**Development of Verdant as a Platform for Rapid Phylogenetic Analysis of Whole Chloroplast Genomes**

Verdant was developed specifically to increase throughput of chloroplast annotation. While DOGMA (Wyman et al., 2004) has been an effective and widely used tool for many years, it requires manual intervention at several points in the annotation process. Other annotation tools for chloroplast genomes have also been developed, including CpGAVAS (Liu et al., 2012) and Plann (Huang and Cronk, 2015). Each of these has novel benefits, with CpGAVAS adding functionality to create datasets from multiple chloroplast genomes and Plann allowing for automated annotation of very similar chloroplast genomes. As plastome datasets now include hundreds of plastomes, we saw a need for a rapid pipeline to go from an assembled plastome to one that is fully annotated and placed in an alignment without any manual curation steps. Verdant, and the incorporated annotation software annoBTD, is such a tool.

**Overview of Annotation Method**

The core of Verdant is annoBTD (annotates better than Daniel).

The annoBTD process is broken into 3 primary steps:

1. *de novo* ORF Identification
2. a) Best Reference Identification for ORFs,

b) Best Reference Identification for tRNAs and rRNAs

1. Feature Boundary Resolution

The annoBTD process is designed to annotate plastomes for phylogenetic or population genetic studies, which include dozens or hundreds of plastomes that need to be annotated and aligned. For this reason, we focus on genes, rRNAs, and tRNAs that are found across most angiosperm chloroplast genomes (Daniell et al. 2016), although we also flag putative ORFs that are not in this set. These flagged ORFs can be provided by Verdant administrators but are not included in annoBTD’s final annotation. During the development of annoBTD and Verdant, we have found that rRNAs and tRNAs are relatively conserved across angiosperm chloroplast genomes and can be identified with high accuracy using nucleotide sequences, a process similar to that found in DOGMA. Due to this conservation, we treat these features separately from protein-coding regions, which are much more variable in sequences and in exon number across lineages. The process of protein-coding gene annotation is more rigorous than that of rRNAs and tRNAs and includes a *de novo* approach for ORF identification and subsequent identity assignment to the ORFs. Importantly, initial ORF identification is done de novo and reference genomes are only brought in subsequently to filter and provide names for the putative ORFs.

**1) *de novo* ORF Identification**

Protein-coding genes can be highly variable across lineages, with variation in nucleotide sequence and exon number between lineages. Variation also arises through partial or pseudogenized protein-coding genes, especially around borders of the inverted repeat. Finally, long genes, such as *rpoC2* (more then 4,000 bp in grasses) and *ycf1* (more than 5,000 bp in eudicots) are difficult to match to reference genes for their entire length and are often represented as partial high-scoring segment pairs (HSPs) in BLAST results.

annoBTD begins by identifying open reading frames (ORFs) of 18 nucleotides or longer in both the forward and reverse strands for all three reading frames. An ORF can begin with a codon for any amino acid, but ORFs always end with a stop codon. The ORF size of 18 nucleotides was chosen based on analysis of ORF size distributions and known exon sizes of protein-coding genes in angiosperm plastomes. This initial ORF search identifies thousands of potential ORFs and is cleaned to remove nested ORFS. Nested ORFs are defined as those found with at least 50% of their length inside the boundaries of a larger ORF. We allow for some overlap to accommodate overlapping in gene space, such as that found in *psbD* and *psbC*.

Given the *de novo* approach for identifying protein-coding genes in annotated chloroplast genomes, we are able to provide a list of potentially novel ORFs to users upon request, but these are not included in the final annotation given by Verdant. We do not include them automatically to avoid populating the database with unverified features.

**2) Identification of Best References**

The taxonomic information for each user-uploaded chloroplast genome is used to identify the five most closely related samples in the database. Verdant begins by choosing congeners and moves through the taxonomic hierarchy until it identifies five potential references. Once identified, sequences for protein-coding genes, tRNAs, and rRNAs are pulled from all references and labeled with an identifier linking the sequences back to their respective sources. annoBTD then determines which of the five representatives for each feature are optimal for using similar but slightly different methods as explained below.

**2a) ORFs**

The previously identified ORFs are BLASTed against the protein-coding gene set from the five determined references using TBLASTX (Camacho *et al.*, 2009) with an e-value cutoff of 0.1. The BLAST results are filtered using a cutoff of 50% identity between the references and the HSPs. TBLASTX is used on the nucleotide sequence instead of an amino acid approach to allow direct comparison of the nucleotide sequence to the genome being annotated during the Feature Boundary Resolution phase of annoBTD. We have also found that TBLASTX was superior in identification of smaller exons, especially *rps16*\_exon1, which is often a within a larger ORF than the typically accepted exon space.We use a very lax cutoff for our BLAST results, which allows more distantly related taxa to be used for annotation.

