CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Synteny Analysis

Synteny analysis is important in the field of comparative genomics. It allows us to understand the structures and functions of groups of genes in the genomes and their roles in gene expression. In the past, the term “synteny” has been used to explain the phenomenon of co-localization of different genes in corresponding chromosomes of different species. Recently, “synteny” has been used to indicate the conservation of co-localized genes in the same order within different genomes (Vergara *et. al.*, 2010).

Synteny analysis is extensively used in plant comparative genomics because genomic microcolinearity in plants is a useful tool for plant gene identification. Plant species are tremendously diverse in their growth habits, environmental adaption and nuclear genome structures. Plant genomes have highly variable sizes but tend to be large and complex. They exhibit extensive conservation of both gene content and gene order. Often, they use homologous genes for very similar functions (Bennetzen, 2000). Many plant genome comparison studies have been carried out by different groups of researchers (Klein *et. al.*, 2003; Zhu *et. al.*, 2003; Choi *et. al*, 2004; Salse *et. al.*, 2004; Timms *et. al.*, 2006; Kumar *et. al.*, 2009) to investigate how conserved are the genomes of different plant species. For example, Ku *et. al.* (2001) investigated the synteny between arabidopsis (*A. thaliana*) and tomato, Salse *et. al.* (2002) investigated the synteny between arabidopsis and rice (*O. sativa*), Shultz *et. al.* (2007) investigated the synteny between soybean and arabidopsis, and McClean *et. al.* (2010) investigated the synteny between common bean and soybean.

Synteny blocks are genomic segments with a range of several kilobases to a few megabases long. They consist of a set of orthologous genes that share the same relative ordering or colinearity on the chromosomes of two species (Zeng *et. al.*, 2008). Generally, synteny blocks are believed to arise as a result of divergence from their last common ancestor and may be functionally important (Vergara *et. al.*, 2009). Genes within synteny block are often co-regulated and share similar functions. They may be under some selective pressure that prevents genes within it from escaping the block due to the functional significance of those genes. A synteny block may become complex when it forms various types of functional clusters and topological arrangements. Identification of a synteny block is therefore crucial because it may provide clues regarding gene and regulatory element arrangements that are essential for biological processes (Zeng *et. al.*, 2008; Vergara *et. al.*, 2009; Vergara *et. al.*, 2010). Synteny blocks may further be classified into conserved and non-conserved blocks. A conserved synteny block is defined as block of genes that preserves the ordering or strandedness and no mismatch within the block. A non-conserved synteny block is defined as block of genes that preserves the ordering or strandedness but has mismatch within the block.

In the early days, ad hoc methods were used to identify synteny blocks. Those ad hoc methods tended to be slow, not fully reproducible, ignore strandedness (conservation of order and orientation), and inappropriate for general applications. In contrast, computational approaches are probably more effective if they implement efficient algorithms. OrthoCluster is one such computational program that is designed for identification of synteny blocks among multiple genomes if the orthologous relationships that exist among the input genomes are given (Zeng *et. al.*, 2008; Vergara *et. al.*, 2009; Vergara *et. al.*, 2010).

OrthoCluster able to handles some challenging application requirements in synteny analysis. For example, it can include strandedness of genes into its analysis, detect gene inversions or duplications, and permit interruptions within synteny blocks by tolerating different degrees of mismatches. It can be used to compare more than two genomes in a single analysis. Moreover, OrthoCluster is also able to resolve one-to-many orthologous relationships and to identify four types of genome rearrangement events: inversions, transposition, insertion or deletion, and reciprocal translocation (Zeng *et. al.*, 2008; Vergara *et. al.*, 2009; Vergara *et. al.*, 2010). Recently, Vergara *et. al.* (2010) reported the successful identification of synteny blocks between the genomes of two closely related hermaphrodite nematodes: *Caenorhabditis elegans* and *Caenorhabditis briggsae*, using OrthoCluster. With the availability of an online platform called OrthoClusterDB (Ng *et. al.*, 2009), which allows user to identify and visualize synteny blocks among multiple genomes using OrthoCluster, it seems that this bioinformatics tool may become important to researchers interested in doing synteny analysis in their comparative genomics project.

1.2 Orthologous Gene Identification

To run OrthoCluster, the user has to first identify orthologous relationships between the genomes of interest. This requires an understanding of the concept of orthology. Briefly, homologous sequences between two genomes share a common ancestry, and can be characterised as either orthologs or paralogs. Orthologs are genes in different species that originate from a single gene in the last common ancestor of these species and often retain identical or similar biological functions (Remm *et. al.*, 2001). Paralogs are homologous genes that are related through duplication within a genome. Orthologs are more likely to share similar function compared to paralogs because paralogs usually have undergone point mutation and domain recombination that lead to changes in substrate or ligand specificity of the protein (Hulsen *et. al*., 2006; Chen *et. al*., 2007). Paralogy can exist between genes in different species because gene duplication events can occur both before and after speciation. Paralogs that duplicate after speciation are called 'inparalogs' while paralogs that duplicate before speciation are called 'outparalogs'. Inparalogs can form a group of genes that have orthologous relationship with a gene in another species; outparalogs can never be orthologs and they can easily be confused with true orthologs (O'Brien *et. al*., 2001; Remm *et. al*, 2001). Figure 1.1 shows the relationship between inparalogs and outparalogs in a hypothetical gene tree.

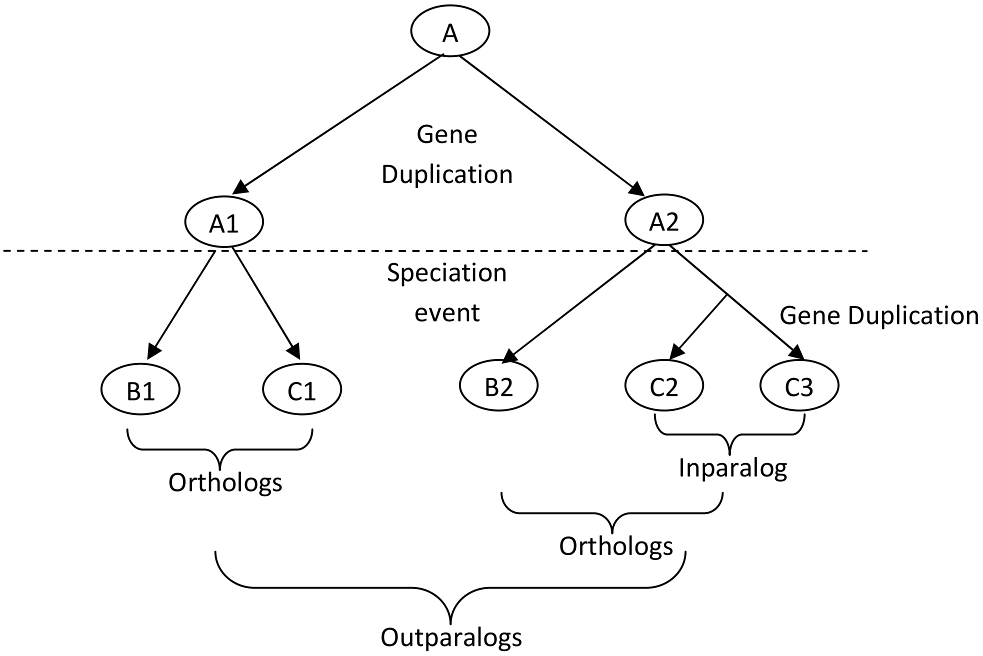


Figure 1.1: In this hypothetical gene tree, Gene A in an ancestral species 'A' undergoes a gene duplication event giving rise to A1 and A2 genes. After that, speciation event occurs leading to two lineages 'B' and 'C'. The genes C2 and C3 in the C genome are inparalogs since their gene duplication occurred after speciation and they are orthologous to the B2 gene because they share a common ancestral gene A2. B1 is an outparalog of the B2, C2 and C3 genes.

Orthology analysis becomes difficult when there are large numbers of paralogs within protein families. Eukaryotic genomes present further challenges to orthology analysis because of their large genome size, difficulty in defining accurate gene models, complexity of protein domain architecture and high number of gene duplication events (Chen *et. al.*, 2007). Phylogenetic methods can be used to detect orthologs and inparalogs but may slow significantly when the number of sequences increases. On the other hand, automatic clustering methods based on two-way best genome-wide matches have so far effectively identified orthologous genes and inparalogs (Remm *et. al*, 2001).

Basic Local Alignment Search Tool (BLAST) is the most common tool for identifying orthologous genes among different genomes. When two sequences are aligned, BLAST produces high-scoring pairs (HSPs) that consist of arbitrary sequence fragments of equal length. HSP is the local sequence alignment where the alignment score passes the user-defined cut-off score. Identification of HSP by BLAST is based on several statistical criteria such as the E-value, alignment score, and percentage of identity (Altschul *et. al.*, 1990).

However, using the default parameters in BLAST can lead to misidentification of orthologous regions due to detection of the gene family's members that are not truly orthologous (Salse *et. al.*, 2008). Furthermore, the alignment of conserved domains in non-orthologous gene in BLAST may complicate the interpretation of orthology analysis results. In order to address the difficulty of inferring orthologous and paralogous relationships from sequence comparisons, stringent alignment criteria are essential to evaluate the reliability of the BLAST sequence alignment results, and to avoid identification of false syntenic regions through the alignment of similar but non-orthologous sequences (Salse *et. al.*, 2002; Salse *et. al.*, 2008; Salse *et. al.*, 2009).

InParanoid is a program that can be used to identify orthologs and inparalogs between any given pair of genomes (Remm *et. al*, 2001). This program was developed specifically to identify clusters of true orthologs while avoiding inclusion of closely related but non-orthologous genes (O'Brien *et. al.*, 2001). The idea behind InParanoid is that if a set of sequences are orthologs, they should score higher with each other than with any other sequence in the genome (Remm *et. al*, 2001). Thus, the methodology of InParanoid can be seen as an extension of all-versus-all sequence comparison technique but with special rules for cluster analysis in order to extract inparalogs (Ostlund *et. al.*, 2010).

The InParanoid algorithm relies on BLAST as the underlying homology detection tool and uses a clustering algorithm to detect inparalogs. At first, it uses BLAST scores to measure the relatedness of proteins and construct orthology groups. An orthology group is initially composed of two seed orthologs that are found by two-way best hits between two proteomes. After that, sequences that are closer to the corresponding seed ortholog will be added to the orthology group to form a paralogous cluster. InParanoid will assign confidence values (relative scale between 0%-100%) for all paralogs in each group to show the degree of relatedness to its seed ortholog.

Orthology analysis is important for annotating function accurately and has been widely used to facilitate comparative and evolutionary genomics studies (Chen *et. al.*, 2007). Unfortunately, there are only a few publications available to assess the quality of different ortholog database by looking into either the accuracy of functional annotation or the inferred accuracy (Hulsen *et. al.*, 2006; Chen *et. al.*, 2007; Altenhoff *et. al.*, 2009). According to Chen *et. al.* (2007), reducing the E-value cutoff for BLAST-based methods improves specificity and decreases sensitivity (lower false positives and higher false negatives). Homology-based detection methods are more sensitive to E-value cutoff especially at low E-values, which reduce false positive rate drastically (Chen *et. al*., 2007). On the other hand, InParanoid ranks within the top three in both assessments conducted by Hudson *et. al*. (2006) and Chen *et. al.* (2007). This suggests that InParanoid was able to balance the false negative and false positive rate.

Altenhoff *et. al.* (2009) evaluated the accuracy of several public available ortholog inference projects such as COG, KOG, Inparanoid, OrthoMCL, Ensembl Compara, Homologene, RoundUp, EggNOG and OMA. They also compared two standard methods, bidirectional best-hit (BBH) and reciprocal smallest distance (RSD), which are popular in orthology analysis. They found that these different databases had their own strengths and weaknesses, depending on user requirement such as the level of specificity or sensitivity. Overall performance of simple BBH achieved good results, and the predicted orthologs showed close functional relatedness.

1.3 Objectives of Study

In this project, we will evaluate the efficacy of two orthologous gene identification methods: InParanoid and ad hoc BLAST. We will compare the synteny blocks returned by OrthoCluster using these two data preparation methods in terms of the total number of blocks identified, percentage of identified synteny blocks, as well as method-specific blocks. This will provide us with some ideas about the sensitivity of OrthoCluster to methods of identifying orthologous genes.

CHAPTER 2

METHODOLOGY

Complete protein sequences and genome annotation data of *Oryza sativa* and *Arabidopsis thaliana* were downloaded from the International Rice Genome Sequencing Project website (IRGSP; http://rgp.dna.affrc.go.jp/IRGSP), the Rice Annotation Project website (RAP; http:// rapdb.dna.affrc.go.jp) and the Arabidopsis Information Resource website (TAIR; http://www.arabidopsis.org). IRGSP is a consortium of publicly funded laboratories. It was established in 1997 to obtain high quality, map-based sequence of the rice genome using the cultivar Nipponbare of *Oryza sativa* ssp. japonica. The RAP aims to provide the scientific community with an accurate and timely annotation of the rice genome sequence. TAIR is a genetic and molecular biology database for *Arabidopsis thaliana*. These two genomes were selected for the present study because of their completeness and extensive curation.

The evaluation of the efficacy of the two orthologous gene identification methods involved the following process:

(i) Identification of orthologous genes between *O. sativa* and *A. thaliana* using ad hoc BLAST and InParanoid.

(ii) OrthoCluster identification of synteny blocks using orthologous gene datasets identified in step (i).

(iii) Comparison of synteny blocks from different orthologous gene identification method.

The representative gene is the longest protein for each gene. It was used in the analysis to prevent different transcripts of the same gene from being assigned to different ortholog groups. In addition, three types of orthologous relationships were considered in the orthologous genes identification analysis. The one-to-one relationship represents a relation that a gene in one species (e.g. *O. sativa*) has only one orthologous gene in another species (e.g. *A. thaliana*). The one-to-many and many-to-one relationships represent relations that a gene in one species has multiple orthologous genes in another species and vice versa.

2.1 Identification of Orthologous Genes by Ad hoc BLAST

Since BLAST is a local alignment algorithm, conserved domain between parts of proteins may report high-scoring matches even though they do not reflect a common origin for the proteins as a whole. In order to increase stringency and significance of BLAST sequence alignment, three additional parameters were used. These were aligned lengths (AL) cumulative alignment length percentage (CALP) and cumulative alignment length identity percentage (CALIP). Briefly, AL corresponds to the sum of all HSP lengths; CALP corresponds to AL divided by the length of the query sequence length or match sequence length; CALIP corresponds to cumulative percentage of sequence identity obtained for all of the HSPs divided by query sequence length or match sequence length.

