

Idotea Assembly Summary

296,173,922 PE reads (Illumina, 125bp)

- Visualized with FASTQC
- Trimmed with TRIMMOMATIC



262,020,548 PE reads

- Re-visualized with FASTQC
- Error corrected with Rcorrector
- In silico normalized with TRINITY



10,099,252 PE reads

- Assembled using TRINITY (20, 25, 30-mers)



115,931 'transcripts' (contigs) (25-mers)

90,663 'genes'

N50: 1,154 Total bases: 79,193,347

- Evaluated and filtered with TRANSLATE



40,122 'transcripts'

30,147 'genes'

N50: 1,710 Total bases: 47,523,534

Raw Assembly

115,931 'transcripts'
90,663 'genes'
N50: 1,154 Total bases: 79,193,347



Completeness determined with BUSCO
* 68% found (many duplicates)



Coding sequences identified with
TRANSDCODER
* 33,731 'peptides' identified



Annotated with PANNZER (UniProtKB)
* 14,288 annotations (many duplicates)
* 1,330,727 GO terms assigned

Filtered Assembly

40,122 'transcripts'
30,147 'genes'
N50: 1,710 Total bases: 47,523,534



Completeness determined with BUSCO
* 61% found (many duplicates)



Coding sequences identified with
TRANSDCODER
* 21,439 'peptides' identified

Are the apparent duplicate genes paralogs or are they alleles?

Next...

- cd-hit will be run on all 3 assemblies to maximize finding potential transcripts
 - This will collapse potential paralogs, but with Idotea's high level of polymorphism many may actually be alleles
 - Duplicate BUSCOs and annotations should be reduced