For each ORF and reference pair, the reference-scoring algorithm is used to determine the best reference sequence for a given ORF. The premise of this algorithm is that for genes in the chloroplast genome, the representative HSPs will not always be complete and the BLAST simply serves provide a potential identity to a given ORF. The reference-scoring algorithm effectively anchors a reference identity through similarity but still requires refinement for boundary identification, which occurs during “Feature Boundary Resolution”. Additionally, ORFs with high bit scores can sometimes be found in the chloroplast genomes that do not represent true gene space. To overcome this problem, we compare similarity of ORF to genes using amino acid triplets.

annoBTD translates the ORF to all six frames and then looks for matching strings of three amino acids between the ORF and the reference, using a sliding window along the entire length of the translated ORF. For each three-residue match, the score is increased by one. Once the entire length of the translated ORF is traversed, the total score is then divided by the length of the reference gene. This provides the percent of the reference gene that shares amino acid triplets to the ORF. If this normalized score is less than 0.70, the ORF-reference pair is removed from consideration. This threshold allows us to ignore instances where genes with inserted introns are matching genes that have lost the introns (or vice-versa), which may impeded annotation. Though rare, this does occur in some parasitic angiosperm lineages. This step does not remove either the ORF or the reference individually, but only their potential relationship. After all putative ORF-reference pairs for a given ORF are scored, the maximum score for that ORF is identified. ORF-reference identities are then compared across the entire genome ultimately assigning a single reference to a single ORF based on position in the genome. This is done by matching strandedness of ORFs to directionality of the reference sequence. The process permits assignment of a single ORF-reference ID for genes repeated in the inverted repeats. The final set of best references per ORF is recorded.

2b) **tRNAs, and rRNAs**

The chloroplast genome sequence being annotated is BLASTed against the tRNA and rRNA datasets from the five reference datasets using BLASTN (Camacho *et al.*, 2009) with an e-value cutoff of 0.1. BLAST results for both tRNAs and rRNAs are filtered to remove HSPs with percent identities less than 50%. For each HSP, the region of the chloroplast genome is compared to the reference RNA using a sliding window of six nucleotides, in a manner similar to that of amino acids triplets for ORFs. A major distinction that it instead of comparing sextets from the HSP to the reference RNA, nucleotide sextets from the reference RNA are searched against the HSP. For each nucleotide sextet that matches to the HSP, the score is increased by 1. The total score is normalized by dividing by the length of the reference RNA. For each tRNA and rRNA, a maximum score is identified across all references for all HSPs. Ultimately, the best scoring reference and HSP for each tRNA and rRNA is identified and recorded.

**3)** Feature Boundary Resolution

Finishing the annotation incorporates the sequence and positional information for ORFs and putative tRNAs and rRNAs from the uploaded chloroplast genome, sequence information for identified best references for all features, and the complete sequence of the chloroplast genome that is being annotated. Unique aspects of different genes have been considered when estimating protein-coding boundaries and various methods, as outlined below, have been implemented to address these problems.

1. **Short exons**

Genes like *petB*, *petD*, and *rpl16* have exons that are below the 18-nucleotide cut-off for ORF detection. These exons are identified from the reference gene set and recorded. The larger exons of these respective genes are identified and the expected position of the short exon, based on the transcriptional direction of the larger exon, is estimated. As these small exons are highly conserved across multiple angiosperm species and are less than 10 base pairs, the exact sequence from the reference is used to search the chloroplast genome being annotated. The search starts at the larger exon and proceeds towards the expected position of the small exon. The first hit, prior to the next gene, is identified as the missing, short exon. This search is based on nucleotide sequence as these small exons are highly conserved.

1. **Identifying Boundaries of Chloroplast Features**

Boundaries of tRNAs and rRNAs are highly conserved and thus reference-based boundary determination is generally sufficient. For tRNAs and rRNAs, the positional information from the previously identified optimal reference is used to mark the boundaries. We find that this method produces very accurate annotations of these features.