These additional parameters can avoid short, domain level matches when using BLAST to infer sequence homology. In particular, the CALP and CALIP parameters allow the identification of the best alignment (i.e., the highest cumulative percentage of identity in the longest cumulative length). BLAST results were parsed using the BioPerl module (http://bioperl.org) in order to calculate AL, CALP and CALIP.

All-against-all intra-species BLASTP search was performed for protein sequences of the species itself to identify putative duplicate or unique genes. Significant matches were claimed only when (a) the CALP value was 70% or more, (b) the CALIP value was 70% or more, (c) the BLAST E-value was less than 10^{-20}, (d) the BLAST output suggested low copy number in the *O. sativa* or *A. thaliana* genome (less than three).

After intra-species BLAST was carried out for both species, the protein sequences that statisfied the filtering criteria were extracted separately and used to perform inter-species BLAST in order to identify orthologous gene between species. Significant matches were claimed only when (a) the CALP value was 50% or more, (b) the CALIP value was 50% or more, (c) the BLAST E-value was less than 10^{-20}. Inter-species BLAST generated three types of orthologous relationships: one-to-one, one-to-many and many-to-one.

2.2 Identification of Orthologous Genes by InParanoid

Protein sequences of *O. sativa* and *A. thalina* were used as input files for stand-alone InParanoid program. InParanoid uses a two-pass BLAST approach to increase specificity and sensitivity for homology detection. All the parameters for the two-pass BLAST approach were followed the default values as published in Ostlund *et. al.* (2010). The BLAST result was considered statistically significant if the BLAST score threshold was more than 40.

The low-complexity filter used in InParanoid provides compositional adjustment and SEG (Wootton *et. al.*, 1993) low-complexity filter in the first step of the two-pass BLAST approach. The low-complexity filter masks the sequence only during seeding phase but not during extension phase of BLAST. This was done to reduce the false matches resulting from unrelated proteins sharing repetitive regions by chance or regions with very biased amino acid composition. However, compositional adjustment often produces short alignments which will affect the result of subsequent steps. In order to overcome this issue, matches accepted in the first pass were realigned using BLAST with SEG and compositional adjustment switched off before subsequent overlap criteria were applied (Ostlund *et. al.*, 2010).

In the second step of two-pass BLAST approach, match area must cover at least 50% of the length of the sequence. Furthermore, the sum of the length of the aligned regions on that sequence must cover at least 25% of the length of the sequence. When there are multiple HSPs, InParanoid requires that they maintain the same relative order on both sequences, and they do not overlap by more than 5% (Ostlund *et. al.*, 2010).

After the two-pass approach of BLAST, InParanoid will apply clustering algorithm to identify inparalogs. Under clustering algorithm, the ortholog detection starts with finding mutually best scoring sequence pairs, bi-directionally best hits between datasets *O. sativa* (A) and *A. thaliana* (B). These mutually best hits are marked as the main ortholog pair of a given ortholog group. The main ortholog pairs serve as central points around which additional orthologs (inparalogs) from both species will be clustered in later steps.

In the case of overlap between two groups, the overlapping groups are merged, deleted, or separated depending on the type and extent of overlap. According to Remm *et. al.* (2001), the rules are applied in the following order: (1) merge groups if main orthologs A_2 and B_2 are already clustered in a stronger group A_1-B_1; (2) merge groups if main ortholog B has equally best hit to two orthologs from species A, A_1 and A_2; (3) delete new group if one of the main orthologs A_2 already belongs to a much stronger group; (4) merge groups if one of the main orthologs already has a high confidence value in another group.

Finally, confidence values were assigned to all paralogs in each group. The confidence value (range from 0% to 100%) indicates how far a given sequence is from the main ortholog of the same species. On this scale, 100% was assigned to the main ortholog and inparalogs of that particular group with assigned confidence value that less than 100%.

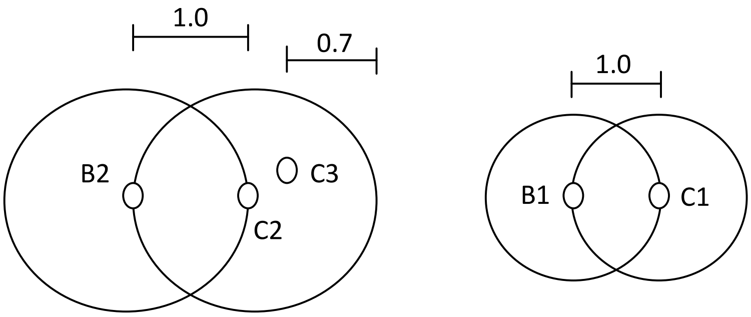


Figure 2.1: Clustering of additional orthologs (inparalogs). Each circle represents sequence from species B or species C. B2 and C2 are the original seed-ortholog pair with an inparalog score of 1.0 (all inparalogs are clustered around this pair). Other inparalogs, C3, are scored according to the relative similarity to the seed-inparalog, C2. The score is the reverse distance between pairwise comparison of sequences, in this case, C2 is relatively more similar to B2 than C3, thus C3 receives a lower inparalog score (0.7). Sequences outside the circle are classified as outparalogs; thus C1 and B1 form a cluster of their own. C1 and B1 are orthologous to each other.

2.3 Identification of Synteny Blocks by Orthocluster

Three datasets were prepared as required by stand-alone version of OrthoCluster to identify synteny blocks: (1) *O. sativa* genes ordered according to their genomic coordinates, (2) *A. thaliana* genes ordered according to their genomic coordinates, (3) a *O. sativa-A. thaliana* gene correspondence file (ad hoc BLAST or InParanoid). The first and second datasets were extracted from *O. sativa* and *A. thaliana* genome annotation data, respectively. Third dataset was extracted from the results of InParanoid or ad hoc inter-species BLAST, which provide the putative orthologous relationships between proteins in these two genomes.

Two types of synteny blocks were considered in this analysis: conserved and non-conserved synteny block. Two values of mismatches were used for non-conserved synteny block analysis: (a) 10% for both in-map mismatch and out-map mismatch; (b) 50% for both in-map mismatch and out-map mismatch. In-map genes are genes with orthologous relationships in the correspondence dataset. In-map mismatch refers to ortholog genes that have no correspondence in the paired genome. Out-map genes are genes without orhologous relationships that are not included in the correspondence dataset. An out-map mismatch refers to non-ortholog genes within the block (Figure 7.5 and 7.6).

For both conserved synteny block and non-conserved synteny block, OrthoCluster only report synteny blocks with sizes do not exceed the defined maximum synteny block size which was 1000 genes per block. In addition, OrthoCluster only searched for synteny blocks where strandedness of genes was enforced and the gene ordering was preserved. These parameters were used in search for synteny block in four conditions: (1) consistent order and consistent strandedness; (2) consistent order and reversed strandedness; (3) inverted order and consistent strandedness; (4) inverted order and reversed strandedness (Figure 7.1 to 7.4).

2.4 Visualisation by Circos

Circos (Krzywinski *et al*., 2009) is a software package for visualising comparative genomic data in a circular layout. The edges linking components of the genomes (e.g. chromosomes) that are compared enable the user to study genome organisation at the macro level. Circos can be used to create publication-quality graphics with high data-to-ink ratio. In this project, Circos plots of gene regions and orthologous genes identified by ad hoc BLAST and InParanoid for both species were made. Synteny blocks identified by OrthoCluster from both orthologous gene datasets were plotted using Circos as well to visualise the distribution of the synteny blocks on each chromosome.

CHAPTER 3

RESULTS

3.1 Orthologous Relationships

Results returned from the identification of orthologous genes between *O. sativa* and *A. thaliana* showed that InParanoid identified more orthologous genes than ad hoc BLAST. InParanoid identified 31% of genes in *O. sativa* and 39% of genes in *A. thaliana* as orthologs; ad hoc BLAST only identified 18% of genes in *O. sativa* and 22% of genes in *A. thaliana* as orthologs (Table 3.1). About 20% of *O. sativa* genes and 27% of *A. thaliana* genes were found to form one-to-one orthologous relationships using InParanoid. In contrast, only 9% of *O. sativa* genes and 12% of *A. thaliana* genes were found to form one-to-one orthologous relationships using ad hoc BLAST. The number of orthologs for one-to-many orthologous relationship inferred using InParanoid method ranged from 2 to 119 in *O. sativa,* and 2 to 23 in *A. thaliana*; for ad hoc BLAST, it was 2 to 19 for both *O. sativa* and *A. thaliana*.

Table 3.1: Orthologous relationships identified from ad hoc BLAST and InParanoid.

	Ad hoc BLAST		**InParanoid**	
	O. sativa	***A. thaliana***	***O. sativa***	***A. thaliana***
Initial number of genes	40,353	35,386	40,353	35,386
Representative genes	33,276	27,416	33,276	27,416
Unique genes	32,265	25,926	NA	NA
Orthologous genes	7,220	7,730	12,475	13,741
One-to-one orthologous relations	3,691	4,317	7,902	9,572
One-to-many orthologous relations	3,529	3,413	4,573	4,169
Total orthologous relations	14,928	14,928	22,124	22,124

Generally, the InParanoid method identified more orthologous gene even at higher stringency value (higher number of confidence value) compared to ad hoc BLAST (Table 3.2). The highest number of one-to-one orthologous gene relationship was between chromosome 3 from *O. sativa* and chromosome 1 from *A. thaliana* for both methods (Figure 3.1, Table 7.1). Chromosome 3 from *O. sativa* and chromosome 1 from *A. thaliana* also had the highest number of orthologous genes for both methods among the rest of the chromosomes from these two species. A Venn diagram (Figure 3.2) shows that a total of 29 795 orthologous genes were found using both methods; 24% were shared, 50% were InParanoid-specific and 26% were ad hoc BLAST specific.

Table 3.2: Number of orthologous gene identify by InParanoid method with different confidence value and above the cutoff. Inparalogs with confidence value less than 0.05% was not shown by default setting of InParanoid.

Confidence value	**Orthologous genes**		**One-to-one**		**One-to-many**	
	O. sativa	***A. thaliana***	***O. sativa***	***A. thaliana***	***O. sativa***	***A. thaliana***
0.05%	12,475	13,741	7,902	9,572	4,573	4,169
10%	12,010	13,282	7,675	9,311	4,335	3,971
20%	11,318	12,477	7,375	8,836	3,943	3,641
30%	10,825	11,784	7,138	8,459	3,687	3,325
40%	10,424	11,156	6,953	8,111	3,471	3,045
50%	10,138	10,566	6,812	7,766	3,326	2,800
60%	9,971	10,127	6,738	7,483	3,233	2,644
70%	9,856	9,819	6,679	7,294	3,177	2,525
80%	9,765	9,668	6,635	7,213	3,130	2,455
90%	9,667	9,575	6,582	7,159	3,085	2,416
100%	9,566	9,535	6,518	7,135	3,048	2,400

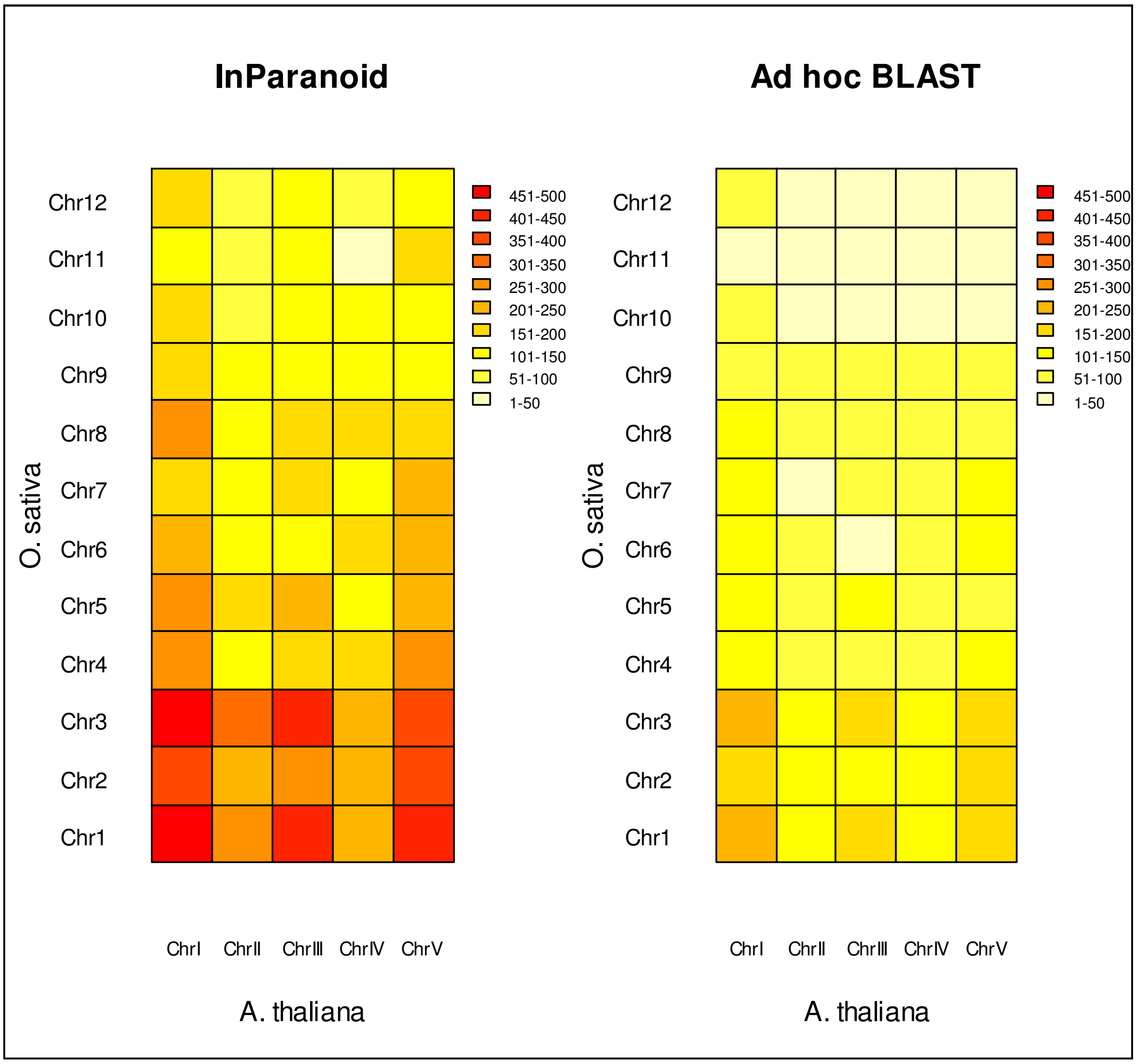


Figure 3.1: Heatplot of one-to-one orthologous relationships between *O. sativa* (row) and *A. thaliana* (column) from InParanoid and ad hoc BLAST. The colours of each block represent the number of orthologous relationships between the chromosome of *O sativa* and *A. thaliana*.

