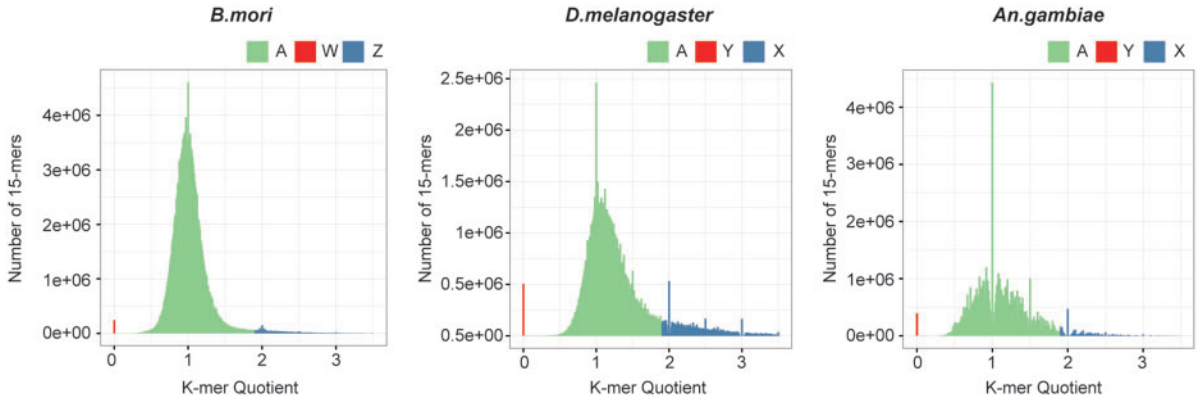

Figure 1. Distribution of calculated KQs. We counted the KQ value for each 15-mer and constructed a distribution histogram for *B. mori, D. melanogaster* and *An. gambiae*. It is apparent that most autosome-derived, Z specific or W unique 15-mers have distinctive KQs. Autosome derived 15-mers have a KQ value around 1 (mainly from 0.3 to 1.8), Z or X derived 15-mers have a KQ value around 2 (mainly from 1.8 to 2.2) and W or Y derived 15-mers have a KQ value around 0. Thus, we can easily identify W specific 15-mers by the KQ value.

heterogametic organisms, particularly for the species in which genomic information is still scarce.

3.2. *De novo* assembled W specific genome contigs and identification of W-BACs

We obtained 93,942 high-frequency 15-mers (frequency >=15 in female reads), where the KQ values were zero in *B. mori*. These 15-mers (W-15-mers) could not be aligned onto a male genomic assembly, which strengthened the likelihood of W-specificity. As reads containing W-15-mers could be W-derived, we aligned all the W-15-mers to female genome sequencing reads and screened the reads containing W-15-mers. We obtained about 1.1 million HiSeq paired-end reads (101 bp, ~216 Mbp) and 0.6 million MiSeq paired-end reads (250 bp, ~300 Mbp) as W-derived reads, which were assembled into 6718 W-specific contigs using the MaSuRCA assembler.[21] We set a threshold AWK (Amount of W-15-mers per Kb sequence) value = 10 and considered contigs with an AWK value greater than 10 as valid W-specific contigs (Supplementary Fig. S2). Eligible contigs of 1281 were selected with a total length of around 1.9 Mb, which covered almost 1/10 of the W chromosome. To confirm these contigs were indeed W-derived, we randomly chose 10 of them (named W_Seq1 to W_Seq10) for a separate PCR test with female and male genomic DNA. The primers were designed to have W-15-mers at the 3' end and used a relatively high annealing temperature to reduce the likelihood of primer mismatch.[22–24] PCR bands appeared only in females for all 10 primer pairs, (Fig. 3A), indicating that all of them were W-specific.