Boundary determination for protein-coding genes is more complex. For protein-coding genes, the type of exon/gene being annotated is first identified as entire gene, first exon, middle exon (for 3-exon genes), and last exon. Identification of gene boundaries is similar for each type of exon, but assumptions change based on expected patterns. The translated ORF sequence is queried using a sliding window of three amino acids starting from the exon boundary of the translated reference sequence and continues until a match between the two ends of the ORF and reference is identified. As the window moves, annoBTD keeps track of the number of amino acids that have been passed over in the reference to identify a matching amino acid triplet in the ORF. When assigning the exon boundary, the position of the matching amino acid triplet is translated to CDS position information and adjusted to reflect the appropriate number of skipped nucleotides, relative the number of amino acids, between the end of the exon in the reference and the designated end of the newly annotated exon. A simple example of this algorithm is demonstrated in Figure S1. Assumptions for exon structure change based on exon type in the following ways:

**Full gene/First Exon**

The first amino acid residue of the reference is ignored and the next three are used to search against the translated ORF as described above. This continues with counting of the number of shifted residues required to make a match. Once identified, if the reference sequence starts with a methionine and a methionine is present upstream of the matching amino acid triplet site, then that methionine is designated the start codon and accepted. If the reference does not start with a methionine or the ORF being annotated does not have a methionine upstream of the site of the matching amino acid triplet, then the start is designated as the number of shifted residues plus one (to represent the initially ignored starting residue of the reference) upstream of the matching site.

The 3’ end of the reference amino acid sequence is identified as either a STOP or any amino acid. If STOP, then gene is assumed to be a single exon gene and the end of the ORF (to the STOP codon) is designated as the end of the gene. If the reference ends in an amino acid and other exons exist, then the same method for boundary identification used for non-methionine start codons is used to identify the boundary. We have found that when present, these exon-intron boundaries are highly conserved across taxa.

**Middle Exon**

A few genes, such as *ycf3*, exist in three exons and possess a middle or second exon that has neither a start nor stop codon. For these exons, the basic boundary identification algorithm outlined above is used for both the 5’ and 3’ ends of the exon. Once amino acid triplets are matched from reference to translated ORF on either end of the exon, the appropriate number of shifted residues is used to adjust the exon boundaries.

**Last Exon**

The 5’ end of the final exon in a multi-exon gene is identified using the same amino acid triplet-matching algorithm as above. The 3’ end of the exon is identified as the end of the ORF, and, as with the end of the full gene example, represents a STOP codon.

**Split Codons at Exon Boundaries**

A few introns split codons (such as *petD*, which has an 8bp short exon and a long exon that completes the split codon). annoBTD uses both amino acid and nucleotide sequence of the reference to identify exon boundaries. The number of nucleotides in a reference exon is divided by three to determine if the exon is split in the reference. If so, annoBTD removes the split-codon nucleotides and records how many were removed to adjust the boundary later. The aforementioned boundary identification algorithm is then used to set the expected boundary based on amino acids, ignoring the amino acid specified by the split codon. Once the boundary has been identified, the split codon is replaced and the boundary adjusted to reflect the removed nucleotides.

1. **Identification of Single Copy and Inverted Repeats**

Inverted repeat regions are identified from the complete chloroplast sequence using a sliding 20-mer window. If found in both the forward and reverse complement in the chloroplast genome, 20-mers are chained together until no reverse complement match of the current 20-mer is found. This identifies IR sequence. The positional information of the IR is then estimated for both IRB and IRA. The positions of the single copy regions relative to the IRB and IRA are used to assign large single copy and small single copy. This method is used instead of size to maintain orthology of these regions.

Once fully annotated, the plastome features are loaded in the Verdant database for user access.

**The reference database**

Verdant relies on a database of whole chloroplast genomes, which are used for finalizing annotations. The initial chloroplast database was initiated with 163 whole chloroplast genomes from GenBank. A full list of these species and their GenBank IDs is found in Table S1. Chloroplast genomes were chosen by first identifying those that exhibit the typical angiosperm genome architecture (large single copy—inverted repeat B—small single copy—inverted repeat A) from the entire set of whole chloroplast genomes. From this reduced set, we then filtered out plastomes that were annotated with non-canonical chloroplast genes. These are typically older entries that had species-specific open reading frames (ORFs) identified as potential genes. Though possibly accurate, these are not tractable as they represent annotations that are simply not updated to currently accept gene identities. We decided to ignore these in favor of consistency across taxa. After our filtering criteria were applied, the remaining chloroplast genomes from GenBank and their accompanying annotations were incorporated directly into Verdant. Since the annoBTD software relies on some reference for feature identification, these GenBank annotations serve as the initial set of chloroplast genomes upon which annotation will be made. Users are able to add additional annotations directly to improve the quality of annotation for their specific projects.

