

Initial Results of Sequencing, Assembly and Annotation of Three Brassicaceae Genomes

Adrian E Platts^{1,2}, Stephen I Wright³, Mathieu Blanchette², Paul M Harrison¹, Eef Harmsen¹, Daniel J Schoen¹, Thomas E Bureau¹
Dept. Biology¹ and School of Computer Sciences², McGill University; Dept Ecology and Evolutionary Biology, University of Toronto³

As part of the Value-directed Evolutionary Genomics Initiative (VEGI) we present early sequencing and assembly results for three Brassicaceae (*Leavenworthia alabamica* (LA), *Sisymbrium irio* (SI) and *Aethionema arabicum* (AA)). These Crucifers were sequenced with 4 lanes of short insert Illumina GAllx data (2x108, 64nt spacer) per species to a depth of x60 (~18GBase). An assembly strategy was optimized a-priori through a series of simulations. Two closely related Brassicaceae genomes (*Arabidopsis thaliana* (AT) and *Arabidopsis lyrata* (AL)) were fragmented in-silico and jointly reassembled to assess quality parameters including contig length, total assembly characteristics and the probability of assembling chimeric sequences from paralogous sub-regions. Variables investigated included read filtering, sequencing depths and Kmer values in three popular assemblers: Velvet, SOAPdenovo and Abyss. The contig-sets generated for the three plants had N₅₀ of 19.1 Kbase (AA), 18.8Kbase (SI) and 21.1Kbase (LA) and a consistent maximum contig size of ~220Kbase. Total assembled libraries ranged in size from 163MBase (AA) to 216MBase (SI). Mapping of highly conserved genes show these contigs appear properly assembled. The presence of repetitive DNA was assessed using a novel de novo repeat assembly approach. Whole-genome alignments to the *A. thaliana* and *A. lyrata* genomes reveals both coding and non-coding functional regions.

This research is supported by Genome Canada/Quebec



Introduction

The aim of the VEGI project is to assess functional non-coding DNA in Brassicaceae through the sequencing and assembly of four crucifer genomes, three of which are reported here. This will permit a six way conservation analysis relative to a reference genome while extensive sequencing of individual plant populations will be used for population genetic studies. The sequencing reported here has been undertaken with short-read, short-insert Illumina GAll data. Short read assembly in plants suffers from several potential problems. Multiple genome duplication events in the Brassicaceae make the assembly of accurate non-chimeric chromosomes challenging. Equally many plants species have seen expansions of repetitive elements that are not yet well characterized. We present here some early results with respect to assembly, repeat characterization and quality metrics for an assembly pipeline.

Methods

Sequencing was conducted using 2x108 paired end GAll reads with a minimal spacer between reads (58-70nt). Assembly parameters were investigated in three popular assemblers using a range of parameters including sequencing depth, k-mer strategy, read depth contigs during the assembly process and read trimming criteria. Prior to assembling actual data we utilized simulated read data to test optimal assembly strategies. This was generated by fragmenting a genome and introducing base errors typical of the GAll platform. Since the most well characterized Brassicaceae genome (*A. thaliana*) is not a recent polyploid, we fragmented this genome together with one chromosome of the recently assembled *A. lyrata* genome. Alignment tests were then undertaken to assess levels of mis-assembly. The results described utilized the Velvet assembler [1] with stringent read trimming (Q<31 -5bp), read depth limits (x5-x180) and a long-Kmer strategy (K:51-61) to assemble contigs approximately three times larger than those generated with an un-optimized strategy [Table 1]. Long K-mers were found to be useful in resolving ambiguity in both repetitive and paralogous domains. The resulting 'contigs' had <1% uncalled (N) bases. Scaffolding with mate-pair libraries to connect these contigs is currently in progress.

Since the AL samples had been previously characterized as highly homozygous, this was used to assess rates of chimeric assembly post-hoc through a SNP rate analysis intended to reveal mis-assembly between polyploid chromosomes. Additional testing used sets of highly conserved genes to assess levels of fragmentation within genic domains. A novel repeat assembler was developed which used a deep k-mer strategy to assemble consensus repetitive sequences from raw read data. This was in turn used to investigate the rates of assembly into intergenic regions from the different assembly strategies.

