Single molecule long read sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*

D.E. Khost[a*], D. Eickbush[a] and A.M. Larracuente[a*]

[a] University of Rochester Department of Biology Rochester, NY 14627

* corresponding authors: dkhost@ur.rochester.edu, alarracu@bio.rochester.edu

337 Hutchison Hall

Department of Biology

University of Rochester, River Campus

Rochester, NY 14627

Keywords: Satellite DNA, SMRT sequencing, PacBio, tandem repeats

ABSTRACT

Highly-repetitive satellite DNA (satDNA) repeats are found in most eukaryotic genomes. SatDNAs are rapidly evolving and have roles in genome stability and chromosome segregation. The repetitive nature of satDNA poses a challenge for genome assembly and makes progress on the detailed study of satDNA structure difficult. Here we experiment with assembly methods using single-molecule sequencing reads from Pacific Biosciences (PacBio) to determine the detailed structure of two complex satDNA loci in *Drosophila melanogaster:* the *260-bp* and *Responder* satellites. We optimized assembly methods and parameter combinations to produce a high quality assembly of these previously unassembled satDNA loci and validate this assembly using molecular and computational methods. We find that satDNA repeats are organized into large arrays interrupted by transposable elements. The repeats in the center of the array tend to be homogenized in sequence, though to a different degree for *Responder* and *260-bp* loci. This suggests that gene conversion and unequal crossovers lead to repeat homogenization through concerted evolution, but the degree of concerted evolution may differ among complex satellite loci. We find evidence for higher order structure within satDNA arrays that suggest recent structural rearrangements.

INTRODUCTION

Satellite DNAs (satDNAs) [1-3] are tandemly repeated DNAs frequently found in regions of low recombination [4] (*e.g.* centromeres, telomeres and Y chromosomes) that can make up a large fraction of eukaryotic genomes [5]. SatDNA families are classified according to their repeat unit

size and composition—simple satellites generally correspond to uniform clusters of small (*e.g* 1-10 bp) repeat units and complex satellites correspond to more variable clusters of larger (*e.g.* >100 bp) repeat units. SatDNAs are highly dynamic over short evolutionary time scales [6, 7]. Changes in satDNA composition and abundance contribute to the evolution of genome structure [4], speciation [8, 9] and meiotic drive [10, 11]. Early studies on satDNA (correctly) assumed that it must have some function in protecting against nondisjunction during chromosome segregation [12] or a structural role in the nucleus [8]. However, subsequent studies suggested that satDNAs were inert "junk" [13] that expand in genomes due to selfish replication [14-16]. In the last 15 years, researchers across the fields of evolutionary biology, cell and molecular biology have accumulated evidence that some satDNAs have important functions [9, 17-22]. However, the highly-repetitive nature of satDNA makes the detailed study of their loci difficult.

Gross-scale techniques such as density-gradient centrifugation and *in situ* hybridization demonstrate that satDNAs are organized into large contiguous blocks of repeats [23, 24]. Molecular assays based on restriction digest mapping indicate that satDNA blocks may be interrupted by smaller "islands" of more complex repeats such as transposable elements in Drosophila mini chromosomes [22, 25]. While these methods have been useful in detailing the overall structure of satDNA loci, detailed sequence-level analysis of satDNA arrays is stymied by the shortcomings of traditional sequencing methods. Highly-repetitive arrays are unstable in BACs and cloning vectors [24, 26, 27]—in some cases they are even toxic to *E. coli* and thus are underrepresented in BAC libraries and among Sanger sequence reads [28]. Next generation short-read sequencing methods such as Illumina or 454 circumvent bacterial-based cloning

related issues but still pose a difficulty for repeat assembly because of PCR biases and short read lengths that result in the collapse of, or assembly gaps in, repetitive regions [28, 29]. However, recent developments in single-molecule real-time (SMRT) sequencing (*e.g.* from Pacific Biosciences; PacBio; [30]) address some of these shortcomings [31-33]. PacBio read lengths are ~16 kb on average but can reach ~50kb, which can bridge repetitive regions that cannot be resolved with short read technology. PacBio reads have a high error rate (~15%) but because these errors appear to be randomly distributed, several approaches can correct the reads for use in *de novo* assembly [33-36]. Hybrid approaches use deep coverage from Illumina reads for error correction of the raw PacBio reads [33]. These methods are not suitable for highly repetitive regions because the Illumina reads cannot be mapped unambiguously to repeats. More promising for the *de novo* assembly of repetitive regions are correction algorithms that use only PacBio reads for self-correction [32, 33]. With sufficiently high read coverage, the longest subset of reads are corrected by overlapping the shorter reads; and the corrected long reads are then used for contig assembly [31]. One popular package for PacBio assembly is the PBcR pipeline included in the Celera assembler. Earlier versions of the assembler (Celera 8.1) used a time-intensive all-by-all alignment step called BLASR to compute overlaps between the uncorrected reads, which accounts for >95% of runtime and is a significant bottleneck for larger genomes [31]. More recent versions (Celera 8.2+) use the newly developed MinHash Alignment Process (MHAP) algorithm to overlap and correct the reads. MHAP is several orders of magnitude faster than BLASR [31].

We aimed to determine if *de novo* long read assembly methods could assemble complex *Drosophila* satDNAs and if so, the optimal assembly methods. We experimented with MHAP

and BLASR-based PacBio *de novo* assembly methods to assemble regions in the pericentric heterochromatin of the *Drosophila melanogaster* genome. We focus on two complex satDNA loci—*Responder (Rsp)* and *260-bp*—and assess assembly quality through molecular and computational validation. *Rsp* is a satDNA that primarily exists as a dimer of two related 120-bp repeats, referred to as *Left* and *Right,* on chromosome *2R* [6, 37-39]. *Rsp* is well-known for being a target of the selfish male meiotic drive system *Segregation Distorter* (reviewed in [40]. *260-bp* is a member of the *1.688* family of satellites located on chromosome *2L* [41]. Using high-coverage (~90X) PacBio data for *Drosophila melanogaster*, we determine the optimal assembly protocols for complex satDNA loci and provide a detailed, base pair-level analysis of the *Rsp* and *260-bp* complex satDNAs.

MATERIALS AND METHODS

The detailed protocols for all molecular methods are available on our website (http://blogs.rochester.edu/larracuente/lab-protocols) and all computational pipelines and intermediate files are available on our lab Github page (https://github.com/LarracuenteLab/Khost_Eickbush_Larracuente2016).

Assemblies

We downloaded raw and error-corrected SMRT PacBio sequence reads from the ISO1 strain [42](raw read SRA accession SRX499318). We downloaded two assemblies constructed using the PBcR pipeline: *1)* "PBcR-BLASR"—an assembly made using Celera 8.1 and a computationally intensive all-by-all alignment with BLASR (Sergey Koren and Adam Phillipy); and *2)* "PBcR-MHAP"—an assembly made using Celera 8.2 and the minhash alignment process (MHAP; [31].

We generated new assemblies using the PBcR pipeline from Celera 8.3 (MHAP) to explore the parameter space that produces the best assembly of repetitive loci (Table 1; S1) . We tested 39 combinations of k-mer size, sketch size, and coverage, as well as with and without the large/diploid genome parameters (http://wgs-assembler.sourceforge.net/wiki/index.php/PBcR#Assembly_of_Corrected_Sequences) that allows a more permissive error rate (Supplementary Files 1 and 2). In addition to the Celera assembler, we tested different parameter combinations in the experimental diploid PacBio assembler Falcon (https://github.com/PacificBiosciences/FALCON). We tested a range of -min_cov lengths, which controls the minimum coverage when overlapping reads in the pre-assembly error correction step, and a range of -min_len sizes, which sets the minimum length of a read to be used in assembly. Overall, we tested 19 different combinations (example spec files are here: https://github.com/LarracuenteLab/Khost_Eickbush_Larracuente2016). All combos produced a highly fragmented *Rsp* locus (Table S2), and thus were excluded from further analysis. We ran all assemblies on a node with a pair of Intel Xeon E5-2695 v2 processors (24 cores) and 124 GB on a Linux supercomputing cluster (Center for Integrated Research Computing at the University of Rochester) using the SLURM job management system (http://slurm.schedmd.com/). Example specification files and SLURM scripts are found at https://github.com/LarracuenteLab/Khost_Eickbush_Larracuente2016. Not all parameter combinations resulted in finished assemblies, as numerous parameter combinations exceeded their allotted memory and failed, and others resulted in impractically long assembly times. For those that did finish, we evaluated assemblies for their ability to generate large contiguous blocks of our complex satDNAs of interest.

