Ultraconserved elements: a plant genome perspective

Ella Yee, Ian Korf[1]
[1]Department of Molecular and Cellular Biology, University of California, Davis, USA.

Abstract

Motivation:
Ultraconserved elements (UCEs) are parts of the genome that have extreme conservation among a clade of organisms. UCEs were first discovered between human, mouse, and rat genomes, and most studies since then have been conducted on animals. Research on plant UCEs is still in the early phases. Studies have identified high conservation in regions of plant genomes, but these regions are generally less similar than the UCEs identified in research on animals. This study aims to explore whether conservation on the level of animal UCEs can be found in plants.
Results:
Comparison of the *Arabidopsis thaliana* genome to the genomes of *Brassica rapa*, *Populus trichocarpa*, and *Oryza sativa* demonstrated that UCEs do not exist in the noncoding regions of these plant genomes. However, they do exist in coding regions as exon outliers, exons with unusually high conservation compared to other exons from the same gene. The results from this study point at a key difference between animal and plant genomes: While animals share both noncoding and coding UCEs, UCEs only exist in coding regions of plant genomes. It is suspected that noncoding UCEs serve as enhancers and gene regulators, whereas coding UCEs control splicing and individual genes or exons. A lack of noncoding UCEs in plants suggests that though some plant and animal genes may be controlled in the same way, large scale genome organization differs significantly between these two kingdoms.
Availability: The FeatMasker and RLSFilter programs from this study are available via GitHub at https://github.com/ellayee33/PlantUCEs
Contact: ellyee@ucdavis.edu, ifkorf@ucdavis.edu

1 Introduction

Ultraconserved elements (UCEs) are sequences that are highly conserved among evolutionarily distant species. First discovered between human, mouse, and rat genomes, UCEs were initially defined as segments of 200 base pairs or longer exhibiting 100 percent identity across different genomes (Bejerano *et al*., 2004). UCEs are suspected to be involved in gene expression and regulation, but their purpose is still unknown. Thus far, most research on UCEs has been conducted on animals, with different studies proposing different requirements for length and similarity (Snetkova *et al.,* 2022). However, the extent of sequence conservation in plants is less clear. Researchers have also identified regions with high conservation in plant genomes, but the regions often fail to meet the requirements for animal UCEs. Instead, they have been referred to as UCE-like elements (ULEs) or conserved elements (Hupalo *et al.*, 2013; Kritsas *et al.*, 2013). Thus, the goal of this study was to characterize the nature of sequence conservation in coding and noncoding regions of plant genomes. *Arabidopsis thaliana* (*A. thaliana*) was used as a model system to investigate whether conservation levels comparable to those of animal UCEs exist within plant genomes.

2 Materials and Methods

2.1 Dataset

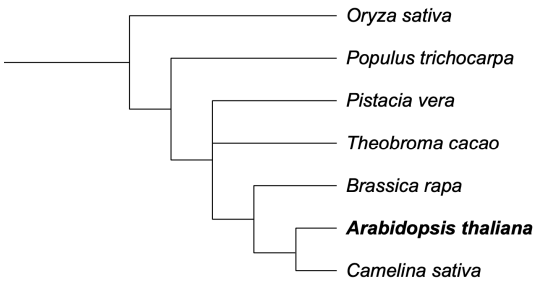

Fig. 1. A phylogenetic tree of the six reference species and *A. thaliana*.

The complete genomes and cDNA sequences of seven plant species were downloaded from Ensembl Plants, a public portal for sequencing data and bioinformatic analysis tools. *Brassica rapa* (*B. rapa*), *Camelina sativa* (*C. sativa*), *Oryza sativa* (*O. sativa*), *Pistacia vera* (*P. vera*), and *Populus trichocarpa* (*P. trichocarpa*) were chosen as reference species to compare against the model system, *A. thaliana*. Since the objective was to identify clades of organisms that diverged long ago but still share UCEs, reference species of varying evolutionary distances to *A. thaliana* were selected (Figure 1). All data was pre-masked by Ensembl Plants for low complexity regions, repetitive sequences that distract from primary findings if left unfiltered (Saini *et al.*, 2016).

2.2 BLAST

Basic Local Alignment Search Tool (BLAST) v. 2.2.26 was used to align the genomes of different species. BLAST identifies similar regions between two sets of sequences by first locating regions of similarity, called seeds, and then identifying optimal alignments near the seeds. Processing large datasets with traditional alignment algorithms such as the Smith-Waterman algorithm is computationally expensive, but the preliminary seeding process of BLAST overcomes this challenge (CLC Bio, 2007). For this study, the scoring scheme was changed from the default +1 match, -3 mismatch to +1 match, -1 mismatch to allow dissimilar sequences to align. Additionally, the costs of opening and extending gaps were set to 20 and 10 to limit alignments containing gaps. Default values were kept for the remaining parameters.

