Peer-reviewed version available at *Genes* **2018**, *9*, 277; doi:10.3390/genes9060277

The size and content of the sex-determining region of the Y chromosome in dioecious *Mercurialis annua*, a plant with homomorphic sex chromosomes

Paris Veltsos[1,4,*], Guillaume Cossard[1], Emmanuel Beaudoing[2], Gensériс Beydon[3], Camille Roux[1,5], Santiago C. González-Martínez[1,6] and John R. Pannell[1]

[1] Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

[2] Faculté de biologie et de médecine, University of Lausanne, Bâtiment Génopode, 1014, Lausanne, Switzerland

[3] Centre National de Ressources Génomiques végétales (CNRGV), 24 Chemin de Borde Rouge - Auzeville - CS52627, 31326 Castanet Tolosan Cedex, France

[4] current address: Department of Biology, Jordan Hall, 1001 East Third Street, Indiana University, Bloomington, IN 47405, USA

[5] current address: CNRS, Univ. Lille, UMR 8198 - Evo-Eco-Paleo, F-59000 Lille, France

[6] current address: BIOGECO, INRA, University of Bordeaux, 33610 Cestas, France

* Author for correspondence: parisveltsos@gmail.com

Running head: Sex chromosomes of *Mercurialis annua*

Abstract

Many dioecious plants have sex chromosomes that are cytologically heteromorphic, but about half of species lack cytological differences between males and females and are thus homomorphic. Very little is known about the size and content of the non-recombining sex-determining region (SDR) in these species. Here, we assess the size and content of the SDR of the diploid dioecious herb *Mercurialis annua*, which has homomorphic sex chromosomes and shows signatures of mild Y-chromosome degeneration. We used RNAseq to identify new Y-linked markers for *M. annua.* Twelve of 24 transcripts with male-specific and male-biased expression could only be PCR-amplified from males and are thus Y-linked. We found a further six Y-linked sequences that were present in males but not females using genome capture data from multiple populations. We used the Y-linked sequences to identify and sequence 17 sex-linked bacterial artificial chromosomes (BACs), which form 11 groups of non-overlapping sequence, covering a total sequence length of about 1.5 Mb. Content analysis of this region suggests it is enriched for repeats, has a low gene density and contains few candidate sex-determining genes. The BACs map to a subset of the sex-linked region of the genetic map, which is estimated to be at least 14.5 Mb. This is substantially larger than estimates for other dioecious plants with homomorphic sex chromosomes, especially given the small genome size of *M. annua*. Our data provide a rare, high-resolution view of the homomorphic Y chromosome of a dioecious plant.

Keywords: bacterial artificial chromosomes, RNAseq, genetic map, transposable element, gene density

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

1. Introduction

Most flowering plants are hermaphroditic, but separate sexes (dioecy) have evolved from hermaphroditism repeatedly. Dioecy is found in only about 7% of flowering plant species but in about half of all families [1, 2]. Although sex in many animal species is determined by environmental triggers [3], sex in almost all dioecious plants studied so far appears to be determined genetically (though see [4, 5]), usually on the basis of segregation of alleles or haplotypes at a single genetic locus within the non-recombining 'sex-determining region' (SDR) on sex chromosomes [6, 7]. Many plant sex chromosomes are cytologically heteromorphic (currently 19 species in four families), but cytological differences between males and females are not evident in others (20 species in 13 families; reviewed in [7, 8]). Indeed, closely related dioecious species may often differ in terms of their degree of heteromorphism [6, 7].

The magnitude of cytological variation between the homologous sex chromosomes of a species might be expected to follow a molecular clock that can be used to estimate their age: strongly heteromorphic sex chromosomes should be older than homomorphic ones as the result of the progressive genetic degeneration of the Y chromosome (in a X/Y system). However, although there is some evidence for this expectation [9], there are many exceptions. For instance, in *Coccinia grandis* (Cucurbitaceae), a species that evolved dioecy about three million years ago [10], the X and Y chromosomes are highly divergent, with an elongation of the Y of 10% compared to the X. In contrast, dioecy evolved in the palm genus *Phoenix* perhaps > 50 million years ago, but the SDR is small and its sex chromosomes are homomorphic [11].

The canonical model for the evolution of heteromorphic sex chromosomes invokes the degeneration of non-recombining regions, either through the loss of genes and the shortening of sequence length, or through the accumulation of repetitive sequences and chromosome lengthening, which appears to be particularly common in plants with heteromorphic sex chromosomes [7]. Both processes can be attributed to the reduced efficacy of purifying selection in the SDR as a result of suppressed recombination [12]. It is thought that the SDR may expand over time as a result of selection for reduced

recombination between the sex-determining locus and putative sexually antagonistic loci (i.e. loci at which segregating alleles differentially benefit either male or female fitness [13]). However, SDR expansion on the Y or W chromosomes may also simply result from the accumulation of repetitive elements themselves [14], without the accumulation of sexually antagonistic genes. Either way, variation among dioecious plants in the relative sizes of their SDR remains puzzling and largely unexplained. Whereas the importance of sexually antagonistic selection in bringing about the suppression of recombination is theoretically plausible [13], there is still little empirical evidence for it. We also do not understand why X and Y (or Z and W) chromosomes of some species diverge rapidly, becoming heteromorphic quickly, while others remain homomorphic for equivalent periods of time.

Flowering plant species vary greatly in their tendency to evolve heteromorphic sex chromosomes. While sex chromosomes in plants have long been described [15], neither species with heteromorphic, nor those with homomorphic sex-chromosomes have been thoroughly studied at the genomic level. Indeed, so far only dioecious plants with an economic interest, such as kiwis (*Actinidia chinensis*), grapes (*Vitis vinifera*), and papaya (*Carica papaya*), have been the object of full genome sequencing. Accordingly, we still know little about variation in the size and content of the SDR, especially for species with putatively large SDR such as *Silene latifolia* [16], but also for species with homomorphic sex chromosomes. The SDR of such species may or may not be small and could include species in which it is restricted to a single gene.

Characterization of the physical size of the SDR is generally challenging because of its often highly repetitive content. It is possible to compare the DNA content of males and females, and attribute the difference to the sex chromosomes. As expected, such comparisons have revealed larger differences in species with heteromorphic sex chromosomes (see Table 1, and [7]). However, to determine the actual size of the non-recombining SDR of the chromosome of the heterogametic sex (Y or W), it is necessary to identify markers on the sex chromosome, and to use them to build genetic maps that estimate the region that does not recombine in one sex [e.g., 17, 18] (Table 1). Alternatively, sex-linked markers may be used to identify and then sequence bacterial

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

artificial chromosomes (BACs) that contain long sections of the SDR [19], with subsequent assembly and potential 'chromosome-walking' [20, 21].

Here, we assess the size and content of the SDR of the Y chromosome of the diploid dioecious plant *Mercurialis annua* based on assemblies of partially overlapping BACs identified from male-specific PCR products. *M. annua* is a polyploid complex that shows striking variation in its sexual system, ranging from diploid populations that have fully separate sexes and a homomorphic XY sex-determination system, to androdioecy (where males co-occur with hermaphrodites) and monoecy [22-25]. The unusual variation in the sexual system of the *M. annua* complex lends itself to testing a number of hypotheses about the selection of combined versus separate sexes [26-28] and the evolution of sex determination and sex chromosomes [23, 29]. Previous work, based on *de novo* sequencing, genetic mapping of SNPs from open reading frames (ORFs) segregating in crossing families, and genome capture from males and females sampled across the species range (representing about 7% of *M. annua* genome; [30]), has allowed the assembly of the diploid *M. annua* genome into eight linkage groups (corresponding to the 2n = 16 chromosomes of the diploid karyotype) and the identification of 568 sex-linked transcripts on the largest linkage group (i.e., chromosome 1), which represent about 33% of the genes on the chromosome [29]. The *M. annua* SDR includes at least one element for a likely sexually antagonistic inflorescence trait (unpublished data), but shows signs of only mild Y degeneration, including the accumulation of transposable elements and other repetitive DNA, and a single gene with a premature stop codon on the Y allele [29].

In the present study, we aimed: (1) to estimate the physical length of the diploid *M. annua* SDR in relation to the size of the rest of the Y chromosome and to the genome; and (2) to further characterise its content and genomic structure, including the identification of additional sex-linked genes that were either not previously mapped due to the absence of suitable variation, or were not among the ORFs used in previous mapping [29]. To estimate the size of the SDR, we identified male-specific PCR products on ORFs, and used them to identify and sequence Y-linked BACs from two males. We inferred the minimum size of the SDR in terms of the size of the sex-linked region in the genetic map of *M. annua* [29] associated with the BACs. Our analysis suggests that the

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

non-recombining SDR of the Y chromosome of *M. annua* is among the largest known for a homomorphic plant sex chromosome.