To determine the step in the assembly process that leads to the most contiguous assembly of repeats, we assembled reads corrected with the Celera 8.1 pipeline by BLASR (from Adam Phillipy and Sergey Koren; http://bergmanlab.ls.manchester.ac.uk/?p=2151) using the MHAP algorithm implemented in Celera 8.3. We used the Celera 8.3 pipeline to sample the longest 25X subset of the BLASR-corrected reads, which we then converted to an .frg file and assembled using Celera 8.3 ("BLASR-corr Cel8.3").

Assembly evaluation

We used custom repeat libraries that we compiled from Repbase and updated with consensus sequences of *1.688* family and *Responder* (*Rsp*) satellites as BLAST (blast/2.2.29+) queries against all assemblies. We created a custom Perl script to annotate contigs containing repetitive elements based on the BLAST output. The gff files containing our repeat annotations for the PBcR-BLASR assembly are in Supplementary Files 6-7 and all annotation files including custom repeat libraries are found here: https://github.com/LarracuenteLab/Khost_Eickbush_Larracuente2016. For *Rsp,* we categorized repeats as either *Left, Right, Variant,* or *Truncated* based on their length and BLAST score. Our cutoff value to categorize *Rsp* repeats as *Left* or *Right* corresponds to the 90th percentile of the BLAST score distribution in reciprocal BLAST searches. We categorized *Rsp* repeats with a score below this cutoff as *Variant* and partial repeats <90 bp as *Truncated*. We evaluated PacBio assemblies based on the copy number and contiguity of *Rsp* and *260-bp* repeats (Table 1; S1). For both the *Rsp* and *260-bp* loci, we imported our custom gff files into the Geneious genome analysis tool (http://www.geneious.com; [43]) and manually annotated repeats that were still

ambiguous. We also compared these assemblies to the *D. melanogaster* reference genome v6.03 [44].

Assembly validation

We used cytological, computational and molecular approaches to validate the PacBio assemblies. Cytological validation: To confirm the higher-order genomic organization of our target satellites, we used fluorescence *in situ* hybridization (FISH). We designed a Cy5-labeled oligo probe to the *Bari1* repeats distal to the *Rsp* locus (*Bari1:* 5'-/Cy-5/ATGGTTGTTTAAGATAAGAAGGTATCCGTTCTGAT-3') (Fig S1). For *Rsp* and *260-bp,* we created biotin- and digoxigenin-labeled probes using nick translation on gel-extracted PCR products from the *Rsp* and *260-bp* repeats, respectively (260F: 5'-TGGAAATTTAATTACGAGCT-3'; 260R: 5'-ATGAAACTGTGTTCAACAAT-3'; [41]; RspF: 5'-CCGATTTCAAGTACCAGAC-3'; RspR: 5'-GGAAAATCACCCATTTTGACCGC-3'; [6]. We conducted FISH according to [45](Fig 1; Fig S1). Briefly, larval brains were dissected in 1X PBS, treated with a hypotonic solution (0.5% Sodium citrate) and fixed in 1.8% paraformaldehyde; 45% acetic acid and dehydrated in ethanol. Probes were hybridized overnight at 30°C, washed in 4X SSCT and 0.1X SSC, blocked in a BSA solution and treated with 1:100 Rhodamine-avadin (Roche) and 1:100 anti-dig fluorescein (Roche), with final washes in 4X SSCT and 0.1X SSC. Slides were mounted in VectaShield with DAPI (Vector Laboratories), visualized on a Leica DM5500 upright fluorescence microscope at 100X, imaged with a Hamamatsu Orca R2 CCD camera and analyzed using Leica's LAX software.

Computational validation: Because we only use a subset of error-corrected PacBio reads to

create *de novo* assemblies, we assessed the computational support for each assembly using independently derived short Illumina reads, Sanger-sequenced BACs and the entire set of raw PacBio reads. We mapped high-coverage Illumina reads from the ISO1 strain [46] to each assembly using "–very-sensitive" settings in bowtie2 [47] to identify regions of low coverage that could indicate mis-assemblies (Fig S2; S3). Similarly, we mapped raw PacBio reads to each assembly using the PacBio-specific BLASR aligner in the SMRT Analysis 2.3 software package available from Pacific Biosciences (Fig S4; S5). We also mapped available BACs sequences (BACN05C06, BACR32B23, CH221-04O17) that localize to the *Rsp* locus [6].

Molecular validation: The two assemblies that ranked highest in contiguity and representation of *Rsp* and *260-bp* repeats —PBcR-BLASR and BLASR-corr Cel8.3—were well supported with the mapped PacBio reads but differed in the structure of the *Rsp* locus (Fig S6). To distinguish between the alternative structures, we designed long PCR primers that could only amplify a ~15kb segment of the distal part of the locus found in one of the possible configurations (Fig 2A; primer pair 3). We digested the PCR product with *HindIII*, *EagI*, *SstI*, and *XmaI* and performed a Southern blot analysis using a biotinylated *Rsp* probe and the North2South kit (ThermoFisher #17175, Fig S7A). Both of these assemblies also had two clusters of Jockey family elements called *G5*, one on each side of the homogenized *Rsp* repeats. We used informative indels that distinguish *G5* repeats to validate the existence of these two nearly identical *G5* clusters and their orientations. We mapped raw PacBio reads to the locus and identified long reads spanning informative sites. Additionally, we confirmed the presence of two distinct *G5* clusters using PCR with primers designed in and around the deletions (Fig 2A; Table S3). To further validate the assembly of the proximal and distal ends of the *Rsp* locus, we digested genomic DNA with four

restriction enzymes (*AccI, EcoRI, FspI* and *SstI*) and performed a Southern blot analysis (below).

Composition and structure of satellite loci

Using maps of the locus based on our BLAST output, we extracted individual repeat units and created alignments using ClustalW [48]. We inspected and adjusted each alignment by hand in Geneious 8.05 [43]. We examined the relationship between genetic distance and physical distance between repeats. We used the APE phylogenetics package in R [49] to construct neighbor-joining trees for all monomers of each repeat family, using the "indelblock" model of substitution (Fig 3). We then collapsed the repeats down to individual unique variants and plotted their distribution across the locus using a custom Perl script to examine any higher-order structures (Fig 4). All scripts are available at https://github.com/LarracuenteLab/Khost_Eickbush_Larracuente2016.

Southern Blot Analyses

We used ~60 adult females in standard phenol-chloroform extractions, spooled the genomic DNA and resuspended in TE buffer. We performed Southern blot analyses on the 15 kb amplicon and genomic DNA. Approximately 1 μg of the 15 kb amplicon or 10 μg of genomic DNA was digested with each restriction enzyme. The digested DNA was fractionated on a 1% agarose gel, and then depurinated by washing the gel in 0.25 N HCl for 20 mins, denatured in 0.5 M NaOH/ 1.5 M NaCl for 30 mins, and neutralized in 0.5 M Tris (pH 7.5)/ 3 M NaCl for 60 mins before being transferred for 16 hrs in high salt (20 X SSC/ 1 M NH_4Acetate) to a nylon membrane (Genescreen PlusR). The DNA was UV crosslinked to the membrane and hybridizations were conducted overnight at 55°C in North2South hybridization buffer

(ThermoScientific). To make the biotinylated RNA probe, we *in vitro* transcribed a 240-bp *Rsp* gel extracted PCR amplicon (primers: T7_rsp1 5'-TAATACGACTCACTATAGGGGAAAATCACCCATTTTGATCGC-3" and rsp2 5'-CCGAATTCAAGTACCAGAC-3') and labeled using the Biotin RNA Labeling Mix (Roche) and T7 polymerase (Promega). The hybridized membrane was processed as recommended for the Chemiluminescent Nucleic Acid Detection Module (ThermoScientific), and the signal recorded on a ChemiDoc XR+ (BioRad).

Slot Blots

Genomic DNA (100 ng to 600 ng) was denatured (final concentration 0.25 N NaOH, 0.5 M NaCl) for 10 mins at room temperature and then quick cooled by pipetting the denatured sample into loading solution on ice. We performed slot blots as recommended using a 48-well BioDot SF microfiltration apparatus (Bio-Rad). Each blot was first hybridized with an rp49 probe generated by PCR amplification (primers: T7_rp49REV 5'-GTAATACGACTCACTATAGGGCAGTAAACGCGGTTCTGCATG-3 and rp49FOR 5'-CAGCATACAGGCCCAAGATC-3') and a biotinylated RNA produced as described above. The rp49 probe was stripped from the membrane by pouring a 100° C solution of 0.1X SSC/ 0.5% SDS shaking 3 times for ~20 mins. The membrane was then re-hybridized with the *Rsp* probe described above. Hybridization criteria and signal detection for each probe were as described for the genomic Southern analysis. Signals captured on the ChemiDoc were quantitated using the ImageLab software (BioRad).