As usage of Verdant increases and identification of problems in existing chloroplasts are found, we will be updating and removing poorly annotated chloroplast genomes that have made it through our current filtering process. In the event of removal, a message will be sent to users so they may verify that the removed chloroplast genome did not affect their research.

When a new plastome is uploaded to Verdant, the annoBTD process is started automatically. This process involves multiple steps that combine *de novo* and referenced-based annotation methods to describe the gene space of chloroplast genomes. As the Verdant database grows, the accuracy of the reference-based methods improves due to increased similarity of references to novel plastomes.

**Best Practices for Annotation**

We highly recommend that users thoroughly check a few annotations for their project prior to completing a full dataset using Verdant. Users are advised to upload a chloroplast genome from their taxon set, annotate using Verdant, download the annotation, and verify the presence of known chloroplast features. We have provided a JBrowse session for each annotation so it can be easily checked. Users can edit either the Verdant or GFF format annotations outside of Verdant and re-upload the corrected annotations directly to Verdant. These annotations will then most likely be used as references for novel plastome annotation due to their taxonomic proximity to other taxa in the user data set.

**Problems for annoBTD**

If the reference plastome is missing genes relative to the plastome being annotated, then annoBTD will continue to miss those genes. For example, orchid chloroplast genomes are missing *ndh* genes, and if an orchid plastome is chosen as the reference for another member of Asparagales, then the annotation will be incomplete. For this reason, we recommend manually annotating and uploading a complete chloroplast annotation for the family in which the user is working if it does not exist.

If members of a lineage are highly variable, then annoBTD may have difficulty accurately annotating. We have only seen this happen in the genus *Cuscuta*, where certain species have highly reduced chloroplast genomes as a consequence of holoparasitism. In addition to gene reduction and intron loss, many protein-coding genes in the lineage have undergone high rates of change at the amino acid level. When this occurs, the default parameters for annoBTD often will not allow these ORFs to be identified to a reference gene. In these cases, we recommend running annoBTD directly with relaxed parameters ([**https://github.com/mrmckain/annoBTD**](https://github.com/mrmckain/annoBTD)**).**

Finally, annoBTD does not need a complete chloroplast genome for annotation. However, if an incomplete chloroplast genome is annotated and not removed by the user, then Verdant may choose this chloroplast as a potential reference. In these cases, genes will be missing from the annotation. To avoid this, users should add incomplete chloroplast genomes last to a project or delete them from their project.

**References:**

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Daniell, H., C-S. Lin, M. Yu, and W-J. Chang. (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology,* 17, 134.

Huang, D. and Q. C. B. Cronk. (2015) Plann: A Command-Line Application for Annotating Plastome Sequences. *Applications in Plant Sciences*, 3, 8.

Liu, C., et al. (2012) CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequence. *BMC Genomics*, 13, 715.

Wyman, S.K., et al. (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, 20(17), 3252–3255.

