**Supplementary Materials 1: Generation of SNPs Used in Sequenom Assays**

**by Srikar Chamala**

**454 EST Sequencing**

Leaf tissues of 13 individuals: 7 *T. dubius*, 3 *T. pratensis*, and 3 *T. porrifolius* were sequenced using 454 FLX and Titanium sequencing technologies. Details of 454 Sequencing runs are provided in table of the following excel spreadsheet

S2\_ TragopogonReadsAssemblyData.xlsx🡪454RunInfo.

*T. dubius*, *T. pratensis*, and *T. porrifolius* constitute 76.5%, 10.9%, and 12.6% respective read sizes of the total read size of 436,374,304 bp.

**Assembly of 454 Reads**

Assembly of 454 transcriptome was performed in two steps. First 454 reads were assembled using Roche 454 Newbler assembler Version 2.0.00.20 (Margulies, et al. 2005). Newbler performs very stringent assembly, so reads with high similarity will get assembled. Reads with relatively high error rate or from diverse alleles might have been thrown out of the assembly because they did not meet the high stringency Newbler assembly criteria. In order to include these thrown out reads, a second step of assembly was done using TGI Clustering tools (TGICL) (Pertea et al. 2003). Newbler assigns six different assembly status flags for each input read, “Assembled”, “Partially Assembled”, “Repeat”, “TooShort”, “Singleton”, and “Outlier”. TGICL assembly was performed on contigs resulting from Newbler assembly and trimmed reads with assembly status other than “Assembled”.

The above two-step assembly is done on four different transcript data sets, reads from *T. dubius*, *T. pratensis*, and *T. porrifolius* (Data Set I), reads from *T. dubius* only (Data Set II), reads from *T. pratensis* only (Data Set III), and reads from *T. porrifolius* only (Data Set IV). The strength of combined assembly is its high coverage, which will help capture all transcripts and low expression transcripts across all 3 individual *Tragopogon* species. This will also help identify transcripts, which are species specific during the down stream analysis. Species-specific datasets will act as quality control to make sure all individual contigs reported in individual species assembly are reported in Data Set I.

Newbler assembly was performed separately on raw SFF files containing 177,959,264 454 FLX and Titanium reads for Data Set I, 160,512,420 454 FLX and Titanium reads for Data Set II, 3,427,518 Titanium reads for Data Set III, and 4,219,544 Titanium reads for Data Set IV. 454 MID tags, low quality, repetitive, and polyA tail portions of reads were trimmed by Newbler during the assembly process.

These assemblies resulted in 41,336 contigs for Data Set I (average size = 509.7 bp, maximum contig length = 7541 bp, minimum contig length = 91, N50 Size  = 687), 32,977 contigs for Data Set II (average size = 559.9 bp, maximum contig length = 7541 bp, minimum contig length = 91, N50 Size  = 758), 7542 contigs for Data Set III (average size = 444.3 bp, maximum contig length = 3101 bp, minimum contig length = 91, N50 Size  = 496), and 7806 contigs for Data Set IV (average size = 525.5 bp, maximum contig length = 3458 bp, minimum contig length = 95, N50 Size  = 608) (Table in the following spreadsheet S2:TragopogonReadsAssemblyInfo.xlsx🡪 454\_Assembly). As expected Data Set I Newbler assembly resulted in greater number of contigs, total contig size, and number of reads assembled. This is because Data Set I assembly picked extra contigs, which would have missed with individual assemblies and larger contigs because of the higher depth of reads.

The second assembly step was done using TGICL assembler on contigs and reads with assembly status other than “Assembled” resulting from Newbler assembly. Reads were trimmed using Newbler software package utilities before feeding them into TGICL. TGICL assembly step made total number of contigs to go down, which is due to collapse of the contigs, Newbler could not have collapsed because of it is very conservative read overlap stringency criteria. Also, TGICL assembly produced greater contig size, average contig length, and N50 size because of its less stringent overlap and alignment criteria of reads during assembly when compared to Newbler.