Nuclei Isolation and Pulse-Field Gel Analysis

Nuclei isolation was performed as described in [50] with some modification. Approximately 100 flies were ground in liquid nitrogen. The powder was suspended in 900 μl of nuclei isolation buffer with 5 mM DTT, filtered first through a 50-μm and then through a 20-μm nitex nylon membrane (03-50/31 and 03-20/14, Sefar America) and pelleted by centrifugation at 3500 rpm for 10 mins. The pelleted nuclei were resuspended in 200 μl of 30 mM Tris, pH 8.0, 100 mM NaCl, 50 mM EDTA, 0.5% Triton X-100. An equal volume of 1% agarose prepared in the same buffer (without Triton X) was added to the nuclei. Using a wide bore pipette tip, 80 μl of the nuclei suspension was placed into the individual wells of a block maker (BioRad). The agarose blocks were placed into 0.5 M EDTA (pH 8.0), 1% sodium lauryl sarcosine, 0.1 mg/ml proteinase K and incubated overnight at 50°C. The plugs were washed for 4 hours in TE at room temperature and then washed overnight in 1 x restriction enzyme buffer. The plugs were digested overnight in 300 μl of fresh buffer, 100 units of enzyme (*EcoRI* and *AccI*), and 100 μg of BSA at 37°C. For pulse field gel electrophoresis, the plugs containing digests from whole fly nuclei were added to the wells of a 1% agarose gel. The gel was run for 21 hours at 8°C, with a voltage of 4.5 V/cm and pulse timing of 0.5-50 seconds. The gel was then subjected to Southern analysis as above using the biotinylated *Rsp* probe.

RESULTS

Rsp and 1.688 FISH

To confirm the gross-scale genomic distribution of *Rsp* and *260-bp* satellites in the sequenced strain (ISO1), we performed multi-color fluorescence *in situ* hybridization (FISH). *Rsp* is located in the pericentric heterochromatin on chromosome *2R* (Fig 1; S1) and *Bari-1* is located just distal to *Rsp* on *2R* (Fig S1), in agreement with previous studies [51] and the PacBio assemblies. The

260-bp satellite is *2L* heterochromatin (Fig 1). The *260-bp* probe cross-hybridizes with other members of the *1.688* family: *353-bp* and *356-bp* on chromosome *3L,* and *359-bp* on chromosome *X* (Fig 1).

Optimal approaches to complex satellite DNA assembly

Our goal was to determine the best method for assembling arrays of complex satellites. We compared *de novo* PacBio assemblies generated using different methods and parameters (both our own and existing assemblies) and evaluated them based on the contiguity of complex satellite sequences. We generated our *de novo* PacBio-only assemblies using the Celera 8.3 and 8.2 PBcR pipelines (referred to as "MHAP") with a range of parameters (Table S1). We generated assemblies with the experimental FALCON diploid assembler that yielded highly fragmented assemblies that we will not discuss further (Table S2). We also generated an assembly using the Celera 8.3 assembler but using error-corrected reads from the computationally-intensive BLASR method (referred to as BLASR-corr Cel8.3). MHAP assemblies that we built using the diploid/large genome parameters were able to recover a 1.3 Mb contig that contains ~230 *260-bp* repeats (~75kb), as were the PBcR-BLASR and the BLASR-corr Cel8.3 assemblies (Table 1; S1). The other contigs containing *260-bp* have < 10 copies, or are short contigs made up of only satellite sequence. We noted that our MHAP assemblies tended to produce these short contigs comprised entirely of *Rsp* or *1.688* family satellites, which were not present in the PBcR-BLASR and the BLASR-corr Cel8.3 assemblies. In contrast to the *260-bp locus,* the *Rsp* locus on *2R* was more variable between the different assembly methods. MHAP assemblies that lacked the diploid/large genome parameters produced a fragmentary locus consisting only of the distal-most repeats (similar to the current release 6

assembly). The PBcR-BLASR and BLASR-corr cel8.3 assemblies each contained a contig with ~1000 *Rsp* repeats, and whose distal end matched the *Rsp* locus in the latest release of the *D. melanogaster* reference genome (Release 6.03, which contains only ~200 copies). This roughly agrees with our estimates of *Rsp* locus size using pulse field gel electrophoresis and Southern blotting (Fig S7B-C). We also estimated the relative abundance of *Rsp* in ISO1 based on three genotypes with previously published estimates of *Rsp* copy number: *cn bw, lt pk cn bw* and *SD* [39]. We believe that while the relative abundances are accurately estimated using hybridization-based approaches like slot blots (based on phenotypic data on the sensitivity to *SD* [39]; and our independent Illumina estimates; data not shown), we believe that these methods underestimate copy number due to variability at the *Rsp* locus (*e.g.* see [37]). *Rsp*-containing BACs mapping to *2R* heterochromatin align with >99% homology to the distal portion of the locus. Several of our MHAP assembly parameter combinations (*e.g*. MHAP 16_1500_20X) also produced a *Rsp* locus with ~1000 repeats, similar to PBcR-BLASR and BLASR-corr cel8.3 (Table 1). However, while the total locus size and number of repeats were roughly consistent between the PBcR-BLASR, BLASR-corr Cel 8.3 and MHAP assembly methods, close examination revealed rearrangements in the central *Rsp* repeats between these assemblies.

Molecular and computational validation of the Rsp locus

To distinguish between the possible configurations of the locus, we mapped high coverage Illumina reads to the assemblies that contained ~1000 *Rsp* copies (PBcR-BLASR, BLASR-corr Cel 8.3 and our example MHAP assembly 16_1500_20X). Each MHAP assembly had dips in coverage across the *Rsp* locus, suggesting that they might be mis-assembled (*e.g.* Fig S2). In contrast, the PBcR-BLASR and BLASR-corr cel8.3 assemblies had uniform coverage across the

contig (*e.g.* Fig S3). For these two assemblies, we mapped raw PacBio reads using BLASR, which also revealed uniform coverage (Figs S4; S5). Aligning the *Rsp* loci from these two assemblies showed that the central segment of the *Rsp* is inverted in one compared to the other (Fig S6). To determine the correct orientation, we designed long PCR primers that should amplify a 15kb product based on the PBcR-BLASR assembly and no product based on the BLASR-corr Cel8.3 assembly (indicated in Fig 2A; primer pair 3). We obtained a 15kb fragment, which we excised and digested with several restriction enzymes; southern analysis of the restriction digest pattern was as predicted from the PBcR-BLASR assembly. In addition, we performed a restriction digest and Southern blot of genomic DNA to look at large segments across the entire *Rsp* locus, which produced a digest pattern that also supported the PBcR-BLASR assembly (Fig S7B-C). Thus, for the *Rsp* locus, the time-intensive BLASR correction step appears to be required for correct assembly of the locus, and we use the PBcR-BLASR assembly for subsequent analysis.

Structure of Rsp and 260-bp loci

While small blocks of *Rsp* are found across the genome, with the largest non-satellite array on chromosome 3L in an intron of *Ago3* [6], we only focus on the main *Rsp* locus in the pericentric heterochromatin on chromosome *2R.* We find that a single 300 kb contig contains most of the main *Rsp* locus. This locus is ~170 kb and contains ~1050 *Rsp* repeats and transposable elements (Fig 2A). The center of the *Rsp* array contains uninterrupted tandem repeats, while the centromere proximal (left) and distal (right) ends are interrupted with transposable element sequences. The presence of the *Bari1* repeats at the distal end of the contig agrees with our FISH analysis (Fig S1) and previous studies [39, 51]. The proximal end of the contig terminates in *Rsp,*

therefore it is likely missing the most centromere-proximal repeats. However, 7 raw uncorrected PacBio reads contain large (up to 6kb) blocks of both tandem *Rsp* repeats and the centromeric AAGAG simple satellite, which suggests that the *Rsp* is centromere-adjacent [52]. The AAGAG+*Rsp* reads were not present in the error-corrected PacBio reads, and due to the high error rate of the uncorrected reads, we could not compare the AAGAG-adjacent *Rsp* repeats to our contig. However, the AAGAG+*Rsp* reads also contain a single *Jockey* element insertion called *G2*, which we used to identify 11 error-corrected reads that link these most centromere-proximal repeats to the rest of the locus (Fig 2A). We created a contig from the 11 error-corrected reads (Fig S8) that, when combined with the AAGAG+*Rsp* raw reads, suggests that our 300kb contig is missing ~22kb of sequence containing ~200 *Rsp* repeats. These *Rsp* elements are most similar to the proximal-most repeats in our 300 kb contig (Fig 3A), suggesting that they indeed correspond to the centromere-proximal repeats.