Table S1. Complete list of GenBank chloroplast genomes in Verdant

|  |  |  |
| --- | --- | --- |
| **Genus** | **Species** | **GenBank ID** |
| *Acidosasa* | *purpurea* | HQ337793 |
| *Agrostemma* | *githago* | KF527884 |
| *Agrostis* | *stolonifera* | NC008591 |
| *Alstroemeria* | *aurea* | KC968976 |
| *Anthriscus* | *cerefolium* | GU456628 |
| *Apios* | *americana* | KF856618 |
| *Arachis* | *hypogaea* | KJ468094 |
| *Aralia* | *undulata* | KC456163 |
| *Ardisia* | *polysticta* | KC465962 |
| *Artemisia* | *frigida* | JX293720 |
| *Artemisia* | *montana* | KF887960 |
| *Arundinaria* | *fargesii* | JX513413 |
| *Asclepias* | *nivea* | KF539844 |
| *Asclepias* | *syriaca* | KF386166 |
| *Azadirachta* | *indica* | KF986530 |
| *Bambusa* | *emeiensis* | HQ337797 |
| *Bambusa* | *multiplex* | KJ722536 |
| *Brassaiopsis* | *hainla* | KC456164 |
| *Buxus* | *microphylla* | EF380351 |
| *Calanthe* | *triplicata* | KF753635 |
| *Camellia* | *crapnelliana* | KF753632 |
| *Camellia* | *grandibracteata* | NC024659 |
| *Camellia* | *leptophylla* | KJ806275 |
| *Camellia* | *petelotii* | KJ806276 |
| *Camellia* | *pubicosta* | KJ806277 |
| *Campanula* | *takesimana* | KP006497 |
| *Capsicum* | *lycianthoides* | KP274856 |
| *Carica* | *papaya* | EU431223 |
| *Castanea* | *mollissima* | HQ336406 |
| *Castanopsis* | *echinocarpa* | KJ001129 |
| *Ceratonia* | *siliqua* | KJ468096 |
| *Ceratophyllum* | *demersum* | EF614270 |
| *Chimonocalamus* | *longiusculus* | JX513415 |
| *Chrysobalanus* | *icaco* | KJ414480 |
| *Citrus* | *sinensis* | DQ864733 |
| *Coffea* | *arabica* | NC008535 |
| *Couepia* | *guianensis* | KJ414482 |
| *Crithmum* | *maritimum* | HM596072 |
| *Cuscuta* | *exaltata* | EU189132 |
| *Cuscuta* | *obtusiflora* | EU189133 |
| *Dioscorea* | *elephantipes* | EF380353 |
| *Drimys* | *granadensis* | DQ887676 |
| *Echinochloa* | *oryzicola* | NC024643 |
| *Elaeis* | *guineensis* | JF274081 |
| *Elodea* | *canadensis* | JQ310743 |
| *Erodium* | *absinthoides* | KJ523886 |
| *Erycina* | *pusilla* | JF746994 |
| *Eucalyptus* | *grandis* | HM347959 |
| *Fargesia* | *nitida* | JX513416 |
| *Fargesia* | *spathacea* | JX513417 |
| *Fargesia* | *yunnanensis* | JX513418 |
| *Fragaria* | *chiloensis* | JN884816 |
| *Fragaria* | *virginiana* | JN884817 |
| *Fritillaria* | *cirrhosa* | KF769143 |
| *Fritillaria* | *hupehensis* | KF712486 |
| *Fritillaria* | *taipaiensis* | KC543997 |
| *Gaoligongshania* | *megalothyrsa* | JX513419 |
| *Gelidocalamus* | *tessellatus* | JX513420 |
| *Glycine* | *max* | NC007942 |
| *Glycine* | *soja* | KF611800 |
| *Gossypium* | *anomalum* | JF317351 |
| *Gossypium* | *areysianum* | JN019795 |
| *Gossypium* | *gossypioides* | HQ901195 |
| *Gossypium* | *hirsutum* | HQ901196 |
| *Gossypium* | *incanum* | JN019792 |
| *Gossypium* | *longicalyx* | JF317354 |
| *Gossypium* | *raimondii* | HQ325744 |
| *Gossypium* | *robinsonii* | JN019791 |
| *Gossypium* | *somalense* | JN019793 |
| *Gossypium* | *stocksii* | JF317355 |
| *Gossypium* | *sturtianum* | JF317356 |
| *Gossypium* | *thurberi* | GU907100 |
| *Haematoxylum* | *brasiletto* | KJ468097 |
| *Helianthus* | *annuus* | NC007977 |
| *Hevea* | *brasiliensis* | HQ285842 |
| *Hibiscus* | *syriacus* | KP688069 |
| *Hirtella* | *physophora* | KJ414485 |
| *Hirtella* | *racemosa* | KJ414479 |
| *Hydrocotyle* | *verticillata* | HM596070 |
| *Hyoscyamus* | *niger* | KF248009 |
| *Indigofera* | *tinctoria* | KJ468098 |
| *Indocalamus* | *wilsonii* | JX513421 |
| *Indosasa* | *sinica* | JX513422 |
| *Iochroma* | *stenanthum* | KP262399 |
| *Ipomoea* | *purpurea* | EU118126 |
| *Jacobaea* | *vulgaris* | HQ234669 |
| *Kalopanax* | *septemlobus* | KC456167 |
| *Licania* | *alba* | KJ414483 |
| *Licania* | *heteromorpha* | KJ414481 |
| *Licania* | *sprucei* | KJ414484 |