To obtain the fine structure of the W chromosome, we used this method to find W-derived bacterial artificial chromosomes (BACs) in silkworm BAC libraries. The silkworm BAC reference library (derived from both sexes) contains 74,717 BACs, for which about 200–500 bp ends have already been sequenced with the Sanger ABI3730 sequencer.[25] Using the threshold AWK value = 10,122 BAC end sequences were obtained to be putatively W derived, and we chose 10 of them at random to confirm their W specificity by PCR test against female *vs.* male genomic DNA (Supplementary Table S2). The presence of PCR bands only in female genomic DNA confirmed (Fig. 3B) that all 10 BAC end sequences were W-specific, i.e. the corresponding BACs were also W-derived.

The results indicated the KQ method could effectively assemble and identify contigs originating from the W chromosome. These contigs will be invaluable in the search for W-specific PCR markers. The length of *B. mori* W chromosome may have the same size with Z chromosome (~20 M) depending on the relative cytogenetic length[26] and if we obtain a sufficient number of W-BACs, it would be possible to construct BAC contigs covering the whole W chromosome.

3.3. Identification of W-linked transcript RNAs

Use of RNA-Seq has improved the identification of a large number of cellularly expressed RNA species. However, little information is reported on how many of these are actually transcribed on the W chromosome. Here, we applied the KQ method to RNA-Seq data derived from silkworm embryos to find W-linked transcripts. We performed deep RNA-Seq for 0-24 h post-oviposition (hpo) mixed embryo samples (each sample contained 5–6 unsexed embryos), which we assembled *de novo* using the Trinity assembler,[27] and obtained 175,492 transcript RNAs including different alternative splicing isoforms. We then screened for valid W-derived RNAs with a strict threshold of AWK value = 50 and thus selected a total of 156 candidate W-linked RNAs. Using an expression heatmap to analyse the data (Fig. 4A), we observed two main expression clusters: in Cluster A RNAs initiated from 16 hpo; in Cluster B RNAs were continuously expressed from 0–24 hpo.

To confirm that these RNAs were indeed W-derived, five candidates were selected for examining their expression in individual embryos. We designed primers using their W-*k*-mers (Supplementary Table S3) in the same way as for W-BAC ends. We isolated both DNA and RNA from single embryos aged 16 hpo, and used DNA to detect the presence of the W chromosome by PCR (Supplementary Fig. S3), while we followed RNA by RT-PCR to examine expression (Fig. 4B). The RT-PCR results showed cluster A RNAs were expressed only in female embryos, whereas cluster B RNAs were expressed in both female and male embryos. For further verification, we performed a separate PCR test of both cluster A and B RNAs with female and male genomic DNA. For these five RNAs, PCR bands appeared only in female genomic DNA (Fig. 4B), indicating both cluster RNAs were transcribed on the W chromosome. These results suggested that cluster B RNAs are very likely to be maternal RNAs deposited in the embryos as they are present throughout early embryos of both females and males. An interesting phenomenon in RT-PCR of cluster B RNAs is that bands were brighter in female than in male, which indicated the expressions of cluster B RNAs are different in female and male embryos. Thus, we did the qPCR of a cluster B RNAs (DN42314_c0) in female and male embryos from 0