Satellites tend to undergo concerted evolution—unequal exchange and gene conversion homogenize repeat sequences within arrays [4, 53, 54]. To test the hypothesis that *Rsp* undergoes concerted evolution, we examined the relationship between genetic and physical distance within the *2R* array. We built neighbor-joining trees for each satellite family using each full-length repeat monomer (Fig 3). We find a pattern consistent with concerted evolution: two large clades of nearly identical repeats corresponding to the *Right* and *Left Rsp* repeats consist mainly of repeats from the center of the array. In contrast, the *Variant* repeats have longer branch lengths and tend to occur toward the proximal and distal ends of the array (Fig 3A). To examine the higher-order structure of the array, we studied the distribution of all unique repeats sequences across the locus according to their abundance (Fig 4A). The ~1050 *Rsp* repeats on the main

contig correspond to ~370 unique variants. Consistent with our phylogenetic analysis, low copy number *Rsp* repeats tend to dominate the ends of the array, while higher copy number variants dominate the center of the array (Fig 3A).

There are several TE insertions within the *Rsp* array located towards the proximal and distal ends of the locus. The homogenized *Rsp* repeats in the center of the array are flanked by two nearly identical clusters of *G5 Jockey* elements (Fig 4A, boxed). These *G5* repeats form their own clade with respect to the other *G5* insertions in the genome and have a high degree of similarity to one another (Fig S9). They have a complicated orientation, with each repeat having a match on the opposite side of the locus ~100 kb away, but in an inverted orientation and near 99% homology (Fig 2A). Despite the similarity between the two clusters, there are several unique configurations of indels in each that allow us to distinguish them. We examined the pileup of raw PacBio reads over sets of long indels found in the *G5* clusters, and identified 8 and 20 individual long reads that spanned the unique configuration of indels in the proximal *G5* cluster (*G5-5* and *G5-6*, Fig 2A) and distal *G5* cluster (*G5-3* and *G5-2*, Fig 2A), respectively. This suggests that the proximal cluster actually exists and is not an erroneous duplication of the distal cluster in the assembly. For further confirmation, we designed PCR primers complimentary to the unique indels in the proximal cluster (Fig 2A), which return products of the expected size (data not shown). The *Rsp* elements surrounding the *G5* elements also show a mirrored structure (Fig 4A). Repeats in between two *G5* elements are >99% identical to the repeats between the partnered *G5*s on the other side of the array (Fig 2A). Interestingly, one 1.7kb stretch of inter-*G5 Rsp* repeats is repeated three times, which suggests a complex series of duplication and inversion within the *G5* cluster. The *Rsp* repeats are oriented on the same strand across most of the array, but they flip to

the opposite strand at the fragmentary *G5* element, mirroring what we see with *G5* elements (Fig 2A). Thus the inversion did not occur only in the local area around the *G5*s, but across the entire proximal end of the contig.

The *260-bp* locus on chromosome *2L* is fully contained within a 1.2 Mb contig and contains 230 repeats interrupted by identical *Copia* transposable elements (Fig 1B). Unlike *Rsp,* the *260-bp* satellite array lacks the homogenized center and has more variant sequences (Fig 3B). The *260-bp* satellite has more unique variants than *Rsp*: the 230 monomers correspond to 153 unique variants, and there are fewer high copy number variants (Fig 4B).

DISCUSSION

Assembly methods for complex satellites

For large complex centromeric repeats, such as human centromeres, the complete assembly of a contiguous stretch of repeats has not been possible with current technologies [55]. Instead, human centromere composition can be inferred using clever graph-based modeling strategies [56]. In contrast, single molecule sequencing produced assemblies of more tractable, but still challenging highly repetitive genomic regions [34, 57, 58], including some plant centromeres [59, 60]. However, validation of these assemblies is difficult. Here, we create accurate *de novo* assemblies of two complex satDNAs in *Drosophila* using single molecule PacBio sequencing reads, allowing us to examine the detailed spatial distribution of elements within these arrays for the first time. We found that assemblers differed in their ability to produce a complete assembly for the two satellites: while the *260-bp* locus assembly was consistent between all PacBio methods, the larger *Rsp* locus required the time-intensive BLASR correction algorithm for an

accurate assembly. We validated the major features of this PBcR-BLASR *Rsp* assembly through extensive molecular and computational validation and, with some manual scaffolding, were able to extend the assembly to what may be the junction between *Rsp* and the chromosome *2* centromeric satDNA. There are four features of the *Rsp* locus that may present a particular challenge for *de novo* assembly, especially for MHAP- and FALCON-based methods: *1)* it is large (more than twice the size of the *260-bp* locus); *2)* it appears to be centromere-adjacent [38], with AAGAG repeats directly proximal to the *Rsp* cluster; *3)* the array center is occupied by a contiguous stretch of nearly identical repeat variants; and *4)* these repeats are flanked by nearly identical TEs in a complex inverted orientation. In addition to struggling with the major satDNA locus, we found that even our most contiguous MHAP assemblies produced short contigs consisting entirely of what we believe are extraneous repeats. Despite these caveats, we recover the gross-scale organization of the locus with our best MHAP parameter combinations, indicating that the faster MHAP approach may offer a starting point for determining the structure of difficult repetitive loci. Therefore, there is a trade-off in speed *vs* accuracy in the correction and assembly of PacBio reads—while MHAP correction is sufficient for smaller, less homogeneous complex satDNA loci, BLASR correction is required for base pair-level resolution of larger loci. Both methods produce larger, more contiguous assemblies of these complex satDNAs than the latest reference genome (release 6 assembly [44]), which offered an impressive improvement in the assembly of pericentric regions over previous releases. All satDNA assemblies require careful, independent validation. We found low coverage junctions between *Rsp* and the adjacent simple AAGAG repeats that occupy the centromere of chromosome *2*. We also find a general reduced representation of simple satellite-rich raw reads, making it difficult to extend our assembly into the centromere. This apparent bias against raw

reads derived from simple repeats has two potential explanations: *1)* PacBio sequencing is subject to a bias that is difficult to measure because it occurs in the most highly repetitive regions of the genome; and/or *2)* the inherent structural properties of some highly repetitive DNAs subject these sequences to misrepresentation in library preparation (*e.g.* non-random chromosome breakage during DNA isolation or library preparation). Therefore, the assembly of some simple tandem repeats still pose a significant challenge for PacBio-based assembly methods.

Structure of complex satDNA loci

Consistent with gross-scale structural analyses of satellite DNA [22, 25-27, 52, 61], we find that *Rsp* and *260-bp* have uninterrupted blocks of homogeneous repeats alternating with "islands" of complex DNA. For both of these complex satDNAs, TE insertions cluster together towards the array ends. The TEs in and around the locus tend to be full-length and similar to euchromatic copies, suggesting recent insertion. What gives rise to this structure? Repetitive tandem arrays are thought to expand and contract via unequal crossing over [62], which along with gene conversion will homogenize the array and lead to a pattern of concerted evolution [4, 53, 54]. The localization of the TEs in islands near the proximal and distal ends of the locus is predicted by the "accretion model", which predicts that repeated unequal exchange over the array should push TEs together and towards the ends of an array [63]. The organization of the sequence variants across the locus and the degree of homogeneity differs between *Rsp* and *260-bp*. The center of the *Rsp* locus is highly homogeneous and dominated by a few high-copy number variants, while the *260-bp* locus is comprised mostly of low-copy number or unique repeats. These differences may simply be because of a difference in size of the two satDNAs, or that

unequal exchange and gene conversion occurred more recently at the *Rsp* locus. As exchange breakpoints are more likely to occur within an element than perfectly at the junction between two repeats, the lack of truncated repeats within the array center suggests that any unequal exchange event would involve a large chunk of the array. One interesting structural feature of the *Rsp* locus is the cluster of *G5* elements on the proximal and distal sides of the array. The clusters are in an inverted orientation and nearly identical. The clusters are not perfectly mirrored, however—one *G5* was duplicated three times and one is fragmented. This indicates that there was a complicated scenario likely involving duplication and an inversion that gave rise to these two clusters. The high degree of similarity between the clusters could be explained by gene conversion, though the clusters are ~100kb distant from one another. Alternatively, the locus could have expanded very recently, subsequent to the duplication and inversion, and differences have not yet had time to accumulate. We are testing these hypotheses by looking at polymorphism in the structure of these loci in natural populations using next generation sequencing technology.

D*e novo* PacBio assembly methods allows for exciting progress in studying the structure of previously inaccessible regions of the genome in unprecedented detail. We show here that some complex satDNA loci are tractable models for determining tandem repeat organization in pericentric heterochromatin. These assemblies provide a platform for evolutionary and functional genomic studies of satDNA.