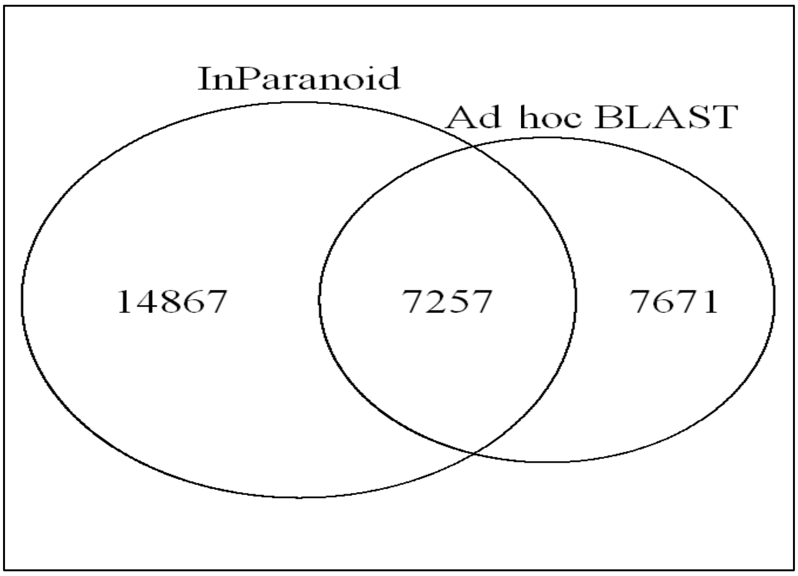


Figure 3.2: Venn diagram for the three types of orthologous relationships: orthologous relationships shared by both methods, InParanoid-specific and ad hoc BLAST-specific orthologous relationships.

3.2 Conserved Synteny Blocks

We identified 423 conserved synteny blocks using OrthoCluster and orthologous relationships from ad hoc BLAST. Out of these, 314 were non-nested blocks and 109 were nested blocks (Figure 3.3, Table 7.2 & 7.3). A nested block consists of subset of genes within a larger synteny block that is found duplicated in different genomic regions in either the same or different chromosomes (Vergara *et. al.*, 2010). The longest conserved synteny block spanned 116.3 kb in *O. sativa* (2 genes in chromosome 12) and 20.5 kb on *A. thaliana* (3 genes in chromosome 1) (Figure 3.4 & 3.5, Table 7.2 & 7.3); the largest gene-rich synteny block contained 3 genes in *O. sativa* (spanned 52 kb in chromosome 7) and 4 genes in *A. thaliana* (spanned 16.1 kb in chromosome 4) (Figure 3.4 & 3.5, Table 7.2 & 7.3). Altogether, conserved synteny blocks covered 427 genes in *O. sativa* and 435 genes in *A. thaliana*, corresponding to 2.3 Mb in *O. sativa* (0.6% of the genomic sequence) and 1.1 Mb in *A. thaliana* (0.9% of genomic sequence). Chromosome 2 in *O. sativa* was the one with highest genomic coverage by synteny blocks (42 synteny blocks, covering around 301 kb); in *A. thaliana*, chromosome 1 had the highest genomic coverage by synteny blocks (60 synteny blocks, covering around 273 kb) (Figure 3.6 & 3.7, Table 7.2 to 7.4).

In the case of orthologous relationships from InParanoid, OrthoCluster identified 1589 conserved synteny blocks. Out of these, 942 were non-nested blocks and 647 were nested blocks (Figure 3.3, Table 7.2 & 7.3). The longest conserved synteny block spanned 149.8 kb on *O. sativa* (5 genes in chromosome 6) and 39.9 kb on *A. thaliana* (7 genes on chromosome 3) (Figure 3.4, 3.5) while the largest gene-rich synteny block contained 7 genes on *O. sativa* (spanned 13.7 kb in chromosome 8) and 7 genes on *A. thaliana* (spanned 39.9 kb in chromosome 3) (Figure 3.4 & 3.5, Table 7.2 & 7.3). Altogether, conserved synteny blocks covered 1234 genes in *O. sativa* and 1403 genes in *A. thaliana,* corresponding to 5.97 Mb in *O. sativa* (1.56% of the genomic sequence) and 2.76 Mb in *A. thaliana* (2.31% of genomic sequence). Chromosome 4 in *O. sativa* had the highest genomic coverage by synteny blocks (87 synteny blocks, covering about 658 kb); chromosome I in *A. thaliana* had the highest genomic coverage by synteny blocks (171 synteny blocks, covering around 800 kb) (Figure 3.6 & 3.7, Table 7.2 to 7.4).

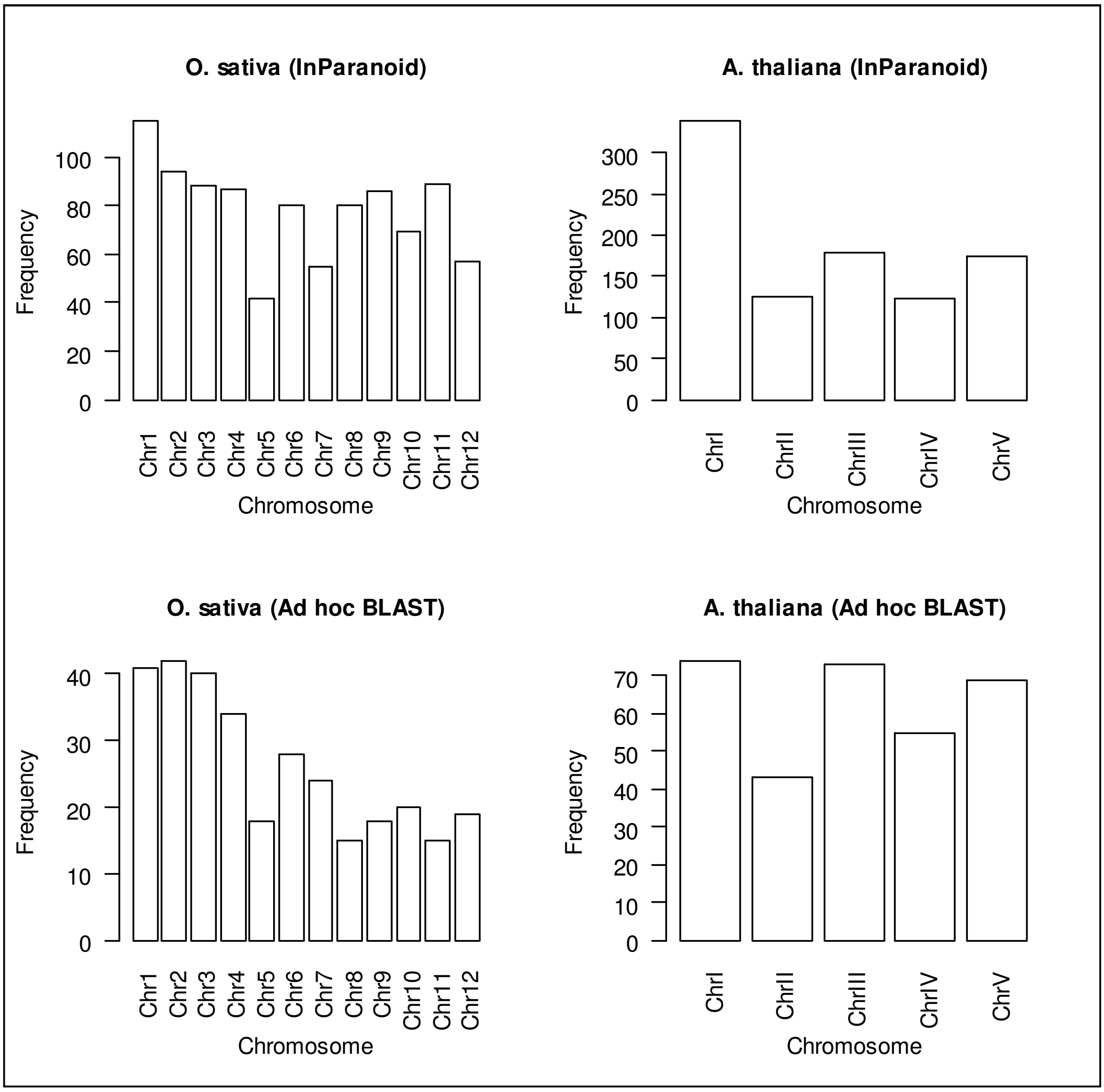


Figure 3.3: Distribution of conserved synteny block numbers for *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Barplots on top panel are distribution of conserved synteny block numbers identify from orthologous gene datasets of InParanoid. Barplots on bottom panel are distribution of conserved synteny block numbers identify from orthologous gene datasets of ad hoc BLAST. The height of the bar represents the number of conserved synteny blocks for each chromosome.

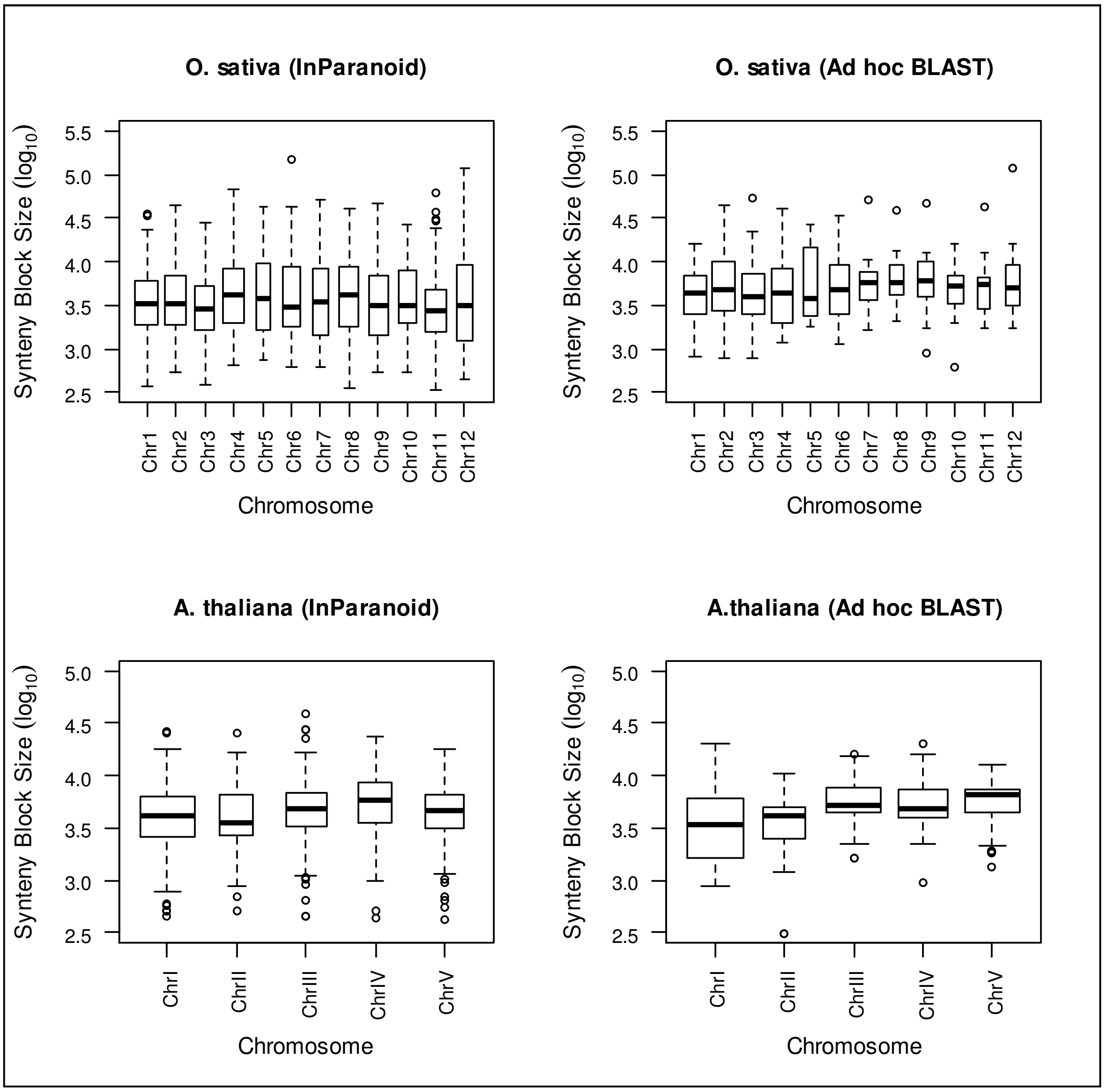


Figure 3.4: Distribution of conserved synteny blocks size in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Boxplots on left panel are distribution of conserved synteny blocks sizes identify from orthologous gene datasets of InParanoid. Boxplots on right panel are distribution of conserved synteny blocks sizes identified from orthologous gene datasets of ad hoc BLAST. The y-axis is plotted in log-scale of base 10.