TGICL assemblies resulted in 37,711 contigs for Data Set I (average size = 598.3 bp, maximum contig length = 7540 bp, minimum contig length = 42, N50 Size  = 790), 31592 contigs for Data Set II (average size = 612.3 bp, maximum contig length = 7540 bp, minimum contig length = 42, N50 Size = 815), 7025 contigs for Data Set III (average size = 487.9 bp, maximum contig length = 3439 bp, minimum contig length = 51, N50 Size  = 527), and 7732 contigs for Data Set IV (average size = 545.7 bp, maximum contig length = 3601 bp, minimum contig length = 52, N50 Size  = 628) (Table in the following spreadsheet S2:TragopogonReadsAssemblyData.xlsx🡪 TGICL\_CAP3).

**Identification and elimination of non-*Tragopogon* and non-EST contigs**

Assembled contigs were screened for non-*Tragopogon* sequences corresponding to bacteria, and insects, which could be a common source of contamination as the plants were greenhouse grown. They were also screened for non-EST sequences resulting from transcripts of chloroplast and mitochondrial sequences. To identify and remove this information from downstream consideration all contig assemblies were aligned by BLAST (WUBLASTN Version 2.0, E=0.000001 B=1 V=1 cpus=1 links topcomboN=1) to bacterial (84,369 sequences), insect (72,328, sequences), chloroplast (17,560 sequences), mitochondrial (76,895 sequences), and fungal sequences (30,110 sequences) collected from Genbank (Table S1). In addition, any 454-transcript assembly that aligned to insect, fungi, or bacterial sequences annotated as “ribosome”, “ribosomal” or “histone” was retained to avoid removing *Tragopogon* representatives of highly conserved ribosomal and histone proteins. Contigs that align (with at least 80% of the length of the subject or query sequence with a minimum of 80% nucleotide identity) to chloroplast, mitochondrion, non-ribosomal or histone, insect, bacterial, and fungal sequences were eliminated from further analysis. The results of these analyses are shown in Table S2.

**Table S1**. The strategy used in removing non-*Tragopogon* and non-EST reads from assembly.

|  |  |  |  |
| --- | --- | --- | --- |
| Sequence Type | Genbank Search Type | Genbank Query | Number of Sequences Obtained |
| Bacterial | Nucleotide | "Bacteria[Organism] AND complete[Title] NOT partial[All Fields] NOT pseudogene[All Fields] NOT EST[All Fields] | 84,369 |
| Insect | Nucleotide | Insecta[Organism] AND complete[Title] NOT partial[All Fields] NOT pseudogene[All Fields] NOT EST[All Fields] NOT mitochondrial[Title] | 72,328 |
| Mitochondrial | Nucleotide | gene\_in\_mitochondrion[PROP] AND "complete"[Title] NOT partial[All Fields] NOT pseudogene[All Fields] NOT EST[All Fields] | 76,895 |
| Chloroplast | Nucleotide | chloroplast[Title] AND "complete"[Title] NOT partial[All Fields] NOT EST[All Fields] NOT partial[Title] NOT pseudogene[All Fields] | 17,560 |
| Fungal | Nucleotide | Fungi[Organism] AND complete[Title] NOT partial[All Fields] NOT pseudogene[All Fields] NOT EST[All Fields] NOT mitochondrial[Title] | 30,110 |

**Table S2.** Qualitative assessment of sequences identified as contamination of the *Tragopogon* EST contigs.

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***T. dubius*** | ***T. pratensis*** | ***T. porrifolius*** |
| **Total Contigs** | 32977 | 7542 | 7806 |
| **Number of Contigs with Hits** | 5597 (16.9%) | 1784 (23.7%) | 2079 (26.6%) |
| **Contigs Passed Criteria** | 179 (.5%) | 23 (.3%) | 30 (.3%) |
| **Mitochondrial Hits** | 55 (.2%) | 3 (0.04%) | 6 (0.08%) |
| **Bacterial Hits** | 27 (.1%) | 0 (0%) | 0 (0 %) |
| **Chloroplast Hits** | 97 (.3%) | 20 (0.26%) | 23 (1.11%) |
| **Insect Hits** | 0 (0%) | 0 (0%) | 0 (0%) |
| **Fungal Hits** | 0  (0%) | 0 (0%) | 0 (0%) |

**Identification of orthologus contig pairs:**

Orthologus contig pairs between *T. dubius* and *T. pratensis*, and *T. dubius* and *T. porrifolius* were identified by doing reciprocal alignments using WU-BLASTN (WUBLASTN Version 2.0, E=0.000001 links topcomboN=1) with E-value at least 1E-10. A contig pair is considered to be orthologous if they are top hits to each other in their reciprocal alignments. *T. dubius* and *T. pratensis* have 3768 ortholog pairs between them whereas *T. dubius* and *T. porrifolius* have 4318. There are 2336 *T. dubius* contigs with both *T. pratensis* and *T. porrifolius* orthologs, and 1432 and 1982 *T. dubius* contigs have only *T. pratensis* and *T. porrifolius* orthologs respectively. Total number of *T. dubius* contigs with at least one *T. pratensis* and *T. porrifolius* ortholog are 5750.