ACKNOWLEDGMENTS

We would like to thank Casey Bergman for helpful conversations about PacBio assembly methods and for sharing assemblies, reads and protocols. We would like the thank the staff of the

Center for Integrated Research Computing at the University of Rochester for maintenance of the computing cluster and access to computational resources. This work was supported by the University of Rochester.

REFERENCES

1. Sueoka N. Variation and Heterogeneity of Base Composition of Deoxyribonucleic Acids - a Compilation of Old and New Data. Journal of Molecular Biology. 1961;3(1):31-&. PubMed PMID: WOS:A19615790B00013.

2. Kit S. Equilibrium Sedimentation in Density Gradients of DNA Preparations from Animal Tissues. Journal of Molecular Biology. 1961;3(6):711-&. PubMed PMID: WOS:A19615795B00007.

3. Szybalski W. Use of cesium sulfate for equilibrium density gradient centrifugation. Methods Enzymol. 1968;12B:330-60.

4. Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 1994;371(6494):215-20. Epub 1994/09/15. doi: 10.1038/371215a0. PubMed PMID: 8078581.

5. Britten RJ, Kohne DE. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science. 1968;161(3841):529-40. Epub 1968/08/09. PubMed PMID: 4874239.

6. Larracuente AM. The organization and evolution of the *Responder* satellite in species of the *Drosophila melanogaster* group: dynamic evolution of a target of meiotic drive. BMC Evol Biol. 2014;14(1):233. doi: 10.1186/s12862-014-0233-9. PubMed PMID: 25424548; PubMed Central PMCID: PMC4280042.

7. Lohe AR, Brutlag DL. Identical satellite DNA sequences in sibling species of Drosophila. J Mol Biol. 1987;194(2):161-70. PubMed PMID: 3112413.

8. Yunis JJ, Yasmineh WG. Heterochromatin, satellite DNA, and cell function. Structural DNA of eucaryotes may support and protect genes and aid in speciation. Science. 1971;174(4015):1200-9. Epub 1971/12/17. PubMed PMID: 4943851.

9. Ferree PM, Barbash DA. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. PLoS biology. 2009;7(10):e1000234. Epub 2009/10/28. doi: 10.1371/journal.pbio.1000234. PubMed PMID: 19859525; PubMed Central PMCID: PMC2760206.

10. Henikoff S, Ahmad K, Malik HS. The centromere paradox: stable inheritance with rapidly evolving DNA. Science. 2001;293(5532):1098-102. Epub 2001/08/11. doi: 10.1126/science.1062939. PubMed PMID: 11498581.

11. Fishman L, Saunders A. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. Science. 2008;322(5907):1559-62. doi: 10.1126/science.1161406. PubMed PMID: 19056989.

12. Walker PM. Origin of satellite DNA. Nature. 1971;229(5283):306-8. Epub 1971/01/29. PubMed PMID: 4925781.

13. Ohno S. So much "junk" DNA in our genome. Brookhaven symposia in biology. 1972;23:366-70. Epub 1972/01/01. PubMed PMID: 5065367.

14. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Nature. 1980;284(5757):601-3. Epub 1980/04/17. PubMed PMID: 6245369.

15. Orgel LE, Crick FH. Selfish DNA: the ultimate parasite. Nature. 1980;284(5757):604-7. Epub 1980/04/17. PubMed PMID: 7366731.

16. Orgel LE, Crick FH, Sapienza C. Selfish DNA. Nature. 1980;288(5792):645-6. Epub 1980/12/25. PubMed PMID: 7453798.

17. Zhu Q, Pao GM, Huynh AM, Suh H, Tonnu N, Nederlof PM, et al. BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. Nature. 2011;477(7363):179-84. Epub 2011/09/09. doi: 10.1038/nature10371. PubMed PMID: 21901007; PubMed Central PMCID: PMC3240576.

18. Hughes SE, Gilliland WD, Cotitta JL, Takeo S, Collins KA, Hawley RS. Heterochromatic threads connect oscillating chromosomes during prometaphase I in Drosophila oocytes. PLoS Genet. 2009;5(1):e1000348. PubMed PMID: 19165317.

19. He B, Caudy A, Parsons L, Rosebrock A, Pane A, Raj S, et al. Mapping the pericentric heterochromatin by comparative genomic hybridization analysis and chromosome deletions in *Drosophila melanogaster*. Genome research. 2012;22(12):2507-19. Epub 2012/06/30. doi: 10.1101/gr.137406.112. PubMed PMID: 22745230; PubMed Central PMCID: PMC3514680.

20. Dernburg AF, Sedat JW, Hawley RS. Direct evidence of a role for heterochromatin in meiotic chromosome segregation. Cell. 1996;86(1):135-46. Epub 1996/07/12. PubMed PMID: 8689681.

21. Csink AK, Henikoff S. Something from nothing: the evolution and utility of satellite repeats. Trends in genetics : TIG. 1998;14(5):200-4. Epub 1998/06/05. PubMed PMID: 9613205.

22. Sun X, Wahlstrom J, Karpen G. Molecular structure of a functional Drosophila centromere. Cell. 1997;91(7):1007-19. Epub 1998/01/15. PubMed PMID: 9428523; PubMed Central PMCID: PMC3209480.

23. Peacock WJ, Brutlag D, Goldring E, Appels R, Hinton CW, Lindsley DL. The organization of highly repeated DNA sequences in *Drosophila melanogaster* chromosomes. Cold Spring Harb Symp Quant Biol. 1974;38:405-16. PubMed PMID: 4133985.

24. Lohe AR, Brutlag DL. Multiplicity of satellite DNA sequences in Drosophila melanogaster. Proc Natl Acad Sci U S A. 1986;83(3):696-700. PubMed PMID: 3080746; PubMed Central PMCID: PMCPMC322931.

25. Le MH, Duricka D, Karpen GH. Islands of complex DNA are widespread in Drosophila centric heterochromatin. Genetics. 1995;141(1):283-303. PubMed PMID: 8536977; PubMed Central PMCID: PMCPMC1206727.

26. Brutlag D, Carlson M, Fry K, Hsieh TS. DNA-Sequence Organization in Drosophila Heterochromatin. Cold Spring Harb Sym. 1977;42:1137-46. PubMed PMID: WOS:A1977FK40800052.

27. Lohe AR, Brutlag DL. Adjacent satellite DNA segments in Drosophila structure of junctions. J Mol Biol. 1987;194(2):171-9. PubMed PMID: 3112414.

28. Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A, Kaminker JS, et al. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome biology. 2002;3(12):RESEARCH0085. Epub 2003/01/23. PubMed PMID: 12537574; PubMed Central PMCID: PMC151187.

29. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res. 2010;20(9):1165-73. doi: 10.1101/gr.101360.109. PubMed PMID: 20508146; PubMed Central PMCID: PMCPMC2928494.

30. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133-8. doi: 10.1126/science.1162986. PubMed PMID: 19023044.

31. Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33(6):623-30. doi: 10.1038/nbt.3238. PubMed PMID: 26006009.

32. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563-9. doi: 10.1038/nmeth.2474. PubMed PMID: 23644548.

33. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012;30(7):693-700. doi: 10.1038/nbt.2280. PubMed PMID: 22750884; PubMed Central PMCID: PMC3707490.

34. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2014. doi: 10.1038/nature13907. PubMed PMID: 25383537.

35. Lam KK, Khalak A, Tse D. Near-optimal assembly for shotgun sequencing with noisy reads. BMC Bioinformatics. 2014;15 Suppl 9:S4. doi: 10.1186/1471-2105-15-S9-S4. PubMed PMID: 25252708; PubMed Central PMCID: PMCPMC4168708.

36. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):R51. doi: 10.1186/gb-2013-14-5-r51. PubMed PMID: 23718773; PubMed Central PMCID: PMCPMC4053816.

37. Houtchens K, Lyttle TW. *Responder* (*Rsp*) alleles in the *segregation distorter* (*SD*) system of meiotic drive in Drosophila may represent a complex family of satellite repeat sequences. Genetica. 2003;117(2-3):291-302. Epub 2003/05/02. PubMed PMID: 12723708.

38. Pimpinelli S, Dimitri P. Cytogenetic analysis of segregation distortion in *Drosophila melanogaster*: the cytological organization of the Responder (Rsp) locus. Genetics. 1989;121(4):765-72. Epub 1989/04/01. PubMed PMID: 2470640; PubMed Central PMCID: PMC1203659.

39. Wu CI, Lyttle TW, Wu ML, Lin GF. Association between a satellite DNA sequence and the *Responder* of *Segregation Distorter* in *D. melanogaster*. Cell. 1988;54(2):179-89. Epub 1988/07/15. PubMed PMID: 2839299.