2. Materials and Methods

2.1 Overview of our approach

We used male-specific gene expression [31] as well as male specific genome sequences, obtained from a large sample of males and females from across the species' range [30], to identify potentially Y-linked markers for *M. annua*. We verified the Y-linked status of these markers by male-specific PCR amplification of samples covering the diploid species range, and verified the amplified sequences are similar to the expected transcripts by Sanger sequencing. We then used the PCR reactions to probe a BAC library constructed from two diploid *M. annua* males, collected near Lausanne in Switzerland. Finally, we used long-read (PacBio) sequencing to assemble contigs of the identified Y-linked BAC sequences, which we then subjected to content and genomic structure analysis.

2.2 Identification and confirmation of Y-linked markers

We used two different approaches to identify Y-linked markers. First, we selected genes with male-specific expression in above- and below-ground vegetative tissues of several males and females grown in the greenhouse from mixed seed from north-eastern Spain. Specifically, our expression analyses were based on two separate experiments, one with five males and five females sampled individually, and the other with three pooled samples (10 individuals of each sex per pool). Male-specific expression was defined as expression in at least two males, with a significant difference in expression between sexes at the 5% threshold, corrected for Benjamini Hochberg false discovery rate, as calculated in DEseq2 [32]. No minimum logFC threshold between the two sexes was employed [31].

From both these gene-expression experiments, we identified 24 genes that we were able to amplify by PCR. Of these 24 genes with male-specific expression, 12 showed male-specific amplification, based on an assay of ten males and ten females sampled

across the species' range, specifically Volos (GR), Bar (ME), Southampton (UK), Saint-Mélaine-sur-Aubance (FR), Jerusalem (IS), Trabzon (TR), as well as sexed DNA from mixed seeds from multiple populations from north-eastern Spain.

Additionally, we identified six more sex-linked markers on the basis of sequences that were obtained by genome capture from males, and not females, sampled from across the species range (details in [30]). In this case, two males and two females were used from the following populations: Southampton (UK), Brouck (FR), Paris (FR), Akçaabat (TR), Antalya (TR), Corinth (GR), Volos (GR), St pere de Ribes (SP), Tarragona (SP), and Barcelona (SP). These additional six sequences were also confirmed by PCR, as above. One of the six sequences identified by this second approach was identical to one of the 24 identified on the basis of gene expression data. In total, therefore, we identified 24 + 6 - 1 = 29 sex-linked markers, which likely represent Y-specific transcripts, or parts thereof (Table 2). All sequences were amplified with the same PCR program, which used 55°C annealing temperature, 1.5 min amplification time and 30 cycles. All sex-linked markers were also Sanger-sequenced to confirm that the PCR product was the expected gene sequence (see Supplementary File 2).

2.3 BAC library construction, sequencing and assembly

M. annua leaves from two males were collected in November 2013 from wild plants growing on the campus of the University of Lausanne (Switzerland). Following storage of leaf material at -80°C, high molecular weight DNA was obtained from these samples using nucleus extractions at the Centre National de Ressources Génomiques Végétales (CNRGV), Toulouse. The DNA was fragmented and ligated into BAC vectors (pIndigoBAC-5) before transformation. The resulting BAC library was screened for the presence of the 29 Y-specific PCR markers described above and we were successful in selecting 18 recombinant colonies amplifying targeted markers. DNA extracted from these colonies was sequenced at the Centre of Integrative Genomics, University of Lausanne, using Pacific Biosciences (PacBio) technology.

Specifically, the BAC DNA was sheared in a Covaris g-TUBE (Covaris, Woburn, MA, USA) to obtain fragments with 6 Kb mean length. After shearing, the DNA size

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

distribution was checked on a fragment analyzer (Advanced Analytical Technologies, Ames, IA, USA). About 300 ng were obtained in 150 µl at 12,000 RPM and were then concentrated by speedvac to 4 µl. Barcoded adapters were added to each BAC during ligation, and BACs were pooled for sequencing. Multiplexing was performed using the SMRTbell Barcoded Adapter Prep Kit #100-465-800 (Pacific Biosciences, Menlo Park, CA, USA). 1.3 µg of the sheared DNA was used to prepare each SMRTbell library with the PacBio SMRTbell Template Prep Kit 1 (Pacific Biosciences, Menlo Park, CA, USA) according to the manufacturer's recommendations. BAC DNA was also sheared to 30 and 50 Kb to obtain two follow-up libraries (pool1 and pool2), without multiplexing.

Each library was sequenced on one SMRT cell with P6/C4 chemistry and MagBeads on a PacBio RSII system (Pacific Biosciences, Menlo Park, CA, USA) at movie lengths of 240 min for the multiplexed library and 360 min for the follow-up libraries. Assembly was performed using the PacBio module "RS_HGAP_Assembly.2" in SMRTpipe version v2.3.0. The multiplexed library assembly did not result in many circularised individual BACs, possibly because of the limited fragment length required for multiplexing. We therefore mixed equal quantities of DNA from the BACs, split into two independent pools (pool1, pool2), and sequenced the corresponding long reads. We used the partial assembly of the multiplexed BAC library to identify reads only mapping to BACs of known origin (from demultiplexing), and assembled those longer reads independently. After successfully circularising the longer read contigs, we removed the vector sequence and linearized the BAC contigs. We used Geneious v9 [33]to assemble these BAC contigs into scaffolds. We used the scaffolds, which contained overlapping BAC sequence, for ORF counting, but the BAC contigs were used for all other analyses, because they allowed detection of subtle differences in their sequence. Table 3 shows a summary of the final assemblies used in downstream analysis.

2.4 Transcript, transposable element (TE) and genomic read mapping

RepeatMasker [34] was run on the genomic contigs and the BAC contigs, using the custom *M. annua* repeat library [29]. This allowed to compare the repeat content of the BACs to that of the full genome and produce a masked BAC assembly. BAC 8, whose

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

ORFs were not sex-linked, was treated separately. The *M. annua* transcriptome ORFs were aligned to the masked BAC assembly with Blat with option "minIdentity 92" [35]. Blat allows large gaps that are compatible with introns. The resulting gff file was used to identify how many of the sex-linked ORFs localise on the BACs. It was also used as a hints file, along with the parameter "--species=arabidopsis" by the gene predictor Augustus v3.2.3 [36], which was also run on the masked BAC contigs, to identify sex linked genes that were not expressed in the samples used for genetic map construction.

3. Results

3.1 Identification of male-specific genes based on expression and genome capture data

Twelve of the 24 genes that had male-biased expression were consistently amplified in males only (and thus were putatively Y-linked), whereas the remaining 12 amplified in both males and females. We also identified six genomic regions present on the Y-chromosome using genome capture data [30], by identifying sequences that were entirely missing from all females, but present in all males; one of these coincided with a sequence found on the basis of gene expression data, so that in total we obtained 17 male-specific PCR products. Three groups of them containing seven transcripts were already expected to be physically closely linked, based on their localization on the same genomic contig (g9930/g9932 and g15325/g15326/g15327, g17561/g17562). We thus conservatively estimate that there are 13 independent parts of the Y chromosome that can be sampled with these PCR products. Details for male-specific PCR amplification primers are summarized in Table 2. Sequences of the products are provided in Supplementary File 2.

3.2 BAC assembly

Our screening approach allowed us to identify 17 BACs containing 11 independent Y-linked DNA sequences, which we were able to confirm via male-specific PCR. A further BAC (BAC 8) was a false positive, which ultimately did not contain a Y-linked sequence. The BACs were aligned to each other and grouped into 11 non-overlapping sets. Four of these groups contained the same male-specific PCR product, indicating either sequence duplication or within-population variation (recall that the sequences were obtained from

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

two different males from Switzerland). Finally, 10/11 independently localising male-specific PCR products were found in these 11 non-overlapping BAC groups (Table 3). Overall, these 11 BAC groups cover a genomic region of 1.5 Mb, corresponding to about 0.47% of the whole haploid *M. annua* genome.

3.3 Functional characterization of genes located on the BACs

By mapping the *M. annua* open reading frames (ORFs) to the BAC scaffolds, we identified 24 broad genomic regions that each contained one or more complete ORFs (i.e., ORFs that mapped over their full length with identity > 90%). These ORFs can be considered functional genes in the non-recombining region of the Y chromosome. The exact gene number matching these BAC regions is difficult to estimate, because some ORFs overlap and may represent alternatively spliced variants of the same gene. Nevertheless, we identified at most 51 ORFs that mapped across their full length to the BACs, 14 of which blasted against sequences in the NCBI's nr database. In addition, we found 87 ORFs that mapped over part of their sequence length to the BACs, and some of these may represent truncated genes on the Y-chromosome; these sequences were located in 53 broad BAC regions. Finally, we found one complete and one truncated copy of the same gene (a sulfate transporter) next to each other on the same BAC (BAC 3), suggesting that it may be the product of a localized, incomplete gene duplication. BAC localisation of ORFs and gene annotation information is provided in Supplementary File 1.