Results

Build	N50 contig length (bases)	Maximum Contig Size (KBase)	Total Contig Library Size (MBase)	Number of Contigs (>500bases)	Estimated Repetitive Element Content in reads (%)	Estimated Repetitive Element Content in Contigs (%)	Σ Sequence used in the assembly (GBase)/input into the assembly (GBase)
Leavenworthia Alabamica	21,050	240	184	17,771	43.4	24.0	6.7/12.4
Sisymbrium Irio	18,824	217	216	26,919	25.8	11.4	7.4/9.8
Aethionema Arabicum	19,100	194	163	20,630	36.1	11.6	8.3/11.9

Table 1: Assembly statistics for three Brassicaceae genomes. Repetitive element content was assessed with RepeatMasker using the species specific consensus repeat sequences merged into the standard *A. thaliana* repeat library (GIRI June 2009).

A-Priori contig quality was assessed through the fragmentation and re-assembly of AT and hybrid AT-AL genomes. This suggested a mis-assembly rate of up to 4% in the assembly of diverged paralogous regions (Fig 1a). *Post-hoc* contig quality was assessed relative to a set of 458 genes that are conserved across Eukaryota. These were detected through BLAST and assessed for protein length using FGenesH+. All the highly conserved genes were present in the contigs with only ~1% of the CDS models less than 80% of their size in *Thaliana* (Fig 1b).

http://biology.mcgill.ca/vegi/

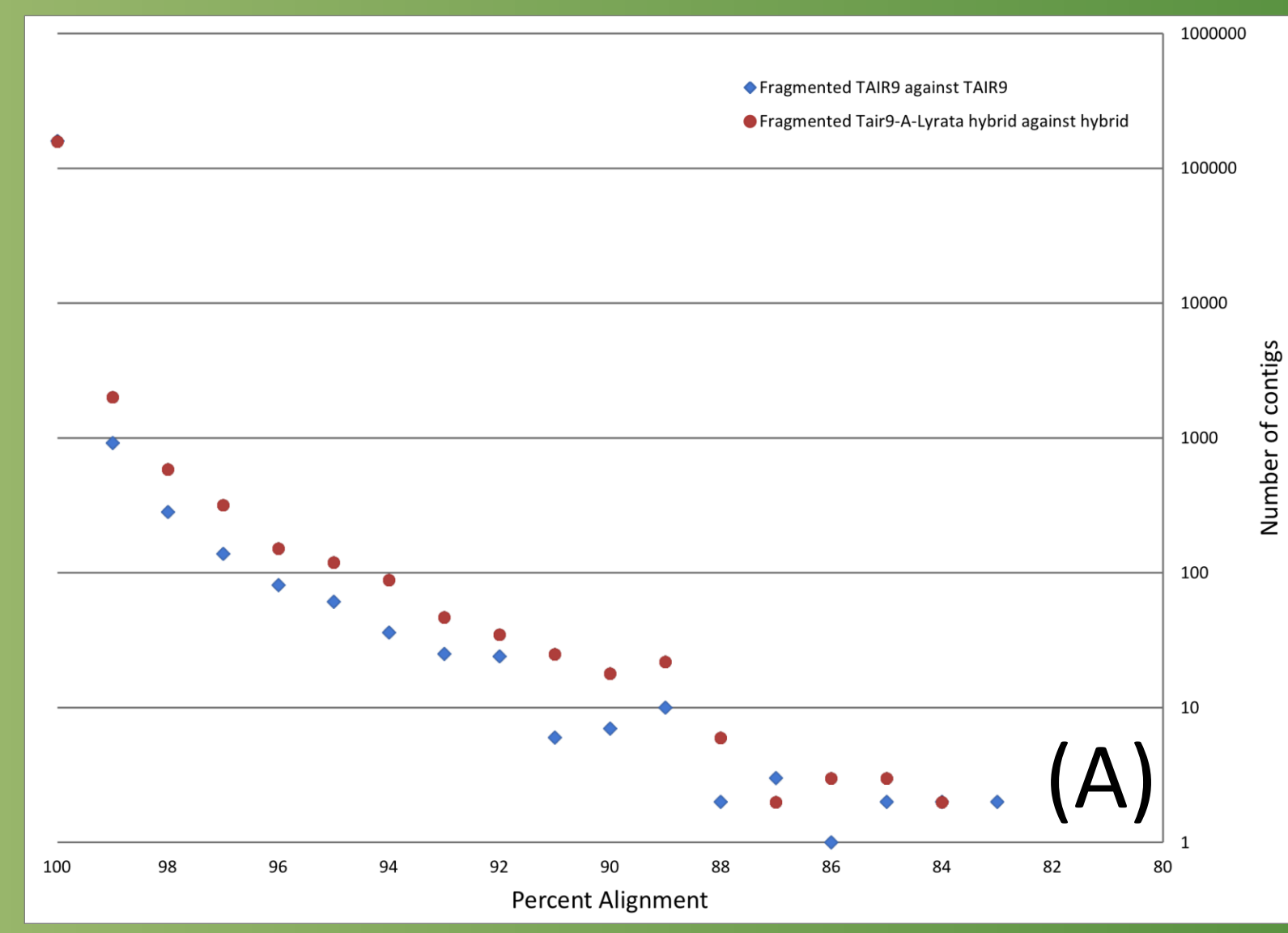
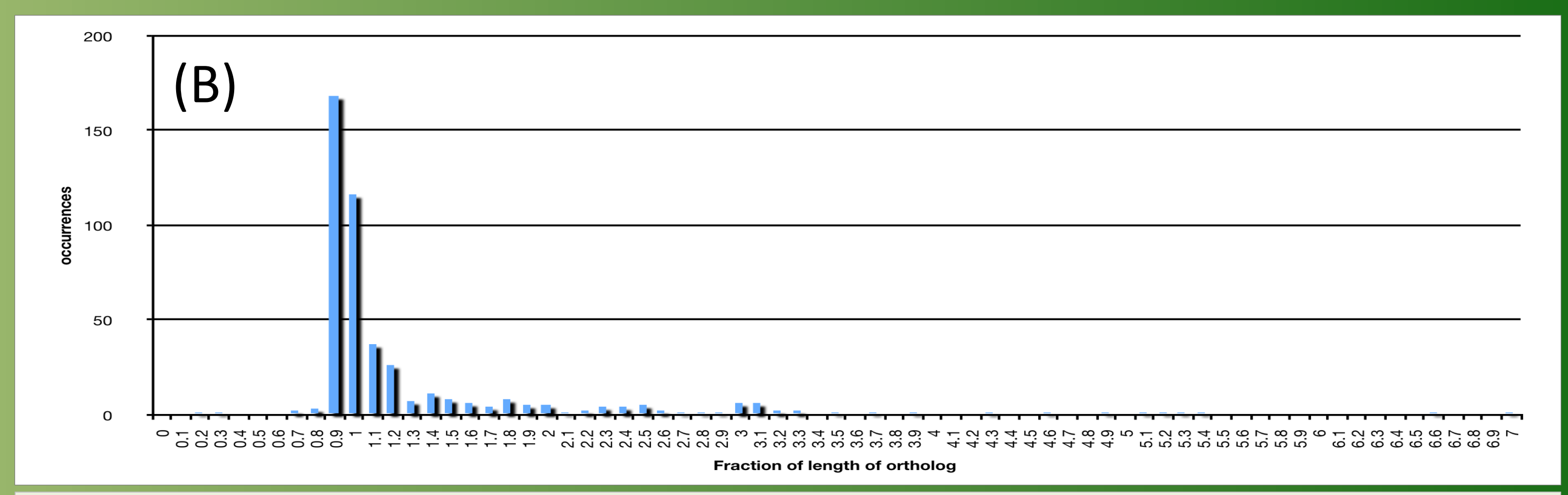