40. Larracuente AM, Presgraves DC. The Selfish *Segregation Distorter* Gene Complex of *Drosophila melanogaster*. Genetics. 2012;192(1):33-53. Epub 2012/09/12. doi: 10.1534/genetics.112.141390. PubMed PMID: 22964836; PubMed Central PMCID: PMC3430544.

41. Abad JP, Agudo M, Molina I, Losada A, Ripoll P, Villasante A. Pericentromeric regions containing 1.688 satellite DNA sequences show anti-kinetochore antibody staining in prometaphase chromosomes of *Drosophila melanogaster*. Mol Gen Genet. 2000;264(4):371-7. PubMed PMID: 11129040.

42. Kim K, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, et al. Long-read, whole-genome shotgun sequence data for five model organisms. Scientific data. 2014;1(140045). Epub 11/25/2014. doi: doi:10.1038/sdata.2014.45.

43. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647-9. Epub 2012/05/01. doi: 10.1093/bioinformatics/bts199. PubMed PMID: 22543367; PubMed Central PMCID: PMC3371832.

44. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6 reference sequence of the *Drosophila melanogaster* genome. Genome Res. 2015. doi: 10.1101/gr.185579.114. PubMed PMID: 25589440.

45. Larracuente AM, Ferree PM. Simple method for fluorescence DNA in situ hybridization to squashed chromosomes. JoVE. 2015;95:e52288. doi: doi:10.3791/52288.

46. Gutzwiller F, Carmo CR, Miller DE, Rice DW, Newton IL, Hawley RS, et al. Dynamics of Wolbachia pipientis Gene Expression Across the *Drosophila melanogaster* Life Cycle. G3 (Bethesda). 2015;5(12):2843-56. doi: 10.1534/g3.115.021931. PubMed PMID: 26497146; PubMed Central PMCID: PMCPMC4683655.

47. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357-9. Epub 2012/03/06. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PubMed Central PMCID: PMC3322381.

48. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23(21):2947-8. doi: 10.1093/bioinformatics/btm404. PubMed PMID: 17846036.

49. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289-90. Epub 2004/01/22. PubMed PMID: 14734327.

50. Kuhn GC, Sene FM, Moreira-Filho O, Schwarzacher T, Heslop-Harrison JS. Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. Chromosome Res. 2008;16(2):307-24. doi: 10.1007/s10577-007-1195-1. PubMed PMID: 18266060.

51. Caizzi R, Caggese C, Pimpinelli S. Bari-1, a new transposon-like family in *Drosophila melanogaster* with a unique heterochromatic organization. Genetics. 1993;133(2):335-45. Epub 1993/02/01. PubMed PMID: 8382176; PubMed Central PMCID: PMC1205323.

52. Lohe AR, Hilliker AJ, Roberts PA. Mapping simple repeated DNA sequences in heterochromatin of Drosophila melanogaster. Genetics. 1993;134(4):1149-74. PubMed PMID: 8375654; PubMed Central PMCID: PMCPMC1205583.

53. Dover G. A Molecular Drive through Evolution. Bioscience. 1982;32(6):526-33. doi: Doi 10.2307/1308904. PubMed PMID: ISI:A1982NX09100015.

54. Dover G. Concerted evolution, molecular drive and natural selection. Current biology : CB. 1994;4(12):1165-6. Epub 1994/12/01. PubMed PMID: 7704590.

55. Miga KH. Completing the human genome: the progress and challenge of satellite DNA assembly. Chromosome Res. 2015;23(3):421-6. doi: 10.1007/s10577-015-9488-2. PubMed PMID: 26363799.

56. Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. Centromere reference models for human chromosomes X and Y satellite arrays. Genome Res. 2014;24(4):697-707. doi: 10.1101/gr.159624.113. PubMed PMID: 24501022; PubMed Central PMCID: PMCPMC3975068.

57. Carvalho AB, Vicoso B, Russo CA, Swenor B, Clark AG. Birth of a new gene on the Y chromosome of *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 2015;112(40):12450-5. doi: 10.1073/pnas.1516543112. PubMed PMID: 26385968; PubMed Central PMCID: PMCPMC4603513.

58. Krsticevic FJ, Schrago CG, Carvalho AB. Long-Read Single Molecule Sequencing to Resolve Tandem Gene Copies: The Mst77Y Region on the *Drosophila melanogaster* Y Chromosome. G3 (Bethesda). 2015;5(6):1145-50. doi: 10.1534/g3.115.017277. PubMed PMID: 25858959; PubMed Central PMCID: PMCPMC4478544.

59. VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. Nature. 2015;527(7579):508-11. doi: 10.1038/nature15714. PubMed PMID: 26560029.

60. Wolfgruber TK, Nakashima MM, Schneider KL, Sharma A, Xie Z, Albert PS, et al. High Quality Maize Centromere 10 Sequence Reveals Evidence of Frequent Recombination Events.

Front Plant Sci. 2016;7:308. doi: 10.3389/fpls.2016.00308. PubMed PMID: 27047500; PubMed Central PMCID: PMCPMC4806543.

61. Carlson M, Brutlag D. Cloning and characterization of a complex satellite DNA from *Drosophila melanogaster*. Cell. 1977;11(2):371-81. PubMed PMID: 408008.

62. Smith GP. Evolution of repeated DNA sequences by unequal crossover. Science. 1976;191(4227):528-35. Epub 1976/02/13. PubMed PMID: 1251186.

63. McAllister BF, Werren JH. Evolution of tandemly repeated sequences: What happens at the end of an array? Journal of molecular evolution. 1999;48(4):469-81. Epub 1999/03/18. PubMed PMID: 10079285.

TABLES

Table 1: Summary of *Rsp* and *260-bp* repeats counts for a subset of assemblies. Counts are for all assembled repeats in any genomic contig.

Assembly name	***# Rsp***	***#Rsp contigs***	***Rsp* score**	***# 260-bp***	***# 260-bp* contigs**	***260-bp* score**
R6.03	343	9	38.1	206	57	3.6
PBcR BLASR	**1088**	**3**	**362.7**	**284**	**13**	**21.8**
BLASR-corr Cel8.3	923	3	307.7	505	46	11.0
MHAP 8.2	251	4	62.8	172	16	10.8
MHAP_16_1500_20X	1260	4	315.0	374	37	10.1

R6.03: The latest reference *D. melanogaster* genome; PBcR BLASR: Assembly from Adam Phillipy and Sergey Koren. This produced the best assembly of both *Rsp* and *260-bp* loci; BLASR-corr Cel8.3: Assembly of BLASR-corrected reads with MHAP in Celera 8.3. MHAP 8.2: Assembly with default parameters (k = 16; sketch = 512; coverage = 25) from [31]; MHAP_16_1500_20X: Our best MHAP assembly with parameters (k = 16; sketch = 1500; coverage = 25). All other MHAP and Falcon assembly statistics and parameters are in Tables S1 and S2.

FIGURE LEGENDS

Figure 1: FISH image of *D. melanogaster* mitotic chromosomes showing *Rsp* and *260-bp* satDNAs. DNA is stained with DAPI (blue), *Rsp* is indicated by an avidin-rhodamine probe (red; arrowhead) and *260-bp* by an anti-digoxigenin probe (green; arrow). The *260-bp* probe also targets other members of the *1.688* satellite family on the X and 3L chromosomes.

Figure 2: Maps of complex satDNAs contigs. Counts for each repetitive element family in our custom Repbase library were plotted in 3kb windows across each contig. **A)** *Rsp* locus on chromsoome *2R*. Blue bars correspond to *Rsp Left, Right* or *Variant/Truncated* repeats, while other colors correspond to various TE families. *Rsp* spans ~170kb of the 300kb contig (thick blue line below x-axis). Above the plot is a schematic showing the orientation of two *G5* clusters

flanking the *Rsp* locus and a separate contig containing *Rsp* and the Jockey element *G2*, which is directly adjacent to centromeric AAGAG repeats. The color of the chevrons indicates which *G5* elements have the highest degree of similarity with one another. Solid and dashed lines within the insertions show the approximate locations of shared insertions or deletions (respectively). Several configurations of indels are unique, such as the two in *G5_5* or the deletion in *G5_1*, which allows verification of the cluster. The *G2* contig links the *Rsp* locus to what appears to be the chromosome *2* centromere (Black circles). **B**) *260-bp* locus on chromosome *2L*. Only the area surrounding the *260-bp* array is shown (300 kb of ~1.1 Mb contig). The *260-bp* locus spans ~70 kb of the 1.1 Mb contig (red line below x-axis) and is interrupted with *Copia* elements.

