Title

Transcriptome resources for three hybrid sunflower species (*Helianthus anomalus, H. deserticola, H. paradoxus*)

Authors

Sébastien Renaut[a,*], Matt G. King[b], Heather C. Rowe[a], Loren H. Rieseberg[a,c]

Affiliations

[a] Biodiversity Research Centre and Department of Botany, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada

[b] DuPont Pioneer, Box 1000, Johnston, IA 50131-0184

[c] Department of Biology, Center for Genomics and Bioinformatics, Indiana University, 1001 East Third Street, Bloomington, IN 47405, USA

* Corresponding author (sebastien.renaut@gmail.com)

Introduction

Natural hybridization between closely related taxa is frequent in many taxonomic groups, yet it has long been perceived as a force preventing diversification and speciation. In recent years, evidence of hybridization facilitating adaptive divergence has accumulated (Mallet 2007; Nolte 2010; Abbott et al. 2013). Homoploid hybrid speciation (the formation of hybrid lineages without changes in chromosome number) occurs when distinct species come into contact, hybridize, and produce hybrid swarms. If hybrid genotypes can colonize areas of the adaptive landscape inaccessible to ancestral species, they may establish new distinct lineages, reproductively isolated from their ancestors.

Much of what we understand about hybrid speciation comes as a legacy of work on annual sunflowers, the "poster children" of this field. Three novel species (*Helianthus anomalus*, *H. deserticola* and *H. paradoxus*), ecologically specialized into extreme habitats, arose via independent hybridization events between *H. annuus* and *H. petiolaris* (Rieseberg *et al.* 2003). Extreme or transgressive values with respect to parental species *H. annuus* and *H. petiolaris*, both for external phenotypes and transcript levels have been observed in populations of these hybrid species. The extreme values are believed to contribute to ecological speciation via enhanced fitness in a novel environment (Lai *et al.* 2006; Rieseberg *et al.* 2006; Donovan *et al.* 2010).

However, little is known about the evolutionary consequences of hybrid speciation from a genomic perspective. By sequencing transcriptomes for these three species, we hope to gain insights about how hybridization affects gene expression in sunflowers. Here, we used next generation Illumina sequencing to sequence the transcriptomes of 18 individuals from the three hybrid species aforementioned. This will serve as an important genome-scale resource for further research on the genomic and phenotypic consequences of hybrid speciation.

Data Access

- *Sequence files* – Sequence files (.fq) can be found on NCBI Sequence Read Archive under project number: PRJNA188794 (see table 1 for individual accession numbers)
- *Reference file* – Reference transcriptome (HA412_trinity_noAltSplice_400bpmin.fa, 51 468 contigs, 51.3 million base pairs) is described in another publication (Renaut *et al.* 2013)

and is accessible on DRYAD (http://dx.doi.org/10.5061/dryad.9q1n4)

- *Sequence alignment files* – Sequence alignments (one .bam file per individual) can be found on NCBI Sequence Read Archive under project number: PRJNA188794
- *SNP file* – SNP tables (one .txt file per species) are accessible on DRYAD (http://dx.doi.org/10.5061/dryad.fj594)
- *Coverage file* – Coverage per gene and per individual (one .txt file) is accessible on DRYAD (http://dx.doi.org/10.5061/dryad.fj594)
- *Adaptor contaminant file* – File containing potential Illumina adaptor contaminants (one .fa file) is accessible on DRYAD (http://dx.doi.org/10.5061/dryad.fj594)
- *Script files* – R (R Core Team 2012) code used to process the data and readme files are accessible on github (https://github.com/seb951/helianthus_hybrid_species_transcriptome)