Gene prediction using Augustus v3.2.3 identified a further 12 putative genes on the BAC sequence (Table 3) that did not overlap to previously identified ORFs. These are likely candidates for further male-specific PCR. None of the functions of those newly identified putative genes (based on blastp matches to the protein NCBI's database) is an obvious candidate for sex determination. However, some of the ORFs mapping to the sequenced BACs might be involved in either sex determination or male-beneficial effects and are good candidates for future study. They include a transcript in BAC 6 similar to the agamous-like MADS-box protein AGL66, which is required for pollen maturation and pollen-tube growth in *Arabidopsis* [37]. Another interesting transcript, from BAC 12,

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

is similar to light-dependent short hypocotyls 6, which is part of a family involved in responses to light and organogenesis [38]. Finally, BAC 13 contained a transcript matching two-component response regulator-like PRR73, which controls photoperiodic flowering response [39]. Identification of such candidate genes is of course only the first step to establishing involvement in sex determination, and thus remains speculative. For instance, BAC 8 revealed a strong candidate sex determiner (similarity to auxin response). However, as noted above, BAC 8 is probably not on the SDR, as its assembled sequence did not contain the male specific PCR product used to identify the BAC, and most of its ORFs mapped to an autosome (LG2; see Fig. 1.). Revealingly, BAC 8 also looks very different to the other BACs in terms of both ORF density and repeat content (Fig. 1.).

Four groups of BACs contained the same ORFs (Fig. 1.). For the group containing BACs 1, 2 and 9, we found that three similar, but different genomic regions had been sequenced. BAC 2 contains a predicted gene not found in the other BACs (possibly a chloroplastic insertion, Fig. 1). However, BAC 1 and 9 also differ, because BAC 9 is missing a predicted gene present in BAC 1 and BAC 2. As only two males were sequenced, each with a single Y chromosome, the results can be interpreted as a duplication of the whole BAC sequence, or assembly error. The remaining three groups of BACs appear to have sampled the same genomic region multiple times.

3.4 Comparison of TE density and type between the BACs and full genome

Using repeatMasker v 4.0.7 [34], we inferred that 76.9% of the BAC assembly comprised of repetitive elements, substantially higher than the 47.9% repetitive content of the full genome [29]. This was true for all categories of repetitive sequence, except for simple repeats. Long terminal repeats (LTRs) showed the largest enrichment on BACs compared to the genomic contigs (25.33% vs 8.45%; Table 4). BAC 8 had a lower repeat content than that across the rest of the genome, boosting our confidence that it is not sex-linked.

4. Discussion

4.1 Identification of Y-linked markers

The combined use of RNAseq and genome capture data permitted us to identify 17 new single or low-copy Y-linked markers, which we confirmed through male-specific PCR amplification. We note, in particular, that 12 of 24 transcripts with male-specific and male-biased expression could only be amplified in males, and are thus Y-linked. Our result confirms that the X and Y chromosomes of *M. annua* are differentiated at the sequence level, even though they appear homomorphic (see below). Moreover, our high rate of success at finding new sex-linked transcripts on the basis of their sex-limited expression suggests that ignoring or filtering sex-limited genes in transcriptome analysis may lead to overlooking loci in the SDR [40, 41] .

As an interesting contrast to our finding of Y-linked genes with male-limited expression, Baker *et al.* [42] found genes only expressed in males to be enriched on the X chromosome of stalk-eyed flies. Their result seems more surprising than ours, and runs counter to the feminization (or demasculization) of expression patterns of X-linked genes reported in other studies [43-45]. Nevertheless, it is unclear how often we ought to expect genes with male-limited expression to be on the Y chromosome. Note that the patterns of gene expression used in our study here were observed for genes expressed in non-reproductive tissues [31], whereas many of those reported by Baker *et al.* [42] were gonadal.

The identification of 17 new potential single or low-copy sex-linked markers in *M. annua* represents a substantial advance on previous work on the species by Khadka *et al.* [46], who identified a ‘Single-sequence Characterized Amplified Region’ (SCAR) marker that was male-specific. This SCAR marker corresponds to a high-copy transposable element that is present in both sexes [29, 46], and is thus of limited utility beyond the identification of the sex of pre-reproductive individuals (see [31]). We used several of the new sex-linked markers to probe a newly constructed BAC library for sex-linked genomic regions, which were the main focus of the present study.

4.2 Size of the SDR

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

We may gain an idea of the size of the SDR of *M. annua* by taking two different approaches, which suggest rather divergent values. First, only six ORFs contained in the BACs described here, which we independently infer to be non-recombining based on the PCR result, were found in the sex-linked region of *M. annua* previously mapped by Ridout *et al.* [29]. These transcripts span a region from 52 to 66.82 cM in the female recombination map, corresponding to 441 ORFs, a length of 14.5 Mb and a proportion of 4.86% of the genome. This estimate is somewhat smaller than the one based on the mapping families alone (568 ORFs, equivalent to 19 Mb; [29]).

Second, given that only 6/441 = 1.3% of sex-linked transcripts from Ridout *et al.* [29] map to the combined, non-overlapping, 1.5 Mb of the sequenced BACs, and if we assume that the sex-linked transcripts are distributed similarly on the rest of the SDR not sampled by our BACs, we might infer the SDR to be 1.5 Mb x 441/6, or about 110 Mb. This second estimate corresponds to about 34% of the haploid genome of *M. annua* and is evidently too large. It would seem likely, therefore, that our BACs substantially under-represent the average gene density across the SDR. We thus suggest that the most reliable estimate is still based on the number of transcripts not recombining in males, i.e., between 14.5 Mb (the length spanned by the sex-linked BACs) and 19 Mb (inferred from recombination in small crossing families).

Plant species with homomorphic sex chromosomes usually have a SDR that does not exceed 1% of the total length of the Y chromosome, e.g., *Vitis vinifera* [77,78], several *Populus* species [72,74], and *Fragaria chiloensis* [64] have SDR estimated to be < 1 Mb. Compared with these other plant species with homomorphic sex chromosomes, our results suggest that *M. annua* is an outlier in having a relatively large SDR. This result adds to the evidence from Ridout *et al.* [29] that the sex chromosomes of *M. annua* have been evolving independently of the rest of the genome, though still without substantial degeneration. In this context, it is worth recalling, for instance, that YY males, which lack the X, are completely viable in *M. annua* [47].

4.3 Content of the SDR

Although none of the *de-novo* predicted genes on the BACs were obvious strong

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

candidates for sex determination, we identified three ORFs located on the BACs that might function in sex determination or male benefit: a circadian gene involved in flowering, a light response gene involved in organ and boundary differentiation [38], and a transcription factor associated with pollen maturation [37]. These are interesting genes to follow up in a future study, for example for surveys of population variation, or gene expression at critical time points in sex determination.

Analysis of the content of the BAC contigs revealed typical features of non-recombining sex-chromosomes, congruent with expectations of partial Y degeneration. For instance, the BACs contained mainly repetitive elements, which are expected to accumulate in non-recombining regions [48]. Comparison of the genomic scaffolds [29] and the sex-linked BACs reveals the BACs to be enriched in transposable elements (TEs). This finding is similar to that for the *Carica papaya* Y^h chromosome, for example, where more than 80% of the non-recombining region that has undergone inversion is composed of repetitive elements, whereas the homologous region on the X contains only around 60% [49]. We also found a clear case of partial gene duplication next to a complete gene and multiple complete ORFs that did not completely map to the BACs. Although we cannot characterize how many of these ORFs are real disrupted genes (because no X-only contigs are available), their high number is consistent with early stages of degeneration of the Y chromosome of *M. annua*.

Finally, we note that the overall gene density in our sex-linked BACs appears to be lower than observed in the rest of the genome [29]. In this context, it is interesting that BAC 8, which turned out not to be sex-linked, had a much higher gene density than the sex-linked BACs and a much lower repeat composition than the sex-linked BACs or the genomic contigs.

5. Conclusions

Dioecious plants offer tremendous scope for examining the evolution of sex chromosomes because separate sexes have evolved independently, often relatively recently. A particularly noteworthy feature of dioecious plants is the degree to which they vary in the relative sizes of their X and Y (or Z and W) chromosomes, with species that

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

have strongly heteromorphic sex chromosomes often closely related to species whose sex chromosomes are homomorphic [6, 7]. Species with homomorphic sex chromosomes might simply be young, or might be subject to processes that maintain relative uniformity between homologues and a small SDR, such as frequent turnover [50] or occasional recombination [51]. However, our study of the Y chromosome of *M. annua* illustrates that the SDR in homomorphic sex chromosomes also be relatively large. Indeed, the SDR of *M. annua* appears to be the largest for all plants with homomorphic sex chromosomes studied so far, whether viewed in absolute terms or relative to the size of the sex chromosomes or the rest of the genome. The BAC sequences analyzed here point to an SDR with low gene density and enriched for repeats, with often incomplete mapping of complete ORFs (see also [29]. In this sense, it may represent a species at a particularly interesting intermediate stage along the path towards sex-chromosome heteromorphism and Y-chromosome degeneration.