Figure 3: Neighbor-joining tree of complex satDNA monomers. **A.)** *Rsp* repeats in the chromosome *2R* locus. Repeats were divided into bins where each bin contains 1/6th of the locus, or ~180 repeats/bin. Tip color corresponds to position in the array (red is most centromere proximal, blue is most centromere distal). The tip symbol indicates if the repeat is *Left* (square), *Right* (triangle), or *Variant/truncated* (circle). Repeats corresponding to the *G2* contig containing the most centromere-proximal repeats are indicated in pink. Note that these repeats cluster with the repeats on the proximal end of the *Rsp* contig (red), supporting their location adjacent to the centromere. **B.)** *260-bp* repeats in the chromosome *2L* locus. Repeats were divided into bins where each bin contains 1/4th of the locus, or ~57 repeats/bin. Tip color corresponds to position in the array (red is most centromere proximal and green is most centromere distal).

Figure 4: Distribution of satDNA sequence variants across loci. Each row corresponds to a unique monomer, while the x-axis shows the position of that monomer sequence in the array.

The color of the point indicates the copy number of each monomer in the array. **A)** The *Rsp* locus on *2R*. Several high copy number *Rsp* variants dominate the center of the array (purple and blue), with the low frequency and unique sequences found more towards the proximal and distal ends (gray and green). One cluster of repeats is duplicated on either side of the array (boxed). **B.)** The *260-bp* locus on *2L*. The majority of repeats occur only once, while a few variants have intermediate copy number.

Figure 1: FISH image of *D. melanogaster* mitotic chromosomes showing *Rsp* and *260-bp* satDNAs.

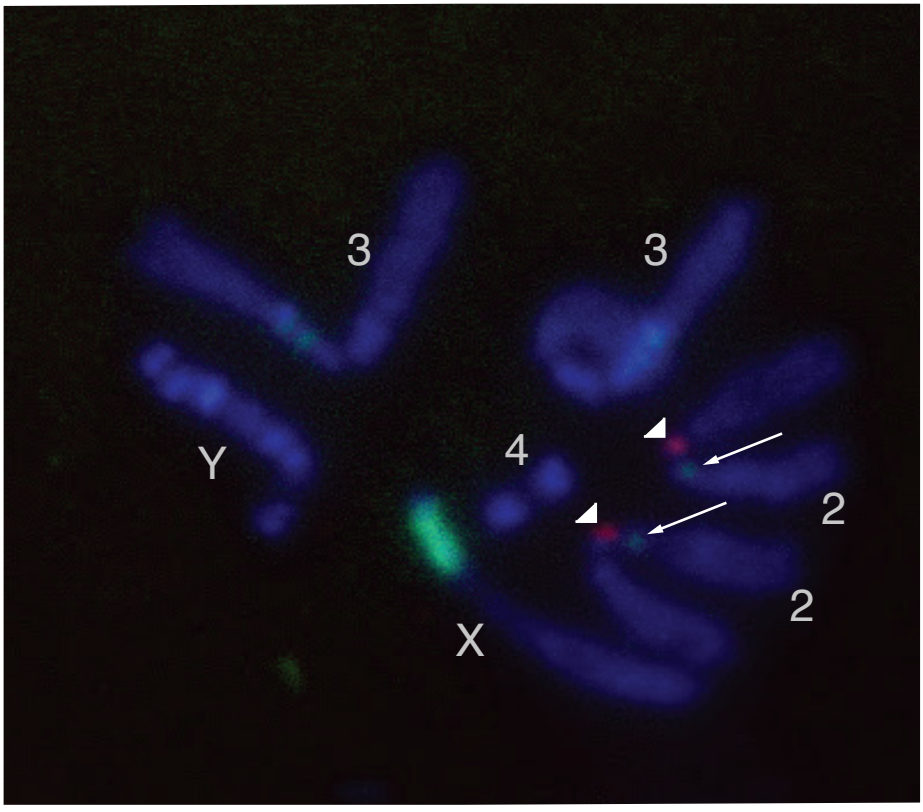

Fig 1. DNA is stained with DAPI (blue), *Rsp* is indicated by an avidin-rhodamine probe (red; arrowhead) and *260-bp* by an anti-digoxigenin probe (green; arrow). The *260-bp* probe also targets other members of the *1.688* satellite family on the X and 3L chromosomes.

Figure 2: Maps of complex satDNAs contigs.

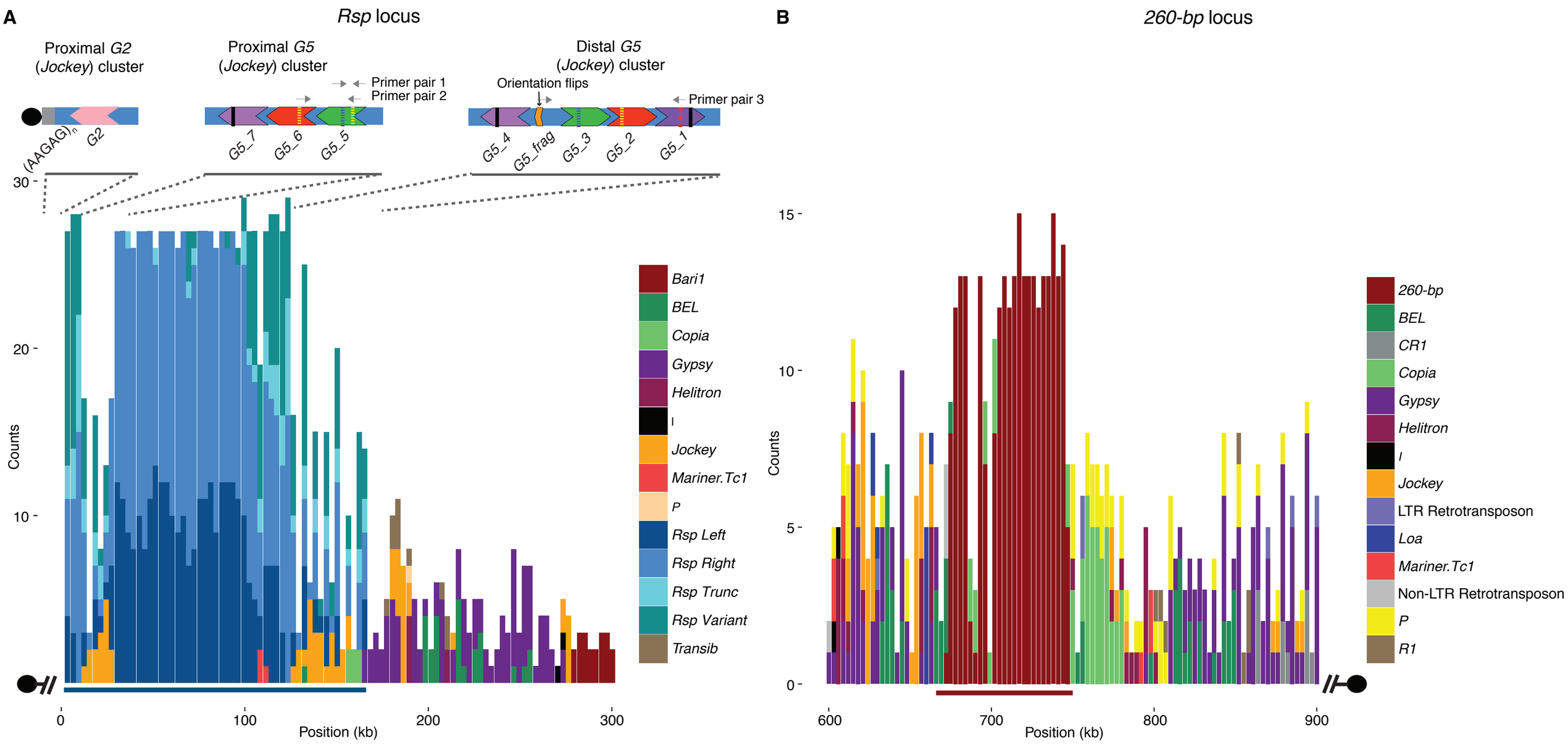

Fig 2. Counts for each repetitive element family in our custom Repbase library were plotted in 3kb windows across each contig. A) *Rsp* locus on chromosome *2R*. Blue bars correspond to *Rsp Left, Right* or *Variant/Truncated* repeats, while other colors correspond to various TE families. *Rsp* spans ~170kb of the 300kb contig (thick blue line below x-axis). Above the plot is a schematic showing the orientation of two *G5* clusters flanking the *Rsp* locus and a separate contig containing *Rsp* and the Jockey element *G2*, which is directly adjacent to centromeric AAGAG repeats. The color of the chevrons indicates which *G5* elements have the highest degree of similarity with one another. Solid and dashed lines within the insertions show the approximate locations of shared insertions or deletions (respectively). Several configurations of indels are unique, such as the two in *G5_5* or the deletion in *G5_1*, which allows verification of the cluster. The *G2* contig links the *Rsp* locus to what appears to be the chromosome 2 centromere (Black circles). B) *260-bp* locus on chromosome *2L*. Only the area surrounding the *260-bp* array is shown (300 kb of ~1.1 Mb contig). The *260-bp* locus spans ~70 kb of the 1.1 Mb contig (red line below x-axis) and is interrupted with *Copia* elements.