Meta Information

- *Sequencing center* – Canada's Michael Smith Genome Science Center (Vancouver, Canada, www.bcgsc.ca/platform/solexa) and Biodiversity NextGen Sequencing Facility (Vancouver, Canada, sites.google.com/site/biodiversitynextgensequencing/home)
- *Platform and model* – Illumina (San Diego, CA, USA) Genome Analyzer IIx (Genome Science Center) and Illumina HiSeq 2000 (Biodiversity NextGen Sequencing Facility)
- *Design description* – We sampled one individual per population, choosing populations that cover most of the established geographic range of each study species. The goals were to identify species-wide polymorphism in coding sequence and transcript abundance, and to compare homoploid hybrid sunflower species with existing datasets generated from progenitor species *H. annuus* and *H. petiolaris*.
- *Run date* – 2011-05-26 (GAIIx) and 2012-11-05 (HiSeq 2000)

Library

- *Strategy* – non-normalized cDNA
- *Taxa – Helianthus deserticola*, *H. petiolaris, H. anomalus*
- *Tissue* – Young leaf/stem tissue from plants approximately two months old
- *Location* – see Table 1
- *Sample handling to prevent possible contamination* – We germinated all achenes at the

University of British Columbia (Vancouver, Canada) and grew them for approximately two months in growth chambers (12 hours of daylight at 22 degrees). Then, we harvested young leaf/stem tissue, flash froze it in liquid nitrogen and kept it at -80 degrees. Once sequencing was performed, sequences were cleaned to remove low quality reads and potential adaptors sequences using TRIMMOMATIC (Lohse *et al.* 2012). Alignment to the reference dataset also reduced contaminating reads (see pipeline description below).

- *Additional sample information* – see Table 1
- *Layout* – Paired end reads (2 X 100 bp or 2 X 101 bp)
- *Library construction protocol* –For each individual, we extracted RNA using a modified TRIzol Reagent protocol (Invitrogen, Carlsbad, CA, USA). We quantified the RNA samples using a NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA) and verified their quality on agarose gels. We stored total RNA in pure water. Libraries were then prepared following standard Illumina Tru-Seq (LT) Protocol (pp. 35-69) with one slight modification. The RNA was not fragmented during the poly-A mRNA purification step, but directly reverse transcribed into cDNA. Upon cDNA purification, samples were then sheared to ~ 400 bp on a Covaris (Woburn, MS, USA) sonicator. These were then sequenced the Illumina GAIIx or HiSeq 2000 platform (see table 1). Base calling was performed via the standard Illumina CASAVA (1.8) pipeline.
- *Nominal size (paired) of fragments sequenced* – 400 bp
- *Nominal standard deviation* – Sizes ranged from 200 - 500 bp

Table 1: Sample Description. Seqeunce Read Archive accession number can be searched here (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=search_obj). "state (USA)", "latitude", and "longitude" refer to original collection locations; additional phenotype data and propagation records for USDA accessions are available at http://www.ars-grin.gov/ and can be specifically referenced using the GRIN # (PI XXXXXX) provided.

species	sample name	tissue	SRA accession number	state (USA)	latitude	longitude	collection date (or GRIN #)
H.anomalus	Ano1495	leaves/stem	SRR696562	AZ	36.97	-109.63	PI 468638
H.anomalus	Sample_Ano1506	leaves/stem	SRR696986	UT	38.37	-110.70	PI 468642
H.anomalus	Sample_Goblinvalley	leaves/stem	SRR696987	UT	38.58	-110.71	2008-10-01
H.deserticola	Des1484	leaves/stem	SRR696563	UT	37.05	-112.53	PI 468703
H.deserticola	des2458	leaves/stem	SRR696571	NV	39.19	-118.19	PI 649873
H.deserticola	Sample_Des2463	leaves/stem	SRR696962	NV	39.02	-118.80	PI 664663
H.deserticola	Sample_des1486	leaves/stem	SRR696963	AZ	36.94	-111.43	PI 468705
H.deserticola	Sample_desA2	leaves/stem	SRR696966	NV	39.22	-118.70	2008-10-01
H.deserticola	Sample_DES1476	leaves/stem	SRR696979	UT	37.20	-113.19	PI 468702
H.deserticola	Sample_desc	leaves/stem	SRR696984	NV	39.55	-118.86	2008-10-01
H.paradoxus	king141B	leaves/stem	SRR710275	NM	34.94	-104.68	2008-10-01
H.paradoxus	king145B	leaves/stem	SRR688268	NM	33.32	-104.33	2008-10-01
H.paradoxus	king147A	leaves/stem	SRR688282	TX	26.25	-98.48	2008-10-01
H.paradoxus	King151	leaves/stem	SRR688286	NM	33.43	-104.47	2008-10-01
H.paradoxus	king152	leaves/stem	SRR696542	NM	33.43	-104.47	2008-10-01
H.paradoxus	King156B	leaves/stem	SRR696561	TX	31.01	-102.92	2009-10-01
H.paradoxus	Sample_king1443	leaves/stem	SRR696989	NM	34.93	-104.67	2008-10-01
H.paradoxus	Sample_king159B	leaves/stem	SRR696991	TX	30.90	-102.89	2008-10-01