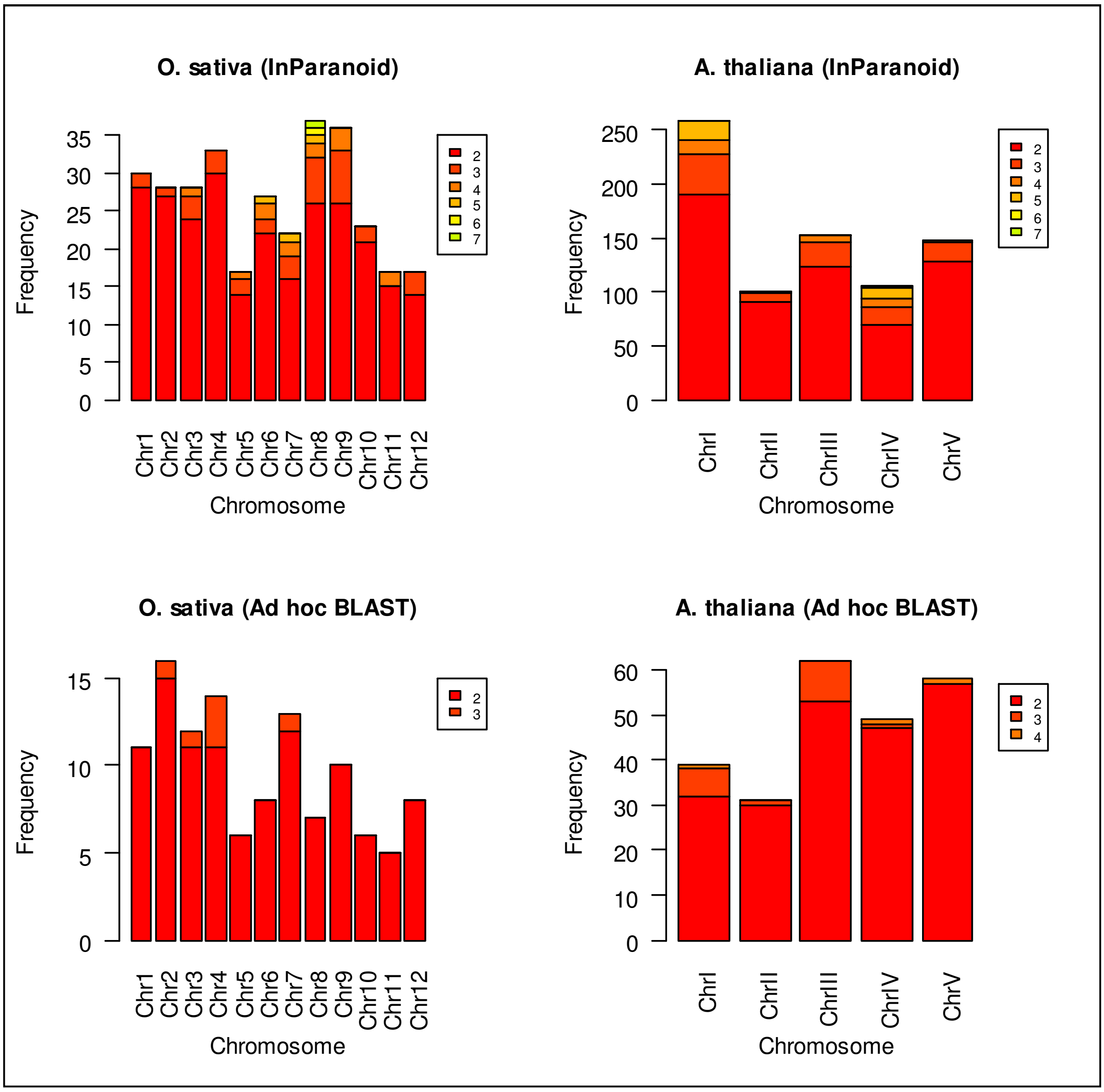


Figure 3.5: Distribution of gene numbers of each conserved synteny block in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Barplots on top panel are distribution of gene numbers of each conserved synteny block identify from orthologous gene datasets of InParanoid. Barplots on bottom panel are distribution of gene numbers of each conserved synteny block identify from orthologous gene datasets of ad hoc BLAST. The colours of the bar represent the number of genes of the conserved synteny block. The height of the bar represents the number of conserved synteny blocks with specified number of genes within the block.

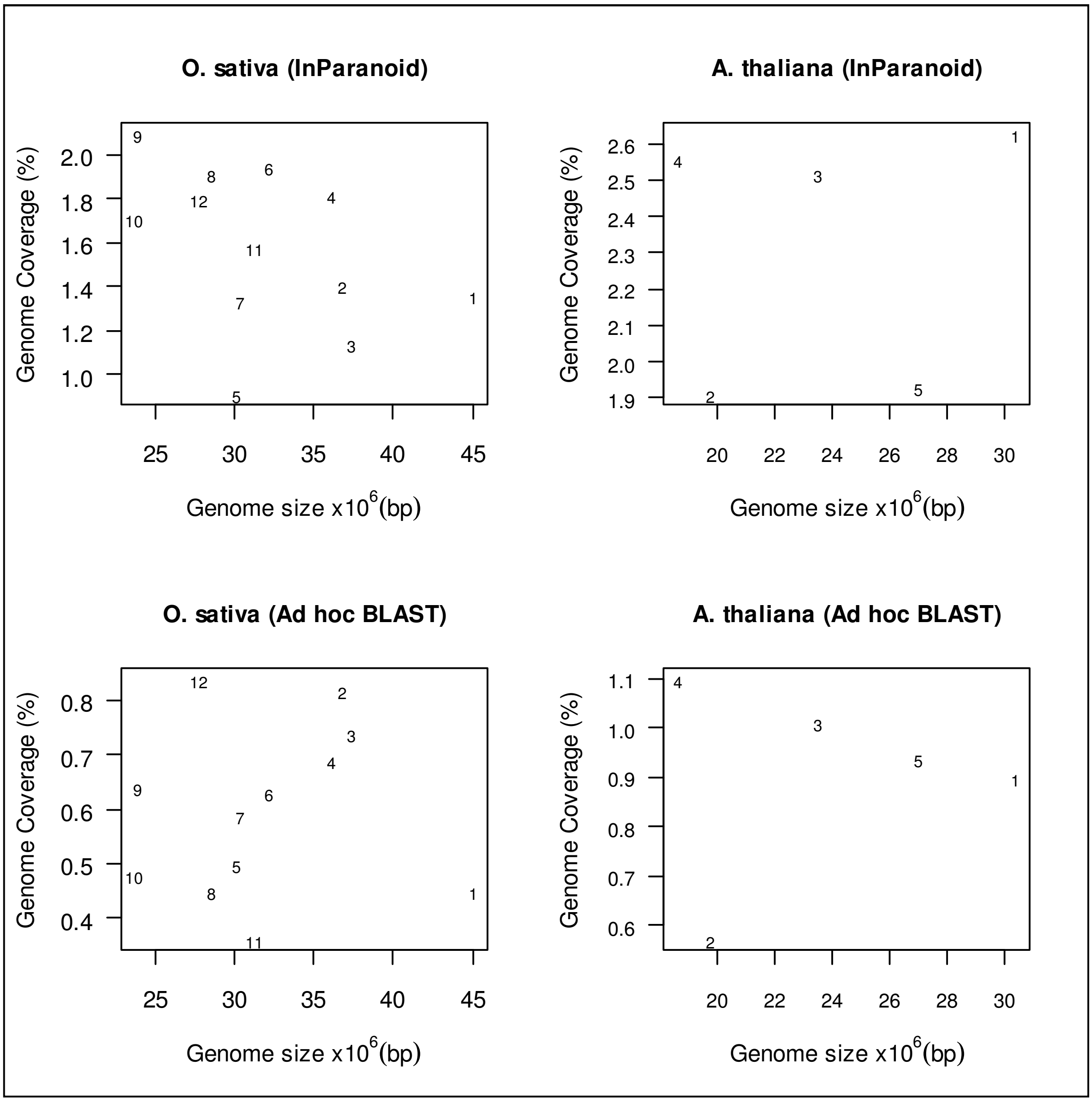


Figure 3.6: Coverage of conserved synteny blocks on each chromosome in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Scatterplots on top panel are coverage of conserved synteny blocks on each chromosome identify from orthologous gene datasets of InParanoid. Scatterplots on bottom panel are coverage of conserved synteny blocks on each chromosome identify from orthologous gene datasets of ad hoc BLAST. The numbers in the scatterplot represent the number of chromosome of each species. The genome sizes of each chromosome were plotted on x-axis while the coverage of conserved synteny blocks on each chromosome was plotted on y-axis.

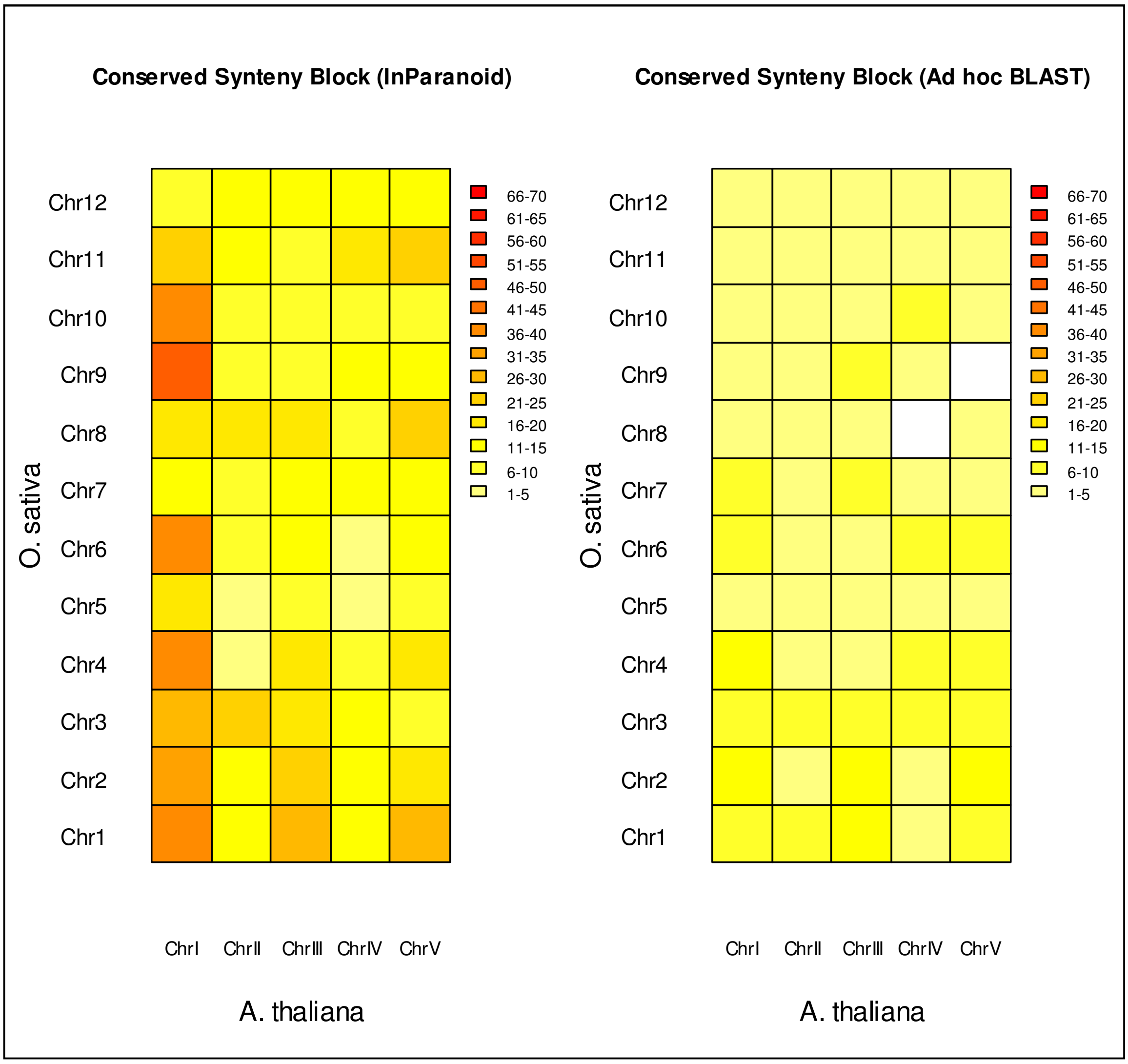


Figure 3.7: Distribution of conserved synteny blocks between *O. sativa* (row) and *A. thaliana* (column) identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. The colours of the each block represent the number of conserved synteny blockd between the chromosome of *O sativa* and *A. thaliana*.

3.3 Non-conserved Synteny Blocks

OrthoCluster identified 423 non-conserved synteny blocks for 10% mismatches and 790 non-conserved synteny blocks for 50% mismatches using orthologous relationships from ad hoc BLAST. There were 314 non-nested blocks and 109 nested blocks for 10% mismatches; for 50% mismatches, there were 589 non-nested blocks and 201 nested blocks (Figure 3.8, Table 7.5 & 7.6). For 10% mismatches, the longest non-conserved synteny block spanned 116.3 kb on *O. sativa* (2 genes in chromosome 12) and 20.5 kb on *A. thaliana* (3 genes in chromosome I) (Figure 3.9 & 3.10, Table 7.5 & 7.6) while the largest gene-rich synteny block contained 3 genes in *O. sativa* (spanned 52 kb in chromosome 7) and 4 genes in *A. thaliana* (spanned 16.1 kb in chromosome 4) (Figure 3.9 & 3.10, Table 7.5 & 7.6). For 50% mismatches, the longest non-conserved synteny block spanned 209.8 kb on *O. sativa* (10 genes in chromosome 2) and 38.6 kb on *A. thaliana* (10 genes in chromosome 3) (Figure 3.9 & 3.10, Table 7.5 & 7.6) while the largest gene-rich synteny block contained 15 genes on *O. sativa* (spanned 102.8 kb in chromosome 3) and 10 genes on *A. thaliana* (spanned 38.6 kb in chromosome 3) (Figure 3.9 & 3.10, Table 7.5 & 7.6).

Altogether, non-conserved synteny blocks at 10% mismatches covered 427 genes in *O. sativa* and 435 genes in *A. thaliana*, corresponding to 2.3 Mb in *O. sativa* (0.6% of the genomic sequence) and 1.1 Mb in *A. thaliana* (0.9% of genomic sequence). Chromosome 2 in *O. sativa* was the one with highest genomic coverage by synteny blocks (42 synteny blocks, covering about 301 kb) while chromosome I in *A. thaliana* was the one with highest genomic coverage by synteny blocks (60 synteny blocks, covering around 273 kb) (Figure 3.11 & 3.12, Table 7.5 to 7.7). Non-conserved synteny blocks at 50% mismatches covered 1335 genes in *O. sativa* and 1257 genes in *A. thaliana*, corresponding to 8.22 Mb in *O. sativa* (2.15% of the genomic sequence) and 3.32 Mb in *A. thaliana* (2.78% of genomic sequence). Chromosome 3 in *O. sativa* was the one with highest genomic coverage by synteny blocks (82 synteny blocks, covers about 1.29 Mb) while chromosome I in *A. thaliana* was the one with highest genomic coverage by synteny blocks (149 synteny blocks, covers about 885 kb) (Figure 3.11 & 3.12, Table 7.5 to 7.7).