Figure 3: Neighbor-joining tree of complex satDNA monomers.

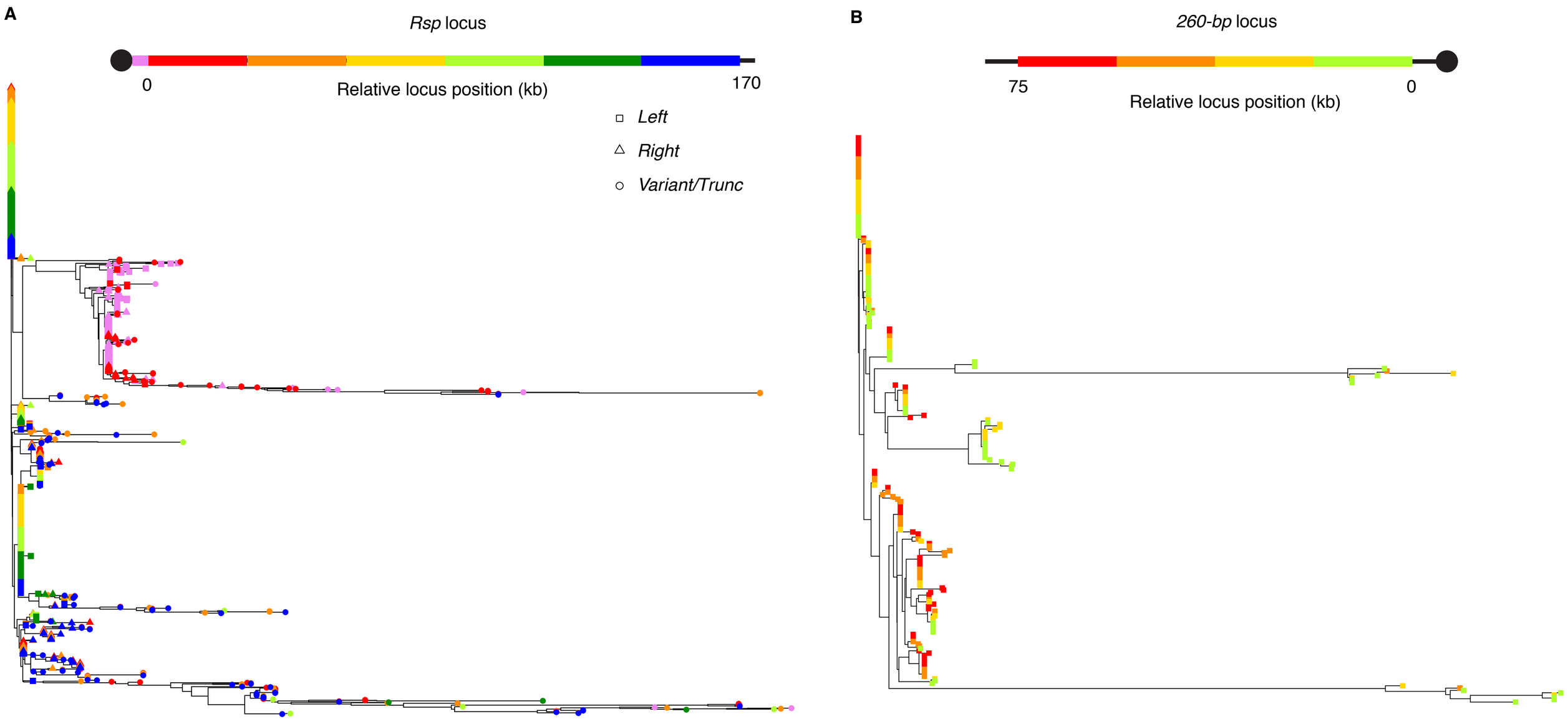

Fig 3. A.) *Rsp* repeats in the chromosome *2R* locus. Repeats were divided into bins where each bin contains 1/6th of the locus, or ~180 repeats/bin. Tip color corresponds to position in the array (red is most centromere proximal, blue is most centromere distal). The tip symbol indicates if the repeat is *Left* (square), *Right* (triangle), or *Variant/truncated* (circle). Repeats corresponding to the *G2* contig containing the most centromere-proximal repeats are indicated in pink. Note that these repeats cluster with the repeats on the proximal end of the *Rsp* contig (red), supporting their location adjacent to the centromere. B.) *260-bp* repeats in the chromosome *2L* locus. Repeats were divided into bins where each bin contains 1/4th of the locus, or ~57 repeats/bin. Tip color corresponds to position in the array (red is most centromere proximal and green is most centromere distal).

Figure 4: Distribution of satDNA sequence variants across loci.

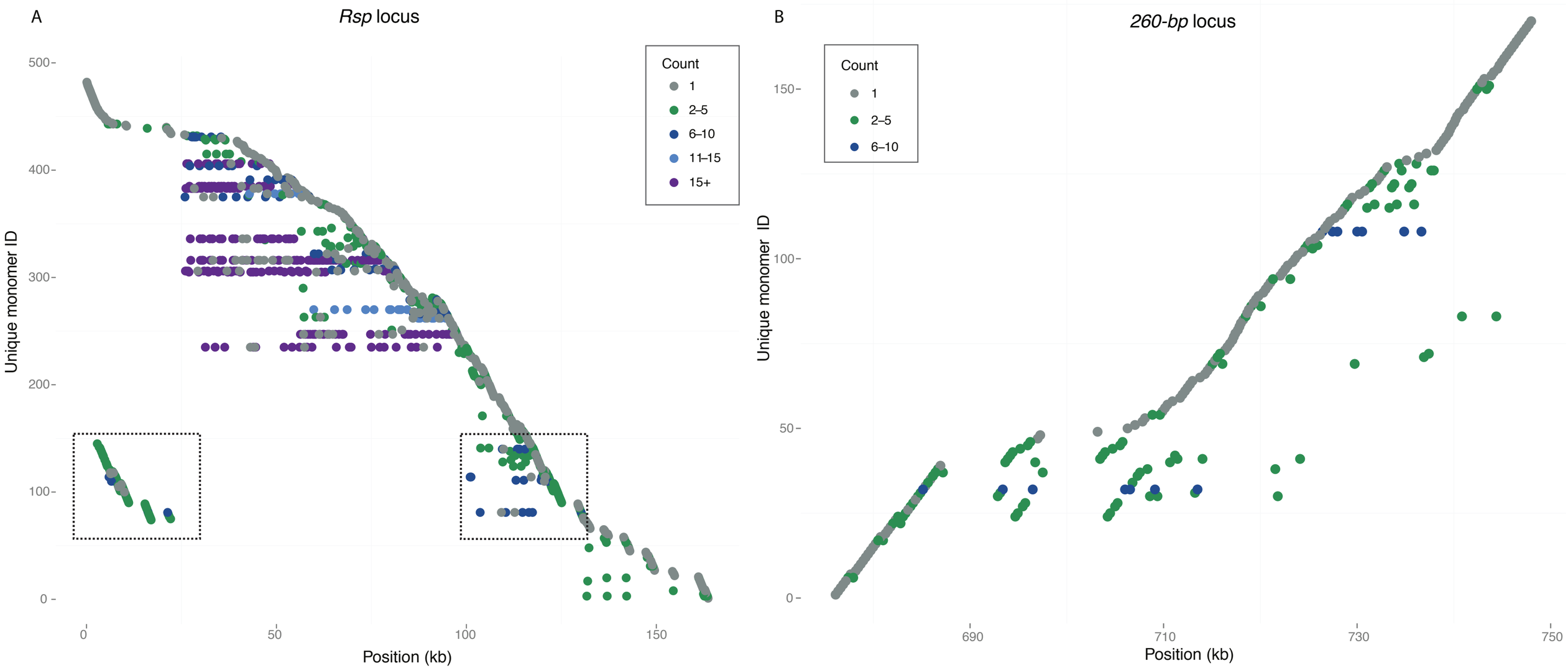

Fig 4. Each row corresponds to a unique monomer, while the x-axis shows the position of that monomer sequence in the array. The color of the point indicates the copy number of each monomer in the array. A) The Rsp locus on chromosome *2R*. Several high copy number *Rsp* variants dominate the center of the array (purple and blue), with the low frequency and unique sequences found more towards the proximal and distal ends (gray and green). One cluster of repeats is duplicated on either side of the array (boxed). B.) The *260-bp* locus on 2L. The majority of repeats occur only once, while a few variants have intermediate copy number.