Acknowledgments

We are grateful to Thomas Chassaing, Yusuf Kurt and Fernando del Caño for providing plant material, to Melanie Dupasquier and Lausanne Genomic Technologies Facility for PacBio library production and sequencing and to Mathias Scharmann and other members of the Pannell lab for valuable discussion. This work was supported by a Sinergia and personal grant to JRP from the Swiss National Science Foundation and a Marie Sklodowska Curie Fellowship grant to SCGM and JRP. The purchase of the Pacific Biosciences RSII instrument at the University of Lausanne was financed in part by the Loterie Romande through the Fondation pour la Recherche en Médecine Génétique.

Author contributions

JRP, GC, PV conceived the project. GC and GB prepared the BAC library. EB assembled the BAC contigs. GC identified male specific transcripts in RNAseq. SCGM designed and provided the genome capture data, which were analysed by CR. JRP, GC and PV wrote the manuscript with input from all authors. All authors have approved the final article.

Competing interests

The authors declare no competing interests.

References

1. Renner, S.S., and Ricklefs, R.E. Dioecy and its correlates in the flowering plants. *Amer. J. Bot.* **1995**. *82*, 596-606.
2. Renner, S.S. The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. *Amer. J. Bot.* **2014**. *101*, 1588-1596.
3. Beukeboom, L.W., and Perrin, N. *The Evolution of Sex Determination*. ed. Oxford University Press. Oxford. 2014.
4. Policansky, D. Sex change in plants and animals. *Annu. Rev. Ecol. Syst.* **1982**. *13*, 471-495.
5. Zimmerman, J.K. Ecological correlates of labile sex expression in the orchid *Catasetum viridiflavum*. *Ecology*. **1991**. *72*, 597-608.
6. Ming, R., Bendahmane, A., and Renner, S.S. Sex chromosomes in land plants. In *Book*., S.S. Merchant, W.R. Briggs and D. Ort, edspp. 485-514.
7. Charlesworth, D. Plant sex chromosomes. *Annual Review of Plant Biology*. **2016**. *67*, 397-420.
8. Sousa, A., Fuchs, J., and Renner, S.S. Cytogenetic comparison of heteromorphic and homomorphic sex chromosomes in *Coccinia* (Cucurbitaceae) points to sex chromosome turnover. *Chromosome Res.* **2017**. *25*, 191-200.
9. Charlesworth, D. Plant contributions to our understanding of sex chromosome evolution. *New Phytol.* **2015**. *DOI: 10.1111/nph.13497*,
10. Sousa, A., Fuchs, J., and Renner, S.S. Molecular cytogenetics (FISH, GISH) of *Coccinia grandis*: A ca. 3 myr-old species of Cucurbitaceae with the largest Y/autosome divergence in flowering plants. *Cytogenet. Genome Res.* **2013**. *139*, 107-118.

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

11. Cherif, E., Zehdi-Azouzi, S., Crabos, A., Castillo, K., Chabrillange, N., Pintaud, J.C., Salhi-Hannachi, A., Glemin, S., and Aberlenc-Bertossi, F. Evolution of sex chromosomes prior to speciation in the dioecious *Phoenix* species. *J. Evol. Biol.* **2016**. *29*, 1513-1522.
12. Bergero, R., and Charlesworth, D. The evolution of restricted recombination in sex chromosomes. *Trends Ecol. Evol.* **2009**. *24*, 94-102.
13. Rice, W.R. The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. *Evolution*. **1987**. *41*, 911-914.
14. Hobza, R., Cegan, R., Jesionek, W., Kejnovsky, E., Vyskot, B., and Kubat, Z. Impact of repetitive elements on the Y chromosome formation in plants. *Genes*. **2017**. *8*, 12.
15. Westergaard, M. The mechanism of sex determination in dioecious plants. *Advances in Genetics*. **1958**. *9*, 217-281.
16. Qiu, S., Bergero, R., Guirao-Rico, S., Campos, J.L., Cezard, T., Gharbi, K., and Charlesworth, D. RAD mapping reveals an evolving, polymorphic and fuzzy boundary of a plant pseudoautosomal region. *Mol. Ecol.* **2016**. *25*, 414-430.
17. Fraser, L.G., Tsang, G.K., Datson, P.M., De Silva, H.N., Harvey, C.F., Gill, G.P., Crowhurst, R.N., and McNeilage, M.A. A gene-rich linkage map in the dioecious species *Actinidia chinensis* (kiwifruit) reveals putative X/Y sex-determining chromosomes. *BMC Genomics*. **2009**. *10*, 15.
18. Mathew, L.S., Spannagl, M., Al-Malki, A., George, B., Torres, M.F., Al-Dous, E.K., Al-Azwani, E.K., Hussein, E., Mathew, S., Mayer, K.F.X., et al. A first genetic map of date palm (*Phoenix dactylifera*) reveals long-range genome structure conservation in the palms. *BMC Genomics*. **2014**. *15*, 10.
19. Hobza, R., Lengerova, M., Svoboda, J., Kubekova, H., Kejnovsky, E., and Vyskot, B. An accumulation of tandem DNA repeats on the Y chromosome in *Silene latifolia* during early stages of sex chromosome evolution. *Chromosoma*. **2006**. *115*, 376-382.

20. Na, J.K., Wang, J.P., Murray, J.E., Gschwend, A.R., Zhang, W.L., Yu, Q.Y., Navajas-Perez, R., Feltus, F.A., Chen, C.X., Kubat, Z., et al. Construction of physical maps for the sex-specific regions of papaya sex chromosomes. *BMC Genomics*. **2012**. *13*, 11.

21. Telgmann-Rauber, A., Jamsari, A., Kinney, M.S., Pires, J.C., and Jung, C. Genetic and physical maps around the sex-determining M-locus of the dioecious plant asparagus. *Molecular Genetics and Genomics*. **2007**. *278*, 221-234.

22. Pannell, J.R., Dorken, M.E., Pujol, B., and Berjano, R. Gender variation and transitions between sexual systems in *Mercurialis annua* (Euphorbiaceae). *Int. J. Pl. Sc.* **2008**. *169*, 129-139.

23. Russell, J.R.W., and Pannell, J.R. Sex determination in dioecious *Mercurialis annua* and its close diploid and polyploid relatives. *Heredity*. **2015**. *114*, 262-271.

24. Durand, B. Le complèxe *Mercurialis annua* L. *s.l.*: une étude biosystématique. *Ann. Sci. Nat. Bot. Paris*. **1963**. *12*, 579-736.

25. Durand, R., and Durand, B. Dioecy, monoecy, polyploidy and speciation in the annual Mercuries. *Bull. Soc. Bot. France Lett. Bot.* **1992**. *139*, 377-399.

26. Obbard, D.J., Harris, S.A., and Pannell, J.R. Sexual systems and population genetic structure in an annual plant: testing the metapopulation model. *Amer. Nat.* **2006**. *167*, 354-366.

27. Eppley, S.M., and Pannell, J.R. Sexual systems and measures of occupancy and abundance in an annual plant: testing the metapopulation model. *Amer. Nat.* **2007**. *169*, 20-28.

28. Pannell, J.R., Eppley, S.M., Dorken, M.E., and Berjano, R. Regional variation in sex ratios and sex allocation in androdioecious *Mercurialis annua*. *J. Evol. Biol.* **2014**. *27*, 1467–1477.

29. Ridout, K., Veltsos, P., Muyle, A., Emery, O., Rastas, P., Marais, G., Filatov, D., and Pannell, J.R. Hallmarks of early sex-chromosome evolution in the dioecious plant *Mercurialis annua* revealed by de novo genome assembly, genetic mapping and transcriptome analysis. *bioRxiv*. **2017**. *doi.org/10.1101/106120*,

30. Gonzalez-Martinez, S.C., Ridout, K., and Pannell, J.R. Range expansion compromises adaptive evolution in an outcrossing plant. *Curr. Biol.* **2017**. *27*, 2544-+.

31. Cossard, G., and Pannell, J.R. Sexual dimorphism and rapid turnover in gene expression in pre-reproductive seedlings of a dioecious herb. **submitted**.

32. Love, M.I., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. **2014**. *15*, 38.

33. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. **2012**. *28*, 1647-1649.