Using orthologous relationships from InParanoid, OrthoCluster identified 1584 non-conserved synteny blocks for 10% mismatches and 2325 non-conserved synteny blocks for 50% mismatches. There were 942 non-nested blocks and 642 nested blocks for 10% mismatches; for 50% mismatches, there were 1510 non-nested blocks and 815 nested blocks (Figure 3.8, Table 7.5 & 7.6). In 10% mismatches, the longest non-conserved synteny block was found to span 149.8 kb on *O. sativa* (5 genes in chromosome 6) and 39.9 kb on *A. thaliana* (7 genes in chromosome 3) (Figure 3.9 & 3.10, Table 7.5 & 7.6); while the largest gene-rich synteny block contains 11 genes on *O. sativa* (spanned 20.9 kb in chromosome 8) and 7 genes on *A. thaliana* (spans 39.9 kb in chromosome 3) (Figure 3.9 & 3.10, Table 7.5 & 7.6). On the other hand, in 50% mismatches, the longest non-conserved synteny block spanned 307.4 kb on *O. sativa* (11 genes in chromosome 6) and 103.8 kb on *A. thaliana* (8 genes in chromosome 1) while the largest gene-rich synteny block contained 30 genes in *O. sativa* (spanned 124.2 kb in chromosome 10) and 24 genes in *A. thaliana* (spanned 61.8 kb in chromosome 5) (Figure 3.9 & 3.10, Table 7.5 & 7.6).

Altogether, non-conserved synteny blocks at 10% mismatches covered 1237 genes in *O. sativa* and 1403 genes in *A. thaliana*, corresponding to 6.03 Mb in *O. sativa* (1.58% of the genomic sequence) and 2.76 Mb in *A. thaliana* (2.31% of genomic sequence). Chromosome 4 in *O. sativa* had highest genomic coverage by synteny blocks (87 synteny blocks, covering about 658 kb) while chromosome I in *A. thaliana* was the one with highest genomic coverage by synteny blocks (171 synteny blocks, covering about 800 kb) (Figure

3.11 & 3.12, Table 7.5 to 7.7). Non-conserved synteny blocks at 50% mismatches covered 3509 genes in *O. sativa* and 3648 genes in *A. thaliana*, corresponding to 25.1 Mb in *O. sativa* (6.55% of the genomic sequence) and 9.06 Mb in *A. thaliana* (7.57% of genomic sequence). Chromosome 3 in *O. sativa* had the highest genomic coverage by synteny blocks (180 synteny blocks, covers about 3.11 Mb) while chromosome I in *A. thaliana* had the highest genomic coverage by synteny blocks (518 synteny blocks, cover around 2.43 Mb) (Figure 3.11 & 3.12, Table 7.5 to 7.7).

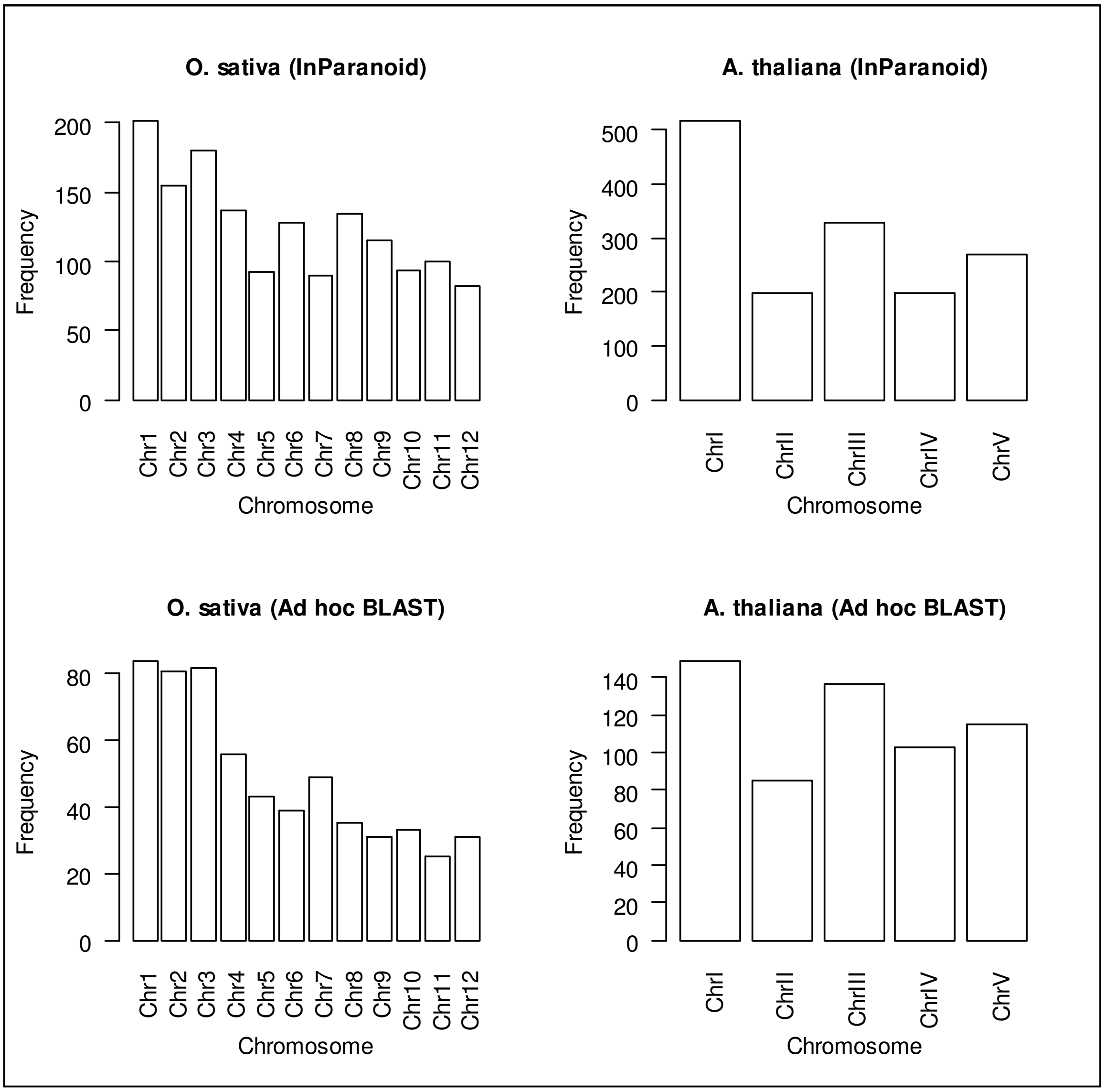


Figure 3.8: Distribution of non-conserved synteny block numbers for *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Barplots on top panel are distribution of non-conserved synteny block numbers identify from orthologous gene datasets of InParanoid. Barplots on bottom panel are distribution of non-conserved synteny block numbers identify from orthologous gene datasets of ad hoc BLAST. The height of the bar represents the number of conserved synteny blocks for each chromosome.

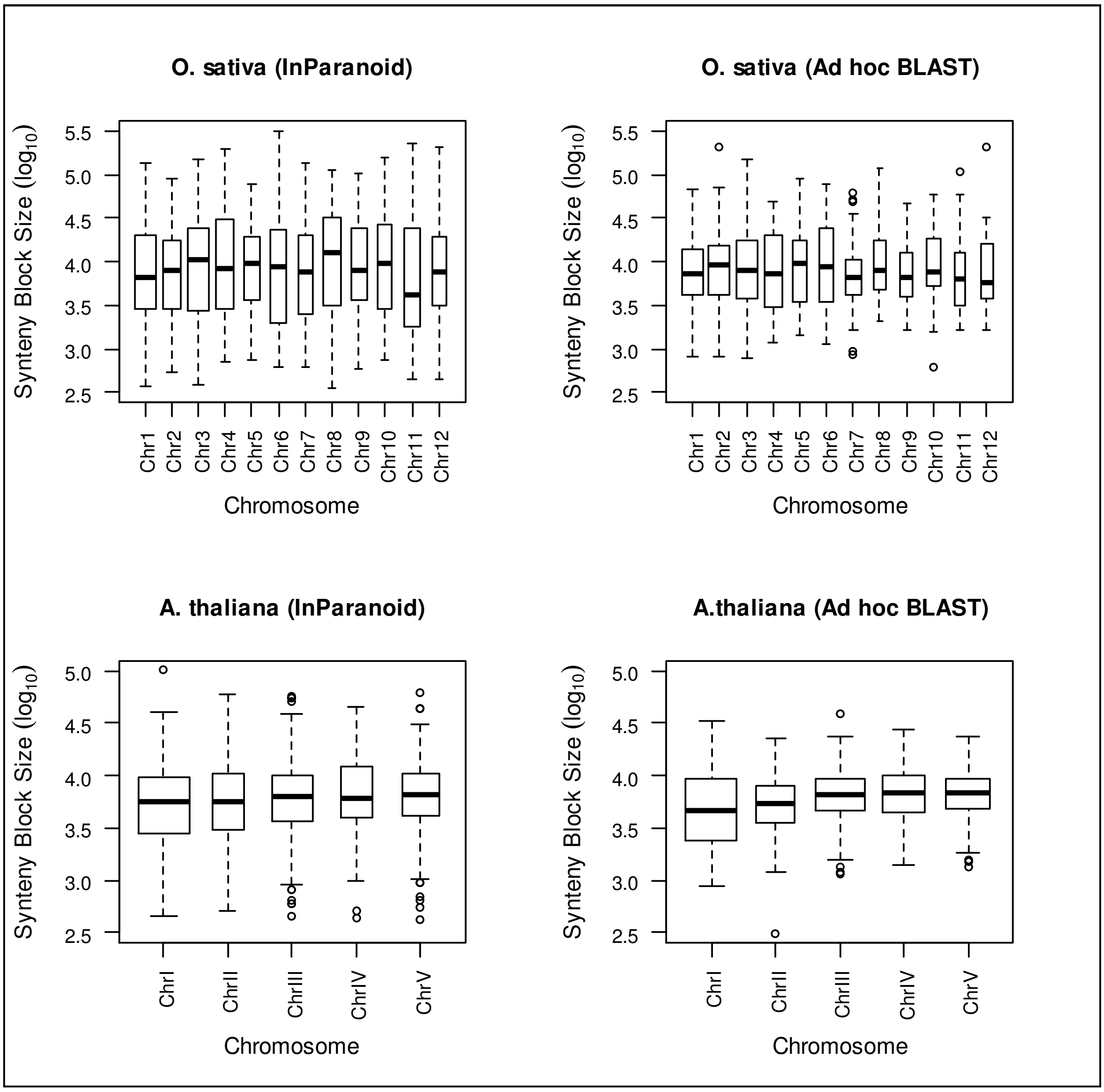


Figure 3.9: Distribution of non-conserved (50% mismatches) synteny blocks size in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Boxplots on right panel are distribution of non-conserved synteny blocks sizes identify from orthologous gene datasets of InParanoid. Boxplots on left panel are distribution of non-conserved synteny blocks sizes identified from orthologous gene datasets of ad hoc BLAST. Result for 10% mismatches was excluded because the result was same with conserved synteny block. The y-axis is plotted in log-scale of base 10.

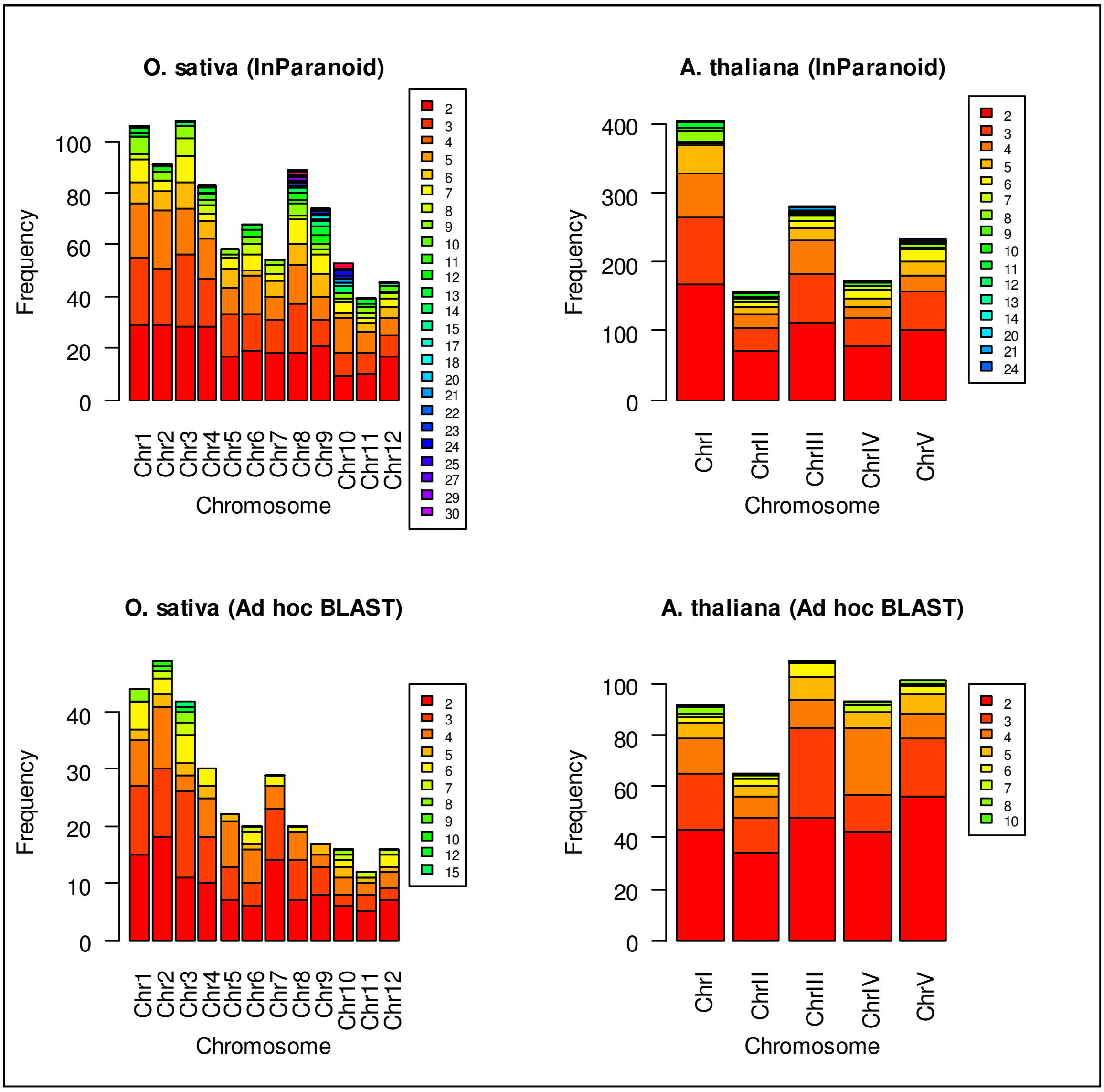


Figure 3.10: Distribution of gene numbers of each non-conserved (50% mismatches) synteny block in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Barplots on top panel are distribution of gene numbers of each non-conserved synteny block identify from orthologous gene datasets of InParanoid. Barplots on bottom panel are distribution of gene numbers of each non-conserved synteny block identify from orthologous gene datasets of ad hoc BLAST. The colours of the bar represent the number of genes of the non-conserved synteny block. The height of the bar represents the number of conserved synteny blocks with specified number of genes within the block. Result for 10% mismatches was excluded because the result was same with conserved synteny block.