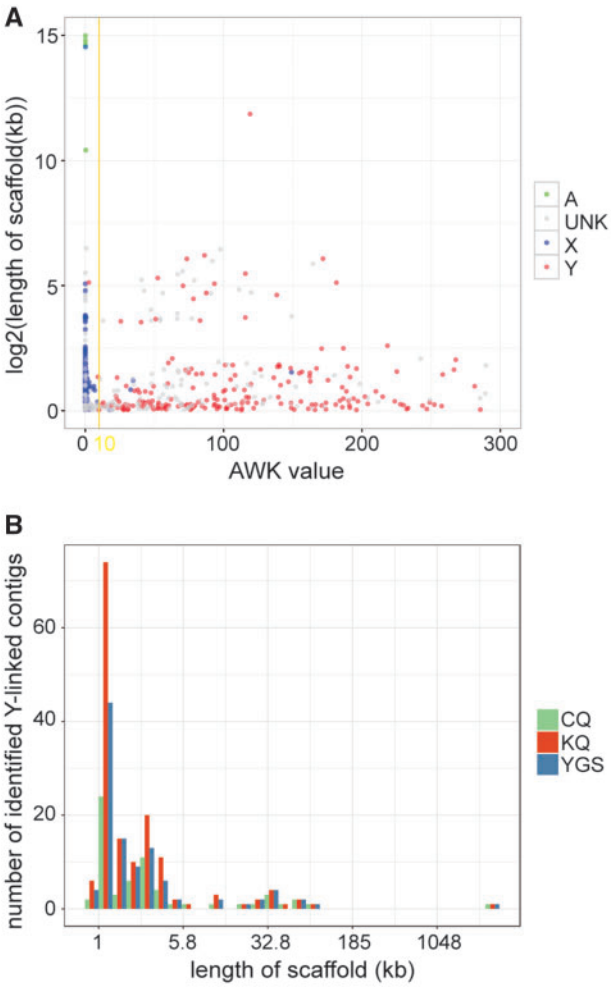

Figure 2. Identification of Y-linked sequences in *D. melanogaster*. (A) We applied KQ method in genome sequences of *D. melanogaster* and counted length and AYK value for each scaffold. For the vast majority of Autosome (green point) and X-derived (blue point) scaffolds, their AYK values are less than 10 (usually close to 0); Y-derived scaffolds (red point) were with AYK value more than 10. Unmapped scaffolds (grey point) have dispersive AYK values, we speculated that scaffolds with AYK value more than 10 are possible Y-linked. (B) Compared with the CQ and YGS methods, KQ method could identify more precise Y-derived scaffolds in *D. melanogaster*. Based on the analysis of scaffold length and number, the results of three methods were similar when the length of scaffold is more than 5 kb, whereas KQ method (red pillar) could identify more precisely Y-linked scaffolds when the length of scaffold is less than 5 kb.

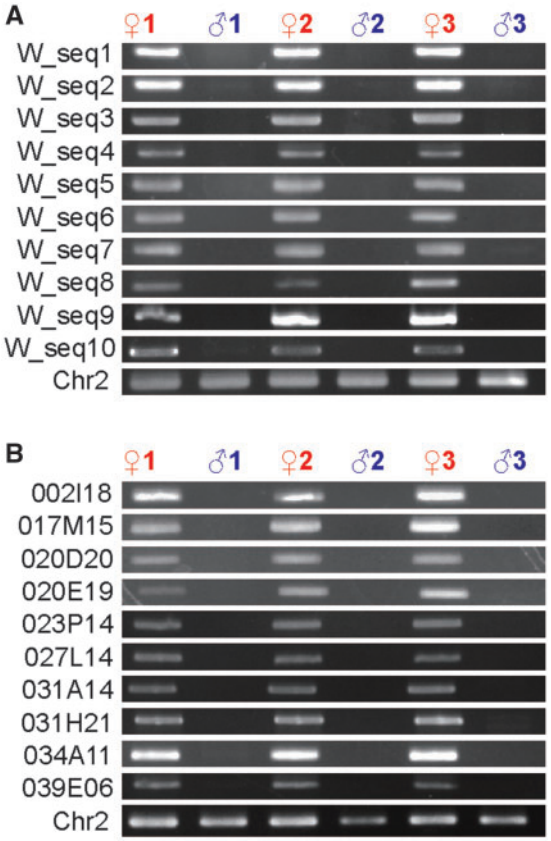

Figure 3. Female-specific PCR amplification of *de novo* assembled contigs (A) and BAC end sequences (B). (A) PCR performed with three female and three male genomic DNA samples shows female-specific amplification of 10 W-contigs. Chr 2 sequence was amplified as a control in both female and male genomic DNA confirming the integrity of the genomic DNA samples. (B) PCR carried out with three female and three male genomic DNA samples shows female-specific amplification of 10 BAC end sequences.

to 36 hpo, and found sex-biased expression starting from 8 hpo (Fig. 4C). The expression levels were almost same in both sex embryos before 8 hpo, but after that, it seemed to fade away in male embryos and *de novo* expressed in female embryos.