34. Smit, A.F.A., Hubley, R., and Green, P. RepeatMasker Open-4.0. <http://www.repeatmasker.org/>. **2013-2015**.

35. Kent, W.J. BLAT - The BLAST-like alignment tool. *Genome Res.* **2002**. *12*, 656-664.

36. Stanke, M., Schoffmann, O., Morgenstern, B., and Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*. **2006**. *7*, 11.

37. Adamczyk, B.J., and Fernandez, D.E. MIKC* MADS Domain Heterodimers Are Required for Pollen Maturation and Tube Growth in Arabidopsis. *Plant Physiol.* **2009**. *149*, 1713-1723.

38. Cho, E.N., and Zambryski, P.C. ORGAN BOUNDARY1 defines a gene expressed at the junction between the shoot apical meristem and lateral organs. *Proceedings of the National Academy of Sciences of the United States of America*. **2011**. *108*, 2154-2159.

39. Murakami, M., Matsushika, A., Ashikari, M., Yamashino, T., and Mizuno, T. Circadian-associated rice pseudo response regulators (OsPRRs): Insight into the control of flowering time. *Biosci. Biotechnol. Biochem.* **2005**. *69*, 410-414.

40. Muyle, A., Zemp, N., Deschamps, C., Mousset, S., Widmer, A., and Marais, G.A.B. Rapid de novo evolution of X chromosome dosage compensation in *Silene latifolia*, a plant with young sex chromosomes. *Plos Biology*. **2012**. *10*,
41. Uebbing, S., Kunstner, A., Makinen, H., and Ellegren, H. Transcriptome Sequencing Reveals the Character of Incomplete Dosage Compensation across Multiple Tissues in Flycatchers. *Genome Biol. Evol.* **2013**. *5*, 1555-1566.
42. Baker, R.H., Narechania, A., DeSalle, R., Johns, P.M., Reinhardt, J.A., and Wilkinson, G.S. Spermatogenesis Drives Rapid Gene Creation and Masculinization of the X Chromosome in Stalk-Eyed Flies (Diopsidae). *Genome Biol. Evol.* **2016**. *8*, 896-914.
43. Sturgill, D., Zhang, Y., Parisi, M., and Oliver, B. Demasculinization of X chromosomes in the *Drosophila* genus. *Nature*. **2007**. *450*, 238-U233.
44. Parisi, M., Nuttall, R., Naiman, D., Bouffard, G., Malley, J., Andrews, J., Eastman, S., and Oliver, B. Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. *Science*. **2003**. *299*, 697-700.
45. Gao, G., Vibranovski, M.D., Zhang, L., Li, Z., Liu, M., Zhang, Y.E., Li, X.M., Zhang, W.X., Fan, Q.C., VanKuren, N.W., et al. A long-term demasculinization of X-linked intergenic noncoding RNAs in *Drosophila melanogaster*. *Genome Res.* **2014**. *24*, 629-638.
46. Khadka, D.K., Nejidat, A., Tal, M., and Golan-Goldhirsh, A. DNA markers for sex: Molecular evidence for gender dimorphism in dioecious *Mercurialis annua L*. *Mol. Breed.* **2002**. *9*, 251-257.
47. Kuhn, E. Selbstbestäubungen subdiöcischer Blütenpflanzen, ein neuer Beweis für die genetische Theorie der Geschlechtsbestimmung. *Planta*. **1939**. *30*, 457-470.
48. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **2013**. *14*, 113-124.
49. Wang, J.P., Na, J.K., Yu, Q.Y., Gschwend, A.R., Han, J., Zeng, F.C., Aryal, R., VanBuren, R., Murray, J.E., Zhang, W.L., et al. Sequencing papaya X and Y-h chromosomes reveals molecular basis of incipient sex chromosome evolution.

Proceedings of the National Academy of Sciences of the United States of America. **2012**. *109*, 13710-13715.

50. van Doorn, G.S., and Kirkpatrick, M. Turnover of sex chromosomes induced by sexual conflict. *Nature*. **2007**. *449*, 909-912.
51. Perrin, N. Sex reversal: a fountain of youth for sex chromosomes? *Evolution*. **2009**. *63*, 3043-3049.
52. Sakamoto, K., Akiyama, Y., Fukui, K., Kamada, H., and Satoh, S. Characterization, genome sizes and morphology of sex chromosomes in hemp (*Cannabis sativa L.*). *Cytologia (Tokyo)*. **1998**. *63*, 459–464.
53. van Bakel, H., Stout, J.M., Cote, A.G., Tallon, C.M., Sharpe, A.G., Hughes, T.R., and Page, J.E. The draft genome and transcriptome of *Cannabis sativa*. *Genome Biology*. **2011**. *12*, 17.
54. Grabowska-Joachimiak, A., Sliwinska, E., Pigula, M., Skomra, U., and Joachimiak, A.J. Genome size in *Humulus lupulus* L. and *H. japonicus* Siebold & Zucc. (Cannabaceae). *Acta Soc. Bot. Pol.* **2006**. *75*, 207-214.
55. Divashuk, M.G., Alexandrov, O.S., Kroupin, P.Y., and Karlov, G.I. Molecular cytogenetic mapping of *Humulus lupulus* sex chromosomes. *Cytogenet. Genome Res.* **2011**. *134*, 213-219.
56. Armstrong, S.J., and Filatov, D.A. A cytogenetic view of sex chromosome evolution in plants. *Cytogenet. Genome Res.* **2008**. *120*, 241-246.
57. Bennett, M.D., and Leitch, I.J. (2012). http://www.rbgkew.org.uk/cval/homepage.html.
58. Blocka-Wandas, M., Sliwinska, E., Grabowska-Joachimiak, A., Musial, K., and Joachimiak, A.J. Male gametophyte development and two different DNA classes of pollen grains in *Rumex acetosa* L., a plant with an XX/XY1Y2 sex chromosome system and a female-biased sex ratio. *Sexual Plant Reproduction*. **2007**. *20*, 171-180.

59. Shibata, F., Hizume, M., and Kuroki, Y. Differentiation and the polymorphic nature of the Y chromosomes revealed by repetitive sequences in the dioecious plant, *Rumex acetosa*. *Chromosome Res.* **2000**. *8*, 229-236.

60. Hough, J., Hollister, J.D., Wang, W., Barrett, S.C.H., and Wright, S.I. Genetic degeneration of old and young Y chromosomes in the flowering plant *Rumex hastatulus*. *Proceedings of the National Academy of Sciences of the United States of America*. **2014**. *111*, 7713-7718.

61. Grabowska-Joachimiak, A., Kula, A., Ksiazczyk, T., Chojnicka, J., Sliwinska, E., and Joachimiak, A.J. Chromosome landmarks and autosome-sex chromosome translocations in *Rumex hastatulus*, a plant with XX/XY1Y2 sex chromosome system. *Chromosome Res.* **2015**. *23*, 187-197.

62. Huang, S.X., Ding, J., Deng, D.J., Tang, W., Sun, H.H., Liu, D.Y., Zhang, L., Niu, X.L., Zhang, X., Meng, M., et al. Draft genome of the kiwifruit *Actinidia chinensis*. *Nature Communications*. **2013**. *4*, 9.

63. Zhang, Q., Liu, C.Y., Liu, Y.F., VanBuren, R., Yao, X.H., Zhong, C.H., and Huang, H.W. High-density interspecific genetic maps of kiwifruit and the identification of sex-specific markers. *DNA Res.* **2015**. *22*, 367-375.

64. Tennessen, J.A., Govindarajulu, R., Liston, A., and Ashman, T.L. Homomorphic ZW chromosomes in a wild strawberry show distinctive recombination heterogeneity but a small sex-determining region. *New Phytol.* **2016**. *211*, 1412–1423.

65. Kafkas, S., Khodaeiaminjan, M., Guney, M., and Kafkas, E. Identification of sex-linked SNP markers using RAD sequencing suggests ZW/ZZ sex determination in *Pistacia vera L. BMC Genomics*. **2015**. *16*, 11.

66. Ming, R., Hou, S.B., Feng, Y., Yu, Q.Y., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., Lewis, K.L.T., et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*. **2008**. *452*, 991-U997.

67. Wang, J.P., Na, J.K., Yu, Q.Y., Gschwend, A.R., Han, J., Zeng, F.C., Aryal, R., VanBuren, R., Murray, J.E., Zhang, W.L., et al. Sequencing papaya X and Yh

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proceedings of the National Academy of Sciences of the United States of America*. **2012**. *109*, 13710-13715.