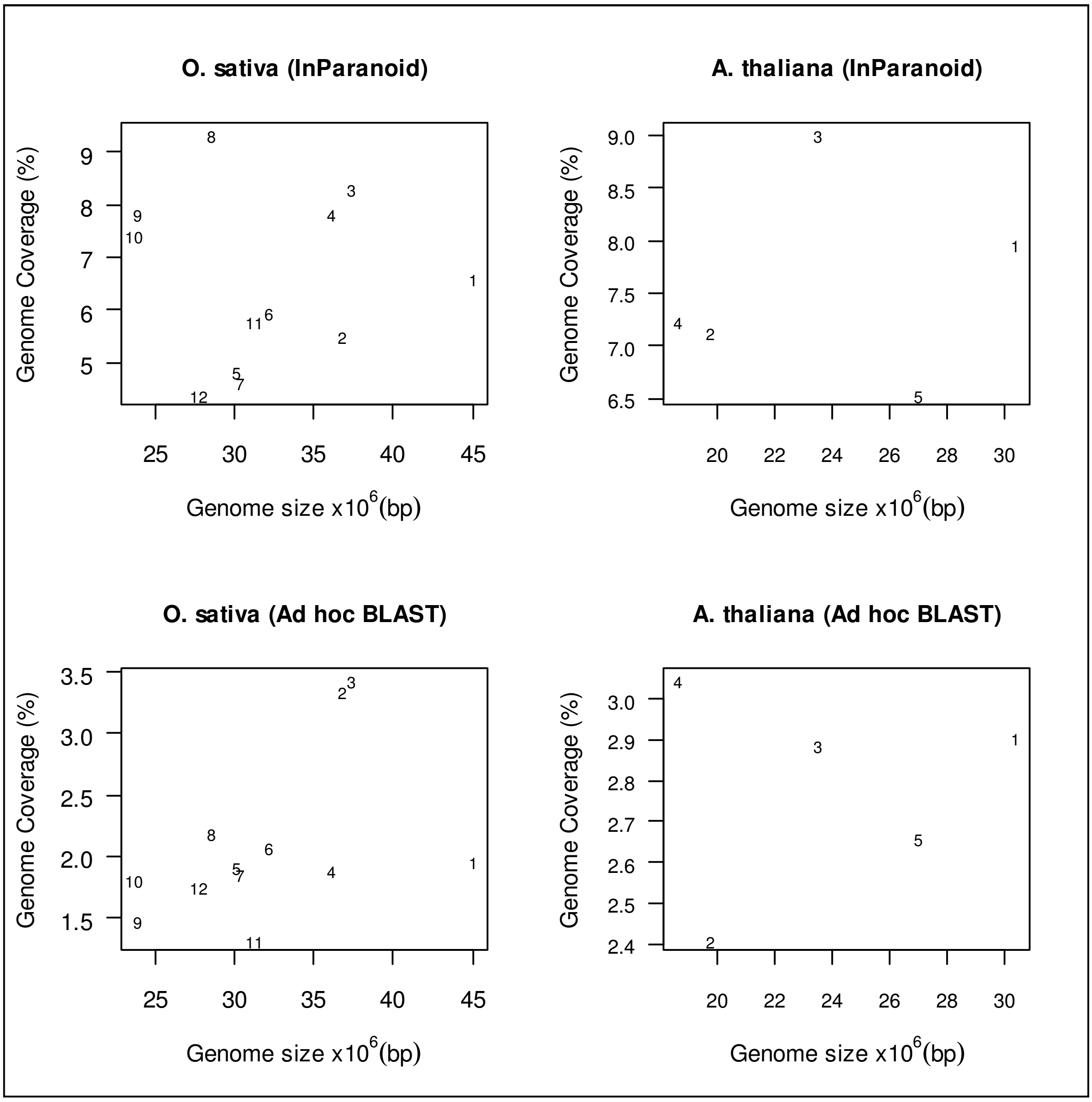


Figure 3.11: Coverage of non-conserved synteny blocks on each chromosome in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Scatterplots on top panel are coverage of non-conserved synteny blocks on each chromosome identify from orthologous gene datasets of InParanoid. Scatterplots on bottom panel are coverage of non-conserved synteny blocks on each chromosome identify from orthologous gene datasets of ad hoc BLAST. The numbers in the scatterplot represent the number of chromosome of each species. The genome sizes of each chromosome were plotted on x-axis while the coverage of non-conserved synteny blocks on each chromosome was plotted on y-axis.

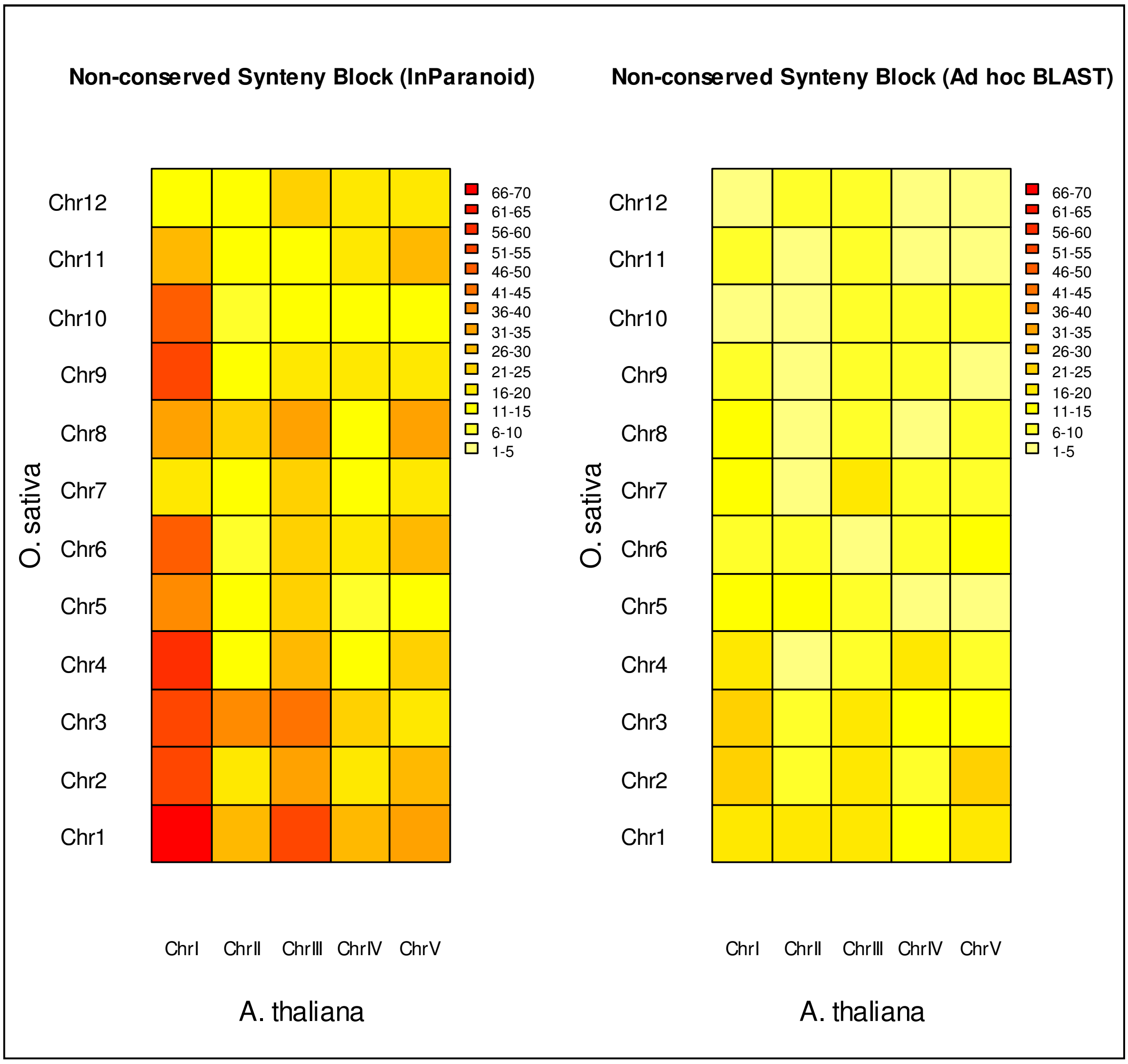


Figure 3.12: Distribution of non-conserved synteny blocks between *O. sativa* (row) and *A. thaliana* (column) identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. The colours of the each block represent the number of non-conserved synteny blockd between the chromosome of *O sativa* and *A. thaliana*.

Table 3.3: Comparison of synteny block returned from OrthoCluster for the orthologous genes dataset identified from InParanoid and ad hoc BLAST. Common block is the block identify by both orthologous gene identification method. There are some blocks in the same genome position from both methods but block size is larger on either one of the method (larger block size in InParanoid or ad hoc BLAST). Some blocks are method-specific block which only appear in either InParanoid or ad hoc BLAST. Result for 10% mismatches was excluded because the result was same with conserved synteny block.

	Conserved Synteny Block		**Non-conserved Synteny Block**	
	O. sativa	***A. thaliana***	***O. sativa***	***A. thaliana***
Common blocks	119	83	231	179
Larger block in InParanoid	5	8	32	31
Larger block in ad hoc BLAST	7	2	28	15
Inparanoid-specific block	801	425	1191	872
Ad hoc BLAST-specific block	188	113	330	253

3.4 Visualisation

Four Circos plots were generated in order to visualise the conserved and non-conserved synteny blocks identified from orthologous gene datasets using InParanoid and ad hoc BLAST methods. Figure 3.14 shows the conserved synteny blocks identified from orthologous gene datasets of ad hoc BLAST; those identified from orthologous gene are shown in Figure 3.15. The non-conserved synteny block identified from orthologous gene datasets of ad hoc BLAST and InParanoid are shown in Figure 3.16 and Figure 3.17, respectively.

The genome of *O. sativa* is drawn at right hand side of the plot while *A. thaliana* is drawn at left hand side. The distribution of annotated gene regions and identified orthologous genes from both ad hoc BLAST and InParanoid are drawn underneath the respective chromosomes. The distribution of conserved and non-conserved synteny blocks identified from orthologous gene datasets of ad hoc BLAST and InParanoid are drawn after the orthologous genes region. Finally, synteny block relationships are drawn as line in the centre of the Circos image to visualise the connection of synteny block between chromosomes of the two species.

Synteny blocks between chromosomes of two species are connected by edges. Those ranging from 50 kb to 100 kb in sizes are highlighted in green; those more than 100 kb in red. The synteny blocks related to chromosome 1 of *O. sativa* and chromosome I of *A. thaliana* are highlighted in brown and purple respectively, because these chromosomes have the highest number of synteny blocks. The synteny blocks on other chromosomes are coloured grey to reduce the complexity of the Circos image.

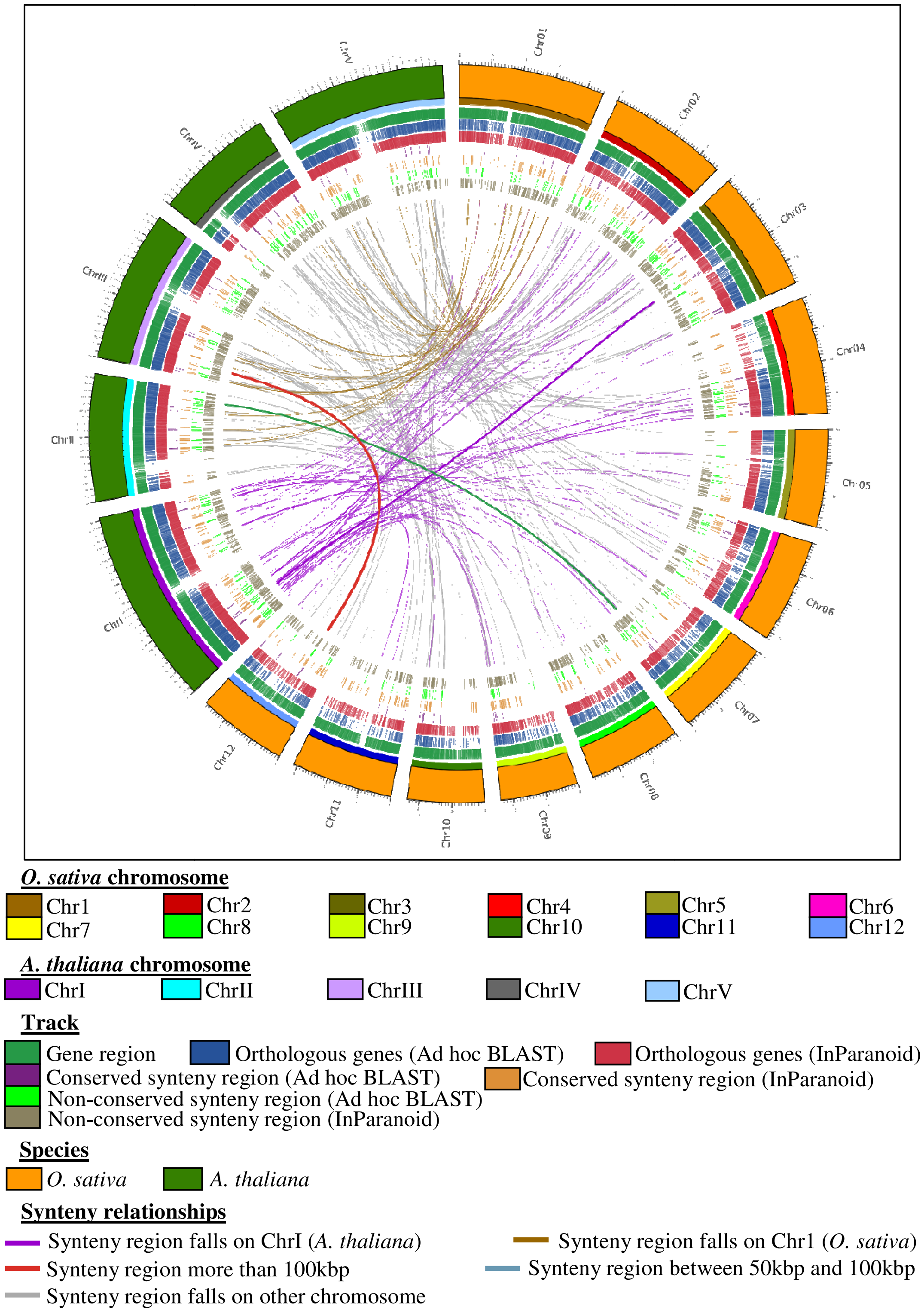


Figure 3.13: Circos image of conserved synteny block identified from orthologous gene datasets of ad hoc BLAST.