3.4. Annotation of W-linked transcript RNAs

We found 156 W-linked RNAs, among which we predicted 134 were non-coding RNAs by their Coding-Non-Coding Index (CNCI).[28,29] Non-coding RNAs, most of which are long ncRNAs (lncRNAs), are the main repertoire of the transcripts expressed from the W chromosome. We found 93 ovarian piRNAs among the lncRNAs through NCBI BLAST, which implied many lncRNAs seemed to be precursors of piRNAs.

With GO annotation and an NCBI BLAST search of 22 candidate coding RNA sequences, we found that many of them had high similarity with some *B. mori* uncharacterized genes located on an autosome or Z chromosome (Supplementary Table S4), suggesting these genes may have been translocated onto the W chromosome through transposons during evolution.[30] The GO analysis showed that these coding transcripts were diverse, possibly involved in DNA integration and metabolic processes, nucleic acid binding, nucleotidyltransferase activity, etc. (Supplementary Fig. S4).

3.5. Identification of W-linked small RNAs

The W chromosome being a source of female-enriched piRNAs[31] indicates that W chromosome function mainly depends on them. For example, Fem piRNA was reported to be the primary sex-determiner in silkworm.[32] As small RNAs are usually around 18–30 nt, it is difficult to determine whether or not they are W-derived. Using shorter length W-15-mers, the KQ method can also be used to identify W-derived small RNAs from the pool of small RNA sequences in *B. mori*.

Kawaoka et al. (2008) identified 38,493 small RNA species from ovarian RNA (pupa day 4) in the silkworm, while we identified 72,439 small RNA species from embryo RNA of day 8 pupa. We merged the two small RNA datasets, which amounted to 106,717 small RNA species and then aligned W-15-mers to all these small

