Using Targeted Sequencing for the Discovery of Intron Sequences with Baits Derived from Transcriptome Data

Yusuf Murgha[1], Rebecca Ward[1], Erdogan Gulari[1], Geoff Gill[2], Jean-Marie Rouillard[1]

[1] MYcroarray, Ann Arbor, MI, USA, info@mycroarray.com
[2] ViaLactia Biosciences, Auckland, New Zealand

MYCROARRAY

MYBAITS

Introduction

- Genome sequences of many organisms of interest are unavailable. However, it is now economically feasible to deep sequence the transcriptome.
- Here, we demonstrate that transcriptome data can be used to generate MYbaits sequence capture kits targeting expressed sequences to enrich samples for exon sequences and flanking intron sequences.
- We illustrate the discovery of exon boundaries, complete genes including full introns as well as the capture of paralog genes.

Experimental design

Perennial ryegrass *Lolium perenne*

- A set of 28,867 bait sequences (100mer baits, tiled every 60bp) were designed from a mix of transcript and genomic sequences (2.25Mb total) from the perennial ryegrass. RNA baits were manufactured into a MYbaits kit.
- 454 Sequencing libraries were prepared from ryegrass genomic DNA with an average fragment length of 600 – 700bp.
- MYbaits kit was used to capture targets following the recommended protocol with the exception that the captured molecules were directly sequenced without any intermediary PCR amplification.
- Sequencing was performed either on a 454 Junior or a 454 FLX system.

Western terrestrial garter snake *Thamnophis elegans*

- A set of 14,468 bait sequences (80mer baits, tiled every 40bp) were designed from transcript sequences (692 Kb) from the western terrestrial garter snake.
- Two 454 Sequencing libraries were prepared from *T. elegans* genomic DNA with an average fragment length of 600 – 700bp.
- MYbaits kit was used to capture targets following the recommended protocol.
- Sequencing was performed on a 454 Junior yielding 180,470 reads (77Mb).
- 57% of reads were directly mapped to the reference transcript sequences.
- The reads were first assembled de-novo into contigs and subsequently mapped to the reference transcript sequences, we obtained a total of 2,556 reference contigs (defined as containing some transcript reference sequences).
- 85% of the reads can be mapped to the reference contigs
- Altogether, the reference contigs contain 615,338 bases mapped to reference transcripts and 2,161,616 unmapped bases. Thus, for 1 base of exon sequence, we gained an average 3.5 bases of new sequence.

2. Uncover exact exon structure

- Exon 9 (421bp) was incorrectly predicted by aligning this snake transcript to other species transcripts. Mapping of reads from captured sample to the reference sequence reveals that exon 9 should be split into 4 distinct exons.

Structure	Size (bp)
Predicted Exon 9	421
Exon 9a	153
Intron	93
Exon 9b	81
Intron	385
Exon 9c	86
Intron	259
Exon 9d	101
# novel bases per known base = 2.57	

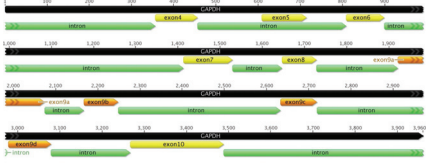


Figure shows delineation of exon 9 into 4 exons (orange) and discovery of novel intron sequences (green).

1. Identify exons and novel introns

- The entire coding region of the **ryegrass** SACDH1 gene (9,288bp) was captured using 44 baits designed from a transcript sequence without prior knowledge of the exon boundaries.
- Statistics:

	Exon	Intron
Number	25	24
Min. size (bp)	21	69
Max. size (bp)	312	1039
# novel bases per known base		1.84

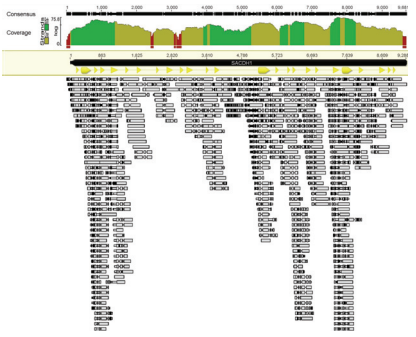
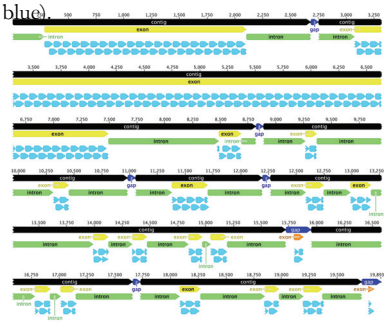


Figure shows the coverage and mapping of 454 reads along the assembled gene sequence. The reference sequence (exons) are shown in yellow.

- 212 baits (light blue) were designed from an uncharacterized **snake** transcript sequence (8,713bp) without prior knowledge of the exon boundaries.
- The reference sequence maps to 6 contigs (black) separated by gaps of unknown length inside introns (royal blue).
  - Identified at least 19 exons.
  - Fully sequenced 9 introns and partially sequenced 8 introns.
  - Largest sequenced intron is 996bp in length. Partially sequenced introns were likely larger.

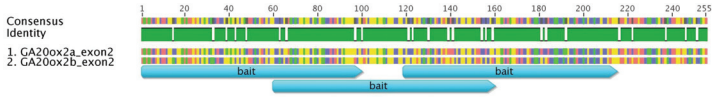


	Exon	Intron
Min. size (bp)	51	75
Max. size (bp)	4415	996
Average size range (bp)	100 - 200	200 - 500
# novel bases per known base		1.19

- Failure to capture two short exons (51 and 87bp) because of their relatively short length and limited sequencing depth.

3. Capture of gene paralogs

- 30 paralogs were identified for a set of 300 ryegrass transcripts targeted in this study.



- In this example, an exon was targeted with three baits and two paralogs were captured.

Conclusion

- It is possible to design baits from transcriptome sequence data to capture genomic sequence, characterize exons, and gain insight into some intron sequences.
- The identification of maximal exon-flanking (introns, upstream and downstream) sequences depends on intron length, exon length, and sequencing library fragment size.

Acknowledgments We thank Tonia Schwartz and Anne Bronikowski (Iowa State University) for permitting the use of *T. elegans* data.