Figure 1. *A-priori* (A) and *Post-hoc* (B) estimates of misassembly. (A) The assembly strategy was simulated by fragmenting *A. thaliana* (blue) and an *A. thaliana/A. lyrata* hybrid genome (red). Mis-assembly was detected in those contigs that could not be fully aligned to the genome from which they originated. Misassembly rates were anticipated to lie in the 2-4% range. (B) Assembly results from a leavenworthia assembly. Mis-assembly rates were estimated by comparing the protein lengths from 458 highly conserved genes with the length of their orthologs in *A. thaliana*. 1% of proteins were found to be truncated to less than 80% of their lengths in *A. thaliana* while over 70% were found to be between x0.9 and x1.1 their lengths in *A. thaliana*.



Assembly parameters were related to repeat content through a repetitive element (RE) assembly assay. The percentage of REs in contigs increased steadily with K-mer value from ~12% at k=31 to ~24% at k=51. Contig data was passed to the annotation pipeline and represented as tracks in a custom build of the UCSC browser [2] (Fig 2). Conservation was investigated and found within exonic domains, with less in UTRs and high intonic/intergenic divergence. Data on whole genome conservation upstream of consensus TSSs (Fig 3) indicates that conservation falls to within background levels beyond ~300bp.

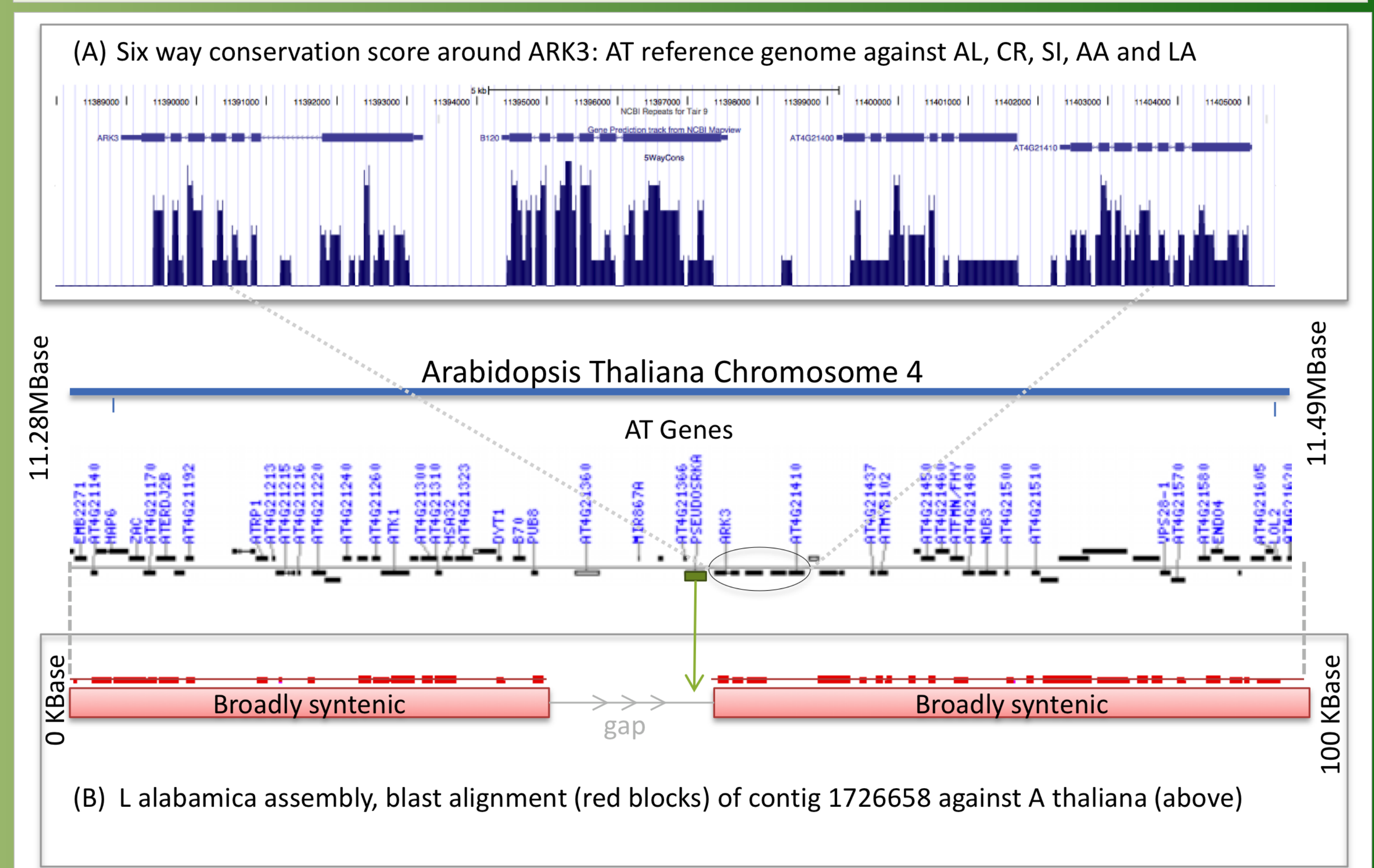
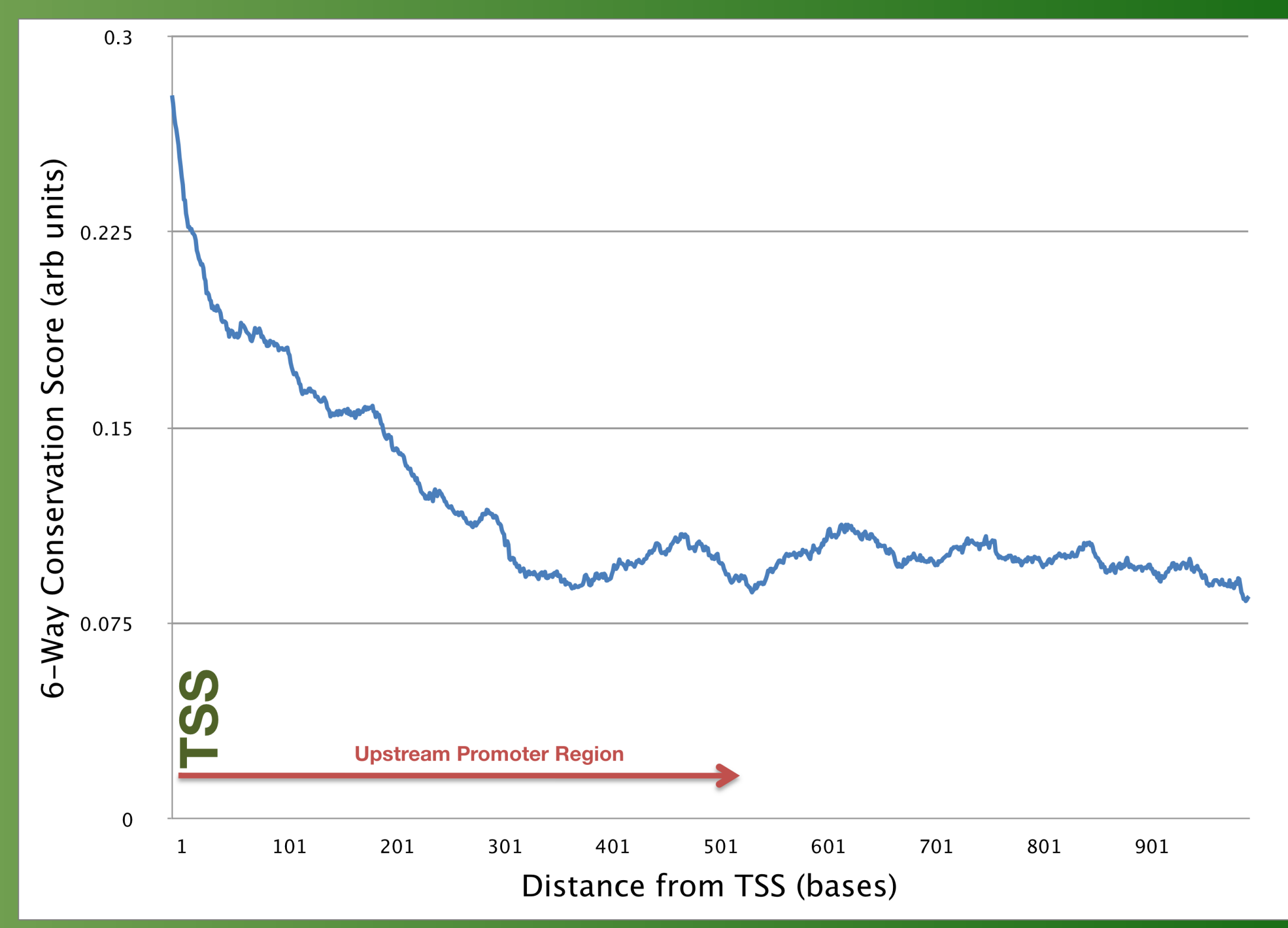