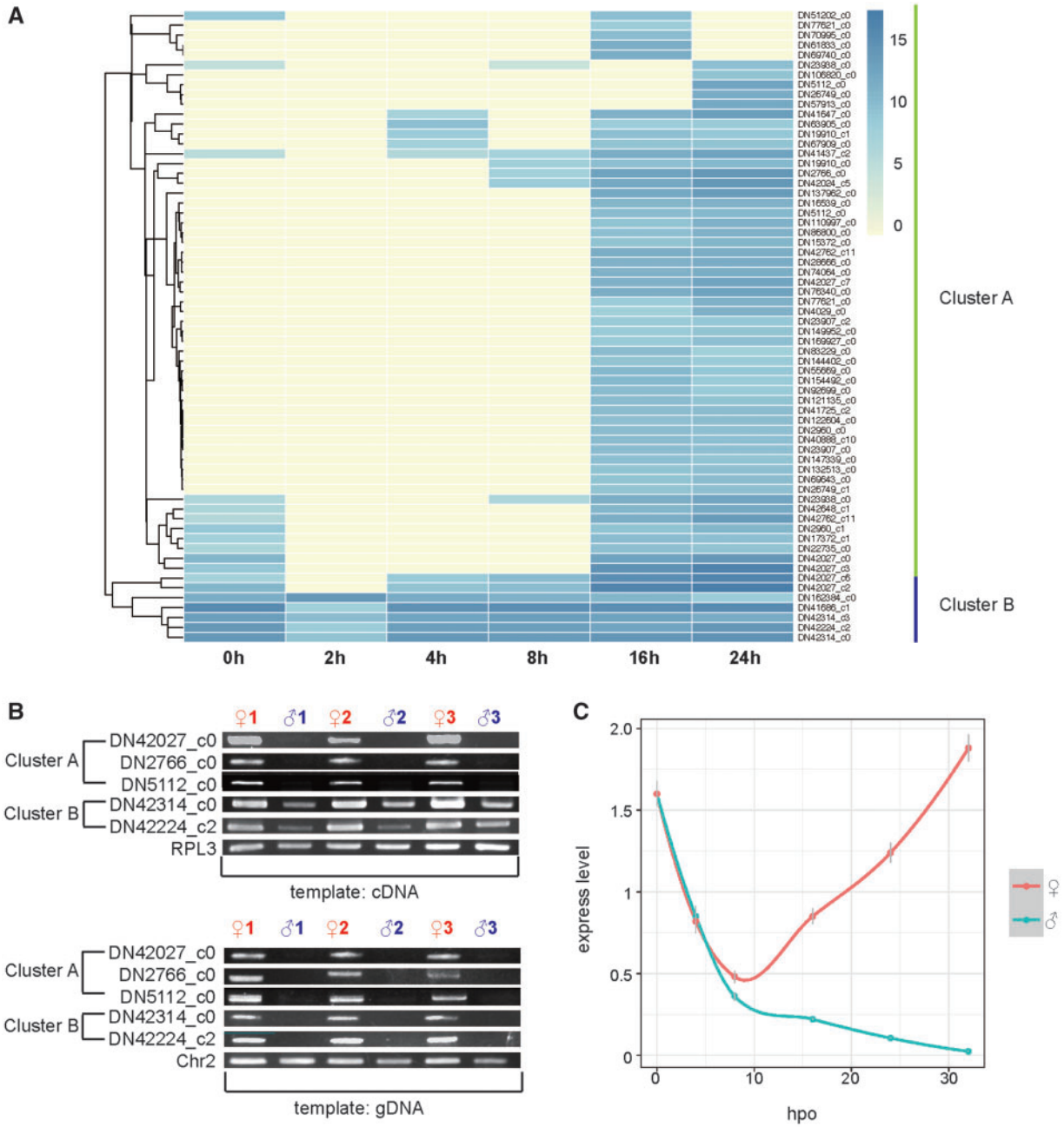

Figure 4. W-derived transcripts, expression profiles and genomic PCR amplification. (A) Hierarchical cluster analysis of W-derived transcripts. Expression profiles of transcripts in mixed embryos (0–24 hpo) were clustered and visualized in a heatmap. (B) PCR amplification of W-derived transcripts in cDNA (Left panel) and gDNA (Right panel). RT-PCR (cDNA) showed expression as cluster A (female embryos specific) or cluster B (in both female and male embryos). However, a PCR test of genomic DNA showed both transcript RNAs were female-specific, indicating that they are W-derived, but cluster A RNAs are zygotically expressed in female embryos whereas cluster B RNAs are probably maternal RNAs deposited into eggs. (C) qPCR of a cluster B RNA (DN42314_c0) from 0 to 36 hpo exhibited its sex-biased expression in embyos.

RNA sequences to screen small RNAs with no fewer than three aligned W-15-mers, resulting in 345 candidates. Their sizes were mostly around 27–29 nt (Fig. 5A) with a strong bias for U(T) at the 5' end, indicating that most of the W-linked small RNAs were piRNAs. We randomly chose 5 small RNAs out of the 345 candidates to check their expression levels using RT-PCR in separate female and male pupal cDNA (day 4, Supplementary Table S5). The bands appeared only in females for all the five small RNAs (Fig. 5B), which demonstrated that they were indeed W-derived.

