Use of Next-Generation Sequencing for the Identification of Y-chromosome Derived Genomic Sequence Scaffolds from the Red Flour Beetle, *Tribolium castaneum*

Nathaniel Grubbs¹, Fu-Chyun Chu¹, Mark Robinson², Daniel Bopp², and Marcé D. Lorenzen¹

¹Department of Entomology, North Carolina State University, Raleigh, NC 27695
²Institute of Molecular Life Sciences, University of Zürich, Zürich, Switzerland CH-8057

ABSTRACT

We previously reported the use of microarray data (male-specific expression) and bioinformatic analyses (male-associated repeat content) to identify Y-chromosome sequence from the red flour beetle, *Tribolium castaneum*. In combination, these approaches allowed us to identify approximately 750 Kb of Y-chromosome sequence. Since the size of the T. castaneum genome is estimated to be approximately 5 Mb in length, we sought new methods for uncovering additional Y-chromosome sequence. Here we report the use of two next generation sequencing methods, Illumina and PacBio, for identification of Y-chromosome derived genomic sequence scaffolds. The first method, Illumina, provided millions of short sequence reads independently from T. castaneum males and females. The sex-specific sequence reads were independently aligned to the published (mixed sex) genome assembly and differences in relative abundance were determined. Candidate scaffolds were screened for Y chromosome counterparts. Y/X size ratios are ~1/3 for human, but 2/1 for *T. castaneum*; this pairing suggests the presence of evolutionary development and pest control, a complete sequence of the *T. castaneum* genome assembly by limiting how reliably contigs can be made. We have sought to overcome this limitation by using PacBio to obtain longer sequence reads from thousands of long sequence reads from *T. castaneum* males. These long male-specific sequence reads are currently being used to extend Y- chromosome sequence scaffolds, as well as to generate super scaffolds. Of the ~1.5 Mb of Y chromosome sequence identified to date, we find that the sequence is enriched in highly repetitive sequence, including transposons and retrotransposons. In addition to repetitive sequences, some Y-specific scaffolds have also been shown to contain expressed, single-copy genes. However, thus far no gene known to be required for maleness has been found.

INTRODUCTION

The Y chromosomes of humans and flies (*Drosophila melanogaster*) have low gene content and are littered with repetitive sequences (e.g. transposons). They also differ dramatically in size from their respective X chromosome counterparts. Y/X size ratios are ~1/3 for human, but 2/1 for *D. melanogaster*. Almost nothing is known about the composition and function of the Y chromosome in the red flour beetle, *Tribolium castaneum*. Although current research indicates sex determination in this species relies on the presence of a male-determining factor most likely found on the Y. In order to better understand this sex chromosome the Massachusetts Institute of Technology (MIT), and the Howard Hughes Medical Institute, are interested in the processes of evolutionary development and pest control, a complete sequence of the *T. castaneum* Y chromosome is necessary.

As with humans and flies, the Tribolium Y consists predominantly of regions of repetitive sequence. These regions interfere with genome assembly by limiting how reliably contigs can be made. We have sought to overcome this limitation by using PacBio to obtain longer sequence reads that may help link repetitive regions to known Y sequence. We are currently in the process of analyzing these data. We have also identified putative Y-specific repeats which have helped identify additional Y-linked contigs and scaffolds.

**Trichosanthes castaneum** karyotype

![Karyotype photo courtesy of Jeff Stuart](Image)

**Figure 1.** XY karyotype observed in *Tribolium castaneum*. In Tribolium and other beetles the large X and much smaller Y chromosomes undergo pairing during metaphase to form a bivalent structure termed the “X Y pairing” (indicated by arrow head). This pairing suggests the presence of evolutionary development and pest control, a complete sequence of the *T. castaneum* genome assembly by limiting how reliably contigs can be made. We have sought to overcome this limitation by using PacBio to obtain longer sequence reads that may help link repetitive regions to known Y sequence. We are currently in the process of analyzing these data. We have also identified putative Y-specific repeats which have helped identify additional Y-linked contigs and scaffolds.

DEFINITIONS

Contigs = assembled fragments of overlapping DNA sequence from the whole genome shotgun sequencing project (*Tribolium Castaneum*, 2008).

Scaffolds = a series of contigs that are in the right order but not connected in one continuous stretch of DNA sequence.

Note: Contigs and scaffolds are numbered as in the *Tcnu* 2.0 assembly, but the names are modified as follows: “s1338” = “contig1830”, while scaffold names have been shortened to designate only the first constituent contig. For example, “contig6668_contig1831” is renamed “s6668” (s = scaffold).