68. Spigler, R.B., Lewers, K.S., Main, D.S., and Ashman, T.L. Genetic mapping of sex determination in a wild strawberry, *Fragaria virginiana*, reveals earliest form of sex chromosome. *Heredity*. **2008**. *101*, 507-517.
69. Tamura, M., Tao, R., Yonemori, K., Utsunomiya, N., and Sugiura, A. Ploidy level and genome size of several *Diospyros* species. *J. Jpn. Soc. Hortic. Sci.* **1998**. *67*, 306-312.
70. Akagi, T., Henry, I.M., Tao, R., and Comai, L. A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. *Science*. **2014**. *346*, 646-650.
71. Yin, T., DiFazio, S.P., Gunter, L.E., Zhang, X., Sewell, M.M., Woolbright, S.A., Allan, G.J., Kelleher, C.T., Douglas, C.J., Wang, M., et al. Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. *Genome Res.* **2008**. *18*, 422-430.
72. Paolucci, I., Gaudet, M., Jorge, V., Beritognolo, I., Terzoli, S., Kuzminsky, E., Muleo, R., Mugnozza, G.S., and Sabatti, M. Genetic linkage maps of *Populus alba* L. and comparative mapping analysis of sex determination across *Populus* species. *Tree Genet. Genomes*. **2010**. *6*, 863-875.
73. Pakull, B., Groppe, K., Meyer, M., Markussen, T., and Fladung, M. Genetic linkage mapping in aspen (*Populus tremula* L. and *Populus tremuloides* Michx.). *Tree Genet. Genomes*. **2009**. *5*, 505-515.
74. Geraldes, A., Hefer, C.A., Capron, A., Kolosova, N., Martinez-Nunez, F., Soolanayakanahally, R.Y., Stanton, B., Guy, R.D., Mansfield, S.D., Douglas, C.J., et al. Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). *Mol. Ecol.* **2015**. *24*, 3243-3256.
75. Pucholt, P., Ronnberg-Wastljung, A.C., and Berlin, S. Single locus sex determination and female heterogamety in the basket willow (*Salix viminalis* L.). *Heredity*. **2015**. *114*, 575-583.

76. Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*. **2007**. *449*, 463-U465.

77. Fechter, I., Hausmann, L., Daum, M., Sorensen, T.R., Viehover, P., Weisshaar, B., and Topfer, R. Candidate genes within a 143 kb region of the flower sex locus in *Vitis*. *Molecular Genetics and Genomics*. **2012**. *287*, 247-259.

78. Picq, S., Santoni, S., Lacombe, T., Latreille, M., Weber, A., Ardisson, M., Ivorra, S., Maghradze, D., Arroyo-Garcia, R., Chatelet, P., et al. A small XY chromosomal region explains sex determination in wild dioecious *V. vinifera* and the reversal to hermaphroditism in domesticated grapevines. *BMC Plant Biol.* **2014**. *14*, 17.

79. Obbard, D.J., Harris, S.A., Buggs, R.J.A., and Pannell, J.R. Hybridization, polyploidy, and the evolution of sexual systems in *Mercurialis* (Euphorbiaceae). *Evolution*. **2006**. *60*, 1801-1815.

80. Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*. **2003**. *423*, 825-U822.

81. Hughes, J.F., Skaletsky, H., Pyntikova, T., Graves, T.A., van Daalen, S.K.M., Minx, P.J., Fulton, R.S., McGrath, S.D., Locke, D.P., Friedman, C., et al. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature*. **2010**. *463*, 536-539.

82. Ellegren, H. Evolution of the avian sex chromosomes and their role in sex determination. *Trends Ecol. Evol.* **2000**. *15*, 188-192.

83. Ayers, K.L., Davidson, N.M., Demiyah, D., Roeszler, K.N., Grutzner, F., Sinclair, A.H., Oshlack, A., and Smith, C.A. RNA sequencing reveals sexually dimorphic gene expression before gonadal differentiation in chicken and allows comprehensive annotation of the W-chromosome. *Genome Biology*. **2013**. *14*, 16.

84. Zhou, Q., Zhu, H.M., Huang, Q.F., Zhao, L., Zhang, G.J., Roy, S.W., Vicoso, B., Xuan, Z.L., Ruan, J., Zhang, Y., et al. Deciphering neo-sex and B chromosome evolution by the draft genome of *Drosophila albomicans*. *BMC Genomics*. **2012**. *13*, 12.

85. Ogawa, A., Murata, K., and Mizuno, S. The location of Z- and W-linked marker genes and sequence on the homomorphic sex chromosomes of the ostrich and the emu. *Proceedings of the National Academy of Sciences of the United States of America*. **1998**. *95*, 4415-4418.

86. Zhou, Q., Zhang, J.L., Bachtrog, D., An, N., Huang, Q.F., Jarvis, E.D., Gilbert, M.T.P., and Zhang, G.J. Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science*. **2014**. *346*, 1332-+.

87. Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*. **2012**. *484*, 55-61.

88. Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., et al. The medaka draft genome and insights into vertebrate genome evolution. *Nature*. **2007**. *447*, 714-719.

89. Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. The genome sequence of *Drosophila melanogaster*. *Science*. **2000**. *287*, 2185-2195.

90. Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., et al. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and cis-element evolution. *Genome Res.* **2005**. *15*, 1-18.

91. Steinemann, M., and Steinemann, S. Enigma of Y chromosome degeneration: Neo-Y and Neo-X chromosomes of *Drosophila miranda* a model for sex chromosome evolution. *Genetica*. **1998**. *102-3*, 409-420.

92. Kaiser, V.B., and Charlesworth, B. Muller's ratchet and the degeneration of the *Drosophila miranda* neo-Y chromosome. *Genetics*. **2010**. *185*, 339-U491.

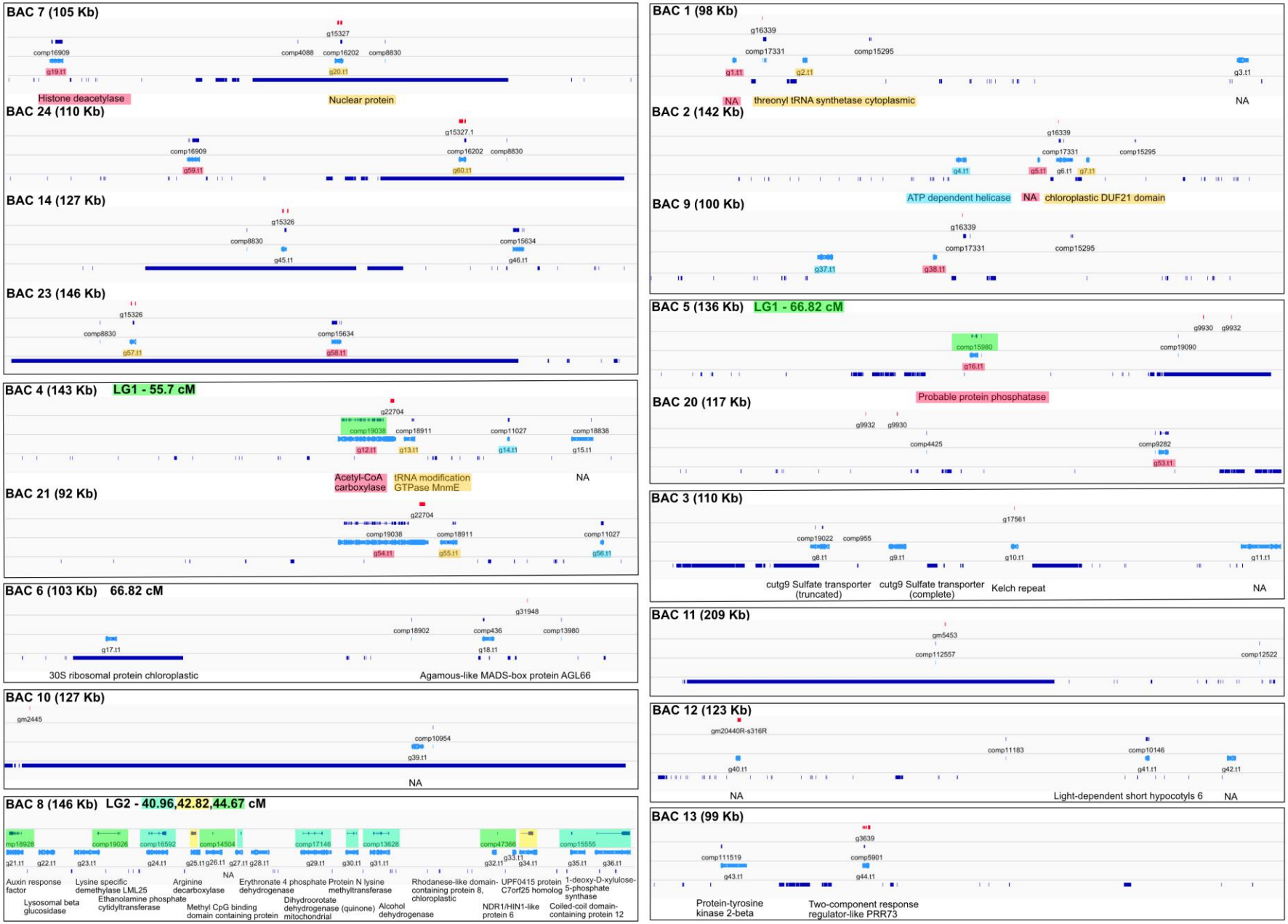
93. Criscione, C.D., Valentim, C.L.L., Hirai, H., LoVerde, P.T., and Anderson, T.J.C. Genomic linkage map of the human blood fluke *Schistosoma mansoni*. *Genome Biology*. **2009**. *10*, 42.