4. Discussion

In this study, we first explored the existence of W or Y-specific 15-mers in *B. mori*, *D. melanogaster* and *An. gambiae*, then developed the KQ method and obtained a successful result in *D. melanogaster*. After that, we applied KQ method in *B. mori* to assemble W-linked genome contigs and to identify W-linked transcript RNAs and small RNAs effectively. In contrast to the previously reported CQ and YGS methods, our KQ method is more effective and accurate

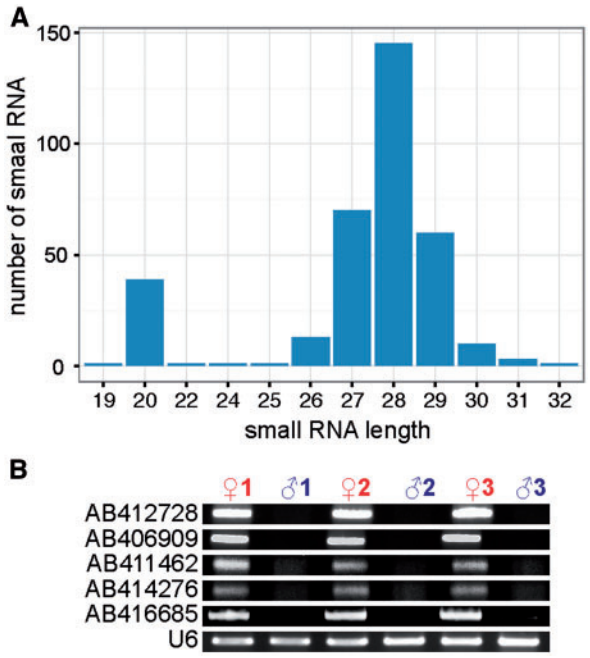

Figure 5. Length distribution graph of W-linked small RNAs and female-specific RT-PCR amplification. (A) The lengths of most W-derived small RNAs were concentrated on around 27–29 nt and their first bases from the 5′ end were preferentially T (U), which indicated most W-derived small RNAs are piRNAs. (B) Female-specific RT-PCR amplifications indicated that all five small RNAs were indeed W-derived.

especially in short sequences. In addition, the KQ method opens up a novel approach for *de novo* assembly of W contigs, and is now the only method available to identify W or Y-specific small RNAs. Also KQ method can be used to identify sequences linked to X or Z chromosome if we used X or Z-specific 15-mers (KQ value around 2). As a genome reference sequence is not necessary, the KQ method can be applied to most species with heterogametic sex chromosomes where genome information is scarce, and it could become an useful tool for obtaining whole genome information of heterogametic sex chromosomes (Y or W).

The complete biological function of the *B. mori* W chromosome is still unclear. Although many scientists have tried to uncover the biological process of sex-determination in this species,[2,26,32] the lack of W chromosome specific sequence information hinders the annotation and molecular studies of the W. Here, we showed that the KQ method could help in drastically changing this situation. We obtained W-linked transcripts and small RNAs without the use of complete genome sequences. These RNA sequences will be helpful in the functional study of the W chromosome. Based on preliminary transcript RNA analysis, we found many lncRNAs and a few possible coding RNAs expressed from 16 hpo when the sex-determination cascade initiates,[33] and which may be involved in the sex determination process. Many lncRNAs seem to be the precursors of piRNAs, and our small RNA analysis showed that most W-linked small RNAs are likely piRNAs, which indicates that piRNA based gene regulation may be the main function of the *B. mori* W chromosome.

piRNA associates with PIWI family proteins whose complexes cleave target sequences to silence genes or transposons (Aravin et al., 2006; Saito et al., 2006; Brennecke et al., 2007). For example, Kiuchi et al. (2014) found that a specific W-linked piRNA (Fem piRNA) interacts with the product of a Z-linked *Masculinizer* (*Masc*) RNA, silencing of *Masc* by Fem piRNA is required for the production of female-specific isoforms of *Bmdsx* in female embryos and that a *Masc*-associated process may control both dosage compensation and masculinization in male embryos.[32,34,35] This suggests that W-linked non-coding RNAs, especially piRNAs, play important roles in sex-determination and dosage compensation, and may be the main functional form encoded on the W. Also many W-derived piRNAs are expressed at the pupal period, which may function to maintain germline stem cells,[36,37] suggesting that the effects of W-linked piRNAs are continuous throughout female life.