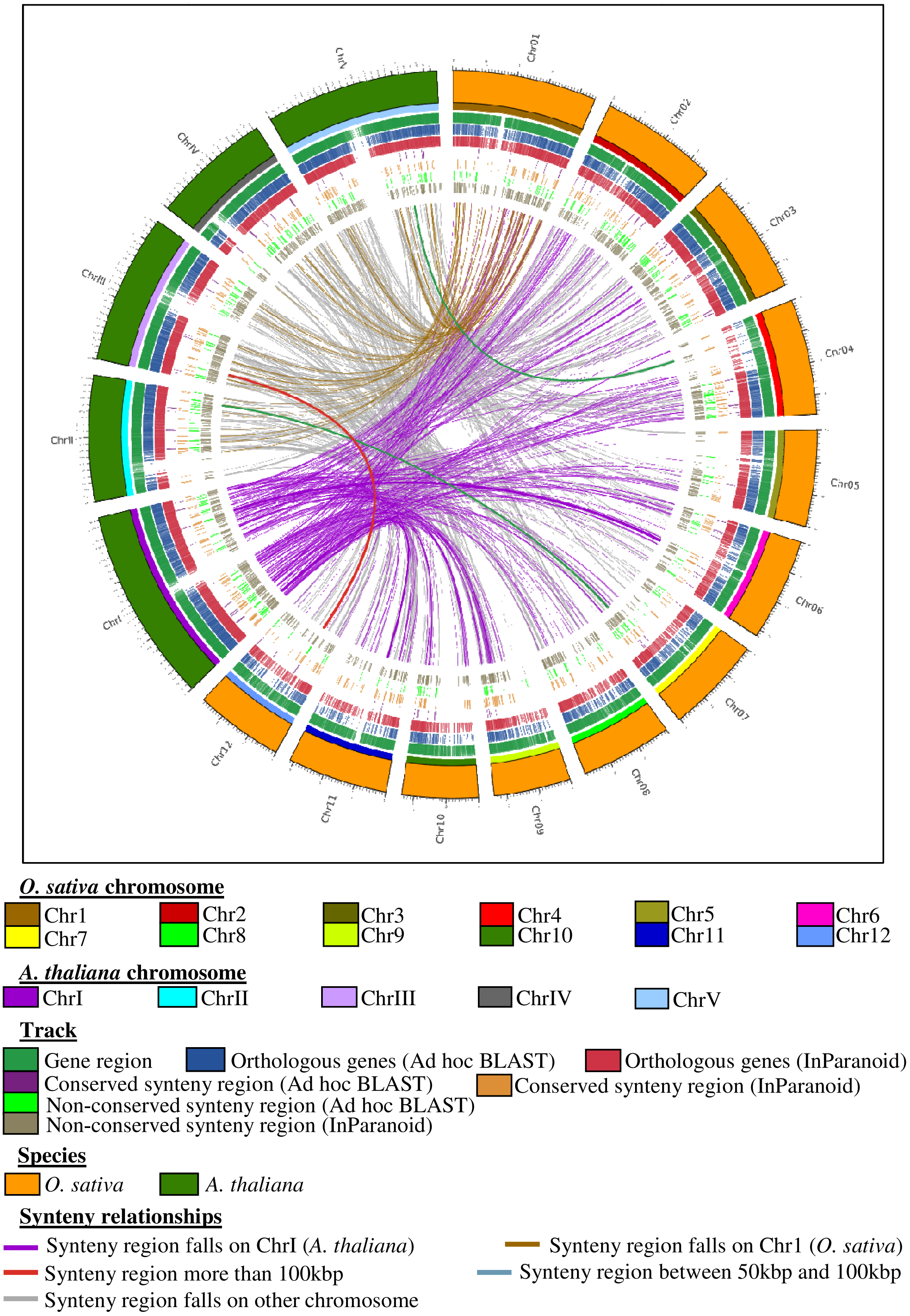
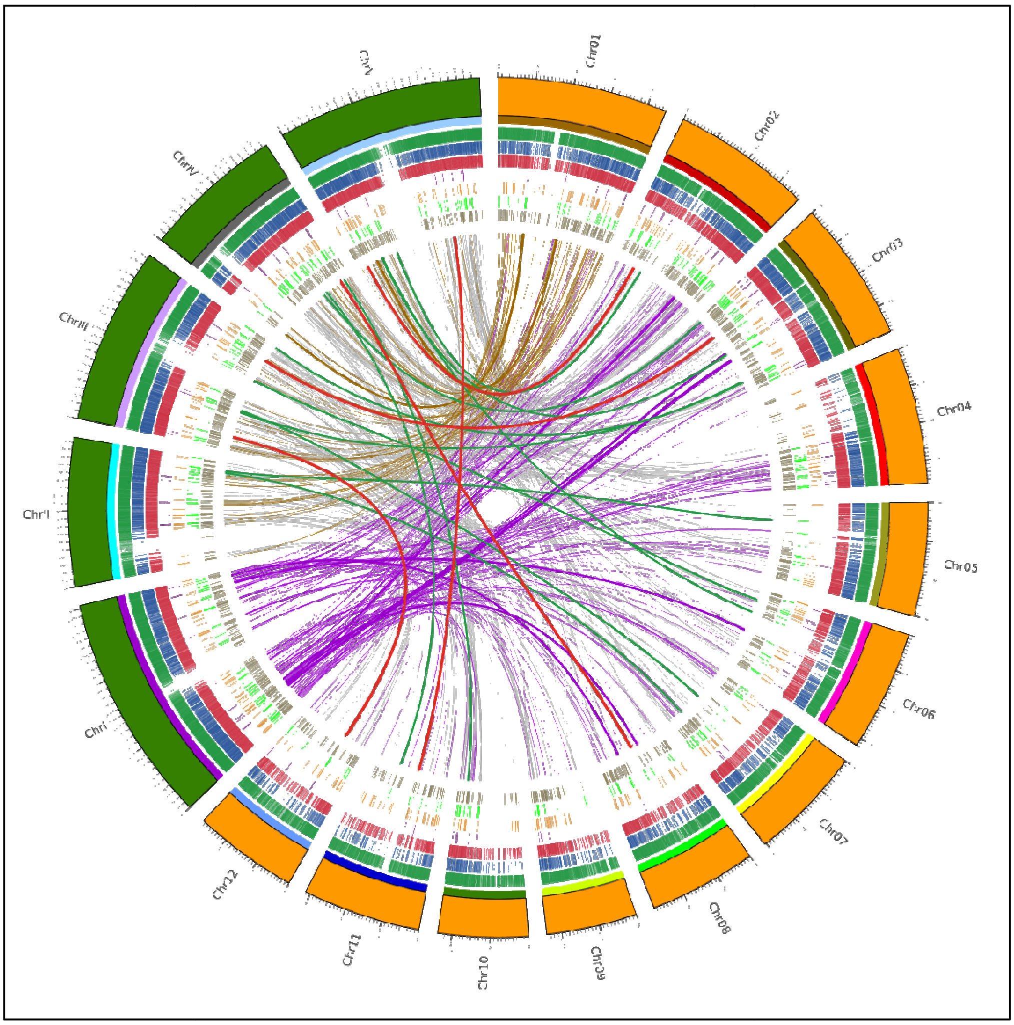


Figure 3.14: Circos image of conserved synteny block identified from orthologous gene datasets of InParanoid.



***O. sativa* chromosome**

🟫 Chr1 🟥 Chr2 🟫 Chr3 🟥 Chr4 🟫 Chr5 🟪 Chr6
🟨 Chr7 🟩 Chr8 🟨 Chr9 🟩 Chr10 🟦 Chr11 🟦 Chr12

***A. thaliana* chromosome**

🟪 ChrI 🟦 ChrII 🟪 ChrIII ⬛ ChrIV 🟦 ChrV

Track

🟩 Gene region 🟦 Orthologous genes (Ad hoc BLAST) 🟥 Orthologous genes (InParanoid)
🟪 Conserved synteny region (Ad hoc BLAST) 🟧 Conserved synteny region (InParanoid)
🟩 Non-conserved synteny region (Ad hoc BLAST)
🟫 Non-conserved synteny region (InParanoid)

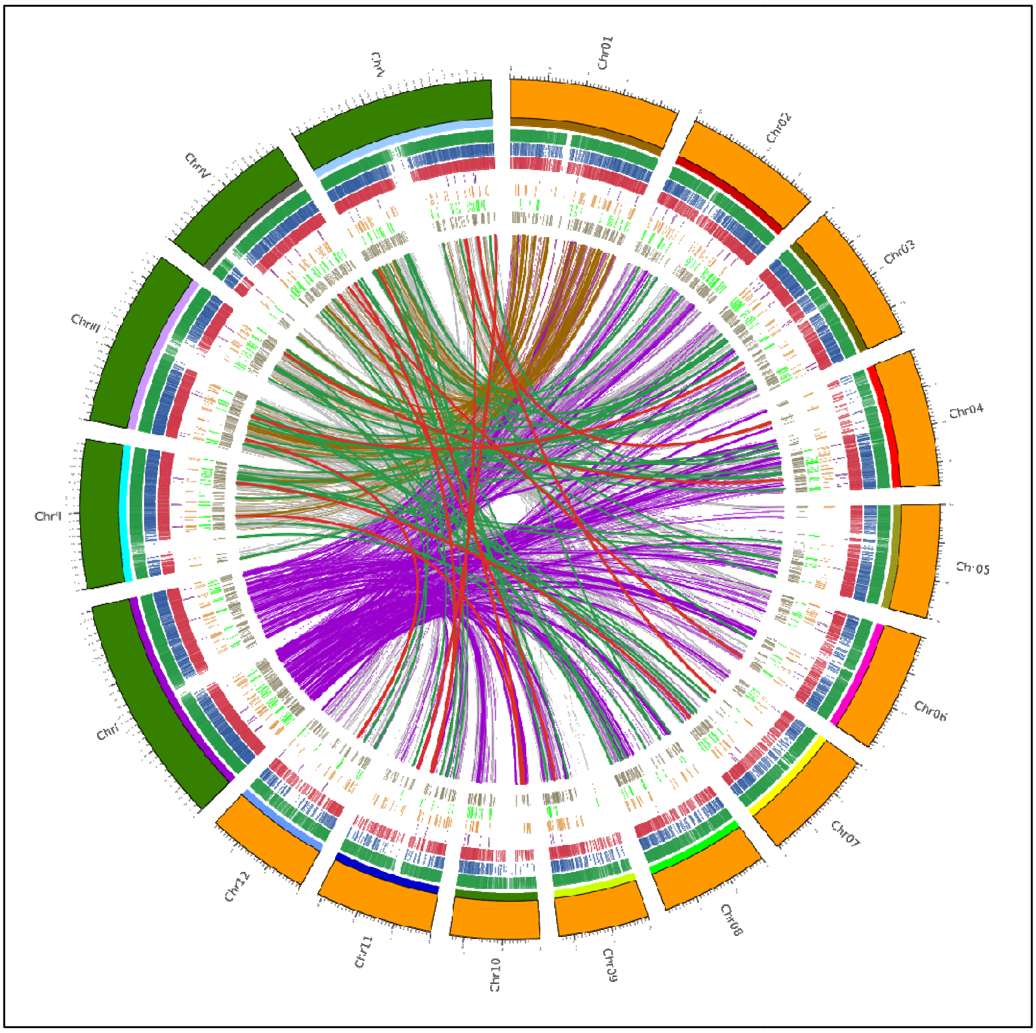
Species

🟧 *O. sativa* 🟩 *A. thaliana*

Synteny relationships

— Synteny region falls on ChrI (*A. thaliana*) — Synteny region falls on Chr1 (*O. sativa*)
— Synteny region more than 100kbp — Synteny region between 50kbp and 100kbp
— Synteny region falls on other chromosome

Figure 3.15: Circos image of non-conserved synteny block identified from orthologous gene datasets of ad hoc BLAST.



O. sativa chromosome

🟫 Chr1 🟥 Chr2 🟫 Chr3 🟥 Chr4 🟫 Chr5 🟪 Chr6
🟨 Chr7 🟩 Chr8 🟨 Chr9 🟩 Chr10 🟦 Chr11 🟦 Chr12

A. thaliana chromosome

🟪 ChrI 🟦 ChrII 🟪 ChrIII ⬛ ChrIV 🟦 ChrV

Track

🟩 Gene region 🟦 Orthologous genes (Ad hoc BLAST) 🟥 Orthologous genes (InParanoid)
🟪 Conserved synteny region (Ad hoc BLAST) 🟧 Conserved synteny region (InParanoid)
🟩 Non-conserved synteny region (Ad hoc BLAST)
🟫 Non-conserved synteny region (InParanoid)

Species

🟧 *O. sativa* 🟩 *A. thaliana*

Synteny relationships

— Synteny region falls on ChrI (*A. thaliana*) — Synteny region falls on Chr1 (*O. sativa*)
— Synteny region more than 100kbp — Synteny region between 50kbp and 100kbp
— Synteny region falls on other chromosome

Figure 3.16: Circos image of non-conserved synteny block identified from orthologous gene datasets of InParanoid.