METHODS

Repeat-based analysis:

Genomic sequences are based on *Tcnu* 2.0 assembly, but the newest mapping data was used to determine chromosomal associations. Contigs and scaffolds were obtained from the Human Genome Sequencing Center, Baylor College of Medicine. The only previously-known Y-chromosome sequence, namely the Y-linked RAPD marker AF2790, was used as query (BLASTN, Altshul et al., 1997) to identify related scaffolds in the Tribolium genome assembly. Sequence scaffolds having an E-value lower than e-50 were pair-wise aligned to identify regions of sequence similarity using both BLAST2 (NCBI) and BlockX (Vector NTI, Inforgen). Blocks of high sequence identity were used as bait (BLASTX against the Tribolium genome sequence) to identify additional sequence scaffolds containing related repeats. Since the majority of the newly-identified scaffolds were unmapped, scaffold-specific primers pairs were designed for each newly-identified scaffold and sex-specific PCR performed.

Each Y-linked scaffold was compared to repeat libraries created with TRF (Brenner, 1999), TGLib (Brenner and Tu, 2003) and RepeatScout (Price et al., 2005). Each scaffold was first tested via PCR for linkage to the Y chromosome, as described above. Then a confirmed Y contig, flanked by sex non-specific sequence, is shown with Illumina alignments. In the center are reads that are predominantly male-specific. Center left are reads found in both males and females that we believe represent a region of sequence common to both X and Y. Center right, are reads that are mostly male-specific, but which are also found in females, suggesting the presence of non-Y-specific repetitive elements in this contig.

DISCUSSION

It is our hope that technologies, such as PacBio, which provide longer reads than traditional Sanger and next-gen methods, will help span these regions, permitting us to complete assembly of the Y chromosome sequence.

ACKNOWLEDGMENTS

We are grateful to Terri O’Leary, and Stephanie Neal for their expert technical assistance.

REFERENCES


Lorenzen, M.D., Dryer, Z., Tur, J., Snow, K., Crumly, L.R., Shippy, T.D., Stuart, J.J., Beeman, R.W. and Brown, S.J. (2005). Analysis of repetitive sequence is enriched in highly repetitive sequence, including transposons and retrotransposons. These regions interfere with genome assembly by limiting how reliably contigs can be made. We have sought to overcome this limitation by using PacBio to obtain longer sequence reads that may help link repetitive regions to known Y sequence. We are currently in the process of analyzing these data. We have also identified putative Y-specific repeats which have helped identify additional Y-linked contigs and scaffolds.

**Trichosanthes castaneum** karyotype

![Karyotype photo courtesy of Jeff Stuart](Image)

**Figure 1.** XY karyotype observed in *Tribolium castaneum*. In Tribolium and other beetles the large X and much smaller Y chromosomes undergo pairing during metaphase to form a bivalent structure termed the “X Y pairing” (indicated by arrow head). This pairing suggests the presence of evolutionary development and pest control, a complete sequence of the *T. castaneum* Y chromosome is necessary.

**Figure 2.** Genetic map of the 10 *T. castaneum* linkage groups showing the association between the mapped markers and genomic sequence scaffolds. Color-coded boxes group mapping markers found within a single sequence scaffold. For example, the DNA sequences of 14 mapped markers are located in Scaffold2,5006, while 18 are found with Scaffold3,587.

**METHODS**

Repeat-based analysis:

Genomic sequences are based on *Tcnu* 2.0 assembly, but the newest mapping data was used to determine chromosomal associations. Contigs and scaffolds were obtained from the Human Genome Sequencing Center, Baylor College of Medicine. The only previously-known Y-chromosome sequence, namely the Y-linked RAPD marker AF2790, was used as query (BLASTN, Altshul et al., 1997) to identify related scaffolds in the Tribolium genome assembly. Sequence scaffolds having an E-value lower than e-50 were pair-wise aligned to identify regions of sequence similarity using both BLAST2 (NCBI) and BlockX (Vector NTI, Inforgen). Blocks of high sequence identity were used as bait (BLASTX against the Tribolium genome sequence) to identify additional sequence scaffolds containing related repeats. Since the majority of the newly-identified scaffolds were unmapped, scaffold-specific primers pairs were designed for each newly-identified scaffold and sex-specific PCR performed.

Each Y-linked scaffold was compared to repeat libraries created with TRF (Brenner, 1999), TGLib (Brenner and Tu, 2003) and RepeatScout (Price et al., 2005). Each scaffold was first tested via PCR for linkage to the Y chromosome, as described above.

Male genome vs. female genome analysis:

DNA was isolated from male beetles and female beetles and sequenced at high depth (Illumina HiSeq). These sex-specific sequence reads were then independently aligned to the published (mixed sex) genome assembly and differences in relative abundance determined. Candidate scaffolds were screened for Y specificity via sex-specific PCR.

**RESULTS**

Male-Specific Genomic Regions

![Graph showing sex-specific regions](Image)

**Figure 3.** Tribolium genomic reads from Illumina sequencing showing sex-specific differences in abundance. Illumina reads were first aligned to existing linkage maps, including the unknown chromosome – a composite of contigs and scaffolds not associated with any other linkage group. Reads on the unknown chromosome were then examined for differential patterns, like those shown in this figure. In the middle are reads found only in males, indicative of sex-specific sequence; sex-specific PCR was then used to confirm the presence of such regions on the Y chromosome. On either side, reads show little difference between males and females, suggesting they originated from an autosome or the X-chromosome.