**SNP discovery**

For all 5750 ortholog sets, read sequences belonging to *T. dubius*, *T. pratensis* (if present), and *T. porrifolius* (if present) contigs reads were extracted from ace assembly file and were tagged by species and individual. For each ortholog set *T. dubius* contig in each ortholog set was used as a reference sequence and reads belonging all contigs were aligned using the following MosaikAligner parameters: -a (alignment algorithm) all; –p (CPUs used) 6;  -mmp (maximum mismatch percent) .05; –m (alignment mode) unique; -hs (hash size) 15;  -mhp (maximum number of hash positions to use) 100. These parameters ensured that each read sequence aligned to a unique position within the T. dubius contig.

SNPs were identified within the alignments with the GigaBayes package (http://bioinformatics.bc.edu/marthlab/GigaBayes). GigaBayes is a reimplementation of the PolyBayes (Marth et al. 1999) SNP discovery tool that has been optimized for next-generation sequences. Arguments to GigaBayes were: --D (pairwise nucleotide diversity) 0.001;  --ploidy (sample ploidy) diploid; –sample (sequence source) multiple;  --anchor; --algorithm banded; --QRL (minimum base quality value) 20; --sampleDel (delimiter) \_;  --indel; --O 4. Custom PERL scripts were written to automate the SNP discovery process on all alignments to reference contigs and to parse the GigaBayes output fies (GFF) (which contain the site identification of each SNP) and store into MySQL database.

SNPs and INDELs between species, good quality SNPs between the species, and SNPs and INDELs within each species were identified by querying MySQL database of parsed GigaBayes output with the following criteria:

I. All SNPs/INDELs between species:

*T. dubius* vs. *T. pratensis*:

1. Criteria: # of distinct alleles per species = 1 and  Two species

should have different allele type

2. # of SNPs/INDELs = 17298

3. File: tragapogan\_dubius\_pratanses\_All\_SNPsINDELs\_data.xls

*T. dubius vs. T. porrifolius:*

1. Criteria: # of distinct alleles per species = 1 and  Two species

should have different allele type

2. # of SNPs/INDELs = 24674

3. File: tragapogan\_dubius\_porrifolius\_All\_SNPsINDELs\_data.xls

II. Good SNPs between species:

*T. dubius Vs. T. pratensis:*

1. Criteria: # of distinct alleles per species = 1, Two species

should have different allele type, and # of Reads >= 3

2. # of SNPs = 6289

3. File: tragapogan\_dubius\_pratanses\_goodSNPs\_data.xls

*T. dubius Vs. T. porrifolius:*

1. Criteria: # of distinct alleles per species = 1, Two species

should have different allele type, and # of Reads >= 3

2. # of SNPs = 12191

3. File: tragapogan\_dubius\_porrifolius\_goodSNPs\_data.xls

III. All SNPs/INDELs within species:

*T. dubius:*

1. Criteria: # of distinct alleles per species > 1, Species=Dubius

2. # of SNPs/INDELs = 43273

3. File: tragapogan\_SNPs\_Indels\_Within\_dubius.xls

*T. pratensis:*

1. Criteria: # of distinct alleles per species > 1, Species=Pratensis

2. # of SNPs/INDELs = 8345

3. File: tragapogan\_SNPs\_Indels\_Within\_prathensis.xls

*T. porrifolius:*

1. Criteria: # of distinct alleles per species > 1, Species=Porrifolius

2. # of SNPs/INDELs = 9307

3. File: tragapogan\_SNPs\_Indels\_Within\_porrifolius.xls

**Diagnostic parental genome SNPs identified are illustrated in the below cartoon:**

**Chart, bar chart

Description automatically generated**

**References**

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380. DOI nature03959 [pii]10.1038/nature03959

Pertea G, Huang XQ, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19:651-652. DOI DOI 10.1093/bioinformatics/btg034