CHAPTER 4

DISCUSSION

In this study, two orthologous gene identification methods: InParanoid and ad hoc BLAST, were applied to generate two different orthologous gene datasets. OrthoCluster used these as input to infer synteny blocks between *O. sativa* and *A. thaliana*. Comparison of orthology detection strategies is useful because agreement between methods enhances confidence in a particular method, while disagreement between methods indicates possible errors (Chen *et. al*., 2007). Since the accuracy of synteny block inference will be affected by the correctness of annotation between species of comparison, *O. sativa* and *A. thaliana* were selected in this study because their comprehensively annotated genome removes annotation errors as a source of variation.

InParanoid identified more orthologous relationships compared to ad hoc BLAST method (Table 3.1). Since the default setting of InParanoid includes low confidence level inparalogs group, we applied a cut-off value for confidence level to examine the relationship between the number of orthologous gene and the confidence level of inparalogs group. Each 10% increase in the confidence level caused the number of identified orthologous genes to decrease about 3% on average (Table 3.2). The number of one-to-many orthologous relationship identified by InParanoid was close to that returned from ad hoc BLAST. Overall, the total number of orthologous gene still remained at least about 10% higher in InParanoid even after cut-off values were applied on confidence level.

The Venn diagram (Figure 3.2) shows that there were many method-specific orthologous relationship which were only identified by either InParanoid or ad hoc BLAST. Examination of these orthologous relationship suggests that ad hoc BLAST probably generates all the possible orthologs groups when highly similar genes are involved, whereas InParanoid generates the unique orthologs groups only. One example of such case is the Os08t0554050-01 gene from *O. sativa* and AT1G76270.1 or AT1G20550.1 from *A. thaliana*. Os08t0554050-01 has orthologous relationship with AT1G76270.1 and AT1G20550.1. However, there is another gene in *O. sativa,* Os04t0563000-01, which is very similar to Os08t0554050-01 and AT4G38390.1 and AT1G76270.1 in *A. thaliana.* In this case, InParanoid will generate two orthologous relationships while ad hoc BLAST will identify six orthologous relationships (Figure 4.1).

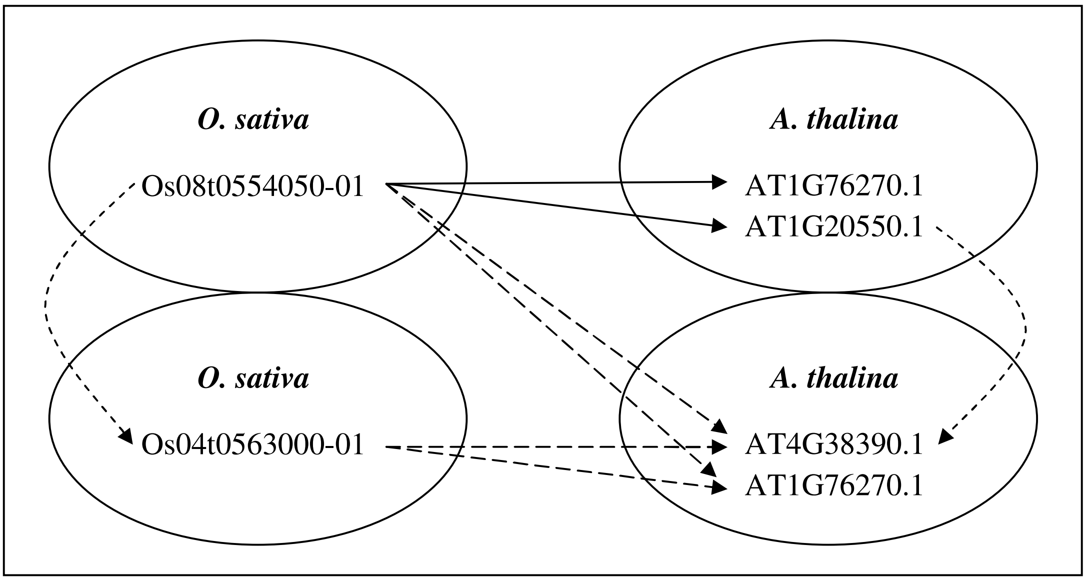


Figure 4.1: Example of orthologous relationships identified by InParanoid (solid lines) and ad hoc BLAST (dotted lines) when highly similar genes are involved.

A lot more InParanoid-specific orthologous relationships than ad hoc BLAST-specific orthologous relationships were found in this study. This was because the filtering criteria of ad hoc BLAST had reduced the number of genes from both species, thus restricting the identification of orthologous relationships. One example of such case is the Os01t0609200-00, Os01t0609300-01, Os01t0609900-02, and Os01t0609000-00 genes from *O. sativa*. These genes have orthologous relationships with AT1G15520.1 genes from

A. thaliana based on InParanoid method. However, ad hoc BLAST method only identified Os01t0609900-02 and Os01t0609000-00 to have orthologous relationships with AT1G15520.1 gene from *A. thaliana*, because Os1t0609200-00 and Os01t0609300-01 genes were excluded due to more than three copies of similar genes in the genome. Further investigation shows that the number of orthologous relationship from ad hoc BLAST was highly dependent on the filtering criteria such as the CALP and CALIP values, which determine the stringency of the method (Figure 4.2).

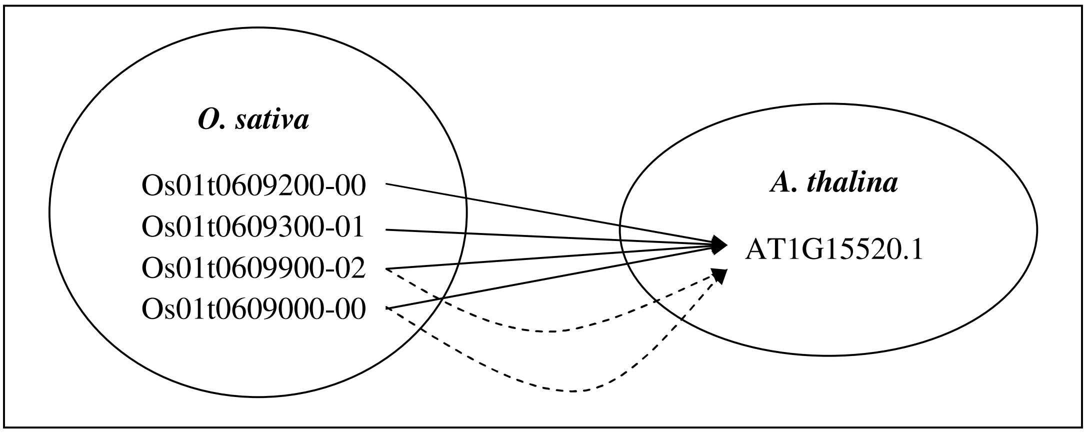


Figure 4.2: Example shows the orthologous relationships identified by InParanoid (solid lines) and ad hoc BLAST (dotted lines). Filtering criteria has cause ad hoc BLAST identified less number of orthologous relationships.

Comparison between synteny block returned from OrthoCluster by using two different orthologous gene datasets showed that the InParanoid dataset resulted in larger number and size of synteny block for both conserved and non-conserved situation (Figures 3.3 to 3.12). This result is not surprising because InParanoid identify higher number of orthologous relationship compared to ad hoc BLAST, thus resulting in higher number and larger size of synteny blocks between *O. sativa* and *A. thaliana*. For *O. sativa*, average median for non-conserved synteny block size with 50% mismatches is higher than conserved synteny block size for both ad hoc BLAST (~8 kbp and ~4 kbp) and InParanoid (~10 kbp and ~3.2 kbp). Most of the synteny block sizes (conserved and non-conserved) have symmetric distribution except chromosome 5, 7, and 11 of ad hoc BLAST from conserved synteny block and chromosome 10 and 12 of ad hoc BLAST from non-conserved synteny block have skewed distribution (Figure 3.4 & 3.9). For *A. thaliana*, average median for non-conserved synteny block size with 50% mismatches is slightly higher than conserved synteny block size for both ad hoc BLAST (~5 kbp and ~3.2 kbp) and InParanoid (~6.3 kbp and ~4 kbp). Most of the synteny block sizes (conserved and non-conserved) have symmetric distribution except chromosome 2, 3, ,4 and 5 of ad hoc BLAST from conserved synteny block have skewed distribution (Figure 3.4 & 3.9).

The results were no different for conserved synteny block and non-conserved synteny block with 10% mismatches for both InParanoid and ad hoc BLAST (Details omitted because the results are similar). This suggests that *O. sativa* and *A. thaliana* had possibly undergone large genome rearrangements since their divergence from the last common ancestor. However, the number and size of block for non-conserved synteny block with 50% mismatches increased tremendously for InParanoid while ad hoc BLAST only increased gradually (Figures 3.3, 3.4, 3.8 and 3.9). Identification of non-conserved synteny block is valuable because they provide a global view of the existing synteny between different species for regions that have been subjected to various types of rearrangement events (Vergara *et. al.*, 2010). In general, relaxing the constraints of synteny block will generate blocks with larger sizes when compared to conserved syntney block. The Circos plot showed the synteny relationships between chromosomes of *O. sativa* and *A. thaliana* (Figure 3.13 to 3.16). The number and size of synteny block increase tremendously when InParanoid dataset was used in non-conserved synteny block inference (Figure 3.16). Circos plot reviewed there are many synteny blocks from different chromosomes in *O. sativa* map to chromosome 1 of *A. thaliana*. This phenomenon might provide useful information for the fields of study in evolution, genome structure, whole genome duplication or species divergence between monocots and dicots.

Synteny blocks identified by both orthologous gene identification methods were likely to be true synteny blocks. However, there are a lot more synteny blocks that were identified only by InParanoid but not ad hoc BLAST, suggesting that the number of orthologous relationships does play an important role in the synteny block inference by OrthoCluster. Higher number of orthologous relationship allows OrthoCluster to identify larger block sizes, especially in non-conserved synteny blocks because there would be more “anchor” points for OrthoCluster to form a larger block. One example of such case is the chromosome 3 on *A.thaliana* which involves 7 genes: AT3G47730.1, AT3G47740.1 AT3G47750.1, AT3G47760.1, AT3G47770.1, AT3G47780.1 and AT3G47790.1. InParanoid identified these 7 genes are having orthologous relationship with 2 genes in *O. sativa,* Os08t0398000-01 and Os08t0398300-01. However, ad hoc BLAST only identified orthologous relationship between Os08t0398000-01- AT3G47730.1 and Os08t0398300-01-AT3G47790.1. There were 5 genes that were not in the ad hoc BLAST datasets; thus the synteny block did not appear in OrthoCluster’s results (Figure 4.3).

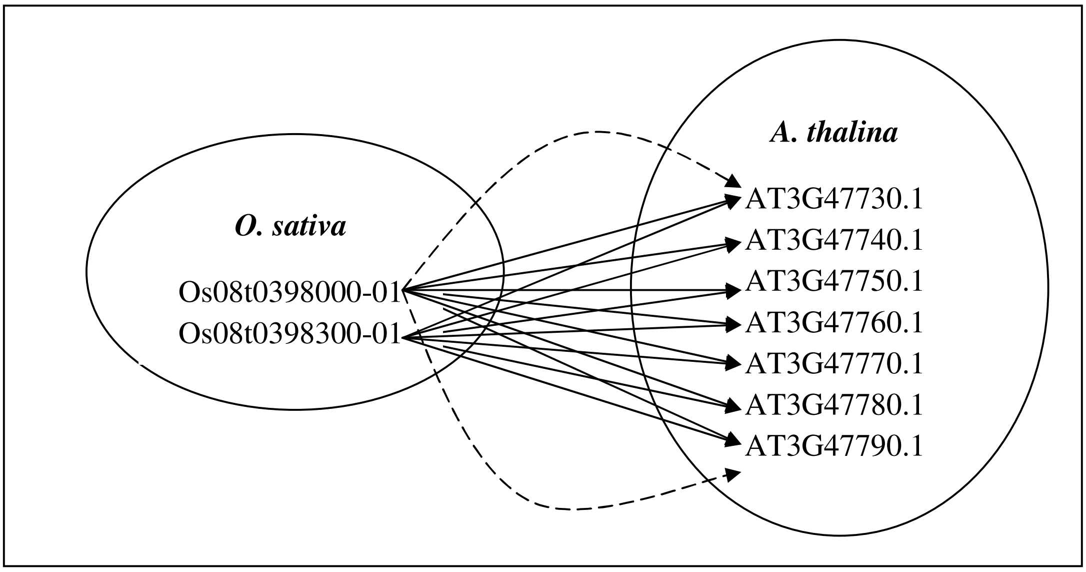


Figure 4.3: Example of orthologous relationships identified by ad hoc BLAST (dotted lines) and InParanoid (solid lines). InParanoid identified a lot more orthologous relationships than ad hoc BLAST.

In an ortholog database assessment study (Altenhoff *et. al.*, 2009), the researchers commented that InParanoid introduce significant biases in the inferred synteny blocks when only two pairs of species are compared. They then concluded that InParanoid was not the overall best performer in terms of specificity or sensitivity as believed in previous studies (Hulsen *et. al.*, 2006; Chen *et. al.*, 2007), when compared to other ortholog databases. Some of the orthologous relationships identified by InParanoid were probably false positives. It would be useful for future work to attempt an estimation of the false positive rate in InParanoid. If specificity is more important than sensitivity, and having only one ortholog per protein is sufficient, the best bidirectional hit approach should give the best results (Hulsen *et. al.*, 2006).

CHAPTER 5

CONCLUSIONS

The present comparative study provided quantitative characterisation of differences in OrthoCluster analysis of the complete genome of *Arabidopsis thaliana* and *Oryza sativa*, using two common approaches of identifying orthologous gene relationships. InParanoid identified more orthologous genes between *O. sativa* and *A. thaliana*, compared to ad hoc BLAST. This was because it categorised low and high confidence level orthologous genes into the same inparalog group if the cut-off value did not satisfy the confidence level. On the other hand, ad hoc BLAST emphasised more on stringency by applying CALP and CALIP cut-off values, which restricted the detection of orthologous relationships. Subsequently, when supplied with InParanoid-derived orthologous gene data, OrthoCluster detected more synteny blocks, both conserved and non-conserved. This finding suggests that synteny blocks returned from OrthoCluster is highly dependent on the method and parameter that are used to detect orthologous gene relationships. Therefore, the orthologous gene identification method that should be used depends on the research objective. If the objective is to detect one or more orthologs for a large number of proteins, methods that allow many-to-many relationships such as InParanoid would be more appropriate.