94. Otto, S., Pannell, J., Peichel, C., Ashman, T., Charlesworth, D., Chippindale, A., Delph, L., Guerrero, R., Scarpino, S., and McAllister, B. About PAR: The distinct evolutionary dynamics of the pseudoautosomal region. *Trends Genet.* **2011**. *27*, 358-367.

Fig. 1. Graphical summary of the assembled BACs. Each assembly is annotated with the male specific PCR product location (red), ORF blat hits (top dark blue), Augustus gene predictions on the repeat masked assembly (light blue) and blat hits from the *M. annua* repeat library (bottom dark blue). Each group of non-overlapping BACs is surrounded by a black box. Identical predicted gene models are indicated by overlaid colored boxes, for easy visual alignment of BACs containing the same genes. Green overlaid boxes indicate mapped ORFs that also map well to the BACs, and their female recombination map position. The names displayed under the gene predictions are from their first blastp hit to the nr protein database. Note the BAC contigs are not illustrated in the same scale.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 April 2018 doi:10.20944/preprints201804.0128.v1

Peer-reviewed version available at *Genes* **2018**, *9*, 277; doi:10.3390/genes9060277



9 **Table 1.** Summary of information on size and divergence in systems with homomorphic and heteromorphic sex chromosomes.

Species	Estimated haploid genome size in Mbp (mean male/female)	Sex chromosome types	Heteromorphic sex chromosomes?	Estimated size of sex chromosomes (Mbp)	Size information on the heteromorphic sex-specific region (Mbp)	Notes	References
Angiosperms							
Cannabis sativa	817 (2n = 20) [a]	XX/XY	Yes (Y>X)	-	Diff M/F = c.a 47	Difference of genome size between males and females	[52]; [53]
Hulmulus lupulus	2836 (2n = 20) [a]	XX/XY	Yes (Y<X)	Y: 186.8	Diff M/F = c.a. 73.4	Difference of genome size between males and females	[54]; [55]
Humulus japonicus	1569 (2n = 16/17) [a]	XX/XY$_1$Y$_2$	Yes (Y>X)	Y$_1$: 259.2 Y$_2$: 238.6 X: 270.9	Diff M/F = c.a. 307.1	Difference of genome size between males and females	[54];
Silene latifolia	2832 (2n = 24) [a]	XX:XY	Yes (Y>X)	350 (X) 524(Y)	349	Non-recombining region estimated to be 2/3 of chr. Y	[56]; [57]; [16]
Coccinia grandis	438 (2n=24) [a]	XX/XY	Yes (Y>X)	-	Diff M/F = c.a. 98	Difference of genome size between males and	[10]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 April 2018 doi:10.20944/preprints201804.0128.v1

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

						females	
Rumex acetosa	1614 (2n = 14/15) [a]	XX/XY$_1$Y$_2$ (X/A sex determination)	Yes (Y$_1$/Y$_2$<X)	-	Diff M/F = c.a. 489	Difference of genome size between males and females	[58]; [57]
Rumex hastatulus *1) Texas* *2) North Carolina*	1) 1864 (2n=10) [a] 2) 1801 (2n = 8/9) [a]	1) XX/XY 2) XX/XY$_1$Y$_2$	1) Yes (Y>X) 2) Yes (Y$_1$/Y$_2$<X)	-	1) Diff M/F = c.a 349.1 2) Diff M/F = c.a 333.5	Difference of genome size between males and females	[59]; [60]; [61]
Actinidia chinensis	758 (2n = 58) [b]	XX/XY	No	-	5	Fine scale mapping	[17]; [62]; [63]
Asparagus officinalis	1323 (2n = 20) [a]	XX/XY	No	-	1-10	Raw estimate	[21]; [57]; [64]
Phoenix dactylifera	670 (2n = 36) [b]	XX/XY	No	-	5 - 13	Mapping of sex-specific scaffolds	[18]
Pistacia vera	585 (2n=30)	ZZ/ZW	No	50 (largest chromosome pair)	28 RAD reads with W allele	Sex linked RAD SNPs	[65]
Carica papaya	372 (2n = 18) [b]	XX/XY^{h}Y^h	No	Y^h: 81	8.1 on Y^h (10 % of chr. Y^h)	BAC sequencing of Y^h	[66]; [67]
Fragaria virginiana (subdioecious)	782 (2n = 8x = 58) [a]	ZZ/ZW	No	-	No non-recombining region	Gentic mapping	([68]
Fragaria chiloensis	400 (2n = 56)	ZZ/ZW	No	-	0.280	QTL and amplicon mapping	[64]6
Dyospyros lotus	1809 (2n = 30) [b]	XX/XY	No	-	~ 1	Assembly of Y-specific contigs	[69]; [70]
Populus deltoides x	480 (2n = 38) [a]	ZZ/ZW	No	-	0.71	Fine scale mapping	[71]; [72]

nigra *Populus alba*							
Populus trichocarpa x tremuloides *P. nigra* *P. balsamifera*	485 (2n = 38) [a]	XX/XY	No	-	0.1 (*P. trichocarpa; P. balsamifera)*	Identifiation of sex-associated SNPs	[73]; [74]
Salix virminalis *S. purpurea*	450 (2n = 38) [a]	ZZ/ZW	No	-	< 2.5 (*S. purpurea*)	Fine scale mapping of SD locus	[75]
Vitis vinifera	487 (2n = 38) [b]	XX/XY	No	-	0.143 - 0.155	Estimated to be less than 1 % of chr. Y	[76]; [77]; [78]
Mercurialis annua	645 (2n = 16)	XX/XY	No	-	14.5 - 19	BAC sequencing	[79]; This study
Vertebrates							
Homo sapiens	2900 (2n = 46) [b]	XX/XY	Yes (Y<X)	X = 160 Y = 60	57	Estimated to be 95% of chr. Y	[80]; [81]
Pan troglodytes (Chimpanzee)	2700 (2n = 48) [b]	XX/XY	Yes (Y<X)	-	25.8	BAC sequencing of chr. Y	[80]; [81]; The
Gallus gallus (Chicken)	1050 (2n = 78) [b]	ZZ/ZW	Yes (W<Z)	Z: 82 W: 55	Very low	Observations of chiasmata during meiosis	[82]; [83]; [84]
Struthio camelus (Ostrich)	1230 (2n = 80) [b]	ZZ/ZW	No	95.6	31.9 (~ 1/3 of Z chr.)	Estimated to be a third of the Z chromosome	[85]; [86]
Gasterosteus aculeatus (Threespine	530 (2n = 42) [b]	XX/XY	No (different centromere positions)	X: 20.2 Y: 13.2	10	BAC sequencing and FISH	[87]

stickleback)							
Oryzias latipes (Medaka)	700 (2n = 48) [b]	XX/XY	No	33.7	3.4	Estimated to be 10 % of chr. 1	[88]
Insects							
Drosophila melanogaster	360 (2n = 8) [b]	XY (X/A sex determination)	Yes (Y<X)	X: 41.8 Y: 40.9	40.9	100 % of neo-Y	[89]; [48]
Drosophila albomicans	366 (2n = 6) [b]	Neo-XY	Evidence for degeneration	73.2 (about 40% of the genome	73.2	100 % of neo-Y	[84]; [48]
Drosophila pseudoobscura	312 (2n = 10) [b]	Neo-XY	Evidence for degeneration	-	-	100 % of neo-Y	[90]; [48]3
Drosophila miranda	NA (2n=9/10)	Neo-XY	Yes (Neo-Y<Neo-X)	~ 4.3 Mb (of coding sequence)	~ 4.3 Mb of coding sequence	100 % of neo-Y	[91]; [92]
Flatworms							
Schistosoma mansoni	300 (2n = 16) [c]	ZZ/ZW	No	Z: 60.7 W: 60 - 70	26.7 (44 %)	Estimated to be 44% of chr. Y	[93]; [94]

[a] genome size estimated by flow cytometry

[b] genome size estimated from sequencing data – haploid genome size

[c] genome size estimate used in study, without reference citation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 April 2018 doi:10.20944/preprints201804.0128.v1

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

Table 2. Information on primers, amplifying ORFs in males only or both sexes. Names indicated by a * amplify the same transcript.