Sequence Processing

- *Pipeline* – The scripts along with all parameters for the different analytical steps and a readme file are described and made available on github (https://github.com/seb951/helianthus_hybrid_species_transcriptome).

Sequencing files were cleaned to remove low quality reads and potential adaptor sequences using TRIMMOMATIC (Lohse *et al.* 2012). The trimming parameters for adaptor removal (ILLUMINACLIP) were as follow: seed mismatch of 2, palindrome clip threshold of 40, simple clip threshold of 15. For trimming based on quality, the parameters were: minimum leading and trailing base quality of 2, minimum length of 36, minimum average base quality of 15 for sliding window of size 10.

Cleaned reads were then aligned against the reference transcriptome (51,468 contigs, 51.3M bp) using the Burrows-Wheeler Aligner (BWA, ALN with -q 20 and SAMPE commands, Li & Durbin 2009). SAMTOOLS (MPILEUP with -C50 and BCFTOOLS, Li *et al.* 2009) was used to call Single Nucleotide Polymorphisms (SNPs) using information from all samples for each species separately. SNPs therefore include both fixed differences from the *H. annuus* reference and intraspecific polymorphisms. Genotypes with Phred-scaled likelihoods below 20 were considered as missing, which corresponds to a genotyping accuracy of at least 99%. Custom R (R Core Team 2012) scripts were used to automate analysis.

- *Runs* – 18 runs were submitted to NCBI SRA. Each run contains two (_1.fq and _2.fq) files. Runs were submitted as two different experiments given that samples were sequenced on two different platforms (see Table 1).

Results

- *Total number of reads, percentage of reads surviving filtering, mean length, number of reads aligned, percentage of reads aligned, mean (median) number of reads aligned per contig* – Table 2
- *Number of contigs with coverage > 0, number of base pairs with coverage > 0, total number of SNPs, number of fixed differences, Mean number of SNPs per 100 aligned base*

pairs – Table 3

- *Quality scoring system* – phred+33
- *Quality scoring ASCII character range* – "!" to "J"