2.2 cDNA alignment as benchmarking

As this study involved different species and different evolutionary distances compared to previous studies, the first task was to investigate what typical conservation looks like with each genome. The average conservation of transcribed regions was measured so that later findings involving noncoding

sequences could be benchmarked against expected conservation levels. Using BLAST, the cDNA sequence of each reference species was segmented and compared to the cDNA sequence of *A. thaliana*. The longest alignment from each segment in a given reference species file was recorded, and results from the first 500 segments were used to create a histogram of percent identity with the *A. thaliana* cDNA sequence. Finally, the histogram distributions were compared to select species for complete genome analysis.

2.2 UCE identification in noncoding regions

2.2.1 Preprocessing and parameter selection

The FeatMasker program was developed to hard mask the *A. thaliana* genome for regions labeled as exons, ribosomal RNA (rRNA), or transfer RNA (tRNA). The purpose of preprocessing the genome was to prevent BLAST alignments in coding regions from distracting from alignments in noncoding regions. A General Feature Format (GFF) file, which lists genomic features and their coordinates, was parsed to identify the regions that required masking. The character "N" was used as a placeholder for nucleotides in these regions. Similar to the process for cDNA alignment, the complete genomes of *B. rapa*, *P. trichocarpa*, and *O. sativa* were segmented and compared to the masked *A. thaliana* genome in a BLAST search. For an accurate comparison between conservation of transcribed and noncoding regions, the parameters from cDNA alignment were used again when aligning the complete genomes. The tabular alignment view was selected to simplify postprocessing steps.

2.2.2 Postprocessing

Although masking the *A. thaliana* genome reduced the number of BLAST alignments in coding regions, postprocessing was used to further filter the results. First, alignments involving mitochondrial or plastid DNA sequences were removed. Then, the RLSFilter tool was created to filter alignments for recurrence, length, and similarity. As noncoding UCEs should be unique within the genome, recurring alignments indicated segments of RNA that were not annotated in the GFF file. The filtering process for recurrence involved two phases: In the first phase, identical alignments were removed. The number of alignments for each tuple of chromosome, start coordinate, and end coordinate was recorded. If a given tuple had more than one alignment, all alignments involving the tuple were removed from the results pool. The second phase of the filtering process involved pinpointing of hot regions. Hot regions are locations in the genome where alignments are clustered. Though the alignments in hot regions do not share identical coordinates, they point to the same genomic feature and are therefore also repetitive. Hot regions were identified from the results pool containing non-identical alignments, and alignments falling within hot regions were removed from the pool. Finally, the results were compared to length and similarity requirements. Based on requirements set in previous plant studies, the minimum length for a sequence to be considered an UCE was set to 50 base pairs. 77, 80, and 87 percent identity were selected as minimum similarity requirements between the genomes of *A. thaliana* and *O. sativa*, *P. trichocarpa*, and *B. rapa*. Different similarity requirements were chosen for the three reference species according to their cDNA alignment distributions and their evolutionary distances to *A. thaliana*.

2.2.3 Verifying BLAST alignments

To check whether the filtered alignments were true UCEs, regions of the *A. thaliana* genome were studied on JBrowse, a genome browser provided by The Arabidopsis Information Resource (TAIR). The tracks for novel transcribed regions, transposable elements, protein coding genes, pseudogenes, and noncoding RNA were selected. If JBrowse displayed a feature that overlapped or was close to a suspected UCE, the feature sequence was compared to sequences in the Nucleotide collection default database. The Nucleotide collection consists