Name	primer_F	primer_R	PCR amplification	Identification method
g15325	CATTGGCAGTGAAACCCTGG	TGGATTTCAGTGCAAAGCCT	Male	RNAseq
g15326	GTGACTCTCTCCCTATGGCC	AAACCTTTCTGCACGAGTCG	Male	RNAseq
g15327	TTTGTTGCACCCCGATCAAG	CATCCTCCCTTGCAACGTTT	Male	RNAseq
g16339	ATTCGGGTTTCTCGAGTGGT	ACTAACTGTGTACCAAAAGCTTG	Male	RNAseq
g17303	GTGCGGCAGTCAACACTAC	GACCGGGCTTGAAGTTGAAG	Male	RNAseq
g17561	TCCAGTCATCCCAACGTTCA	TGAACAGAAGGCAGAGACGA	Male	RNAseq
g17562	ACAGTCGGCCTTCATCTTCA	TGAGTCAGAAGAAGAACAAGCT	Male	RNAseq
g22704	TCCGGGAAGCCAGAAATAGT	CGAAGCCCATCCATCAACTG	Male	RNAseq
g31948	TGGAGACGATGGATGTTGCT	AACAGACGGCTCACCCATC	Male	RNAseq
g3639*	ACTGCTGGGACTATCACCTC	TGCATTCGAAGGAGTTTGGAC	Male	RNAseq
g9930	TGCTGGAAATGATGGTTGCC	ACAACTTCTCTCCAGCTGCT	Male	RNAseq
g9932	TGCTGGAAATGATGGTTGCC	ACAACTTCTCTCCAGCTGCT	Male	RNAseq
gm1362	AGGACGTTGTAGAGGTAGACC	GATGGGTCGACATAAGGCAT	Male	Exon capture
gm20440	GGTGTAGCCTTCCCCTTCTT	ACCACTGCCCTGAGAGAATC	Male	Exon capture
gm2445	CTAGTTGGAAGTTGGCGTGG	CCCTTTGCCAAACCGTGTAA	Male	Exon capture
gm44415	AAGTGTCGGCAGTCTTAGGT	GCCTCCATCATGAAGGCTTT	Male	Exon capture
gm5321	GAGCACCCCGACAGATAGAA	GTTTGGAGGACTTAGGGGCT	Male	Exon capture
gm56331*	TCACTACTAGCACAGCCACC	CTGAGAGTTGAGGTTGCACAG	Male	Exon capture
g12424	GCGTGTGAGTGGGCTAATAG	GCACACCATTTTCTTCCTCCT	Both	RNAseq
g13020	TTGATCGGAGCAGAGAGTGG	GGTGTAGCCTTCCCCTTCTT	Both	RNAseq
g17779	TCCTGTTCTGACTTCGACGT	CGAAGAGGCCATGTTAAATCCA	Both	RNAseq
g20091	ATTGAGGAGCTTGTGGACCC	AGTGTCATGACTGGGTTCCC	Both	RNAseq
g22703	TCGCCTACTAGCCATGTTGT	AGAAAAGAAGAAGCCAGCCTG	Both	RNAseq
g25224	GCGACCAAAAGAGGCAGAAT	TGTTGCTGCTATCATCGTGC	Both	RNAseq
g26252	TCCCGATTCCTTCCAGTGAA	TGTTACGTATAGGGCAGCCA	Both	RNAseq
g28106	GGCTGGAATTGCTTTGAACG	TCAATTTGTGGACGCAGCAA	Both	RNAseq
g28854	TGGGGCATACTGATTTGATGTG	CTTCTGAGCTTCTGTCACCTT	Both	RNAseq
g30868	AAGAGTTTGAGGCTGCATCC	GGCTAATACACATGCGGTAGG	Both	RNAseq

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 April 2018 doi:10.20944/preprints201804.0128.v1

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

g31096	GGTAATCCAGCTTCAGTGTGC	ACCACAGGAATCGATTGCAG	Both	RNAseq
g9937	TGGAGGATTATCATGTTGCAAAG	AGCCTCCTGATTCGACAACA	Both	RNAseq

18

19

Table 3. BAC contig information. Overlapping BACs are displayed in the same row and their approximate cumulative single copy length is shown. Candidate sex determining genes are speculated based on the description of the first protein blast hit.

BAC contig	length (Kb)	Male specific PCR hit	ORFs	Additional predicted genes	Candidate genes
1, 2, 9	170	g16339	3	3	
3	110	g17561	1	3	
4, 21	143	g22704	4	0	
5, 20	200	g9930, g9932	3	0	
6	103	g31948	3	1	Agamous-like MADS-box protein AGL66
8	146	NO	13	3	Auxin response
10	127	g2445	1	0	
11	209	g5453	2	0	
12	123	gm20440	2	2	Light-dependent short hypocotyls 6
13	99	g3639	2	0	Two-component response regulator-like PRR73
7, 14, 23, 24	215	g15326, g15327	5	0	

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 April 2018 doi:10.20944/preprints201804.0128.v1

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

Table 4. Summary output from RepeatMasker using the *M. annua* repeat library.

		Percentage of sequence			Number Elements			Length (bp)		
		Y BACs	BAC 8	genome	Y BACs	BAC 8	genome	Y BACs	BAC 8	genome
SINEs	Other	0	0	0.03	0	0	1,127	0	0	163,941
	ALUs	0	0	0	0	0	0	0	0	0
	MIRs	0	0	0	0	0	0	0	0	0
LINEs	Other	4.8	4.28	2.66	134	7	33,781	101,114	6,295	14,537,810
	LINE1	2.87	2.22	2.01	64	5	22,157	60,373	3,264	10,979,834
	LINE2	0	0	0	0	0	0	0	0	0
	L3/CR1	0	0	0	0	0	0	0	0	0
LTR elements	Other	26.63	4.61	8.45	708	14	129,707	560,624	6,782	46,164,934
	ERVL	0	0	0	0	0	0	0	0	0
	ERVL-MaLRs	0	0	0	0	0	0	0	0	0
	ERV_classI	0	0	0	0	0	51	0	0	13,842
	ERV_classII	0	0	0.01	0	0	94	0	0	32,526
DNA elements	Other	6.49	4.01	2.65	254	24	63,665	136,533	5,900	14,467,832
	hAT-Charlie	0	0	0	0	0	0	0	0	0
	TcMar-Tigger	0	0	0	0	0	0	0	0	0
Unclassified		38.14	26.72	31.14	2,605	182	813,812	803,031	39,296	170,128,904
Total interspersed repeats		76.06	39.62	44.93				1,601,302	58,273	245,463,421
Small RNA		0.1	0	0.06	6	0	1,397	2,064	0	352,034
Satellites		0	0	0	0	0	0	0	0	0
Simple repeats		0.67	0.83	2.88	275	28	99,695	14,158	1,228	15,743,639
Low complexity		0.25	0.34	0.15	94	9	16,600	5,336	505	838,784
Bases masked		76.91	40	47.85	1,619,127	60,006	261,415,926	1,619,127	60,006	261,415,926

Peer-reviewed version available at Genes 2018, 9, 277; doi:10.3390/genes9060277

Supplementary files

Supplementary File 1. Blast information on ORF hits to the BAC contigs.

Header

Name - ORF name

Annotation - Blastn information on ORFs whose length fully maps to a BAC

type - Information on whether the ORF represents a complete transcript

length_nt - The length of the ORF in basepairs

sum_hits - The length of all BLAST hits of a given ORF to each BAC

complete_hit - Yes if the ORF length and the sum of all its hits to the BAC closely match (indicates full genes on a BAC)

Multiple gene hits - Indicates the ORF hits more than one BAC or multiple locations within a BAC

minBAC/maxBAC - Rough location of the ORF on the BAC

BAC – BAC contig/scaffold name the ORF hit

Other BAC - Name of other BAC contigs/scaffolds the ORF hit

rough identity - Average identity of each region of the ORF with the BAC region it BLASTed to (not corrected for region size)

BAC_block - Manual subdivision of BAC into regions with independent ORFs mapping to them

Tabs

Summary - Summarises all combinations of ORF-BAC hits

Complete, full mapping - Information and location of complete ORFs, fully mapping to BACs, indicating likely functional transcripts on the BAC contigs/scaffolds

Complete, truncated mapping - Information and location of complete ORFs mapping to BACs for only a subset of their length, indicating likely truncated genes on the BAC assembly.

Remaining tabs provide detailed BLAST information per BAC contig/scaffold, that is summarized in the preceding tabs.

Supplementary File 2. Sanger sequence chromatograms of single copy PCR products.

Supplementary File 3. BAC assembly sequences and mapping files (gff) used to produce Fig. 1., to be used in a genome browser for interactive exploration of the assemblies.

Supplementary File 4. The aminoacid sequence of predicted genes on the BACs. The names of predicted genes with the same sequence have been combined.