Table 2: Alignment Statistics

sample name	species	Sequencing Platform	Total number of reads (M)	Percentage of reads surviving filtering	Mean read length before filtering	Mean read length after filtering	Number of reads aligned (M)	Percentage of reads aligned	Mean (median) number of reads aligned per contig
H.anomalus	Ano1495	GAII	31.8	87.95	100	97.9	18.9	59.3	366.2 (4)
H.anomalus	Sample_Ano1506	HiSeq 2000	46.7	92.77	101	84.0	28.2	60.4	547.4 (10)
H.anomalus	Sample_Goblinvalley	HiSeq 2000	65.2	93.42	101	84.0	39.4	60.5	765.6 (17)
H.deserticola	Des1484	GAII	29.4	88	100	97.9	17.0	57.7	329.7 (6)
H.deserticola	des2458	GAII	53.4	81.76	100	97.0	28.6	53.5	555.9 (11)
H.deserticola	Sample_Des2463	HiSeq 2000	38.2	93.4	101	84.0	22.5	58.9	436.7 (12)
H.deserticola	Sample_des1486	HiSeq 2000	34.9	92.51	101	83.9	21.4	61.3	415.6 (6)
H.deserticola	Sample_desA2	HiSeq 2000	39.6	93.18	101	84.0	22.7	57.4	441.5 (14)
H.deserticola	Sample_DES1476	HiSeq 2000	42.6	92.71	101	84.0	24.8	58.3	482.5 (11)
H.deserticola	Sample_desc	HiSeq 2000	46.9	93.36	101	84.0	27.4	58.5	532.9 (13)
H.paradoxus	king141B	GAII	35.7	88.51	100	95.1	19.5	54.6	378.5 (8)
H.paradoxus	king145B	GAII	18.6	84.34	100	94.6	10.4	55.8	201.8 (2)
H.paradoxus	king147A	GAII	44.0	83.57	100	94.0	22.9	52.0	444.2 (14)
H.paradoxus	King151	GAII	32.7	84.1	100	93.5	17.9	54.8	348.1 (6)
H.paradoxus	king152	GAII	27.6	84.97	100	93.6	15.6	56.4	302.5 (5)
H.paradoxus	King156B	GAII	41.4	86.78	100	95.0	22.8	55.0	442.9 (8)
H.paradoxus	Sample_king1443	HiSeq 2000	58.0	93.21	101	84.0	33.1	57.0	642.2 (17)
H.paradoxus	Sample_king159B	HiSeq 2000	43.7	92.38	101	84.0	25.5	58.5	496.2 (12)

Table 3: SNP statistics

	Number of contigs with coverage > 0	Number of base pairs with coverage > 0 (M)*	Total number of SNPs (K)	Number of fixed differences (K)**	Mean number of SNPs per 100 bp
H. deserticola	31212	27.9	708.0	316.1	2.53
H. anomalus	31985	29.9	769.1	310.5	2.57
H. paradoxus	33809	31.4	446.1	307.6	1.42

K = 1,000; M = 1,000,000; bp = base pairs

*This is compared to the total number of base pairs in the reference dataset (51.3M bp)

**These are fixed differences compared to the *H. annuus* reference dataset

Acknowledgments

We thank M. Stewart for help in performing lab work, D. Adam (GSC) and A. Kuzmin (Biodiversity NGS Facility) for preparing libraries and performing sequencing and C. Grassa, N. Kane and T. Nguyen for help with processing samples. This work was supported by funding from Genome Canada, Genome BC, and a postdoctoral scholarship from the Natural Sciences and Engineering Council of Canada to SR.

References

Abbott R, Albach D, Ansell S *et al.* (2013) Hybridization and speciation. *Journal Of Evolutionary Biology*, **26**, 229–246.

Donovan LA, Rosenthal DR, Sanchez-Velenosi, Rieseberg LH, Ludwig F (2010) Are hybrid species more fit than ancestral parent species in the current hybrid species habitats? *Journal Of Evolutionary Biology*, **23**, 805–816.

Lai Z, Gross BL, Zou YI, Andrews J, Rieseberg LH (2006) Microarray analysis reveals differential gene expression in hybrid sunflower species. *Molecular Ecology*, **15**, 1213–1227.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lohse M, Bolger AM, Nagel A *et al.* (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, **40**, W622–W627.

Mallet J (2007) Hybrid speciation. *Nature*, **446**, 279–283.

Nolte A (2010) Understanding the onset of hybrid speciation. *Trends In Genetics*, **26**, 54–58.

Renaut S, Grassa CJ, Yeaman S *et al.* (2013) The number and size of genomic islands of divergence do not vary with geography of speciation. *Nature Communications*. *In Press*

Rieseberg LH, Kim S-C, Randell RA *et al.* (2006) Hybridization and the colonization of novel habitats by annual sunflowers. *Genetica*, **129**, 149–165.

Rieseberg LH, Raymond O, Rosenthal DM *et al.* (2003) Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, **301**, 1211–1216.

R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.