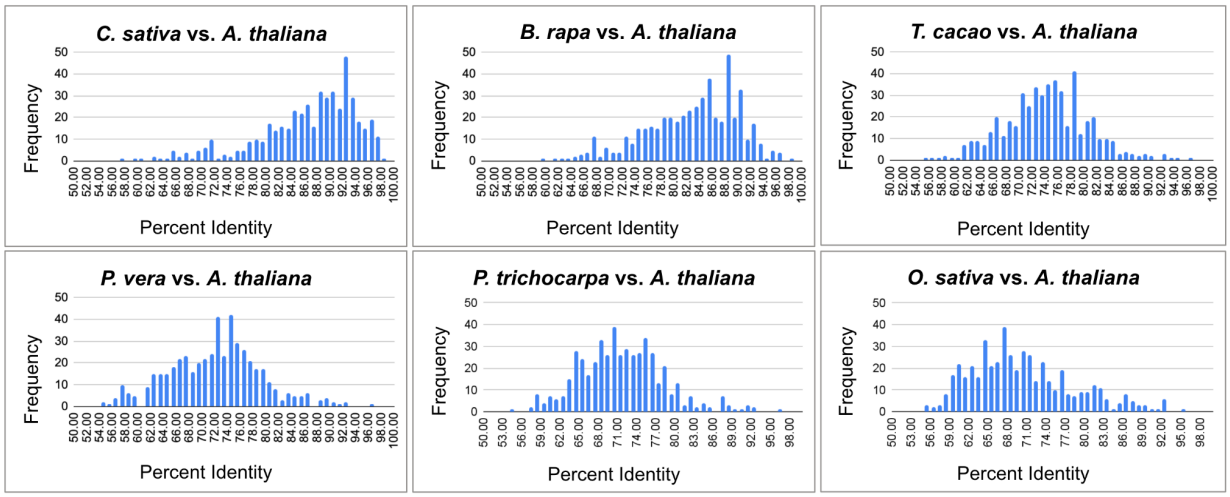

Fig. 2. Percent identity distributions of top alignments between the cDNA sequences of six reference species and *A. thaliana*. The x-axis represents percent identities between sequences, while the y-axis represents frequency.

of 84511849 sequences compiled by GenBank, European Molecular Biology Laboratory, DNA Data Bank of Japan, Protein Data Bank, and RefSeq. If the BLAST search revealed that the feature was part of mitochondrial DNA, plastid DNA, or an RNA coding sequence in other species, the feature was not regarded as an UCE.

2.3 UCE identification in coding regions

2.3.1 Preprocessing and parameter selection

To identify UCEs in coding regions, the FeatMasker tool was used to mask the *A. thaliana* genome for coding sequences (CDS). Then, the masked genome was compared to the original genome to construct an *A. thaliana* genome with only CDS preserved. In initial tests with *B. rapa*, *P. trichocarpa*, and *O. sativa*, the BLAST parameters from identifying UCEs in noncoding regions were used again. The costs of opening and extending gaps were then lowered to 2 and 1 to view alignments containing gaps.

2.3.2 Investigating top alignments

The top ten alignments for each reference species were studied in greater detail using JBrowse. Alignments that approximated one exon from an *A. thaliana* gene were considered exon outlier candidates. If the *A. thaliana* gene had an ortholog with the reference species, exons from the gene were compared to the reference species genome in a secondary BLAST search. Alignments where the *A. thaliana* gene did not have an ortholog with the reference species were attributed to random chance instead of biological conservation. Finally, the percent identity between the exon outlier candidate and the reference genome was compared to the average percent identity for all exons in the same gene.

3 Results

3.1 cDNA alignment

Results of the BLAST search between cDNA sequences corresponded to the evolutionary distances of the reference species to *A. thaliana* (Figure 2). Based on the percent identity distributions, the

reference species were categorized into three groups: The group exhibiting the most conservation with *A. thaliana* included *B. rapa* and *C. sativa*, the intermediate group included *P. trichocarpa*, *P. vera*, and *T. cacao*, and the group exhibiting the least conservation with *A. thaliana* included *O. sativa*. *B. rapa*, *P. trichocarpa*, and *O. sativa* were selected to represent their respective groups in further investigations.

3.2 UCE identification in noncoding regions

Table 1. BLAST alignments before and after custom filtering

Reference Species	Hits (E < 1e-10)	Hits (After Filtering)
B. rapa	2214290	11
P. trichocarpa	14215627	17
O. sativa	619572	22

BLAST searches against the masked *A. thaliana* genome resulted in a significant number of alignments with expectation values less than 10^{-10}. However, the custom filters for repetitive, short alignment length, and low similarity sequences greatly reduced the suspected UCE pool (Table 1). Upon further investigation with JBrowse, alignments could be classified into the following categories:

3.2.1 Transposable elements

Transposable elements are DNA sequences that move throughout the genome. BLAST searches against the Nucleotide collection database revealed that transposable elements in the *A. thaliana* genome were highly similar to predicted mRNA and ncRNA coding sequences in other species.

3.2.2 Pseudogenes

Pseudogenes are mutated genes that no longer code for proteins. Pseudogenes in the *A. thaliana* genome were highly similar to mitochondrial DNA, chloroplast DNA, and transporter genes in other species.

3.2.3 Unannotated regions

Some alignments appeared in regions with no features on JBrowse. However, adjusting the viewing window to include nearby regions revealed that the alignments were in gaps between transposable element clusters. Thus, these alignments were likely parts of transposable elements not labeled on JBrowse.

These alignment categories suggest that the UCEs identified in animal studies do not exist in the noncoding regions of plant genomes.

3.3 UCE identification in coding regions

Table 2. Percent identities from exon outlier alignments compared to average percent identities for all exons in the gene.