Figure 2. **Output of the genome annotation pipeline.** Following repeat characterization and masking, the assemblies are processed as a custom annotation 6-way conservation track on the genome browser (A) to show general levels of conservation in intergenic regions. Long contigs are aligned against reference genomes (*thaliana* and *lyrata*) to suggest areas of rearrangement (B). Contig 1726658 covers the majority of the AT self incompatibility locus with the exception of the gene *SRK* (green) that does not appear in this contig (gap) suggesting a possible rearrangement in LA [3].

Conclusions

- Results to date indicate a promising pipeline for the rapid assembly of high quality contig libraries. Further sequencing to scaffold the contigs using Illumina mate-pair sequencing with a 2-5kb insert size is underway
- Initial conservation results indicated conservation in exonic regions as well as an expected bias towards conservation in proximal promoter regions

Figure 3 (right), 6-way conservation between AT, AL, CR, LA, SI and AA in the upstream promoter regions of *A. thaliana* genes. Gene coordinates were taken from the NCBI mapview database with upstream regions encroaching on other genes or pseudogenes terminated prior to the most proximate feature. Work is currently underway to extend this to introns and UTRs.



References:
[1] DR Zerbino, E Birney, Genome Res. 2008. 18: 821-829
[2] D Karolchik et al, Nucleic Acids Res. 2003 Jan 1;31(1):51-4
[3] JW Busch et al, Genetics. 2008 Apr;178(4):2055-67