Acknowledgements

The authors thank Youbing Guo, Duolian Liu and Li Peng for assistance with the information analysis. They also appreciate Zha XF for providing the embryo RNA-Seq data.

Conflict of interest

The authors declare no competing financial interests.

Funding

This work was supported by the grant from the One Thousand Foreign Experts Recruitment Program of the Chinese Government (No. WQ 20125500074).

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Matsuda, M., Nagahama, Y. and Shinomiya, A. 2002, DMY is a Y-specific DM-domain gene required for male development in the medaka fish, *Nature*, **417**, 559–63.
2. Ayers, K.L., Davidson, N.M., Demiyah, D., et al. 2013, RNA sequencing reveals sexually dimorphic gene expression before gonadal differentiation in chicken and allows comprehensive annotation of the W-chromosome, *Genome Biol.*, **14**, R26.
3. Chen, N., Bellott, D.W., Page, D.C. and Clark, A.G. 2012, Identification of avian W-linked contigs by short-read sequencing, *BMC Genomics*, **13**, 183.
4. Schartl, M., Schmid, M. and Nanda, I. 2016, Dynamics of vertebrate sex chromosome evolution: from equal size to giants and dwarfs, *Chromosoma*, **125**, 553–71.
5. Consortium, I.C.G.S. 2004, Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution, *Nature*, **432**, 695–716.
6. Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., et al. 2003, The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes, *Nature*, **423**, 825–37.
7. Foote, S., Vollrath, D., Hilton, A. and Page, D. 1992, The human Y chromosome: overlapping DNA clones spanning the euchromatic region, *Science*, **258**, 60–6.
8. Hall, A.B., Papathanos, P.A., Sharma, A., et al. 2016, Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes, *Proc. Natl. Acad. Sci. USA*, **113**, E2114–23.
9. Krsticevic, F.J., Schrago, C.G. and Carvalho, A.B. 2015, Long-read single molecule sequencing to resolve tandem gene copies: the Mst77Y region on the drosophila melanogaster Y chromosome, *G3: Genes|Genomes|Genetics*, **5**, 1145–50.
10. Traut, W., Vogel, H., Glockner, G., Hartmann, E. and Heckel, D.G. 2013, High-throughput sequencing of a single chromosome: a moth W