Gene	Reference	Outlier Identity (%)	Average Identity (%)
AT5G60040.2	*B. rapa*	99.29	88.32
AT3G15990.1	*B. rapa*	99.21	89.52
AT4G21710.1	*B. rapa*	99.17	91.55
AT3G05970.1	*B. rapa*	99.16	90.87
AT1G21650.1	*B. rapa*	99.15	93.01
AT3G53570.1	*P. trichocarpa*	97.96	81.96
AT5G64070.1	*P. trichocarpa*	95.52	80.09
AT3G03810.1	*P. trichocarpa*	95.51	78.23
AT2G20190.1	*P. trichocarpa*	95.45	80.26
AT2G07641.1	*O. sativa*	96.30	86.97

Inspection of the top ten alignments for each reference species revealed five ultraconserved exons with *B. rapa*, four ultraconserved exons with *P. trichocarpa*, and one ultraconserved exon with *O. sativa* (Table 2).

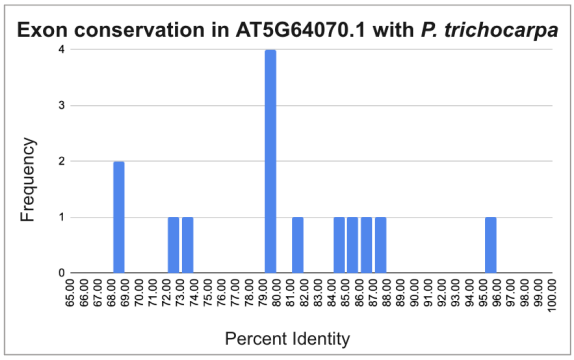

Fig. 3. Percent identity distributions of exons in an *A. thaliana* gene when compared to an orthologous gene in *P. trichocarpa*. The exon exhibiting the highest percent identity is considered an exon outlier.

Each ultraconserved exon was an outlier within its gene: Though other exons from the same gene had varying degrees of conservation, they had significantly lower percent identities when aligned to the reference species (Figure 3). In addition to alignments of ultraconserved exons, the secondary BLAST search also returned partial segments of exons and sequences containing both exonic and intronic sequences. These alignments were disregarded for the purpose of this study.

4 Discussion and Conclusion

While the results of this study demonstrate that UCEs do not exist in the noncoding regions of plant genomes, the identification of several exon outliers proves that certain coding sequences are under extreme conservation. It can be inferred that these exons have purposes beyond coding for proteins: Because different triplets of nucleotides can correspond to the same amino acid, protein coding does not constrain sequences to such extreme conservation. These findings are particularly consequential when analyzed alongside previous research: In past studies, animal genomes were found to contain UCEs in intronic, exonic, and intergenic regions. (Bejerano *et al.*, 2004). Furthermore, while noncoding UCEs are suspected to act as enhancers or regulators of nearby genes, coding UCEs are believed to play a role in splicing (Lareau *et al.*, 2007; Pennacchio *et al.*, 2006; Sandelin *et al.*, 2004). In summary, noncoding UCEs likely control functions at the level of the genome, while coding UCEs likely control functions at the level of individual genes or exons. Thus, the shared existence of coding UCEs in plant and animal genomes suggests that certain genes may be regulated in the same manner. In contrast, the absence of noncoding UCEs in plants suggests that the organization and regulation of large regions differs significantly between the genomes of the two kingdoms. In the future, studies should investigate additional plant species to determine whether noncoding UCEs can be identified by restricting evolutionary distance and to explore the limits of coding UCEs. Additionally, differences in the genomes of plants and animals suggest that differences in cellular structure may exist as well. From a genomic perspective, plants lack the macroscale organization present in animals. Whether or not this is counterbalanced by a greater degree of structural organization is a question demanding deeper examination.

5 Acknowledgments

Thank you to the UC Davis Young Scholars Program for the opportunity to conduct this study in the UC Davis Genome Center.

References

Bejerano, G. *et al.* (2004). Ultraconserved elements in the human genome. *Science*, **304**(5675), 1321–1325.

CLC Bio. (2007). *Bioinformatics explained: BLAST versus Smith-Waterman*.

Hupalo, D. *et al.* (2013). Conservation and functional element discovery in 20 angiosperm plant genomes. *Molecular Biology and Evolution*, **30**(7), 1729–1744.

Kritsas, K. *et al.* (2012). Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. *Genome Research*, **22**(12), 2455–2466.

Lareau, L. *et al.* (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, **446**, 926-929.

Pennacchio, L. *et al.* (2006). *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499-502.

Saini, H. *et al.* (2016). Gene masking - a technique to improve accuracy for cancer classification with high dimensionality in microarray data. *BMC Medical Genomics*, **9**(3), 74.

Sandelin, A. *et al.* (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, **5**, 99.

Snetkova, V. *et al.* (2022). Perfect and imperfect views of ultraconserved sequences. *Nature Review Genetics*, **23**, 182–194.