| *Lilium* | *longiflorum* | KC968977 |
| *Liquidambar* | *formosana* | KC588388 |
| *Liriodendron* | *tulipifera* | NC008326 |
| *Lolium* | *perenne* | NC009950 |
| *Lonicera* | *japonica* | KJ170923 |
| *Lupinus* | *albus* | KJ468099 |
| *Lupinus* | *luteus* | KC695666 |
| *Luzuriaga* | *radicans* | KM233640 |
| *Macadamia* | *integrifolia* | NC025288 |
| *Magnolia* | *denudata* | NC018357 |
| *Magnolia* | *kwangsiensis* | HM775382 |
| *Magnolia* | *officinalis* | JN867579 |
| *Masdevallia* | *coccinea* | KP205432 |
| *Megaleranthis* | *saniculifolia* | FJ597983 |
| *Metapanax* | *delavayi* | KC456165 |
| *Morus* | *mongolica* | NC025772 |
| *Nandina* | *domestica* | DQ923117 |
| *Nelumbo* | *lutea* | FJ754269 |
| *Nelumbo* | *nucifera* | FJ754270 |
| *Neyraudia* | *reynaudiana* | KF356392 |
| *Nicotiana* | *undulata* | JN563929 |
| *Nuphar* | *advena* | DQ354691 |
| *Oryza* | *meridionalis* | NC016927 |
| *Oryza* | *rufipogon* | JN005832 |
| *Pachyrhizus* | *erosus* | KJ468100 |
| *Panax* | *notoginseng* | KJ566590 |
| *Panicum* | *virgatum* | NC015990 |
| *Parinari* | *campestris* | KJ414486 |
| *Pelargonium* | *alternans* | KF240617 |
| *Pentactina* | *rupicola* | JQ041763 |
| *Penthorum* | *chinense* | JX436155 |
| *Petroselinum* | *crispum* | HM596073 |
| *Phaseolus* | *vulgaris* | EU196765 |
| *Phragmites* | *australis* | KF730315 |
| *Phyllostachys* | *edulis* | HQ337796 |
| *Phyllostachys* | *propinqua* | JN415113 |
| *Piper* | *cenocladum* | NC008457 |
| *Platanus* | *occidentalis* | DQ923116 |
| *Pleioblastus* | *maculatus* | JX513424 |
| *Populus* | *euphratica* | KJ624919 |
| *Premna* | *microphylla* | KM981744 |
| *Prosopis* | *glandulosa* | KJ468101 |
| *Quercus* | *aliena* | KP301144 |
| *Ranunculus* | *macranthus* | NC008796 |
| *Raphanus* | *sativus* | KJ716483 |
| *Ricinus* | *communis* | NC016736 |
| *Robinia* | *pseudoacacia* | KJ468102 |
| *Salvia* | *miltiorrhiza* | JX312195 |
| *Sapindus* | *mukorossi* | KM454982 |
| *Sarocalamus* | *faberi* | JX513414 |
| *Schefflera* | *delavayi* | KC456166 |
| *Sedum* | *sarmentosum* | JX427551 |
| *Setaria* | *italica* | KJ001642 |
| *Silene* | *chalcedonica* | KF527886 |
| *Silene* | *conica* | JF715054 |
| *Silene* | *conoidea* | KF527885 |
| *Silene* | *latifolia* | JF715055 |
| *Silene* | *noctiflora* | JF715056 |
| *Silene* | *paradoxa* | KF527887 |
| *Silene* | *vulgaris* | JF715057 |
| *Solanum* | *lycopersicum* | NC007898 |
| *Spirodela* | *polyrhiza* | NC015891 |
| *Tamarindus* | *indica* | KJ468103 |
| *Tetracentron* | *sinense* | KC608752 |
| *Thamnocalamus* | *spathiflorus* | JX513425 |
| *Theobroma* | *cacao* | HQ336404 |
| *Trachelium* | *caeruleum* | EU090187 |
| *Trochodendron* | *aralioides* | KC608753 |
| *Typha* | *latifolia* | GU195652 |
| *Utricularia* | *gibba* | KC997777 |
| *Vigna* | *radiata* | NC013843 |
| *Wolffia* | *australiana* | NC015899 |
| *Wolffiella* | *lingulata* | NC015894 |



**Figure S1.** Graphical representation of amino acid triplet algorithm for boundary identification for genes in annoBTD. Example shown is for a full gene that required two iterations. In the first iteration, the amino acid triplet from the reference ranged from positions 2-4 and had no match in the ORF being annotated. Codon 2 codes for K instead of N in the ORF being annotated so only residues 3 and 4 match. In this case annoBTD records a non-match because the entire triplet is not present. After failing to find a match, annoBTD slides the amino acid triplet window 1 position downstream in the reference. In the second iteration, the amino acid triplet from positions 3-5 was identified in the ORF. Since the reference gene begins with a start codon, a start codon upstream from this matching position sets the 5’ boundary of the gene.