chromosome, *Chromosome Res.: Int. J. Mol. Supramol. Evol. Aspects Chromosome Biol.*, **21**, 491–505.
11. Carvalho, A.B. and Clark, A.G. 2013, Efficient identification of Y chromosome sequences in the human and Drosophila genomes, *Genome Res.*, **23**, 1894–907.
12. Hall, A.B., Qi, Y., Timoshevskiy, V., Sharakhova, M.V., Sharakhov, I.V. and Tu, Z. 2013, Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females, *BMC Genomics*, **14**, 273.
13. Carvalho, A.B., Dobo, B.A., Vibranovski, M.D. and Clark, A.G. 2001, Identification of five new genes on the Y chromosome of Drosophila melanogaster, *Proc. Natl. Acad. Sci. USA*, **98**, 13225–30.
14. International Silkworm Genome, C. 2008, The genome of a lepidopteran model insect, the silkworm Bombyx mori, *Insect Biochem. Mol. Biol.*, **38**, 1036–45.
15. Malone, J.H., Cho, D.Y., Mattiuzzo, N.R., et al. 2012, Mediation of Drosophila autosomal dosage effects and compensation by network interactions, *Genome Biol.*, **13**, r28.
16. Tomaszkiewicz, M., Medvedev, P. and Makova, K.D. 2017, Y and W chromosome assemblies: approaches and discoveries, *Trends Genet.*, **33**, 266–82.
17. Marcais, G. and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics*, **27**, 764–70.
18. Langmead, B. 2010, Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinformatics*, **32**, 11.17.11–11.17.14.
19. Kelley, D.R., Schatz, M.C. and Salzberg, S.L. 2010, Quake: quality-aware detection and correction of sequencing errors, *Genome Biol.*, **11**, R116.
20. Li, R., Zhu, H., Ruan, J., et al. 2010, *De novo* assembly of human genomes with massively parallel short read sequencing, *Genome Res.*, **20**, 265–72.
21. Zimin, A.V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S.L. and Yorke, J.A. 2013, The MaSuRCA genome assembler, *Bioinformatics*, **29**, 2669–77.
22. Fan, X.Y., Hu, Z.Y., Xu, F.H., Yan, Z.Q., Guo, S.Q. and Li, Z.M. 2003, Rapid detection of rpoB gene mutations in rifampin-resistant mycobacterium tuberculosis isolates in shanghai by using the amplification refractory mutation system, *J. Clin. Microbiol.*, **41**, 993–7.
23. Promboon, A., Shimada, T., Fujiwara, H. and Kobayashi, M. 1995, Linkage map of random amplified polymorphic DNAs (RAPDs) in the silkworm, *Genet. Res.*, **66**, 1.
24. Abe, H., Seki, M. and Ohbayashi, F. 2005, Partial deletions of the W chromosome due to reciprocal translocation in the silkworm Bombyx mori, *Insect Mol. Biol.*, **14**, 339–52.
25. Suetsugu, Y., Minami, H., Shimomura, M., et al. 2007, End-sequencing and characterization of silkworm (Bombyx mori) bacterial artificial chromosome libraries, *BMC Genomics*, **8**, 314.
26. Traut, W., Sahara, K. and Marec, F. 2007, Sex chromosomes and sex determination in Lepidoptera, *Sexual Dev.: Genetics, Mol. Biol. Evol. Endocrinol. Embryol. Pathol. Sex Determination Differentiation*, **1**, 332–46.
27. Grabherr, M.G., Haas, B.J., Yassour, M., et al. 2011, Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data, *Nat. Biotechnol.*, **29**, 644–52.
28. Wu, Y., Cheng, T., Liu, C., et al. 2016, Systematic identification and characterization of long non-coding RNAs in the silkworm, Bombyx mori, *PloS One*, **11**, e0147147.
29. Sun, L., Luo, H., Bu, D., et al. 2013, Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts, *Nucleic Acids Res.*, **41**, e166.
30. Wang, J., Long, M. and Vibranovski, M.D. 2012, Retrogenes moved out of the z chromosome in the silkworm, *J. Mol. Evol.*, **74**, 113–26.
31. Kawaoka, S., Kadota, K., Arai, Y., et al. 2011, The silkworm W chromosome is a source of female-enriched piRNAs, *RNA*, **17**, 2144–51.
32. Kiuchi, T., Koga, H., Kawamoto, M., et al. 2014, A single female-specific piRNA is the primary determiner of sex in the silkworm, *Nature*, **509**, 633–6.
33. Sakai, H., Aoki, F. and Suzuki, M. G. 2014, Identification of the key stages for sex determination in the silkworm, Bombyx mori, *Dev. Genes Evol.*, **224**, 119–23.
34. Brennecke, J., Aravin, A.A., Stark, A., et al. 2007, Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila, *Cell*, **128**, 1089–103.
35. Saito, K., Nishida, K.M., Mori, T., et al. 2006, Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome, *Genes Dev.*, **20**, 2214–22.
36. Lin, H. and Spradling, A.C. 1997, A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary, *Development (Cambridge, England)*, **124**, 2463–76.
37. Pal-Bhadra, M., Leibovitch, B.A., Gandhi, S.G., et al. 2004, Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery, *Science*, **303**, 669–72.